UNIVERSITÀ DI PISA

Corso di laurea magistrale in Artificial Intelligence and data engineering

Statistica

Progetto 2

Giuseppe Martino

13 dicembre 2020

Indice

1 Introduzione al problema

L'obiettivo di questa analisi prende spunto dal paper "Atypical Estuaries in NSW: Implications for management of Lake Wollumboola"[1], che dimostra l'esistenza in Australia di una tipologia di foce di corsi d'acqua la quale differisce da tutte le altre tipologie chimicamente e biologicamente. Questo tipo di foce è chiamato "Back Dune Lagoon" ed è importante distinguerlo da tutti gli altri, perchè molto più sensibile ai cambiamenti ambientali e questo aumenta il rischio di comprometterne il valore ecologico. Con questa analisi quindi, si cerca di fornire un modello di classificazione agli enti amministratori di territori in cui sono presenti foci di fiumi,laghi o altri corsi d'acqua.Tale modello può essere utile per individuare foci di tipo BDL, in modo da adottare un approccio precauzionale nella valutazione di proposte di sviluppo dei bacini di questi corsi d'acqua.

2 Dati

Il dataset utilizzato, disponibile al seguente link:

https://datasets.seed.nsw.gov.au/dataset/nsw-estuary-temperature-ph-and-salinity-data/resource/16cff0b8-b490-40d4-91b9-d9fd91801692

comprende dati di foci situate nello stato australiano del New South Wales, raccolti in 12 anni come parte del programma "estuary health Monitoring, Evaluating and Reporting".

Dopo aver eliminato attributi non numerici,non rilevanti per l'analisi, fattori percentuali e osservazioni con valori mancanti, il dataset risulta così strutturato:

Numero di osservazioni:2968
Numero di fattori: 23

- Estuary_type: Tipo di foce
- Disturbance_class: Grado di alterazione (molto basso, basso, medio, alto o molto alto)
- Summer: Anno di rilevazione
- Temp: temperatura in °C
- pH
- Salinity: salinità (PSU)
- Turbidity: torbidità (NTU)
- Days_Elapsed: periodo di campionamento in giorni
- Latitude: Latitudine
- Longitude: Longitudine
- Seagrass_area: area coperta da alghe (km^2)
- Mangrove_area: area coperta da mangrovie (km^2)
- Total_estuary_area: area totale della foce (m^2)

- Estuary_volume: volume foce (mL)
- total_sa_vol: rapporto tra area della superficie della foce e volume
- Total_area_saltmarsh: area palude salmastra (km^2)
- Average_depth: profondità media (m)
- Perimeter: perimetro (km)
- openwater_SA.volume: volume acque aperte
- Total_flush_time: tempo (in giorni) di ricambio dell'acqua
- Retention_factor: rapporto tra volume totale della foce e il volume di deflusso
- Propn_increase_N: aumenti proporzionali dei carichi di azoto
- Catchment_area: area del bacino (km^2)

L'attributo "Disturbance_class" è stato trasformato in numerico, assegnando ai valori un numero da 1 a 5, dove 1 è "molto basso" e 5 "molto alto".

Al fine dello scopo dell'analisi di rilevare le foci di tipo BDL, ho modificato l'attributo "Estuary_type", assegnando il valore 1 quando il tipo di foce è BDL, 0 altrimenti. Il problema diventa quindi un problema di classificazione binaria, in cui lo scopo è quello di distinguere foci di tipo BDL da quelle che non lo sono.

3 Analisi dei dati

3.1 Analisi Discriminante Lineare

Provando ad eseguire una classificazione mediante l'analisi discriminante lineare, il software lancia un'eccezione, evidenziando la presenza di collinearità tra alcune variabili.

Ho quindi cercato di risolvere il problema controllando le correlazioni tra gli attributi, per cercare attributi estremamente correlati.

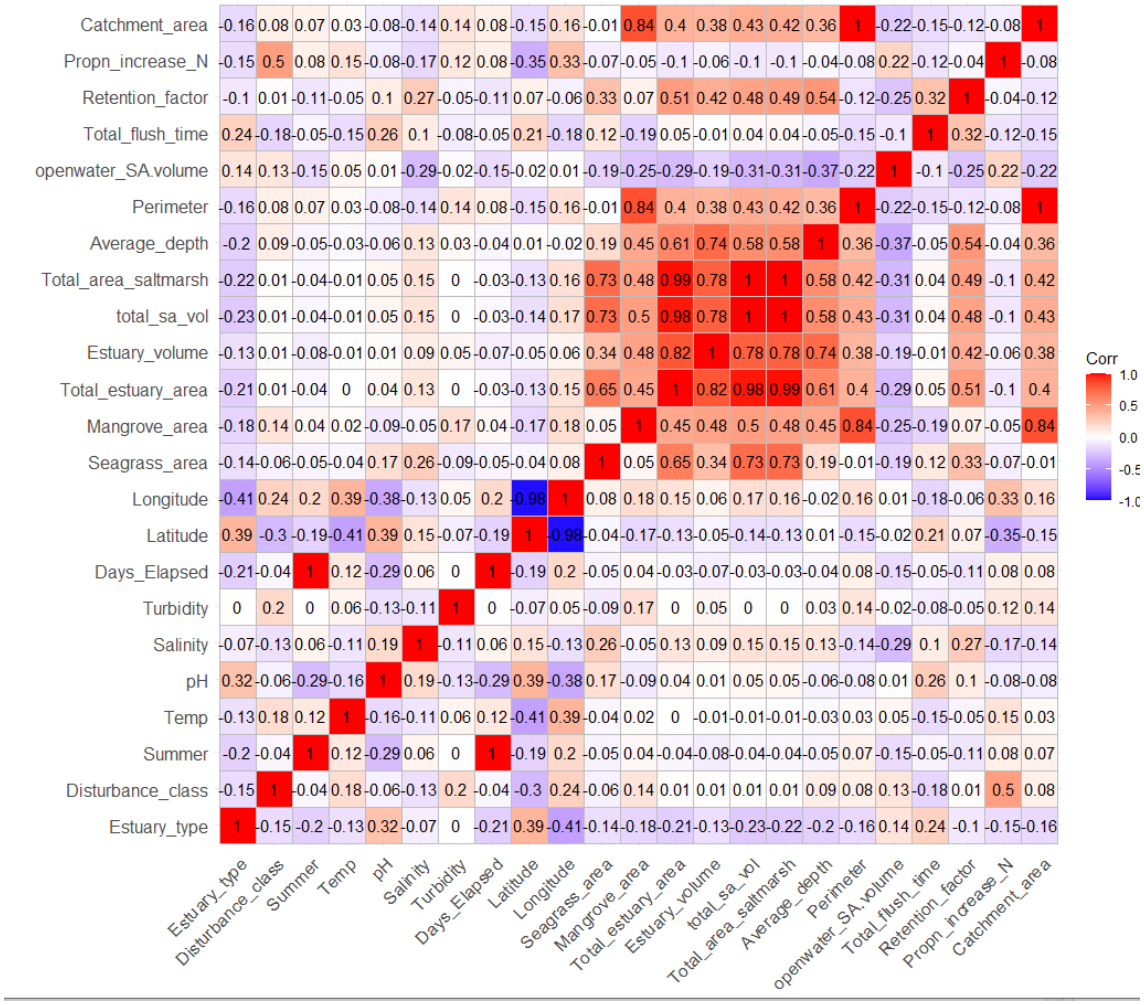

Figura 1: Correlazioni tra gli attributi

Effettivamente notiamo una correlazione perfetta (uguale a 1) tra i fattori: "Catchment_area" e "Perimeter", "Days_Elapsed" e "Summer", "Total_sa_vol" e "Total_area_saltmarsh", per cui, ho eliminato i fattori "Catchment_area", "Summer" e "Total_area_saltmarsh".

L'ipotesi di collinearità di questi fattori è stata verificata anche controllando le variazioni dei p-value dei coefficienti, applicando un modello lineare ed eliminando passo passo i fattori. Il p-value di "Days_Elapsed" è infatti calato drasticamente dopo aver eliminato l'attributo "Summer", mentre per quanto riguarda il fattore "Catchment_area", il suo p-value non è stato proprio definito dal software, in quanto è stata intercettata chiaramente la sua dipendenza lineare.

```
> estu.lda=lda(Estuary_type~.,data = estu,CV=F)
> estu.lda.p=predict(estu.lda)
> estu.lda.post=estu.lda.p$posterior[,2]
>
> #accuratezza
> sum((estu.lda.post>0.5)==(estu$Estuary_type>0.5))/length(estu$Estuary_type)
[1] 0.8908356
>
> #matrice di confusione
> s2_confusion(estu$Estuary_type,estu.lda.post)
            actual 1 actual 0
predicted 1      243      106
predicted 0      218     2401
>
> #area sotto la curva
> estu.lda.roc=s2_roc(estu$Estuary_type,estu.lda.post)
> s2_auc(estu.lda.roc)
[1] 0.9339893
```

Figura 2: Accuratezza e matrice di confusione LDA

Con l'analisi discriminante lineare, si ottiene una accuratezza dell'89%, e l'area sotto la curva ROC, che ci da un'indicazione numerica della capacità di distinguere le due classi, è 0,93, ma a dispetto di una accuratezza non bassa, il modello sbaglia a predire il 47,28% delle foci di tipo BDL e ai fini dell'analisi questo non può essere un risultato soddisfacente.

3.2 Analisi Discriminante Quadratica

```
> estu.qda=qda(Estuary_type~.,data=estu,CV=F)
> estu.qda.p=predict(estu.qda)
> estu.qda.post=estu.qda.p$posterior[,2]
>
> #accuratezza qda
> sum((estu.qda.post>0.5)==(estu$Estuary_type>0.5))/length(estu$Estuary_type)
[1] 0.9053235
>
> #matrice di confusione qda
> s2_confusion(estu$Estuary_type,estu.qda.post)
            actual 1 actual 0
predicted 1      461      281
predicted 0        0     2226
>
> #area sotto la curva
> estu.qda.roc=s2_roc(estu$Estuary_type,estu.qda.post)
> s2_auc(estu.qda.roc)
[1] 0.9533814
```

Figura 3: Accuratezza e matrice di confusione QDA

L'accuratezza del modello è del 90,5% e anche se è superiore di poco rispetto a quella ottenuta con l'analisi discriminante lineare, il risultato è nettamente migliore, perchè dall'analisi della matrice di confusione possiamo vedere che il modello sbaglia soltanto a predire le foci non BDL, mentre riesce a predire perfettamente le foci di tipo BDL. Ai fini del problema è quindi un ottimo modello perchè possiamo anche accettare un certo errore nel predire le foci non BDL se però il modello non sbaglia mai a predire quelle BDL. L'area sotto la curva è 0.95, di poco superiore a quella ottenuta con l'analisi discriminante lineare.

3.3 Regressione Logistica

```
> #accuratezza
> sum((estu.glm.p>0.5)==(estu$Estuary_type>0.5))/2968
[1] 0.9316038
>
> #matrice di confusione
> s2_confusion(estu$Estuary_type,estu.glm.p)
            actual 1 actual 0
predicted 1      343       85
predicted 0      118     2422
>
> #curva ROC
> glm.roc<-s2_roc(estu$Estuary_type,estu.glm.p)
> s2_auc(glm.roc) #area sotto la curva
[1] 0.9681517
```

Figura 4: Accuratezza e matrice di confusione regressione logistica

Con la regressione logistica si ottiene l'accuratezza più alta e il valore più alto dell'area sotto la curva ROC, rispettivamente 93,26% e 0.96. Nonostante ciò, il modello comunque sbaglia nel predire le osservazioni di tipo BDL nel 25,6% dei casi.

4 Confronto tra i modelli e autovalutazione

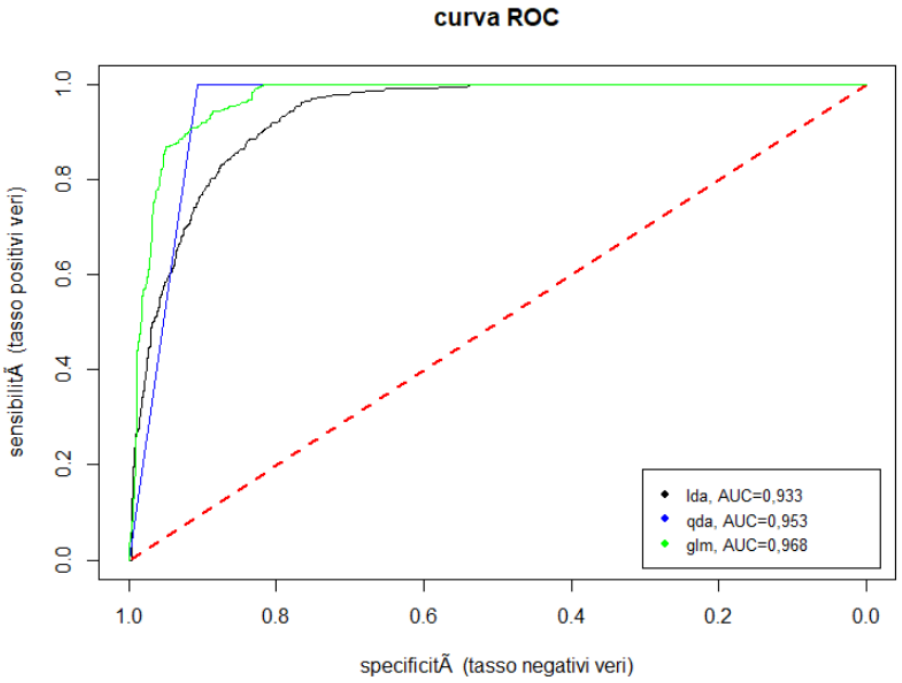

Figura 5: Confronto curve ROC modelli

Considerando i dati ottenuti su accuratezze, matrici di confusione e curve ROC, possiamo affermare che i modelli ottenuti mediante regressione logistica e analisi discriminante

quadratica sono migliori rispetto al modello della LDA, almeno sui dati usati per calibrare i modelli. In particolare, quello ottenuto mediante regressione logistica ha migliore accuratezza e valore dell'area sotto la curva ROC più alto ma, anche se vengono predette più osservazioni in modo corretto, la maggior parte delle predizioni sbagliate riguardano proprio quelle sui tipi di foci BDL, per cui il modello dell'analisi discriminante quadratica sarebbe quello da preferire.

Per capire però quanto i modelli siano buoni a prevedere nuovi risultati e quindi capire quale modello sia migliore nella predizione, è necessaria un'autovalutazione. Sono stati tolti dati dalla tabella iniziale, sono stati calibrati i modelli sui soli dati rimasti e sono state calcolate le risposte dei modelli sui nuovi dati. L'esperimento è stato ripetuto più volte (500) su campioni casuali per avere un risultato statisticamente significativo.

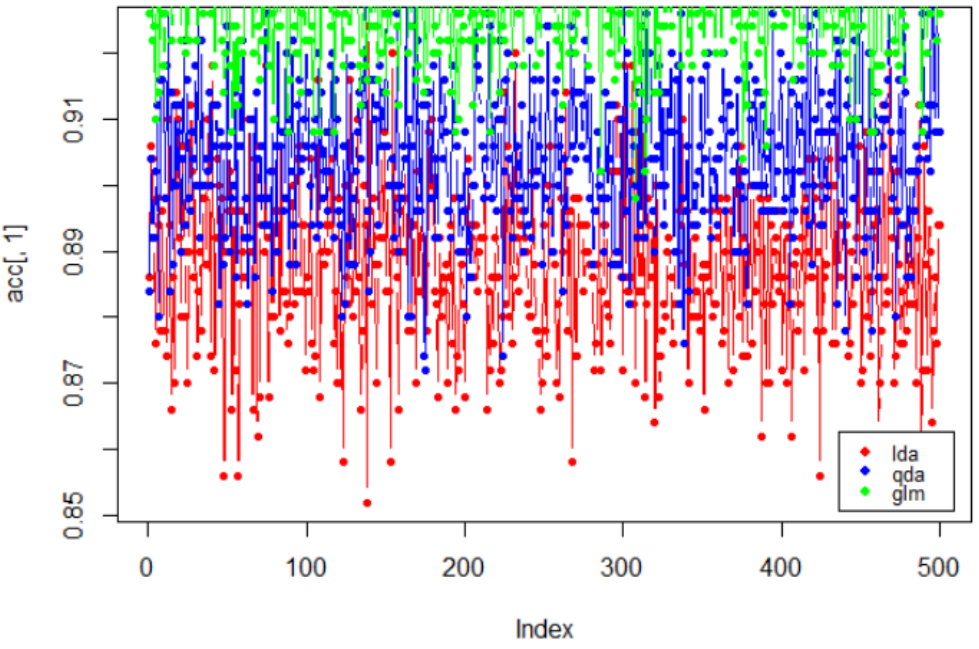

Figura 6: Accuratezze dei modelli su dati nuovi

```
> #lda
> mean(acc[,1])
[1] 0.887916
> sd(acc[,1])
[1] 0.01275768
> #qda
> mean(acc[,2])
[1] 0.904616
> sd(acc[,2])
[1] 0.01213636
> #glm
> mean(acc[,3])
[1] 0.929304
> sd(acc[,3])
[1] 0.0104179
```

Figura 7

5 Conclusioni

L'autovalutazione conferma che tutti i modelli hanno una buona capacità di predizione, infatti la loro accuratezza su dati nuovi non è significativamente minore di quella sui dati su cui erano stati calibrati i modelli, con il modello calcolato tramite regressione logistica che, con una accuratezza media del 92,9%, riesce a predire meglio degli altri in generale. Tuttavia il modello da preferire è quello calcolato con l'analisi discriminante quadratica, dato che, come visto dalla matrice di confusione, è molto preciso nel prevedere foci di tipo BDL, come da obiettivo dell'analisi.

Riferimenti bibliografici

[1] Jaimie Potts Dr Peter Scanes Dr Angus Ferguson. *Atypical Estuaries in NSW: Implications for management of Lake Wollumboola*. URL: https://www.coastalconference.com/2014/papers2014/Peter%20Scanes%20Full%20Paper.pdf (cit. a p. 2).