Trieste 20/11/2024

# CHALLENGE 0

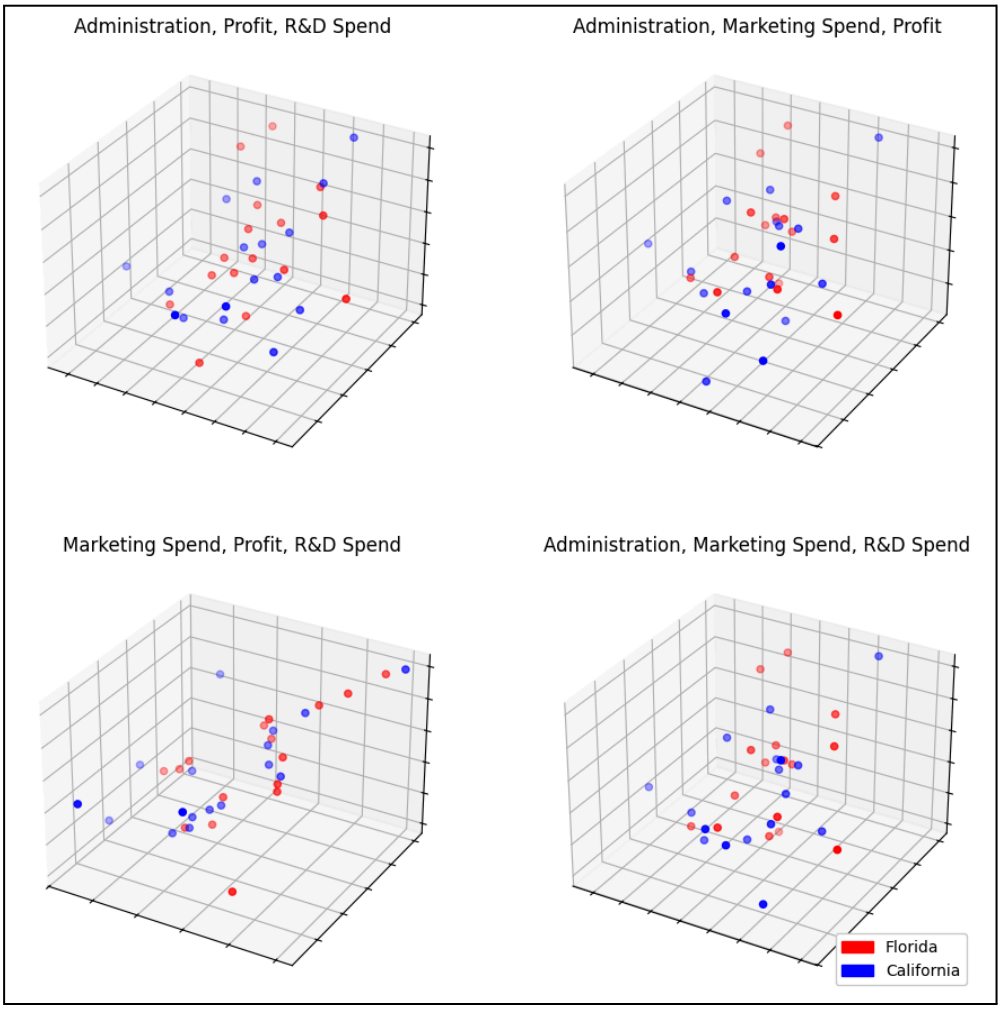
Bredariol Francesco

## INTRODUZIONE

In questa nostra prima Challenge ci è stato chiesto di andare ad eseguire una regressione logistica su un dataset contenente dati relativi a startup americane. La regressione logistica aveva come target la classe *State* che indicava lo stato geografico di una determinata startup. Il dataset è stato semplificato riducendo il dominio della classe *State* da 3 valori a 2 valori. D'ora in avanti ci riferiremo al dataset intendendo i dati filtrati (senza la presenza di *New York* tra le variabili target) e normalizzati.

## DESCRIZIONE DEL DATASET

Il primissimo approccio al problema è stato quello di analizzare, per quanto possibile, la composizione del dataset. Ho quindi eseguito sia i pair plot che i triplet plot delle varie componenti evidenziando le variabili target con colori diversi. Allego i grafici ottenuti dei triplet plot poiché più significativi. I titoli dei grafici indicano, in ordine: la componente lungo l'asse x, la componente lungo l'asse y, la componente lungo l'asse z. Si noti come nel triplet plot in basso a sx (Marketing Spend, Profit, R&D Spend) sembra apparire una relazione lineare tra i dati.



# DETTAGLI TECNICI

Qui riporto in maniera tabulare i parametri da me usati nei vari tipi di modelli di regressione logistica di cui mostrerò i risultati. Da qui in avanti i modelli prenderanno il nome dal tipo di regolarizzazione che sfruttano (il modello senza regolarizzazione sarà il *Plain Model*). Poiché ogni modello può essere visto come un modello di Elastic Net con valori particolari, la tabella sfrutterà questa cosa per compattare la descrizione dei parametri. In particolare : *Iter* è il numero massimo di iterazioni, *Gamma* è il learning rate, *Lambda* è la penalità associate alla regolarizzazione, *Teta* è la penalità associata alla regolarizzazione L2, (*1-Teta*) è la penalità associata alla regolarizzazione L1. La *threshold* per la classificazione è fissata a 0,5.

| ***MODELLO*** | ***ITER*** | ***GAMMA*** | ***LAMBDA*** | ***TETA*** |
|---|---|---|---|---|
| *Plain* | 500 | 0.0001 | 0 | 0 |
| *Lasso* | 500 | 0.0001 | 5 | 0 |
| *Ridge* | 500 | 0.0001 | 5 | 1 |
| *Elastic* | 500 | 0.0001 | 2 | 0.5 |

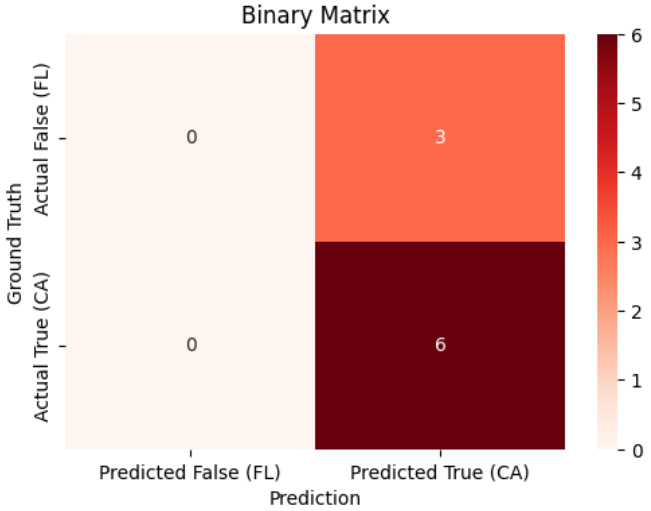
# RISULTATI

I principali risultati sono riportati tramite metrica di *accuracy* e tramite plot di *Confusion Matrix* e *Roc Curve*.

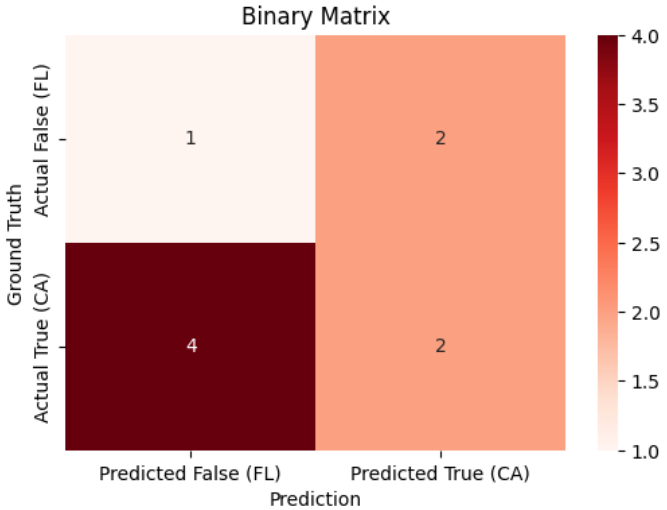
## ACCURACY

| ***MODELLO*** | ***ACCURACY*** |
|---|---|
| *Plain* | 0.66 |
| *Lasso* | 0.66 |
| *Ridge* | 0.33 |
| *Elastic* | 0.33 |

## CONFUSION MATRIX

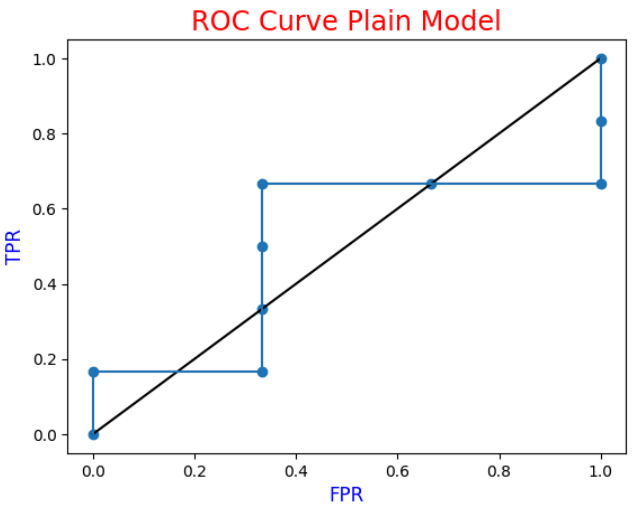



Confusion Matrix per il *Plain Model*

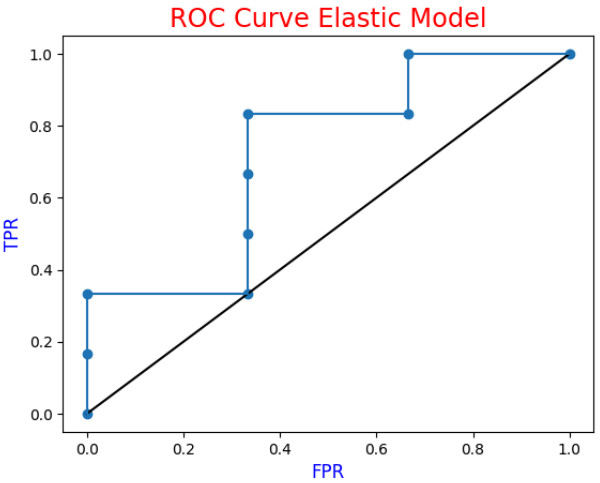



Confusion Matrix per l' *Elastic Model*

## ROC CURVE




Roc Curve per il *Plain Model*




Roc Curve per l' *Elastic Model*

# CONCLUSIONE

La conclusione che deriva dall'analisi dei risultati è che il modello, per quanto si possa provare a perfezionare la scelta dei parametri ed il tipo di regolarizzazione, non è in grado di ottenere un buon tipo di classificazione nel nostro dataset. Questa cosa secondo me era già precedentemente intravedibile dai grafici del dataset, che nonostante non fossero completi (si fermavano ai triplet plot) potevano già mostrare come non sembrava esserci un vero iperpiano separatore delle variabili target. Alcune combinazioni di componenti non solo non sembravano avere un iperpiano separatore ma addirittura sembravano possedere una forte relazione lineare.