

Forest Fires - Gruppo Capri

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

15-01-2024

Contents

1	Dataset	2
1.1	Istogrammi area bruciata	3
1.2	Grafici densità area bruciata	3
2	Domande specifiche	5
3	Metodologia	5
4	Analisi dati e discussione risultati	5
4.1	Scelta modello	10
4.2	Previsioni puntuale	10
4.3	LASSO regression	10
4.4	Previsioni con Lasso	11
5	Conclusioni	11

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:

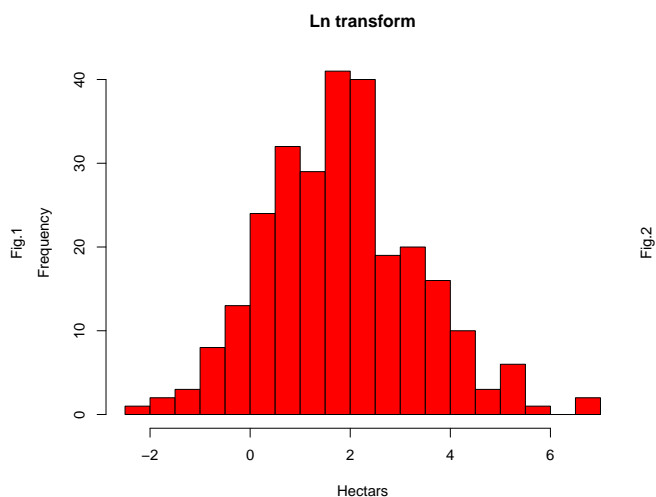
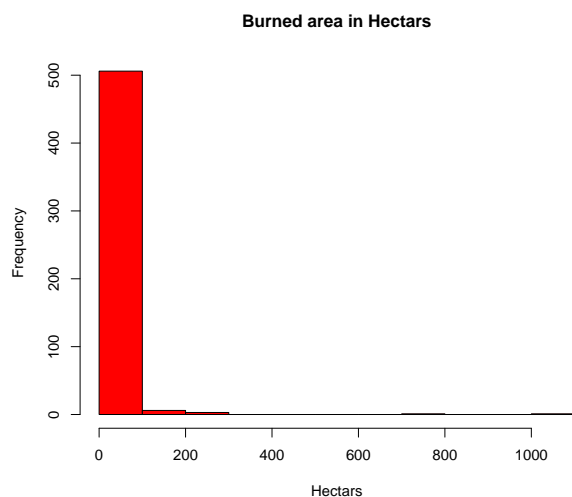
1. Coordinata spaziale dell'asse X all'interno della mappa del parco Montesinho: da 1 a 9
2. Y coordinata spaziale dell'asse y all'interno della mappa del parco Montesinho: da 2 a 9
3. mese dell'anno: da "jan" a "dec"
4. giorno della settimana: da "mon" a "sun"
5. Indice FFMC dal sistema FWI: da 18,7 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 dispersione iniziale (ISI) nel nostro caso indica la velocità della diffusione del fuoco, indice di accumulo (BUI) e FWI

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

1.1 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)
```

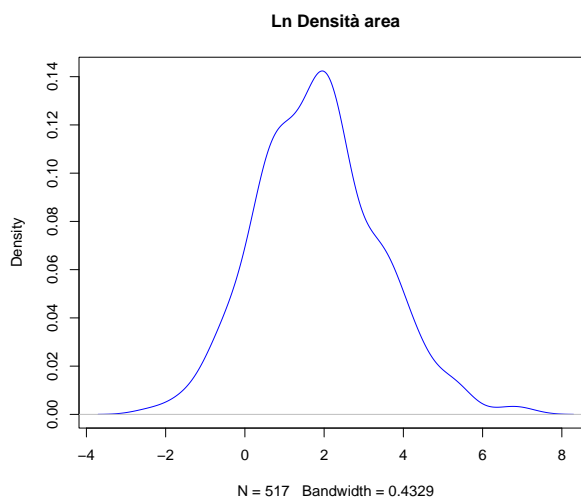
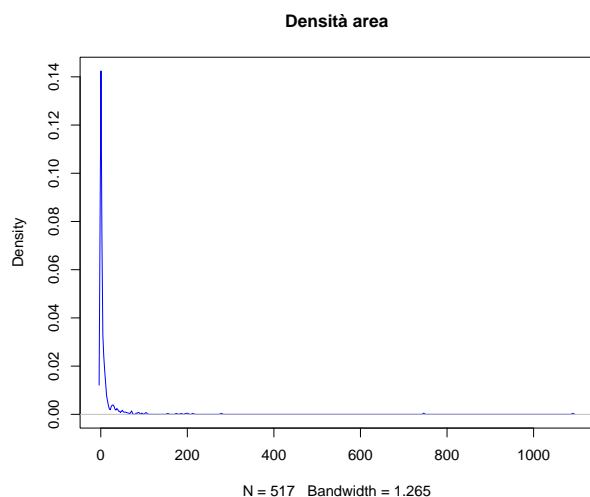
```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.00   0.00   0.52   12.85   6.57 1090.84
```

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.

1.2 Grafici densità area bruciata

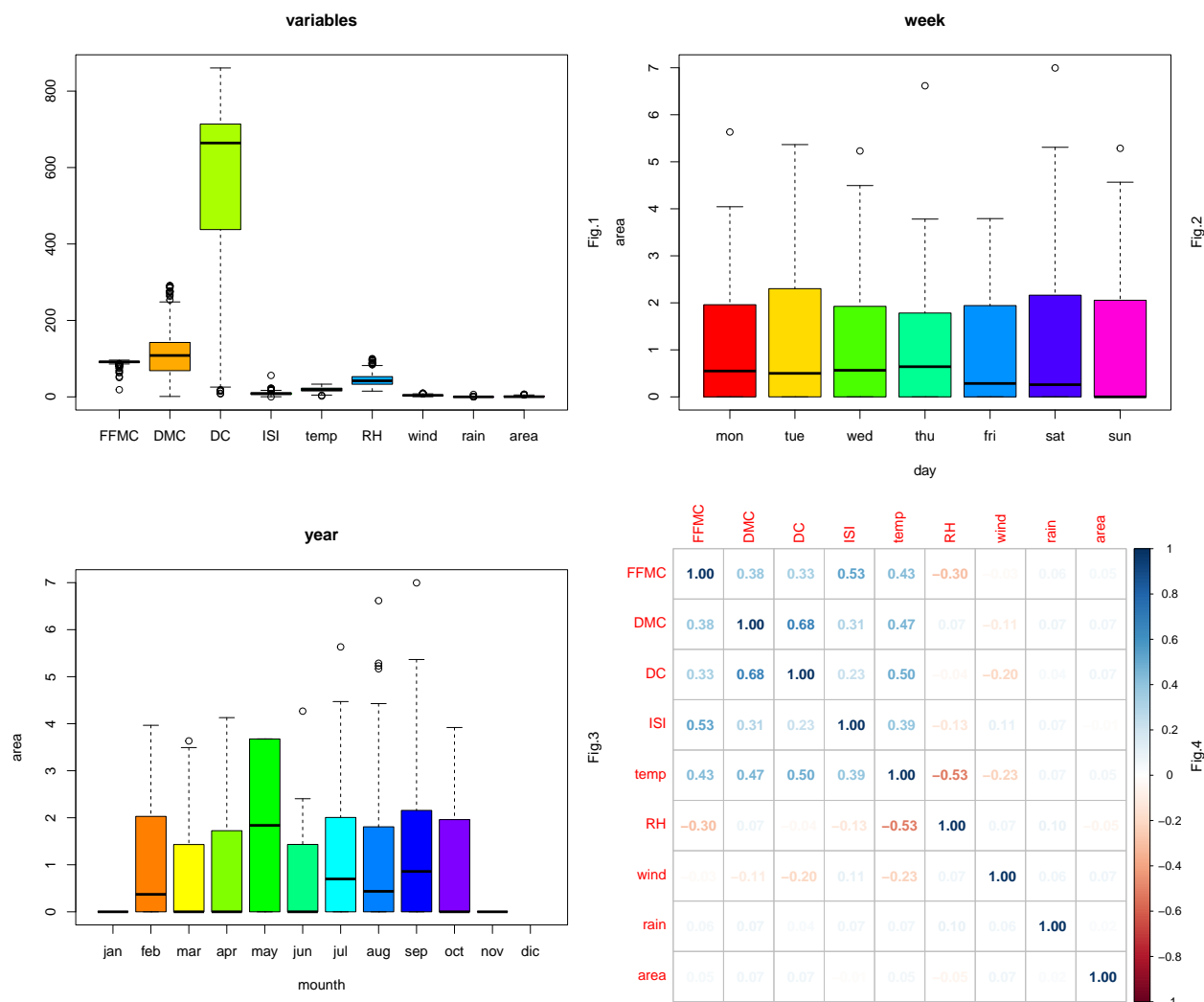


Sembra esserci anche una forte distorsione rispetto ai residui che non sono normalmente distribuiti. Ciò sembra essere dovuto al gran numero di 0 nel database. Quando questi 0 vengono rimossi, possiamo vedere che i residui diventano più normalmente distribuiti.

```
##      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. Concentrandoci sulla variabile risposta (area) non ci sono correlazioni forti con nessuna delle covariate.

-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 covariata 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. La covariata DC ha valori molto più elevati rispetto alle altre covariate, perciò non terremo conto di questa covariata.

-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 Domande specifiche

In questo dataset siamo interessati a modellare l'area bruciata della foresta come funzione delle altre variabili. Siamo in particolare interessati a capire nel mese di agosto come possiamo spiegare l'area bruciata in funzione delle altre variabili, vogliamo capire anche come si comportano i vari modelli(semplice, polinomiale e interazioni tra variabili) e interpretare i loro risultati. Nello specifico vogliamo scoprire da quali variabili dipende l'area bruciata maggiormente.

3 Metodologia

I metodi usati per raggiungere gli obiettivi sono dei metodi di regressione lineare, ci permettono di spiegare una variabile risposta (Y=area bruciata) in funzione delle altre variabili esplicative a disposizione nel dataset(covariate, X). Useremo due metodi di regressione:

-LASSO regression: metodo usato per trovare una best-lambda che ci permette di penalizzare alcuni regressori, questo per rendere il modello più semplice(facciamo model selection).

-K-fold cross validation: metodo usato per trattare un modello più complesso(visto che abbiamo pochi dati), questo per non andare in overfitting cioè il modello performa bene in train, ma male in test.

4 Analisi dati e discussione risultati

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%(cioè con il 70% dei dati alleno il modello, mentre con il restante 30% faccio le previsioni con i metodi).

Proviamo a fare un modello lineare con i dati del mese di agosto. Le variabili che secondo noi ha più 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 *
## FPMC         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

##
## Call:
## lm(formula = area ~ (rain^2) + (RH^2) + (temp^2) + (FFMC^2) +
##      (DMC^2) + (ISI^2), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.5369 -0.8491 -0.5510  0.6542  5.5757
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.643330   4.341081  -0.148   0.8824
## rain         0.046308   0.213925   0.216   0.8290
## RH           0.014371   0.012639   1.137   0.2578
## temp         0.081531   0.040037   2.036   0.0439 *
## FPMC        -0.006684   0.046293  -0.144   0.8854
## DMC         -0.001998   0.002643  -0.756   0.4512
## ISI          0.009492   0.036421   0.261   0.7948
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.305 on 118 degrees of freedom
## Multiple R-squared:  0.04736,    Adjusted R-squared:  -0.001083
## F-statistic: 0.9776 on 6 and 118 DF,  p-value: 0.4436

##
## Call:
## lm(formula = area ~ temp + (rain * temp) + (temp * FPMC) + (RH *
##      FPMC) + (train$day == "tue") + (train$day == "wed") + (train$day ==
##      "thu"), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.3428 -0.8996 -0.4608  0.7033  5.2668
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -79.417342   41.424874  -1.917   0.0577 .
## temp           2.473365    1.204105   2.054   0.0423 *
## rain          -4.701083    3.246733  -1.448   0.1504
## FPMC           0.841054    0.450289   1.868   0.0644 .
## RH             0.524843    0.318901   1.646   0.1026

```

```
## train$day == "tue"TRUE    0.155754    0.357349    0.436    0.6638
## train$day == "wed"TRUE    0.066586    0.366138    0.182    0.8560
## train$day == "thu"TRUE    0.244584    0.342235    0.715    0.4763
## temp:rain                 0.179877    0.119921    1.500    0.1364
## temp:FFMC                 -0.025822    0.013036   -1.981    0.0500 .
## FPMC:RH                   -0.005501    0.003473   -1.584    0.1160
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.294 on 114 degrees of freedom
## Multiple R-squared:  0.09437,    Adjusted R-squared:  0.01493
## F-statistic: 1.188 on 10 and 114 DF,  p-value: 0.3063

## Warning: il pacchetto 'ggplot2' è stato creato con R versione 4.3.2

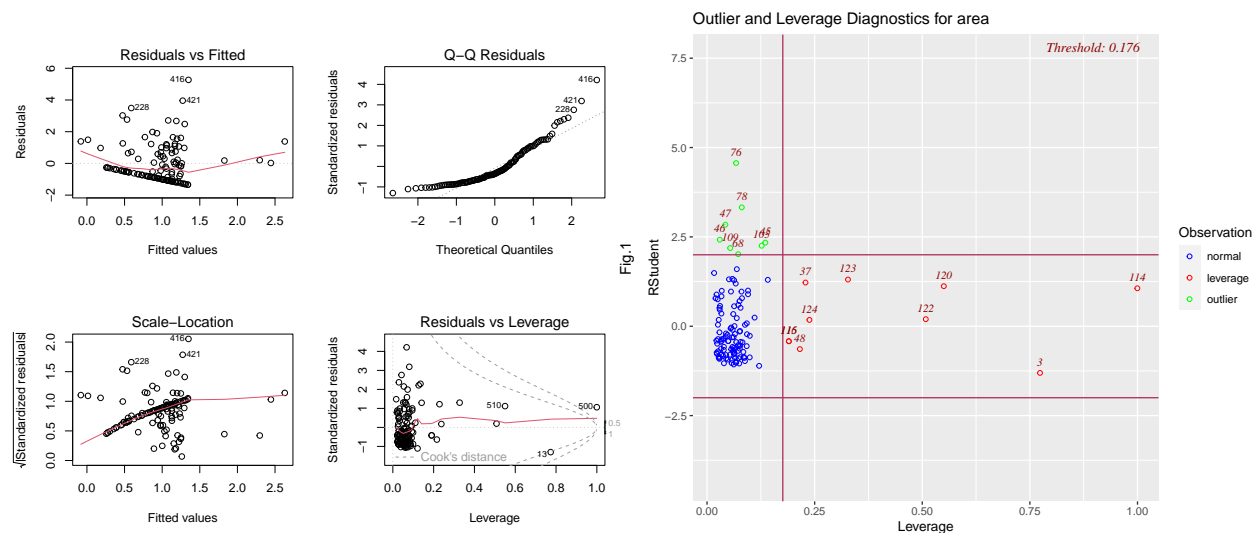
## Warning: il pacchetto 'hrbrthemes' è stato creato con R versione 4.3.2
```

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 più 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 in sqrt(crit * p * (1 - hh)/hh): Si è prodotto un NaN
```

```
## Warning in sqrt(crit * p * (1 - hh)/hh): Si è prodotto un NaN
```

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



Possiamo vedere come nei 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 soprattutto a quelle che sono Outlier&Leverage e le rimuoviamo.

```
##
## Call:
## lm(formula = area ~ rain + RH + temp + DMC, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.7785 -0.7302 -0.4259  0.6797  3.0669
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.238822   1.080414  -2.072  0.04056 *
## rain        -1.614104   0.756201  -2.134  0.03500 *
## RH           0.028180   0.010600   2.658  0.00901 **
## temp         0.113035   0.033566   3.368  0.00104 **
## DMC         -0.004020   0.002037  -1.974  0.05092 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.052 on 111 degrees of freedom
## Multiple R-squared:  0.1136, Adjusted R-squared:  0.0817
## F-statistic: 3.558 on 4 and 111 DF,  p-value: 0.009039
##
## Call:
```



```

## lm(formula = area ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.5748 -0.7234 -0.2893  0.6794  2.9778
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.381401   7.150537   0.053 0.957567
## daymon      -0.392316   0.486475  -0.806 0.421880
## daysat       0.158941   0.445745   0.357 0.722155
## daysun       0.148947   0.421071   0.354 0.724276
## daythu       0.250259   0.436419   0.573 0.567623
## daytue       0.464012   0.474776   0.977 0.330742
## daywed      -0.106543   0.464229  -0.230 0.818940
## FFMC        -0.039730   0.072147  -0.551 0.583067
## DMC         -0.005469   0.003433  -1.593 0.114224
## DC           0.001363   0.002019   0.675 0.501031
## ISI          0.011979   0.044861   0.267 0.789990
## temp         0.130202   0.036525   3.565 0.000558 ***
## RH           0.030606   0.012122   2.525 0.013130 *
## wind        -0.080184   0.071348  -1.124 0.263739
## rain        -1.742869   0.812430  -2.145 0.034330 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.066 on 101 degrees of freedom
## Multiple R-squared:  0.1721, Adjusted R-squared:  0.05729
## F-statistic: 1.499 on 14 and 101 DF,  p-value: 0.1249

##
## Call:
## lm(formula = area ~ temp + (rain * temp) + (temp * FFMC) + (RH *
##      FFMC) + (day) + wind, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.7355 -0.7320 -0.3281  0.6366  3.0175
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -52.318832  36.358093  -1.439  0.1532
## temp         1.249270   1.109021   1.126  0.2626
## rain        -7.511592 108.774565  -0.069  0.9451
## FFMC         0.547833   0.395198   1.386  0.1687
## RH           0.614768   0.283902   2.165  0.0327 *
## daymon      -0.292010   0.471243  -0.620  0.5369
## daysat       0.176281   0.441573   0.399  0.6906
## daysun       0.189648   0.403740   0.470  0.6396
## daythu       0.134305   0.424026   0.317  0.7521
## daytue       0.622127   0.440004   1.414  0.1605
## daywed      -0.025061   0.436334  -0.057  0.9543
## wind        -0.078022   0.069752  -1.119  0.2660
## temp:rain     0.287790   5.002589   0.058  0.9542

```

```
## temp:FFMC      -0.012562    0.011991   -1.048    0.2973
## FPMC:RH        -0.006468    0.003093   -2.091    0.0390 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.06 on 101 degrees of freedom
## Multiple R-squared:  0.1816, Adjusted R-squared:  0.06813
## F-statistic: 1.601 on 14 and 101 DF,  p-value: 0.09169
```

Il modello presenta un leggero miglioramento, l' r^2 aumenta del 7% rendendo il modello un minimo più preciso. La covariata più significativa è la temperatura, quindi all'aumentare unitario dell'area, la temperatura media aumenta di 0.12. I coefficienti associati alle variabili rain, FPMC, 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.

4.1 Scelta modello

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 6 BIC lm 6 AIC lm 5 BIC lm 5 AIC lm 3 BIC lm 3 AIC lm 2 BIC lm 2
## 358.6993 402.7567 430.0576 452.6842 360.0412 404.0986 347.9498 364.4713
## AIC lm 1 BIC lm 1
## 428.1296 447.9278
```

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

4.2 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 (x_1, x_2, \dots, x_n) 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] 1.541404
```

4.3 LASSO regression

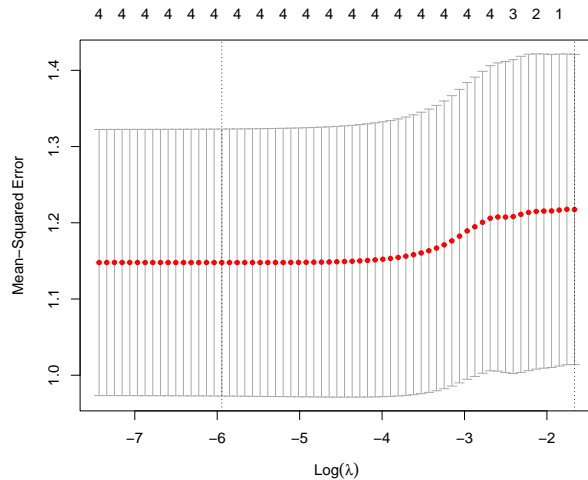
La tecnica LASSO(least absolute shrinkage and selection operator) è un metodo di analisi della regressione che esegue sia la selezione delle variabili sia la regolarizzazione per migliorare l'accuratezza della previsione e l'interpretabilità del modello statistico risultante. Il metodo lasso presuppone 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] 0.002620917
```

```
## 5 x 1 sparse Matrix of class "dgCMatrix"
##              s0
## (Intercept) -2.140629719
## temp        0.109817478
## rain        -1.570693718
## RH          0.027139870
## DMC         -0.003903726
```



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

4.4 Previsioni con Lasso

```
#use lasso regression model to predict response value
new = data.matrix(test[,c("temp", "rain", "RH", "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] 1.539185
```

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

5 Conclusioni

Concentrandosi sul mese di agosto, spiegare l'area bruciata in funzione delle altre variabili risulta non facile visto la poca quantità di dati a disposizione. Poiché il valore r^2 del modello è molto basso, probabilmente significa che i predittori avranno un errore significativo e potrebbero non essere affidabili.

```

# package to compute
# cross - validation methods
library(caret)

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

set.seed(125)

# defining training control
# as cross-validation and
# value of K equal to 10
train_control <- trainControl(method = "cv",
                              number = 10)

# training the model by assigning sales column
# as target variable and rest other column
# as independent variable
model <- train(area ~ temp+(rain*temp)+(temp*FFMC)+(RH*FFMC)+(day)+wind, data = train,
               method = "lm",
               trControl = train_control)

# printing model performance metrics
# along with other details
print(model)

## Linear Regression
##
## 116 samples
## 6 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 105, 104, 105, 104, 105, 104, ...
## Resampling results:
##
## RMSE      Rsquared    MAE
## 1.137352  0.1062959  0.9302166
##
## Tuning parameter 'intercept' was held constant at a value of TRUE

#use lasso regression model to predict response value
best_model <- glmnet(x, y, alpha = 1, lambda = 1.137352)
previsioni = predict(best_model, s = 1.137352, newx = new)

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

## [1] 1.505979

```

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, Portugal, 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>