CHALLENGE 0

Pietro Mihelj

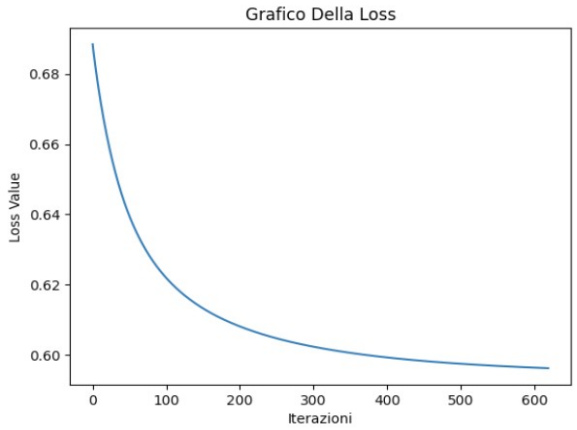
Lo scopo della challenge era trovare un modello che ci permettesse di classificare correttamente delle start up provenienti dalla California e dalla Florida, indicate rispettivamente come 1 e 0.
I modelli da studiare erano logistici con regressione Ridge, Lasso e ElasticNet.
Il dataset da usare è però problematico, vedremo infatti che ha grandi problemi sia di bias che di varianza.
Nota: per tutti i calcoli è stato usato il seed 123 nell'inizializzazione dei pesi in quanto è stato individuato, dopo diverse prove, come quello che non finiva in minimi locali per i parametri usati.

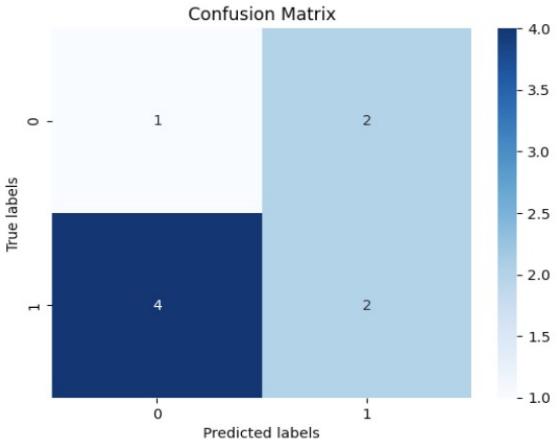
BIAS

Per mostrare i problemi di bias ho scelto di usare un modello con regolarizzazione Ridge e iperparametri

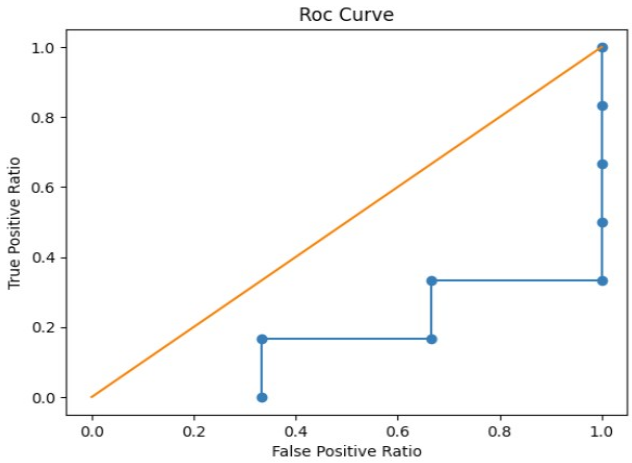
- λ = 0.0001
- γ = 1
- soglia = 0.5
- random_state = 0
- iter = 620

Il modello restituisce una accuracy del 33% e anche precision e recall non superano il 50%.
Dai grafici sottostanti possiamo vedere che la Loss rimane molto alta.
La confusion matrix conferma la scarsa qualità del modello.





Infine dalla Roc Curve ricaviamo le stesse conclusioni.



Nota: i risultati mostrati sono solo sul modello con regolarizzazione Ridge poiché Lasso ed ElasticNet restituiscono risultati pressochè identici.
Nota2: i parametri λ e γ sono stati scelti poiché minimizzano la Loss per la soglia ed il random_state utilizzati. La scelta è stata fatta tramite della validazione a mano.

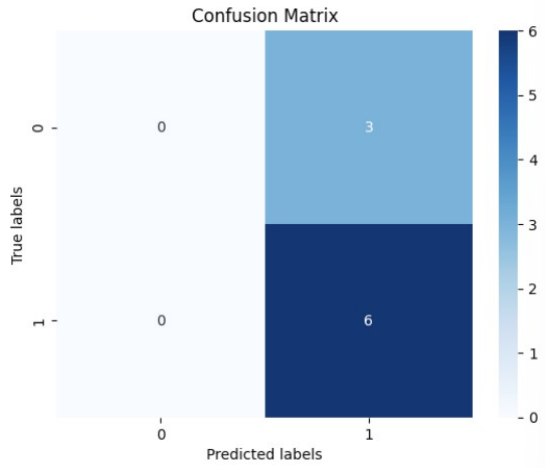
VARIANZA

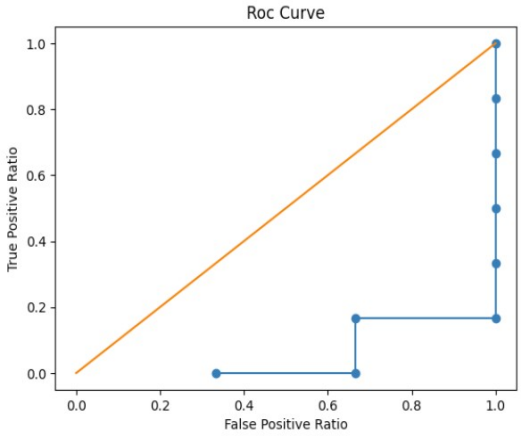
La ristrettezza del dataset ci causa un evidente problema di varianza.
Segue un confronto tra i risultati trovati con un modello Ridge con parametri:

- λ = 1
- γ = 0.0001
- soglia = 0.5
- iter = 620

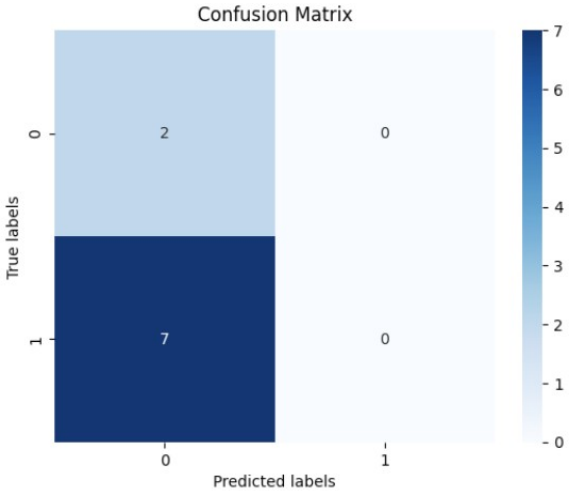
per random state = {0,4}.
Notiamo come i risultati cambino completamente al variare del random state.

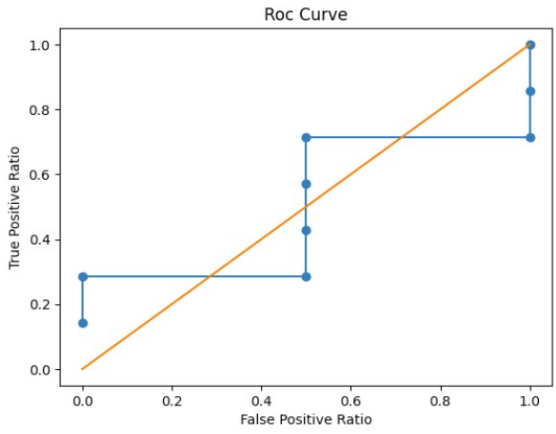
RANDOM_STATE = 0
-accuracy 67%
Florida
-precision 67%
-recall 100%
California
-precision 0%
-recallo 0%





RANDOM_STATE = 7
-accuracy 22%
Florida
-precision 0%
-recall 0%
California
-precision 22%
-recall 100%





Nota: I risultati restituiti da modelli con regressione Lasso e ElasticNet sono comparabili

CONCLUSIONI

I dati forniti non permettono di trovare un modello accettabile. Questo dipende da 2 fattori.
Il primo è la forma dei dati, la qui influenza si vede nello studio del bias. Si può, infatti, dedurre che non è possibile trovare un iper-piano che separi efficaciemente i dati.
Il secondo è la ristrettezza del data set, la qui influenza si vede nello studio della varianza. Si può dedurre che, cambiare anche solo pochi elementi nel testset e nel trainset, provoca uno squilibrio importante nelle proporzioni trai dati di train modificando profondamente le nostre previsioni.