Data Analysis

Faiella Ciro, Giannino Pio Roberto, Scovotto Luigi and Tortora Francesco

{c.faiella8, p.giannino, l.scovotto1, f.tortora21}@studenti.unisa.it

Feb, 2023
Department of Computer Engineering, Electrical Engineering and Applied Mathematics (DIEM), University of Salerno, Fisciano, Italy

1 Pt. 1 - R

1.1 Analisi preliminare

1.1.1 Dataset

Il dataset è composto da 80 osservazioni *(n = 80)* e di 45 regressori *(p = 45)*. La divisione che viene effettuata è la seguente: 80% del dataset sarà assegnato al training set (n di train = 64), mentre il restante 20% andrà al test set (n di test = 14).

1.1.2 Correlazione

La correlazione è una misura della relazione lineare tra due variabili quantitative.

Nel nostro caso i dati hanno una correlazione non troppo alta; ciò può essere notato dalla seguente matrice di correlazione:

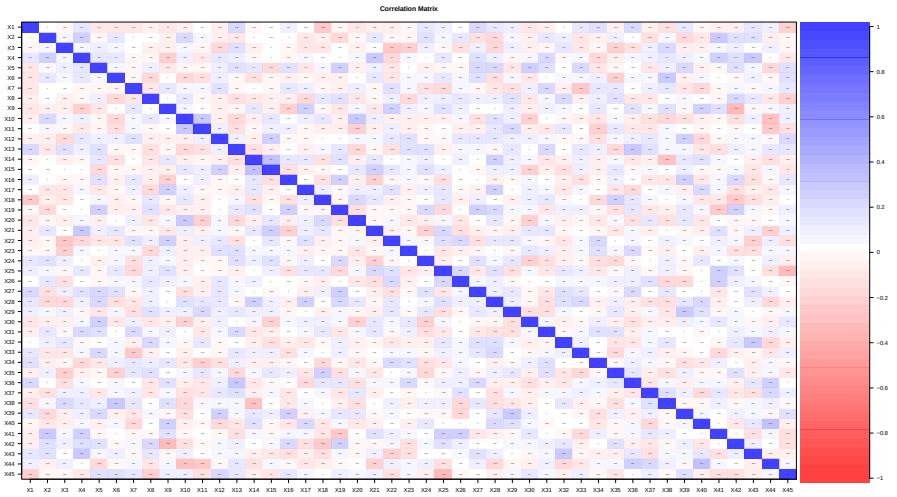


Figure 1: Matrice di Correlazione

1.1.3 Tecniche analizzate

Linear Model Applicare un modello di regressione lineare a dati ad alta dimensionalità e poca correlazione può essere problematico in quanto il modello potrebbe avere difficoltà a catturare le relazioni significative tra le variabili. In queste situazioni, è probabile che molti coefficienti siano molto piccoli o nulli, rendendo il modello meno preciso.

Best Subset Selection L'approccio BSS risulta non applicabile in contesti dove i dati sono ad alta dimensionalità; in particolare, l'algoritmo di BSS non è computazionalmente efficiente con insiemi di dati che hanno una p maggiore di 30 o 40 circa. Il nostro dataset ha 45 regressori (p=45), quindi l'analisi verrebbe effettuata su 2^{45} modelli, ciò porta a scartare il metodo.

Stepwise approach In generale, un enorme spazio di ricerca può portare a un overfitting e a un'elevata varianza delle stime dei coefficienti, per questo al BSS vengono preferiti approcci stepwise che esplorano un insieme di modelli molto più ristretto. Per questo motivo, e per ovviare ai problemi di efficienza del BSS, valutiamo alcuni approcci stepwise: possiamo andare ad utilizzare il forward selection o il backward selection, oppure un ibrido tra i due. I due metodi nel nostro caso andranno a ricercare il migliore tra un numero di modelli finito che sarà: $1 + \sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2$, essendo p=45, possiamo dire che i due approcci ricercheranno tra 991 modelli. Non è garantito però che entrambi i metodi producano il miglior modello contenente un sottoinsieme di predittori p.

Ridge La regressione Ridge è un modello di regressione lineare che introduce una penalità ℓ_2 sui coefficienti per prevenire l'overfitting. Tuttavia, se i dati hanno una alta dimensionalità con poca correlazione tra le variabili, la penalità

ℓ_2 può non essere efficace nel ridurre la complessità del modello. In questo caso, il modello potrebbe avere una penalizzazione troppo bassa e perdere informazioni importanti dai dati. Quando il λ risulta essere prossima allo 0, avremo che il modello Ridge andrà a comportarsi come il modello lineare. In queste situazioni, potrebbero essere preferibili altri modelli di regressione come la Lasso o Stepwise selection.

LASSO La regressione Lasso è un modello di regressione lineare che introduce una penalità ℓ_1 sui coefficienti, invece che una penalità ℓ_2 come nella Ridge Regression. La penalità ℓ_1 ha la proprietà di produrre coefficienti di regolarizzazione nulli per alcune variabili, effettuando quindi *variable selection*. Questo può essere utile quando si ha una alta dimensionalità con molte variabili che non sono correlate o che hanno un effetto debole sulla variabile dipendente. In questo caso, la penalità ℓ_1 può essere più adatta a selezionare solo le variabili più rilevanti, riducendo la complessità del modello e migliorandone la capacità di generalizzazione.

Elastic NET È un metodo di regressione regolarizzato che combina linearmente le penalità ℓ_1 e ℓ_2 dei metodi Lasso e Ridge, quindi fondamentalmente mette insieme caratteristiche di Lasso e di Ridge. Con i dati a nostra disposizione, possiamo ipotizzare che il modello con MSE più basso andrà a tendere verso Lasso, mentre Ridge, come abbiamo già sottolineato, potrebbe avere un termine di penalizzazione ℓ_2 troppo basso per via della composizione dei dati.

1.2 Linear Model

Una volta effettuato il calcolo del Linear Model, possiamo verificare dal summary che vi sono 3 predittori signficativi, con un p-value prossimo allo 0 (quindi <1%). Il risultato ottenuto è ciò che ci aspettavamo già in analisi preliminare (1.1.3), quindi l'MSE sul test set risulta essere più elevato rispetto alle tecniche analizzate di seguito.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.815e-01  1.205e+01  -0.040    0.969
X1          -4.762e+00  1.450e+01  -0.328    0.746
X2          -7.619e+00  1.121e+01  -0.680    0.505
X3           2.612e+00  1.400e+01   0.187    0.854
X4           4.527e+00  1.155e+01   0.392    0.700
X5          -4.871e+00  1.240e+01  -0.393    0.699
X6          -5.408e+00  1.171e+01  -0.462    0.650
X7           4.526e+00  8.679e+00   0.521    0.608
X8           1.327e+00  1.158e+01   0.115    0.910
X9          -2.087e+00  1.043e+01  -0.200    0.844
X10          5.918e+00  1.467e+01   0.403    0.691
X11         -3.462e+00  1.006e+01  -0.344    0.735
X12         -8.854e+00  1.147e+01  -0.772    0.450
X13         -7.270e+00  9.620e+00  -0.756    0.460
X14         -4.420e+00  1.651e+01  -0.268    0.792
X15          1.496e+01  1.554e+01   0.963    0.348
X16         -7.014e+00  1.239e+01  -0.566    0.578
X17          1.400e+01  1.414e+01   0.990    0.335
X18         -1.312e+01  1.398e+01  -0.938    0.360
X19         -8.638e+00  1.399e+01  -0.618    0.545
X20         -1.645e+01  1.173e+01  -1.403    0.178
X21         -1.743e-02  1.024e+01  -0.002    0.999
X22         -7.005e+00  1.277e+01  -0.548    0.590
X23          7.095e+03  1.124e+01 631.237   <2e-16 ***
X24          1.111e+04  1.325e+01 838.321   <2e-16 ***
X25          8.385e+03  1.478e+01 567.269   <2e-16 ***
X26          3.548e+00  1.311e+01   0.271    0.790
X27         -1.448e+01  1.148e+01  -1.261    0.224
X28          8.674e+00  1.094e+01   0.793    0.438
X29          7.480e+00  8.390e+00   0.892    0.384
X30          9.047e+00  9.994e+00   0.905    0.377
X31         -2.631e+00  1.131e+01  -0.233    0.819
X32         -2.162e+00  8.355e+00  -0.259    0.799
X33         -1.647e+01  1.176e+01  -1.401    0.178
X34         -4.707e+00  1.097e+01  -0.429    0.673
X35         -2.354e+00  1.204e+01  -0.196    0.847
X36          2.102e+01  1.397e+01   1.504    0.150
X37          9.495e+00  1.046e+01   0.908    0.376
X38         -4.395e+00  1.138e+01  -0.386    0.704
X39          1.579e+01  1.099e+01   1.437    0.168
X40         -1.049e+00  1.193e+01  -0.088    0.931
X41         -1.177e+01  1.443e+01  -0.816    0.425
X42         -6.545e+00  1.212e+01  -0.540    0.596
X43         -7.588e+00  1.063e+01  -0.714    0.485
X44         -1.205e+01  1.340e+01  -0.900    0.380
X45         -4.026e+00  1.347e+01  -0.299    0.768
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 52.33 on 18 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 9.004e+04 on 45 and 18 DF,  p-value: < 2.2e-16
```

Figure 2: Linear model

1.3 Stepwise approach

La selezione stepwise è un approccio per la selezione automatica di variabili in un modello statistico. In questo approccio, le variabili vengono aggiunte o eliminate dal modello in base a criteri statistici.

1.3.1 Backward selection

I risultati ottenuti in questo caso sono i più promettenti; la scelta dei regressori avviene attraverso il calcolo del minimo MSE. Oltre ad avere il minimo MSE, il modello scelto dalla backward selection riesce a prendere come migliori predittori i 3 che risultano poi essere i regressori che vanno a formare l'indizio.

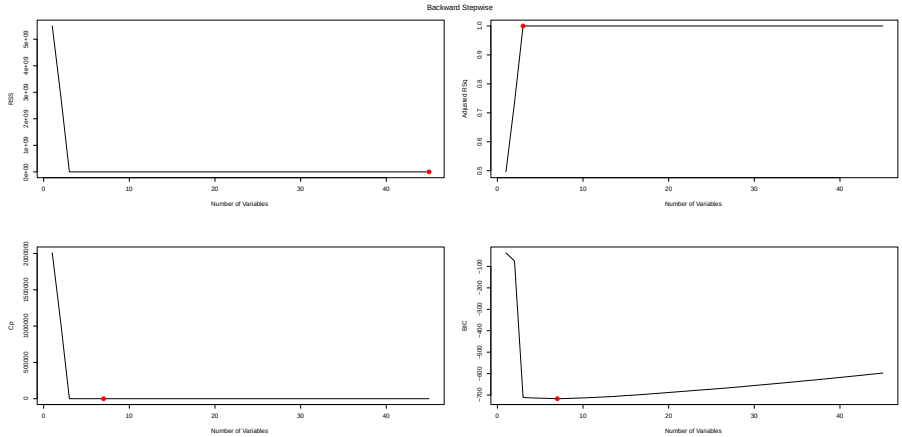


Figure 3: Valutazione Backward Stepwise Selection

Inoltre, si può notare da questi grafici che il modello con 3 regressori risulta essere ottimo anche nel caso in cui si considerino BIC, Cp, Adjusted R2 ed RSS.

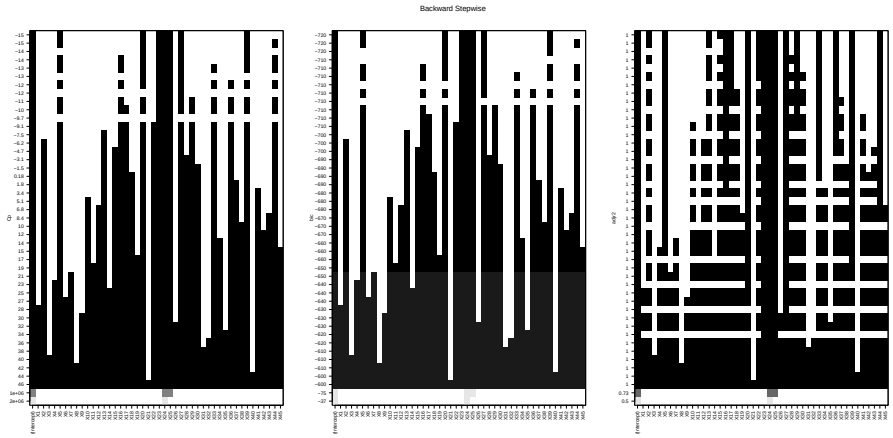


Figure 4: Scelta del numero di regressori

1.3.2 Forward selection

Nel caso forward i risultati combaciano con quanto definito con backward: l'MSE risulta basso e i regressori scelti sono quelli d'interesse. Vengono, di seguito, riportati i grafici relativi a BIC, Cp, Adjusted R2 ed RSS, così da poter mostrare le similarità con il precedente.

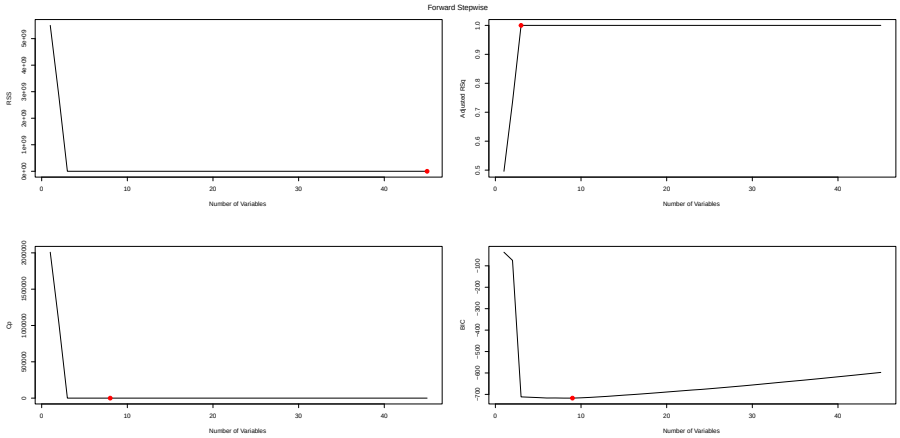


Figure 5: Valutazione Forward Stepwise Selection

La scelta del numero di regressori cade nuovamente sul 3; questo ci porta a pensare che i metodi (forward e backward selection) nel nostro caso, con questa conformazione di dati e con la grandezza limitata del dataset, hanno risultati simili e/o equivalenti, pur avendo metodi di risoluzione differenti.

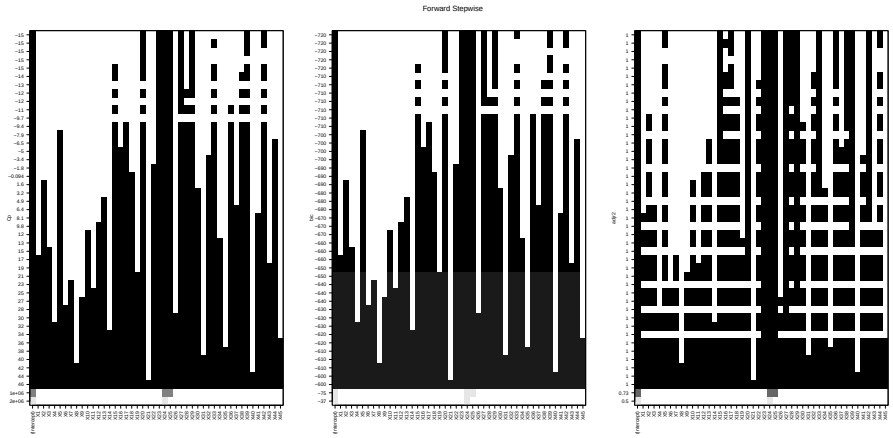


Figure 6: Scelta del numero di regressori

1.3.3 Hybrid selection

Essendo la Hybrid selection un'approccio ibrido delle già citate Backward e Forward selection, ed essendo i risultati delle due molto simili, il risultato in questo caso porta agli stessi risultati delle precedenti, infatti la hybrid selection trova il miglior MSE e determina i 3 regressori di interesse. In questo caso, l'approccio ibrido non risulta apportare miglioramenti alle predizioni effettuate dalle altre due selection, ma non va nemmeno a deteriorarle; per questo motivo e per avere un quadro generale dei tre metodi di Stepwise selection, mostriamo di seguito anche i grafici relativi alla Stepwise hybrid selection.

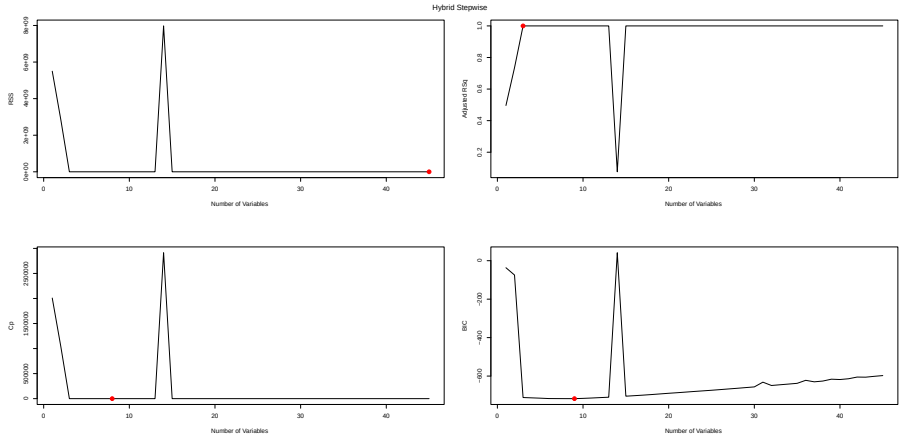


Figure 7: Valutazione Hybrid Stepwise Selection

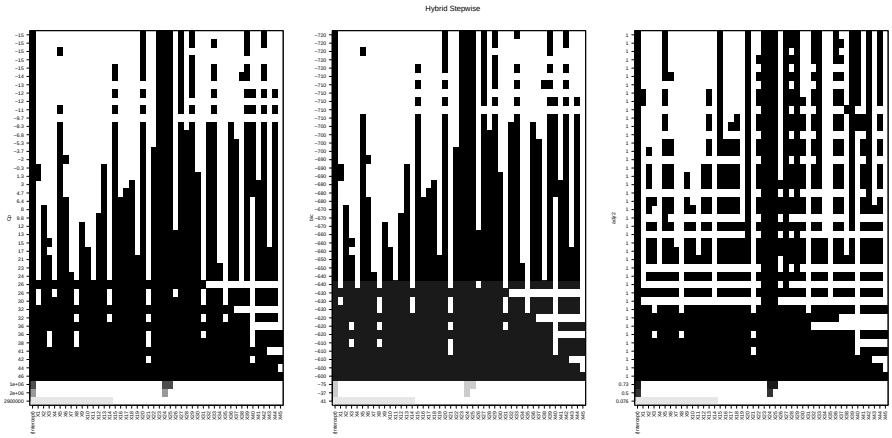


Figure 8: Scelta del numero di regressori

1.4 Metodi di Shrinkage

In questo caso si vanno a determinare modelli più robusti che saranno caratterizzati da un termine di penalizzazione (shrinkage), il quale assegnarà delle penalità ℓ_1, ℓ_2 o una combinazione di essi.

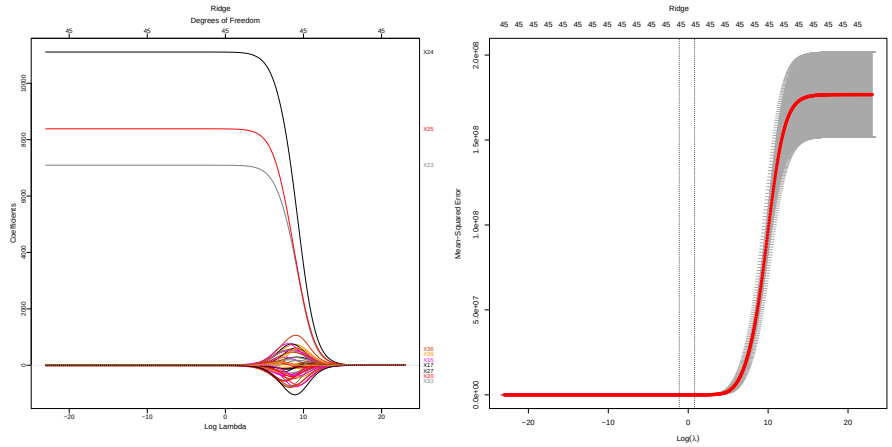
1.4.1 Ridge

Ricordiamo che il caso Ridge andrà ad utilizzare la penalizzazione ℓ_2. I risultati in questo caso sono peggiori rispetto alle altre tecniche utilizzate, avendo un MSE maggiore. Ciò è dovuto al fatto che il valore ottimo di lambda, scelto mediante la tecnica di cross validation, è molto basso, quindi il termine di penalizzazione raggiunge valori prossimi allo 0, e quindi non viene effettuata una penalizzazione sufficiente. Il comportamento di Ridge, dunque, è molto simile a quello del modello lineare. In questo caso, i regressori che avranno valori significativi saranno i 3 di nostro interesse, mentre i rimanenti saranno valori molto bassi, diversi da zero.

In conclusione, nel nostro caso, applicare Ridge non risulta essere l'approccio ottimale, in quanto definisce un termine di penalizzazione prossimo allo 0 e il suo comportamento può essere ridotto ad un modello più semplice, ovvero il modello lineare. Andare, quindi, ad utilizzare un modello più complesso per avere lo stesso risultato non risulta essere utile ed inoltre può portare ad una lettura più difficile dei risultati.

1.4.2 Lasso

Lasso, a differenza di Ridge, va ad effettuare variable selection, andando a porre a 0 i valori dei coefficienti dei regressori che non sono significativi. I risultati sono stati abbastanza superiori a Ridge: si hanno la totalità dei regressori non



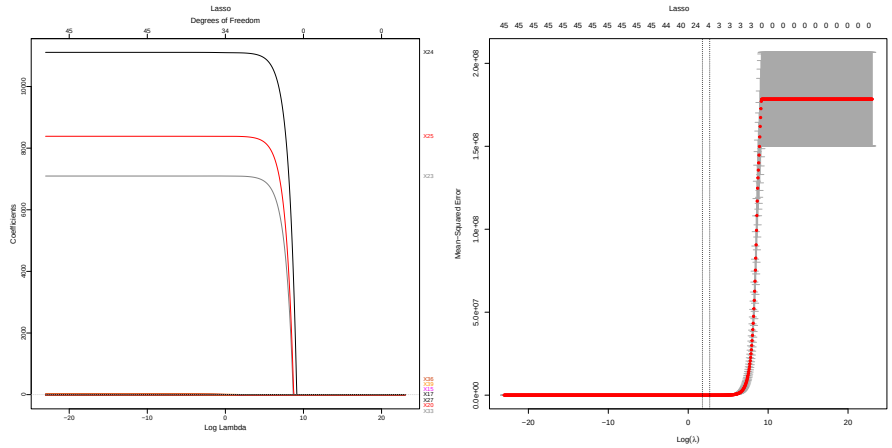
(a) Plot coefficients values/log lambda (b) Cross validation

Figure 9: Ridge

significativi uguali a 0, mentre i 3 regressori che formano l'indizio hanno valori significativamente grandi, portando quindi a un MSE molto più basso rispetto all'utilizzo di Ridge.

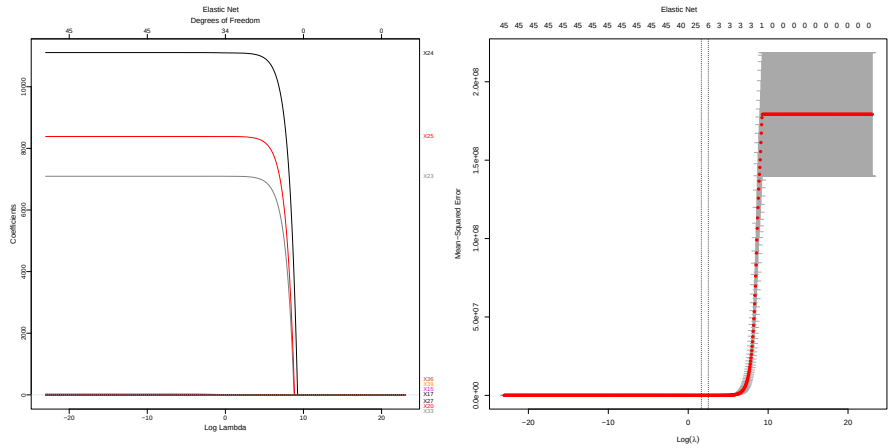
1.4.3 Elastic Net

Con Elastic Net si effettua una combinazioe di Ridge e Lasso, attraverso l'utilizzo di un valore α che pesa la percentuale di utilizzo dei due modelli. Con una $\alpha = 0$ avremo un modello completamente tendente a Ridge, mentre $\alpha = 1$ rappresenta Lasso. Effettuando varie prove, abbiamo notato che il modello migliore utlizzando Elastic Net ha una α molto tendente a 1, quindi a Lasso. L'MSE risulta poco più alto di Lasso di una quantità irrisoria: ciò ci porta a pensare che l'utilizzo del solo Lasso possa essere migliore dell'utilizzo di una combinazione tra Ridge e Lasso.



(a) Plot coefficients values/log lambda (b) Cross validation

Figure 10: Lasso



(a) Plot coefficients values/log lambda (b) Cross validation

Figure 11: Elastic Net

1.5 Conclusioni

L'MSE migliore è stato trovato tramite approcci Stepwise (con l'utilizzo di qualsiasi selection method). Questi ultimi trovano perfettamente i 3 regressori che formano la frase e il loro MSE è abbastanza minore dell'MSE risultante dagli altri approcci visti. L'approccio Lasso, sebbene abbia un MSE maggiore, risulta comunque ottimo nel nostro caso; inoltre, l'MSE trovato non si discosta enormemente dal MSE ottimo trovato e azzera tutti i regressori non significativi.

2 Pt. 2 - Python

L'obiettivo è la realizzaione del seguente sistema, realizzata sfruttando il linguaggio Pyhton: Nel primo blocco del sistema si prendono in input i dati dal

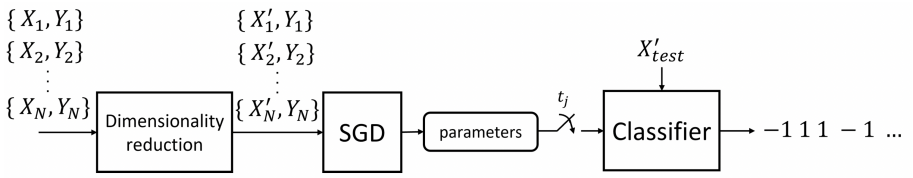


Figure 12: Sistema

dataset e si effettua una operazione di riduzione della dimensionalità applicando la PCA (Principal Component Analysis). In uscita da questo blocco abbiamo delle componenti che sono un sottoinsieme delle componenti iniziali. In seguito vengono calcolati le predizioni di $\hat{\beta}_j$. Come ultimo blocco abbiamo un classificatore, che classifica gli elementi del test set fornendo come output i valori 1 o -1.

2.1 Analisi preliminare

2.1.1 Dataset

Il dataset fornito è composto da un training set e da un test set. Il training set è composto da 24 000 righe e 21 colonne, mentre il test set è composto da 80 righe e 21 colonne dove nella 21esima sono presenti gli istanti di tempo da utilizzare per prendere i β d'interesse. L'intero dataset viene standardizzato (sottraendo la media e dividendo per la varianza), ottenendo la seguente distribuzione dei dati $X_{st} \sim \mathcal{N}(0, 1)$, per far sì che la PCA possa essere applicata correttamente.

2.1.2 Correlazione

Per valutare la correlazione tra i dati di training, viene fornita la relativa matrice:

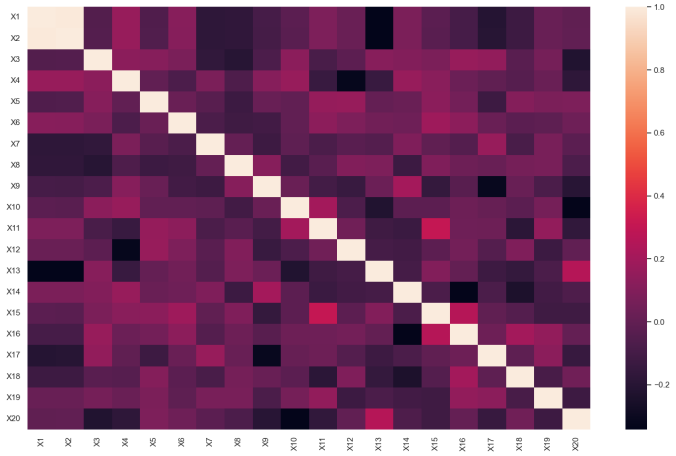


Figure 13: Matrice di correlazione sui dati scalati

Dall'immagine notiamo un'alta correlazione tra *X1* e *X2*, mentre una correlazione relativamente bassa caratterizza gli altri regressori.

2.2 Riduzione della dimensionalità - PCA

Valutata la correlazione tra i dati, andiamo ad effettuare una PCA senza ridurre la dimensionalità: in questo modo avremo i regressori non correlati tra loro e quindi una rappresentazione differente. A titolo di prova, di seguito vi è la matrice di correlazione calcolata dopo la PCA, mantenendo tutte le componenti (figura 14).

A questo punto, possiamo valutare il seguente grafico a barre (figura 15) ed andare a scegliere il numero di Principal Component da considerare così da evitare di perdere informazioni importanti.
Per mantenere una percentuale adeguata, si è deciso di prendere 16 componenti principali in modo da conservare una percentuale di poco inferiore al 95%, per l'esattezza pari a 94.835%, della varianza cumulativa: in questo modo assicuriamo una rappresentazione dei dati affidabile. Possiamo inoltre notare che la componente principale 20 risulta dare un apporto significativamente basso alla varianza comulativa, in quanto, la sua varianza è pari a 0.000462. Questo valore di varianza, come si evince dalla figura 13, è relativamente basso rispetto agli altri valori dando un contributo quasi nullo alla varianza comulativa. Per l'ottenimento del risultato finale anche con una varianza comulativa di circa il 90%, ottenuta andando a prendere le prime 14 componenti principali, ci permette di ottenere lo stesso risultato a minore dimensionalità.

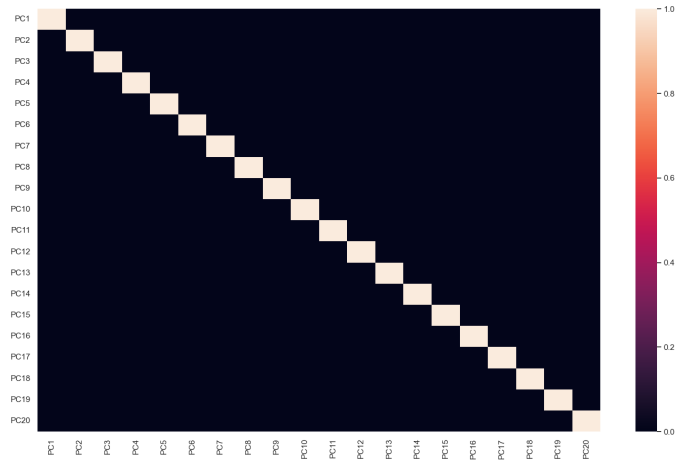


Figure 14: Matrice di correlazione successiva alla PCA

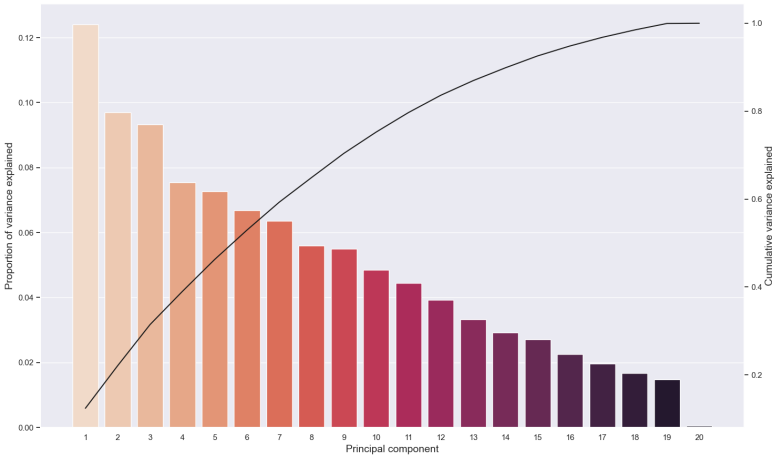


Figure 15: Varianza spiegata da ogni componente principale

2.3 Training - SDG

L'algoritmo del gradiente stocastico viene utilizzato per aggiornare i pesi del nostro modello. La scelta della funzione di loss è stata presa in quanto, secondo le specifiche, doveva essere addestrato un classificatore binario e la loss più comune in questi casi è la *logistic loss*. Come parametri di addestramento abbiamo settato un learning rate pari a 0,001 e un numero di epoche pari a 50. In aggiunta, quest'ultima si adatta meglio alle esigenze del problema, in quanto il nostro output dovrebbe contemplare i soli valori 1 e -1, quindi $Y \in \{-1, 1\}$, a differenza di altre loss inizialmente considerate (es. MSE), che invece considerano l'insieme $\mathbb{R}$. L'andamento trovato con l'utilizzo di questo algoritmo è stato il seguente:

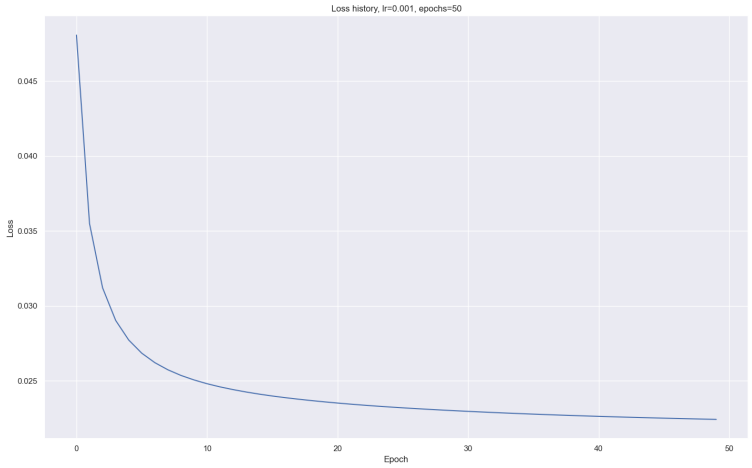


Figure 16: Valore della loss per ogni epoca

2.4 Classificatore

Il classificatore esegue la seguente operazione:

$$\begin{cases} 1 \text{ se } X'_{test}(j)\hat{\beta}(t_j) > 0 \\ -1 \text{ altrimenti} \end{cases}$$

Con $X'_{test}(j)$ si indica la feature a ridotta dimensionalità, con t_j l'instante di tempo e con $\hat{\beta}(t_j)$ il parametro stimato dall'algoritmo del gradiente stocatico al tempo t_j.

2.5 Conversione in ASCII

Infine, dato l'output del classificatore, tutti i valori -1 vengono posti a 0, e si prosegue alla conversione in ASCII.

3 Risoluzione enigma

Gli indizi trovati sono i seguenti:

- Indizio 1: GoT
- Indizio 2: WntrComing

Quindi la frase originale è: The winter is coming, della celeberrima serie TV "Game of Thrones". La scena: https://www.youtube.com/watch?v=5Y6TqxLmxIo