Challenge 0 ML

Matteo Vicenzino

November 2024

Dataset e introduzione

Nel seguente report si vuole analizzare il dataset `50_startups.csv` con diversi algoritmi di Regularized Logistic Regression e comparare i risultati ottenuti.

Il dataset contiene un totale di 50 osservazioni e 5 variabili, di cui 4 variabili quantitative che rappresentano la spesa in in R&D, Amministrazione, Marketing e Profit, E una variabile categoriale che indica lo stato in cui la startup si trova. Alle variabili quantitative è stata apportata una normalizzazione, dopo aver tenuto solo le osservazioni riguardanti startup dello stato della Florida e della California.

In primo luogo, per analizzare le potenziali relazioni tra le features si è deciso di stampare la matrice di correlazione (Figura 1) tra ogni coppia di variabili quantitative, ed è emersa una forte correlazione positiva tra quasi tutte le variabili, tranne che con la variabile Administration in cui la correlazione è più vicina allo zero.

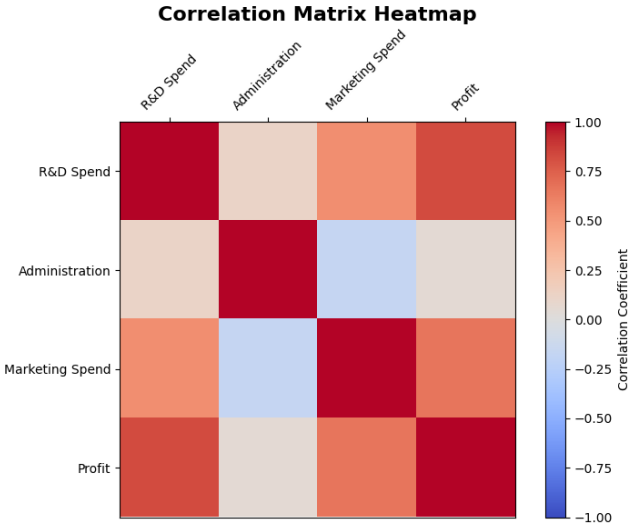

Figure 1:

Regressione Logistica senza regloarizzazione

Il primo modello creato è una regressione logistica standard in cui stampando la matrice di confusione (Figura 2), per confrontare le predizioni ottenute con la ground truth, si nota una bassa accuratezza nel predirre la variabile State (circa 33%). Questo anche per la dimensione del dataset che conta solo 33 osservazioni.

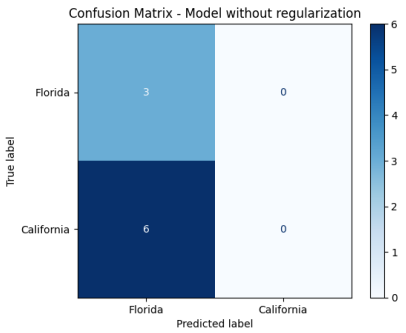

Figure 2:

Anche la curva ROC del modello (Figura 4) indica una scarsa performance nel riconoscere correttamente le labels, infatti la curva si trova sotto la diagonale, indicando una performance peggiore del modello casuale.

Regolarizzazione Ridge, Lasso ed Elastic Net

Si sono poi implementati i modelli di regressione logistica regloarizzata attraverso l'algortimo di Gradient Descent con i seguenti parametri: iterazioni = 1000, learning rate = 0.001, lamda = 0.01 e alfa = 0.5 dove lamda è il parametro di regolarizzazione e alfa è il mixing factor di L2 e L1. Per ogni modello viene visualizzato la funzione di Loss, per ogni iterazione di gradient descent (Figura 3).

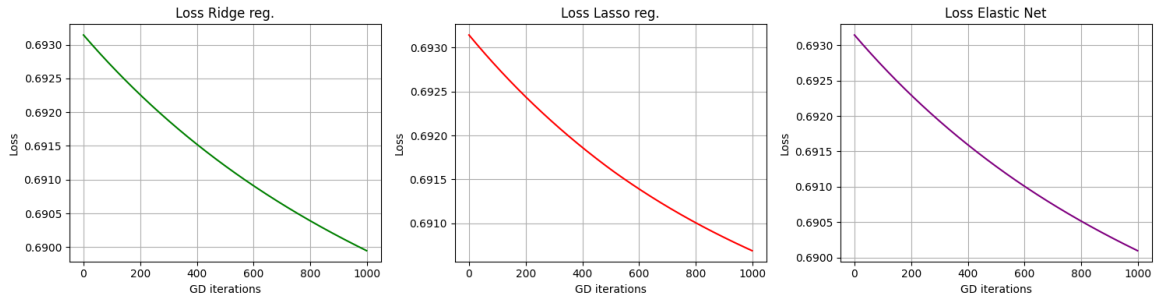

Figure 3:

Curva ROC

Inoltre si stampa la Curva ROC (Figura 4) per valutare le prestazioni dei diversi modelli di classificazione binaria in termini di compromesso tra tasso di veri positivi (True Positive Rate, TPR) e tasso di falsi positivi (False Positive Rate, FPR).

Dalla figura si nota che la curva ROC che si avvicina maggiormente al punto in alto a sinistra è quella relativa al modello con la regolarizzazione Ridge, che sembra avere le performance migliori, molto simile è anche la curva ROC relativa al modello con la regolarizzazione Lasso.

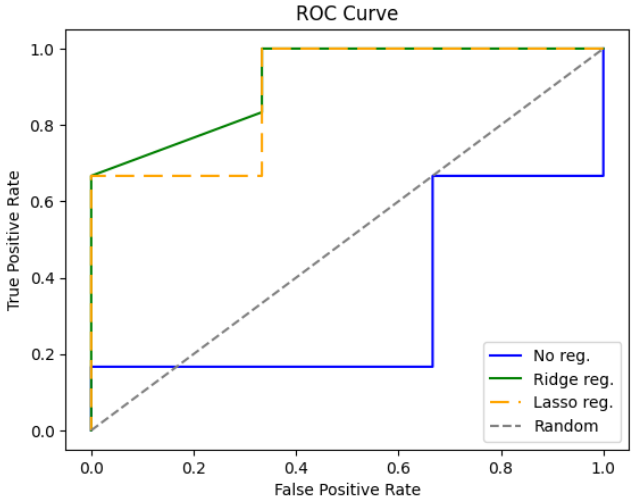

Figure 4:

In generale si nota che la regolarizzazione migliora i tassi di TPR e FPR per certi valori di threshold , e andando a calcolare il valore per il quale il modello ha un'accuratezza migliore, si trova un threshold di circa `0.4805`.

Se si va a predirre le labels del dataset utilizzando un threshold di `0.4805` si ottiene un'accuratezza dell'`88,9%` e la matrice di correlazione mostra che tutte le osservazioni con label "California" sono state predette correttamente, mentre delle osservazioni con label "Florida" sono state assegnate correttamente il `67%` delle volte (Figura 5).

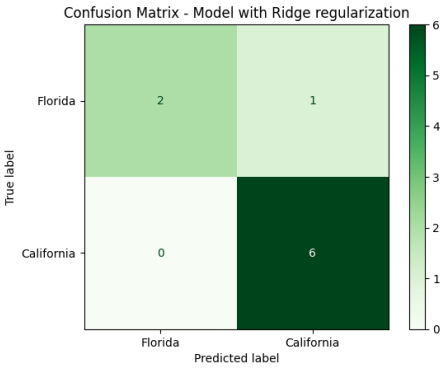

Figure 5:

Conclusioni: Anche malgrado la ridotta dimensione del dataset si nota come introducendo una regolarizzazione, questa possa in parte migliorare le prestazioni e l'accuratezza, e riducendo il rischio di overfitting, più pronunciato in dataset di piccole dimensioni.