Forest Fires - Gruppo Capri

Andrei Petrisor 1085993, Antonio Radu 1085992, Lorenzo Medici 1085852, Andrea Rusconi 1086646

15-01-2024

Contents

1 Dataset

Questo dataset è pubblico e a disposizione per la ricerca. I dettagli sul dataset possono essere trovati in Cortez e Morais (2007). Il dataset è composto dalle seguenti variabili:

1. Coordinata spaziale dell'asse X all'interno della mappa del parco Montesinho: da 1 a 9
2. Y coordinata spaziale dell'asse y all'interno della mappa del parco Montesinho: da 2 a 9
3. mese dell'anno: da "jan" a "dec"
4. giorno della settimana: da "mon" a "sun"
5. Indice FFMC dal sistema FWI: da 18,7 a 96,20
6. Indice DMC dal sistema FWI: da 1,1 a 291,3
7. Indice DC dal sistema FWI: da 7,9 a 860,6
8. Indice ISI del sistema FWI: da 0,0 a 56,10
9. temperatura temporanea in gradi Celsius: da 2,2 a 33,30
10. Umidità relativa RH in %: da 15,0 a 100
11. velocità del vento in km/h: da 0,40 a 9,40
12. pioggia in mm/m2: da 0,0 a 6,4
13. area della superficie bruciata della foresta (in ettari): da 0,00 a 1090,84.

Il Forest Fire Weather Index (FWI) è il sistema canadese per la classificazione del pericolo di incendio e comprende sei componenti: Indice di umidità del combustibile (FFMC), indice di umidità (DMC), indice di siccità (DC), indice di spread iniziale (ISI), indice di accumulo (BUI) e FWI

1.1 Obiettivo

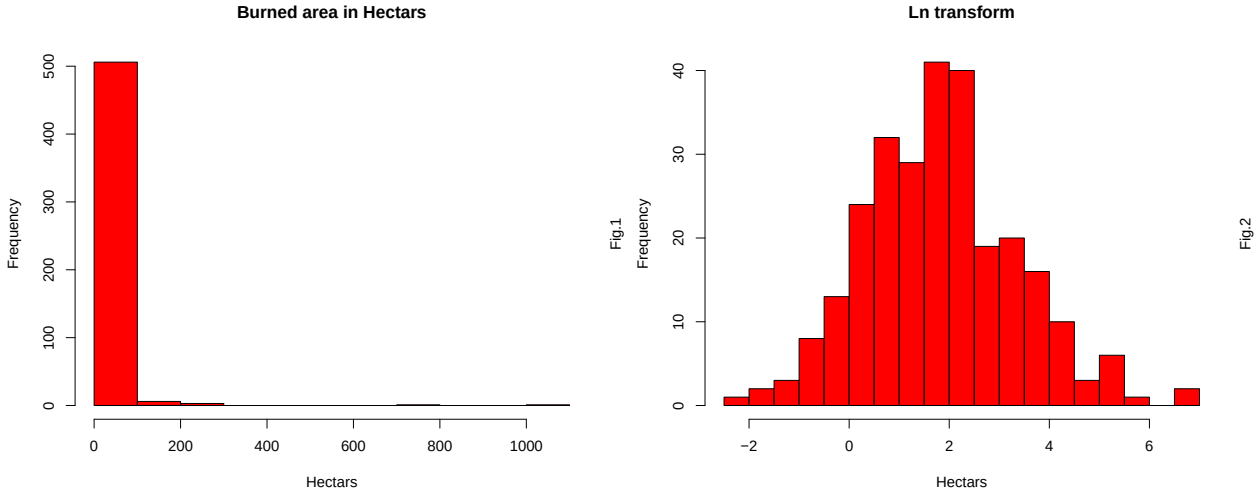
In questo dataset siamo interessati a modellare l'area bruciata della foresta come funzione delle altre variabili. Siamo in particolare interessati a rispondere alla domanda: "Concentrandosi sul mese di agosto, come si può spiegare l'area bruciata in funzione delle altre variabili?".

Cortez P. e Morais A. "Un approccio di data mining per prevedere gli incendi boschivi utilizzando dati meteorologici." In J. Neves, MF Santos e J. Machado Eds., "Nuove tendenze nell'intelligenza artificiale", Atti della 13a EPIA 2007 Conferenza portoghese sull'intelligenza artificiale, dicembre, Guimaraes, Por- tugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9. 18-0-9. Disponibile a: http://www3.dsi.uminho.pt/pcortez/fires.pdf

Il dataset è composto dalle seguenti rilevazioni:

```
##       Days
## Months mon tue wed thu fri sat sun
##    jan   0   0   0   0   0   1   1
##    feb   3   2   1   1   5   4   4
##    mar  12   5   4   5  11  10   7
##    apr   1   0   1   2   1   1   3
##    may   0   0   0   0   1   1   0
##    jun   3   0   3   2   3   2   4
##    jul   4   6   3   3   3   8   5
##    aug  15  28  25  26  21  29  40
##    sep  28  19  14  21  38  25  27
##    oct   4   2   2   0   1   3   3
##    nov   0   1   0   0   0   0   0
##    dic   0   0   0   0   0   0   0
```

1.2 Istogrammi area bruciata



Possiamo vedere come l'istogramma(Fig.1) presenti una asimmetria positiva(obliqua a destra) , di conseguenza valutiamo i dati facendo la trasformata grazie al logaritmo, cosifacendo otteniamo un grafico più simile ad una normale.

```
summary(forest$area)
```

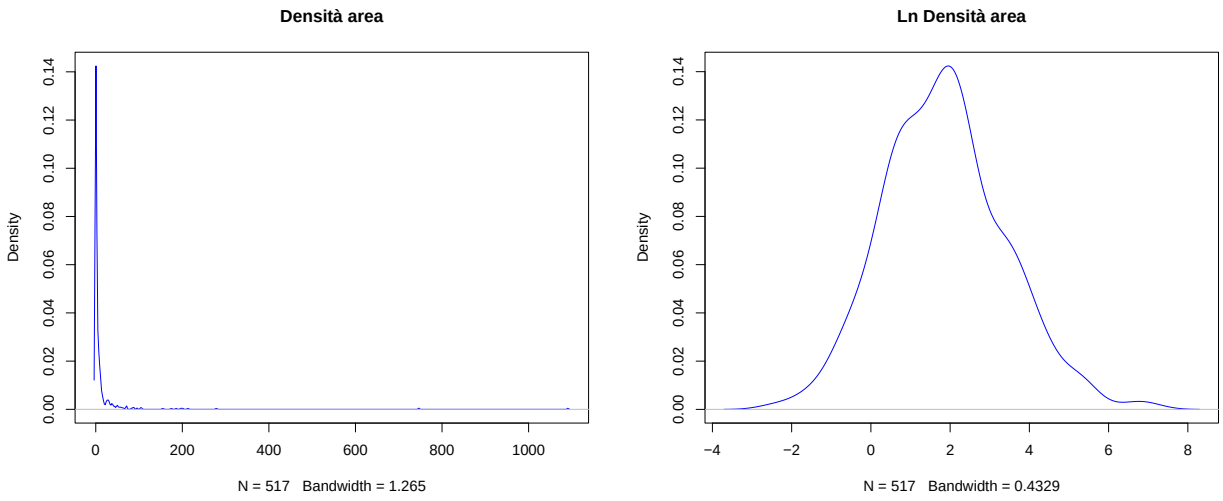
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.52   12.85    6.57 1090.84
```

Facendo il summary otteniamo i valori di: min, max, media, mediana, possiamo considerare i nostri dati molto vicini allo zero.

```
## [1] 0.4777563
```

Si evidenzia che il dataset ha il 48% dei valori che valgono 0.

1.3 Grafici densità area bruciata

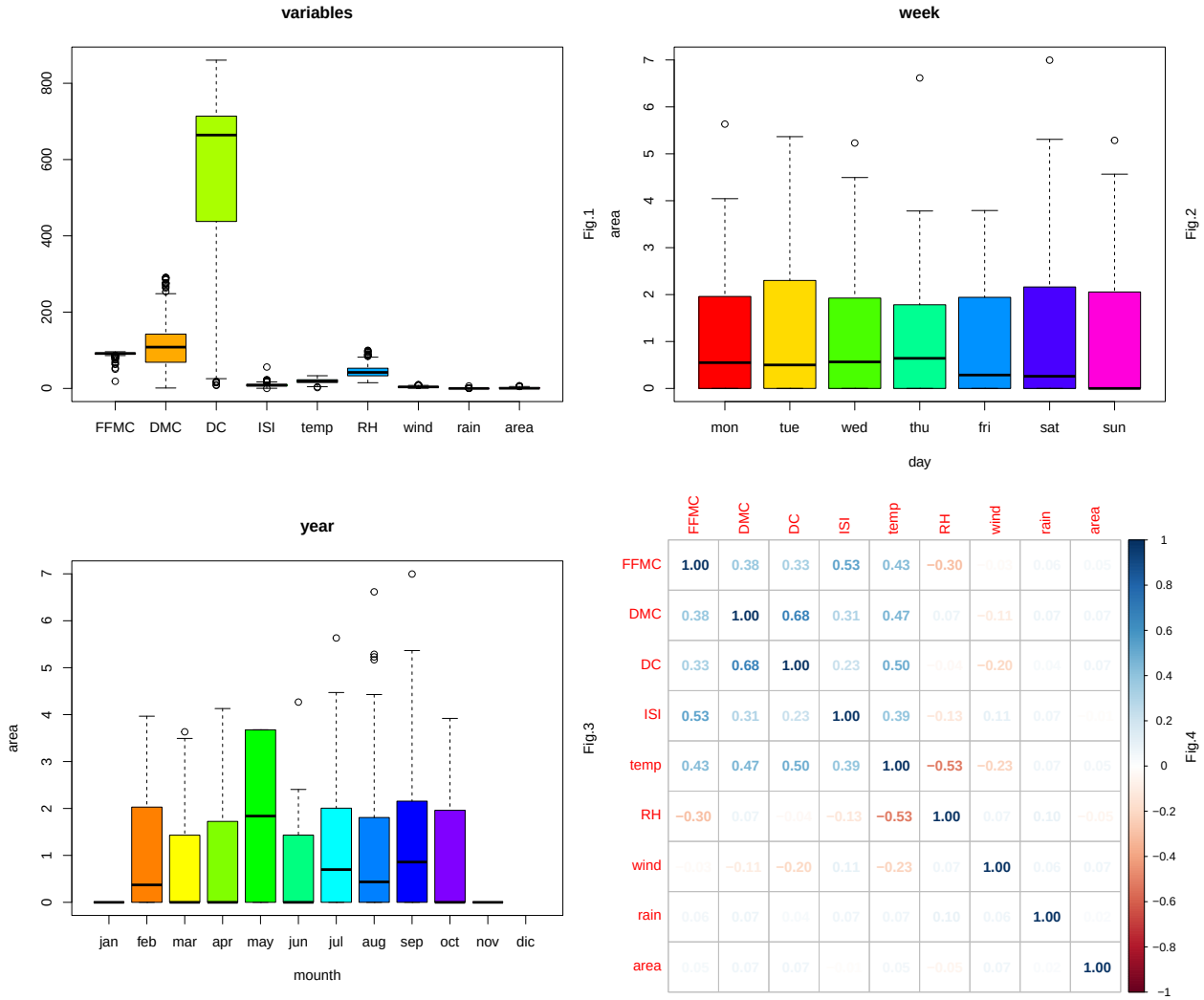


Sembra esserci anche una forte distorsione rispetto ai residui che non sono normalmente distribuiti. Ciò sembra essere dovuto al gran numero di 0 nel database. Quando questi 0 vengono rimossi, possiamo vedere che i residui diventano più normalmente distribuiti.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.4187  1.1110  2.0242  6.9956
```

come possiamo vedere la mediana e la media sono diminuite.

```
## Warning: il pacchetto 'corrplot' è stato creato con R versione 4.3.2
```



-Come possiamo vedere da questo correlation plot(fig.4) le variabili che sono correlate positivamente di più sono: DC con DMC, ISI e FFMC, temp e DC. Le variabili correlate negativamente di più sono, invece: RH e temp.

-Dal box plot delle variabili(fig.1) possiamo determinare che le mediane di quasi tutte le variabili sono più o meno simili quindi ci sono poche differenze,inoltre possiamo notare come in tutte le variabili ci siano degli outliers sia al minimo che al massimo. La variabile DMC, invece, presenta dei baffi più lunghi il che implica che tale variabile ha valori più incoerenti rispetto alle altre, la mediana è lontana da tutte le altre ed è spostata molto verso il terzo quartile.

-Dal box plot dei giorni della settimana(fig.2) possiamo vedere come le mediane siano simili tra di loro, come anche i baffi, questo dimostra come i giorni della settimana abbiano tutti valori coerenti tra di loro.

-Dal box plot dei mesi dell'anno(fig.3) notiamo come il mese di dicembre e quello di maggio siano quelli più incoerenti rispetto agli altri mesi. Il mese di agosto presenta molti più outliers rispetto agli altri mesi.

2 Testing

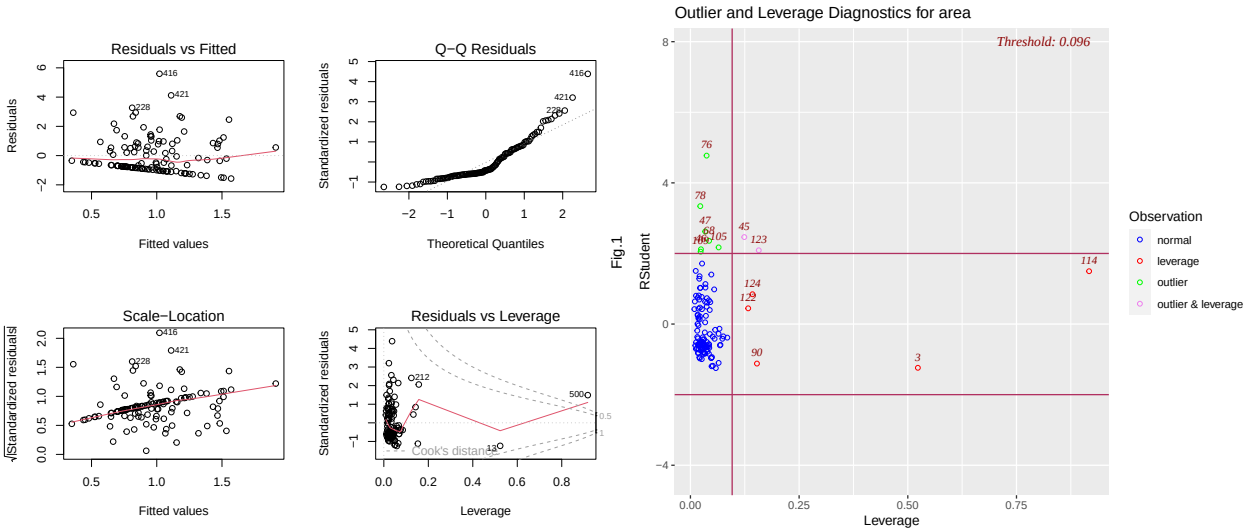
Per la fase di testing prendiamo in considerazione solamente il mese di agosto, per poter spiegare come varia l'area bruciata in relazione alle altre variabili. Dividiamo la fase di Testing in: train e test, con il train al 70% e il test al 30%.

Proviamo a fare un modello lineare con i dati del mese di agosto. Le variabili che secondo noi ha piu senso considerare nel modello sono: rain ,RH ,temp, FFMC, e DMC. RH-FFMC-DMC sono degli indici che tengono conto dell'umidità. Nessuno di questi parametri agisce in modo diretto sullo sviluppo dell'incendio, ma sono tutti fattori predisponenti, perciò consideriamo rilevante studiare come influenzano l'area incendiata.

```
##
## Call:
## lm(formula = area ~ rain + RH + temp + FFMC + DMC, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5702 -0.8345 -0.5287  0.6288  5.5955
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.292937   3.540255  -0.365   0.7156
## rain         0.044054   0.212911   0.207   0.8364
## RH           0.014859   0.012451   1.193   0.2351
## temp         0.083263   0.039326   2.117   0.0363 *
## FFMC         0.001123   0.035157   0.032   0.9746
## DMC         -0.002174   0.002545  -0.854   0.3947
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.3 on 119 degrees of freedom
## Multiple R-squared:  0.04681,    Adjusted R-squared:  0.006758
## F-statistic: 1.169 on 5 and 119 DF,  p-value: 0.3286
```

Come possiamo vedere il nostro modello non performa benissimo, con un 4% come r^2 il modello ha un grande margine di errore. Tra le varie covariate scelte notiamo come la temperatura abbia il valore piu grande, quindi, all'aumentare unitario dell'area, la temperatura media aumenta di 0.08. Proviamo a vedere se ci sono degli Outliers(valori estremi) nel modello.

```
## Warning: il pacchetto 'olsrr' è stato creato con R versione 4.3.2
```



Possiamo vedere come nel grafici (Fig.1), un numero considerevole di osservazioni non segue l'andamento desiderato, poniamo maggiore attenzione sul grafico Q-Q Res. dove possiamo vedere che i vari valori alle code tendono a spostarsi di molto, l'andamento non tende per niente alla normale. Per quanto riguarda il grafico a destra possiamo notare come alcune rilevazioni siano parecchio "lontane" rispetto alle altre, poniamo attenzione soprattuto a quelle che sono Outlier&Leverage e le rimuoviamo.

```
##
## Call:
## lm(formula = area ~ rain + RH + temp + FFMC + DMC, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5217 -0.6957 -0.3932  0.7101  3.0223
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.800705   3.873662   0.465 0.642927
## rain        -0.876780   0.568884  -1.541 0.126057
## RH           0.025039   0.010737   2.332 0.021466 *
## temp         0.115512   0.033134   3.486 0.000699 ***
## FFMC        -0.043793   0.040497  -1.081 0.281822
## DMC         -0.003424   0.002091  -1.638 0.104255
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.049 on 113 degrees of freedom
## Multiple R-squared:  0.1131, Adjusted R-squared:  0.07389
## F-statistic: 2.883 on 5 and 113 DF,  p-value: 0.01732
```

Il modello presenta un leggero miglioramento, l'r^2 aumenta del 7% rendendo il modello un minimo piu preciso. La covariata piu significativa è la temperatura, quindi all'aumentare unitario dell'area, la temperatura media aumenta di 0.12. I coefficienti associati alle variabili rain, FFMC, DMC non sono significativi, quindi supponiamo non influiscano sull'area. Passando da non avere la rain ad averla, a parità di temperatura, l'area diminuisce di 0.87.

2.1 Confronto tra modelli

Per valutare quale dei 3 modelli (1o modello con variabili scelte da noi, il 2o modello è uguale al primo, ma senza outliers, il 3o modello presenta tutte le variabili) sia meglio li mettiamo a confronto tramite la tecnica AIC(Akaike information criterion). Consiste in un metodo per la valutazione e il confronto tra modelli statistici. Fornisce una misura della qualità della stima di un modello statistico tenendo conto sia della bontà di adattamento che della complessità del modello. La regola è quella di preferire i modelli con l'AIC più basso.

```
## AIC lm 3 BIC lm 3 AIC lm 2 BIC lm 2 AIC lm 1 BIC lm 1
## 369.1109 413.5769 356.9427 376.3966 428.1296 447.9278
```

I risultati ci mostrano come il modello 2 abbia l'AIC piu basso, useremo questo modello per fare le previsioni.

2.2 Previsioni puntuale

Per stima puntuale s'intende l'insieme dei metodi inferenziali che permettono di attribuire un valore ad un parametro della popolazione, utilizzando i dati di un campione casuale osservato (x1, x2,. . . ,xn) ed elaborandoli.Per valutare la bontà di uno stimatore è necessario considerare le stime ottenute ripetendo un grande numero di volte il processo impiegato per eseguire una stima.

```
## [1] 1.560072
```

2.3 LASSO regression

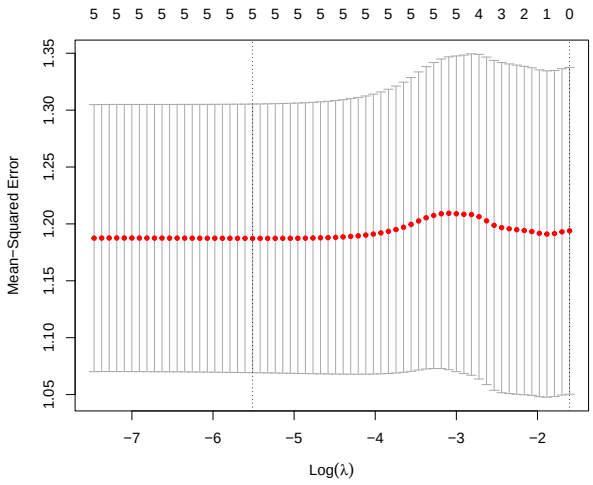
La tecnica LASSO(least absolute shrinkage and selection operator) è un metodo di analisi della regressione che esegue sia la selezione delle variabili sia la regolarizzazione per migliorare l'accuratezza della previsione e l'interpretabilità del modello statistico risultante. Il metodo lasso presuppone che i coefficienti del modello lineare siano sparsi, ossia che pochi di essi siano non nulli.

```
## Warning: il pacchetto 'glmnet' è stato creato con R versione 4.3.2
```

```
## Warning: il pacchetto 'Matrix' è stato creato con R versione 4.3.2
```

```
## [1] 0.004038118
```

```
## 6 x 1 sparse Matrix of class "dgCMatrix"
##                        s0
## (Intercept)  1.882385457
## RH           0.023413676
## rain        -0.822658435
## temp         0.110387520
## FFMC        -0.042988008
## DMC         -0.003248449
```



Il grafico dovrebbe assomigliare ad una curva esponenziale, nel nostro caso il mean-squared error assomiglia ad una retta.

2.4 Previsioni con Lasso

```
#use lasso regression model to predict response value
new = data.matrix(test[,c("RH","rain","temp","FFMC","DMC")])
previsioni = predict(best_model, s = best_lambda, newx = new)

# Root Mean Squared Error (RMSE): è una misura dell'errore che compiamo
sqrt(mean((test$area - previsioni)^2))#errore minimo
```

```
## [1] 1.556394
```

Come possiamo vedere la previsione con la Lasso è migliore rispetto alla previsone puntuale, decidiamo quindi di usare la Lasso regression per prevedere i dati.

3 Conclusioni

Concentrandosi sul mese di agosto, spiegare l'area bruciata in funzionde delle altre variabili risulta non facile visto la poca quantità di dati a disposizione. Poiché il valore r2 del modello è molto basso, probabilmente significa che i predittori avranno un errore significativo e potrebbero non essere affidabili.