Bortolussi Lorenzo SM3201370 18/11/2024

CHALLENGE 0 – REPORT

Il presente report contiene un resoconto di un progetto di machine learning, nel quale si utilizza la regressione logistica per affrontare un problema di classificazione binaria su un dataset, contenente informazioni relative a 50 startup americane. Per eventuali riferimenti al codice si faccia riferimento al notebook relativo.

Preparazione dei dati

Il dataset è stato importato e lavorato attraverso funzioni di pandas, ad esempio sostituendo i **valori nulli** con la **media** dei valori che la stessa feature assume attraverso tutto il dataset.

Già da subito si può notare la bassa quantità di dati a disposizione, il che probabilmente porterà a risultati non troppo buoni nell'allenamento dei modelli

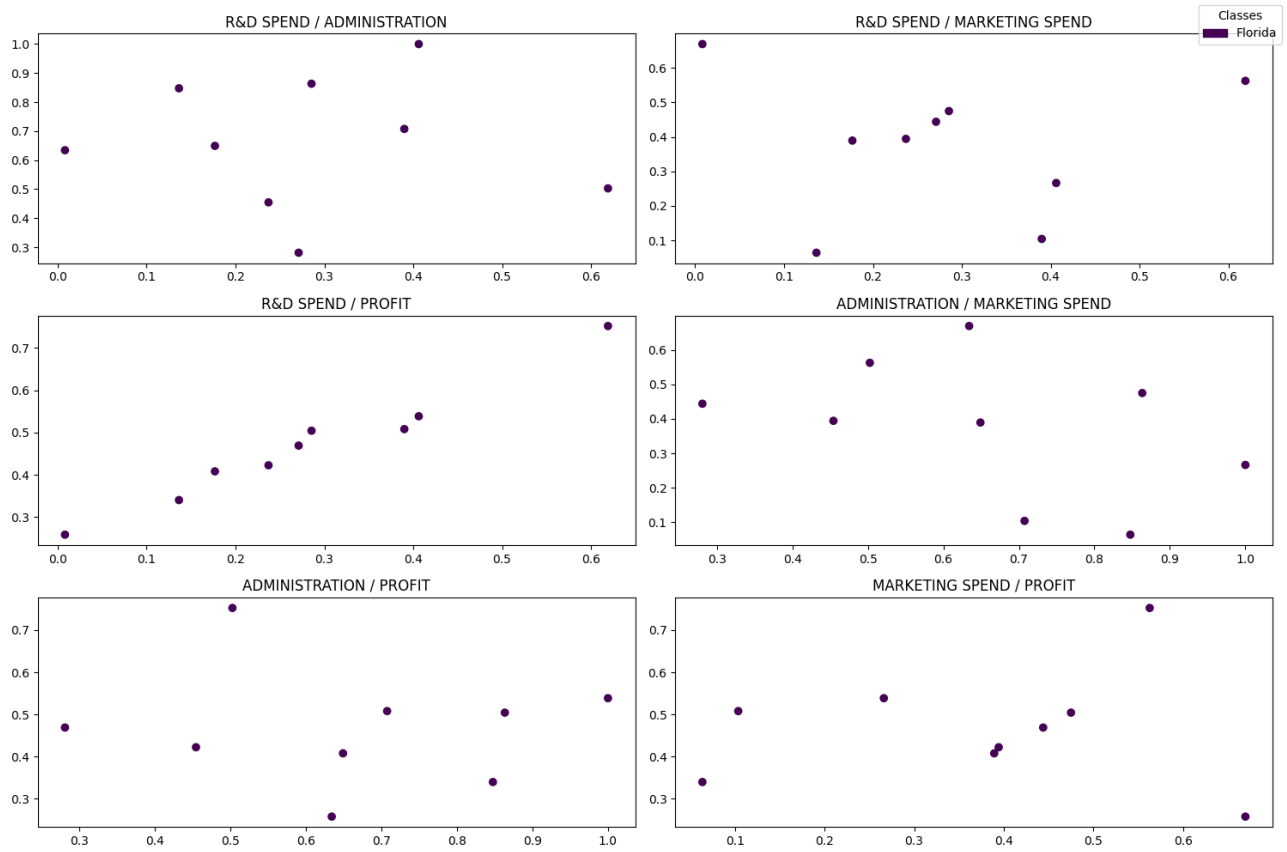
Per proseguire con la classificazione binaria sono stati selezionati gli stati di Florida e California. Il dataset è stato modificato rimuovendo la colonna contenente gli stati e inserendo i due stati selezionati tramite **One Hot Encoding**, mantenendo poi una singola colonna come **variabile target**: è stata scelta la colonna contenente '1' per i dati relativi alla California, '0' per i dati relativi alla Florida.

I valori delle altre categorie (ad esempio Profit, R&D spend …) sono stati **normalizzati** nell'intervallo [0,1] dividendo per il valore massimo all'interno delle singole categorie, per uniformare le scale delle variabili ed evitare penalizzazioni scorrette nella regolarizzazione nei passaggi successivi.

Classificazione

In un primo momento è stata utilizzata la libreria scikitlearn per dividere i dati a disposizione in maniera semi-randomica in dati di training (75%) e dati di test (25%), per poi utilizzare un modello di classificazione binaria tramite regressione logistica disponibile nella libreria. Come si sospettava già nella preparazione dei dati, questo primo modello performa piuttosto male classificando tutti i punti nella stessa categoria. Questa classificazione piatta, assieme alla bassa accuracy, fa pensare che il modello non sia riuscito a imparare a classificare la variabile da noi desiderata attraverso i soli dati che abbiamo messo a sua disposizione.

Di seguito si può vedere come i punti del test set si distribuiscono nello spazio, considerandoli a coppie. Benché alcuni grafici mostrino relazioni lineari tra coppie di feature, limitandoci a due dimensioni non è possibile riconoscere iperpiani di separazione tra i nostri punti.



Dati presenti nel test set e plottati a coppie

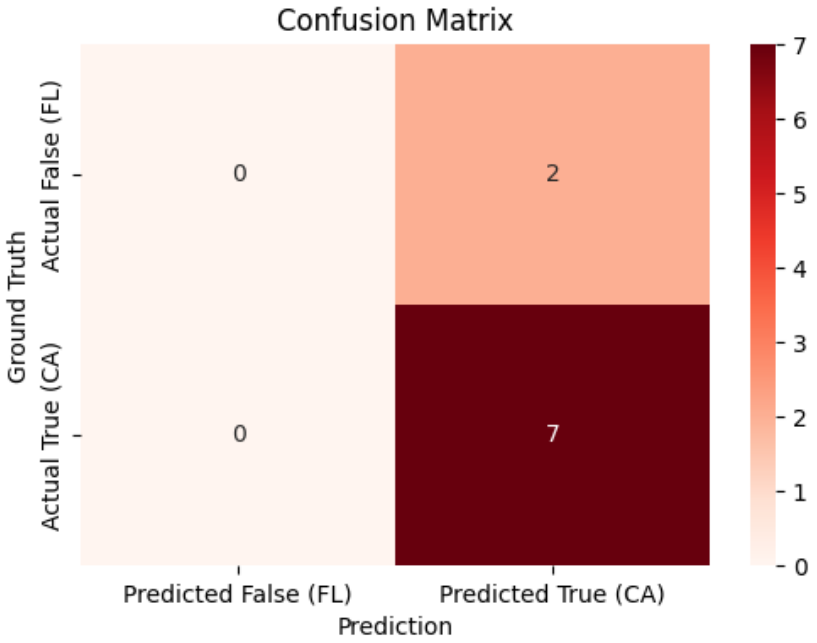
A questo punto è stata implementata da zero la regressione logistica, associando il modello di base a tre tipi di normalizzazione: Ridge, LASSO, Elastic Net.

Questi tre nuovi modelli sono stati allenati con gli stessi dati, per provare a imparare a classificare i punti del test set, ottenendo in tutti e tre i casi gli stessi pesi e lo stesso valore della loss alla fine dell'allenamento.

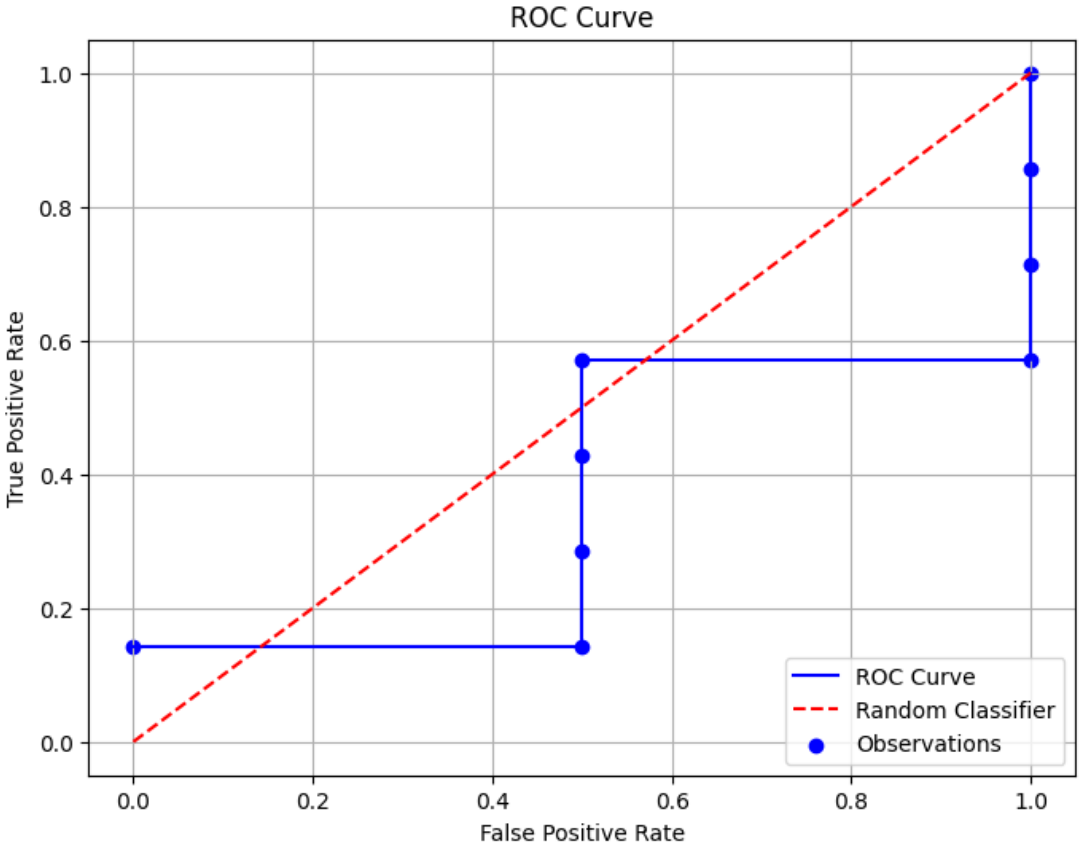
Valutazione del modello

Per valutare la qualità di questi modelli, sono state usate alcune delle metriche più comuni ottenendo i seguenti risultati, identici per tutti e tre i modelli considerati:

	precision	recall	f1-score	support
Florida	0.00	0.00	0.00	2
California	0.78	1.00	0.88	7
Accuracy :	0.78			



Dalla confusion matrix si nota come questi tre modelli, fissata una **treshold** di **0.5** , classificano tutti i punti del test set come appartenenti alla categoria 1, ovvero la California. Diventa quindi di interesse valutare il modello al variare della treshold utilizzando una ROC curve. Di seguito si mostra i risultati ottenuti sul modello utilizzante Elastic Net come regolarizzazione.



Si nota subito la scarsa qualità del risultato ottenuto, ma ciò non sorprende considerando ad esempio la bassa quantità di dati utilizzata per allenare il modello e soprattutto per il test dello stesso.