REPORT CHALLENGE 0

Introduzione al Machine Learning

Sunto

Utilizzando il modello di regressione logistica si vuole provare a classificare l'appartenenza di una startup ad uno stato Florida o California in base alle features presenti nel dataset

KALUS ABDULA [SM3201269]

Introduzione

Dati

Il dataset è 50_startups e proviene da kaggle, ed è relativo a 50 startups di tre stati americani (New York, California e Florida). L'obbiettivo della classificazione è prevedere se una startup appartiene a uno di questi stati (variabile y).
Si è interessati solo alle startup appartenenti a California e Florida, quindi a una classificazione binaria (1/0), California = 1 e Florida = 0.

Il dataset fornisce informazioni sulle startup degli stati. Comprende 33 record e 5 campi.
Dopo un'analisi preliminare sul dataset i valori anomali ovvero quelli mancanti sono stati sostituiti con il valore medio di quella colonna, successivamente i valori sono stati normalizzati in modo tale da avere i valori di tutte le colonne comprese nel range [0,1].
La variabile target invece è stata trasformata in una variabile binaria ove il valore 1 rappresenta l'appartenenza della startup allo stato California e 0 rappresenta la non appartenenza ovvero quella startup appartiene allo stato Florida.

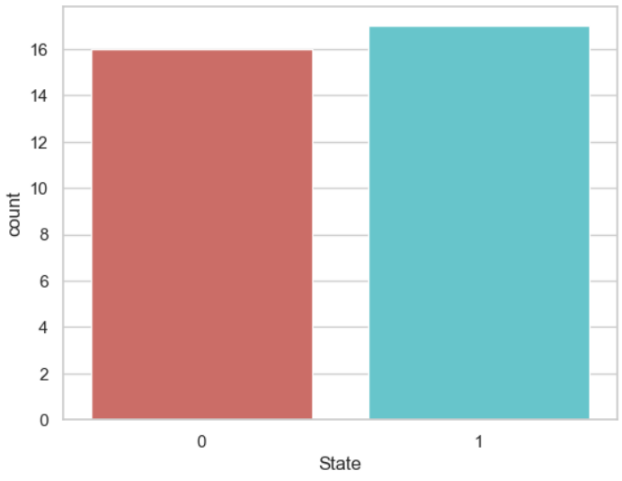
Bisogna considerare inoltre che solo il 75% dei dati verrà utilizzato per costruire il modello il restante sarà utilizzato per testare la sua bontà.

Nella *Tabella 1* è rappresentato il dataset finale sul quale si lavorerà.

	State	R&D Spend	Administration	Marketing Spend	Profit
1	1	1.000000	0.761972	1.000000	1.000000
2	0	0.943229	0.379579	0.913457	0.995812
4	0	0.872953	0.305328	0.812953	0.855434
6	1	0.826501	0.730161	0.239150	0.798603
7	0	0.799733	0.717457	0.711183	0.796514

Tabella 1

Esplorazione dei dati



Etichetta 0 = 48.48%

Etichetta 1 = 51.52%

Dal grafico possiamo notare che i dati sono ben bilanciati, metà delle startup sono in Florida e l'altra metà in California come si può vedere anche dalle percentuali sopra.

Visualizzazione dei dati

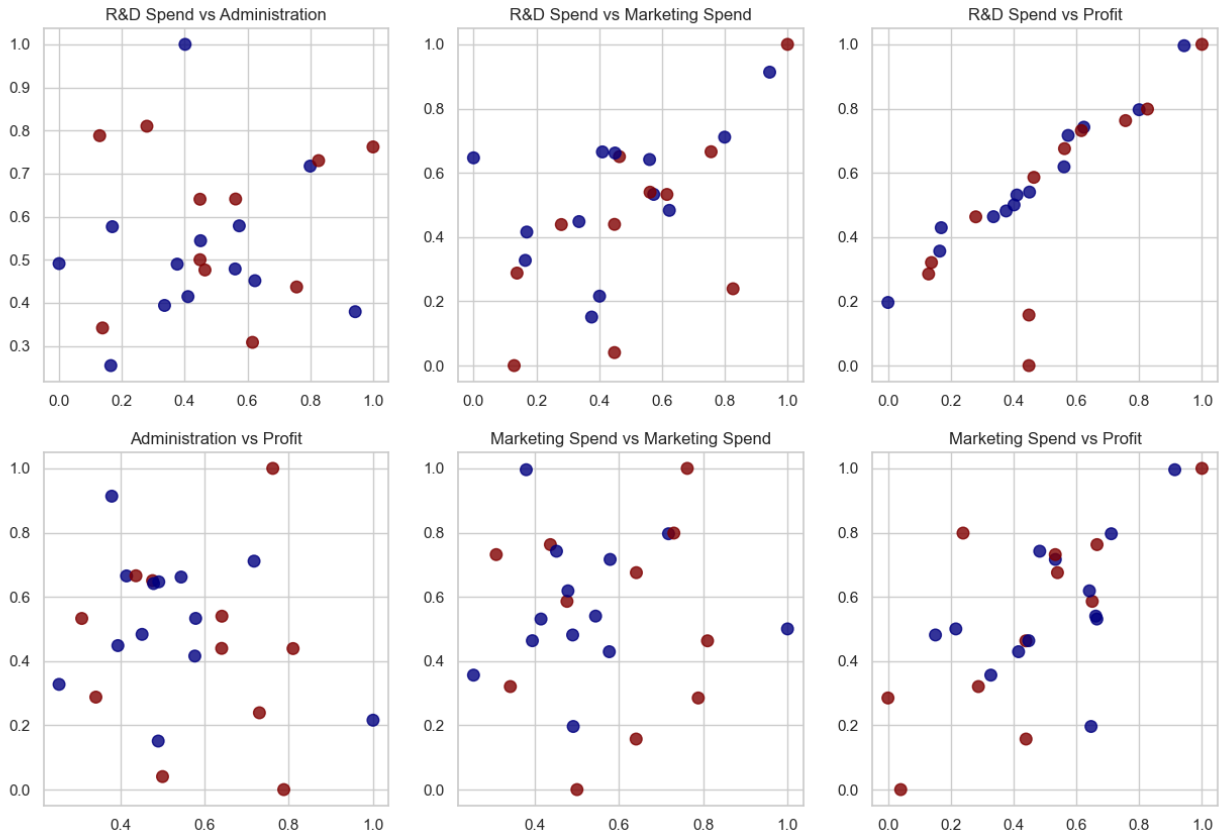
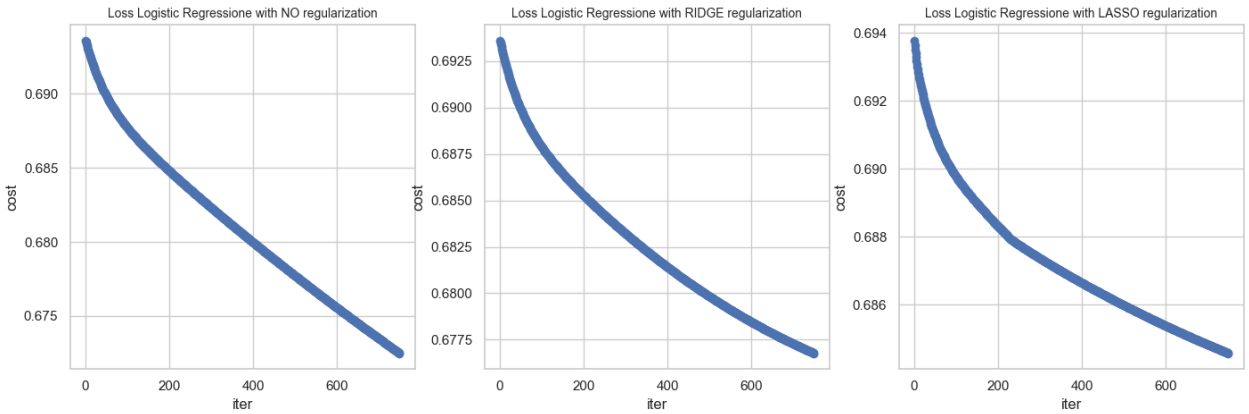


Figura 1

Dagli scatterplot della *Figura 1* si può notare che i dati sono molto mischiati tra di loro, e quindi non è visibile ad occhio una classificazione, inoltre si può notare che la features **R&D Spend e Profit** hanno una correlazione lineare ciò permetterebbe magari di ottenere una riduzione di dimensione del problema da 4 a 3 dimensioni utilizzando PCA oppure se siamo fortunati potremmo ottenere una dimensione minore.

Bisogna tenere conto inoltre che si ha un numero limitato di dati quindi probabilemnte il modelello underfitta.

Creazione dei modelli



Si può notare che la **loss** dei tre modelli *senza regolarizzazione, regolarizzazione ridge* e *regolarizzazione lasso* con:

- **750 iterazioni**
- **learning rate** pari a **0.031**
- **lambda** pari a **0.015**

convergono localmente verso lo stesso partendo dallo stesso punto inziale risultato, nella tabella seguente sono riportati i coefficienti per i vari modelli.

Modello	**w1**	**w2**	**w3**	**w4**
Senza regolarizzazione	0.335698	0.125521	-0.469524	-0.267472
Regolarizzazione RIDGE	0.332117	0.123846	-0.465596	-0.265464
Regolarizzazione LASSO	0.325373	0.120185	-0.461994	-0.259643
SKLEARN	0.339465	0.119216	-0.439438	-0.262364

Si può notare la convergenza dei pesi con un'incertezza minima.

Bontà del modello

Per valutare la bontà del modello partiamo considerando il caso base ovvero quello del **base line model**, il quale rappresenta l'appartenenza di tutti le startup alla classe maggioritaria.

Consideriamo quindi la **confusion matrix** del **base line model**.

Totale = P + N = PP + PN = 9	**Predetti Positivi (PP)**	**Predetti Negativi (PN)**
Positivi Veri (P)	6	0
Negativi Veri (N)	3	0

il quale ci fornisce le seguenti misure

	Accuracy	Precision	Recall	F1-score
Base-Line-Model	0.666667	0.666667	1.000000	0.80

successivamente considerando le **confusion matrix** degli altri modelli e le relative misure possiamo dedurre l'effettiva bontà dei modelli.

Modello senza regolarizzazione

Totale = P + N = PP + PN = 9	**Predetti Positivi (PP)**	**Predetti Negativi (PN)**
Positivi Veri (P)	1	5
Negativi Veri (N)	0	3

Modello con regolarizzazione RIDGE

Totale = P + N = PP + PN = 9	**Predetti Positivi (PP)**	**Predetti Negativi (PN)**
Positivi Veri (P)	1	5
Negativi Veri (N)	0	3

Modello con regolarizzazione LASSO

Totale = P + N = PP + PN = 9	**Predetti Positivi (PP)**	**Predetti Negativi (PN)**
Positivi Veri (P)	1	5
Negativi Veri (N)	0	3

ovviamente dato che tutti i modelli hanno circa gli stessi coefficenti, le predizioni saranno circa uguali anche se in questo caso sono esattamente le stesse. Infatti, guardando le misure di tutti i modelli insieme possiamo notare che sono le stesse.

	Accuracy	Precision	Recall	F1-score
Base-Line-Model	0.666667	0.666667	1.000000	0.800000
Logistic regression without regularization	0.444444	1.000000	0.166667	0.285714
Logistic regression with RIDGE regularization	0.444444	1.000000	0.166667	0.285714
Logistic regression with LASSO regularization	0.444444	1.000000	0.166667	0.285714

Concludendo che il **Base Line Model** sembrerebbe il migliore tra tutti; infatti, dal plot successivo possiamo vedere con la **ROC curve** che l'accuratezza dei vari modelli non è buona.

REPORT CHALLENGE 0