Analisi della valutazione del valore dei giocatori di calcio

Progetto "Laboratorio di Big Data"

di Marco Cossu

Descrizione Dataset

Il dataset scelto per questo progetto espone le statistiche di alcuni giocatori di calcio della stagione 2020/2021, alcune di queste sono di tipo qualitativo ed altre di tipo quantitativo. Il dataset risulta essere composto da 17954 osservazioni, articolate in 97 variabili (successivamente ridotte a 46); tra tutte queste variabili l'analisi si concentra su quella denominata "*value_euro*" che indica, appunto, in valore in euro del singolo giocatore, poi riformulata in scala logaritmica ed in Milioni di euro.

Analisi Grafica Dataset

Le prime analisi effettuate sono state la ricerca di correlazione tra le variabili e la verifica se si evidenziassero segni di multicollinearità, ciò viene effettuato attraverso una Matrice di Correlazione ed una HeathMap, da cui si evidenziano problemi di eccessiva collinearità con la variabile "release_clause_euro" (motivo per cui essa è stata estromessa dall'analisi).

La variabile "Log Market Value" risulta essere maggiormente correlata con le variabili che indicano il Rating Generale, la reputazione internazionale, la reputazione del club, la reputazione della Nazionale, la Capacità di Reazione ed il temperamento ("composure"). Per la variabile Rating Generale siamo quasi alla collinearità ma non viene estromessa dal processo di analisi.

Infine, vengono tracciati un grafico a barre accostate ed un box plot, riguardanti il focus dell'analisi (la variabile "Log_Market_Value"), che evidenziano come i valori più comuni siano quelli compresi tra 4.5 e circa 7.5 Milioni di euro in scala logaritmica, inoltre si notano dei valori alti (realistici anche essi), che non vengono eliminati dall'analisi.

Un ultimo grafico (questa volta un Pie Chart) mostra come il 23,2% dei giocatori utilizzi prevalentemente il piede sinistro, ed il 76,8% il destro, questi valori sono abbastanza realistici e quindi vengono interpretati come corretti.

Algoritmi e metodi utilizzati nell'analisi del Dataset

Prima di procedere all'analisi del dataset attraverso i modelli vengono utilizzati, attraverso una pipeline, lo *String Indexer*, il *One Hot Encoder* ed il *Vector Assembler* sul data frame (quello contenente la variabile in scala logaritmica e, successivamente, in Milioni di Euro con le feature scalate).

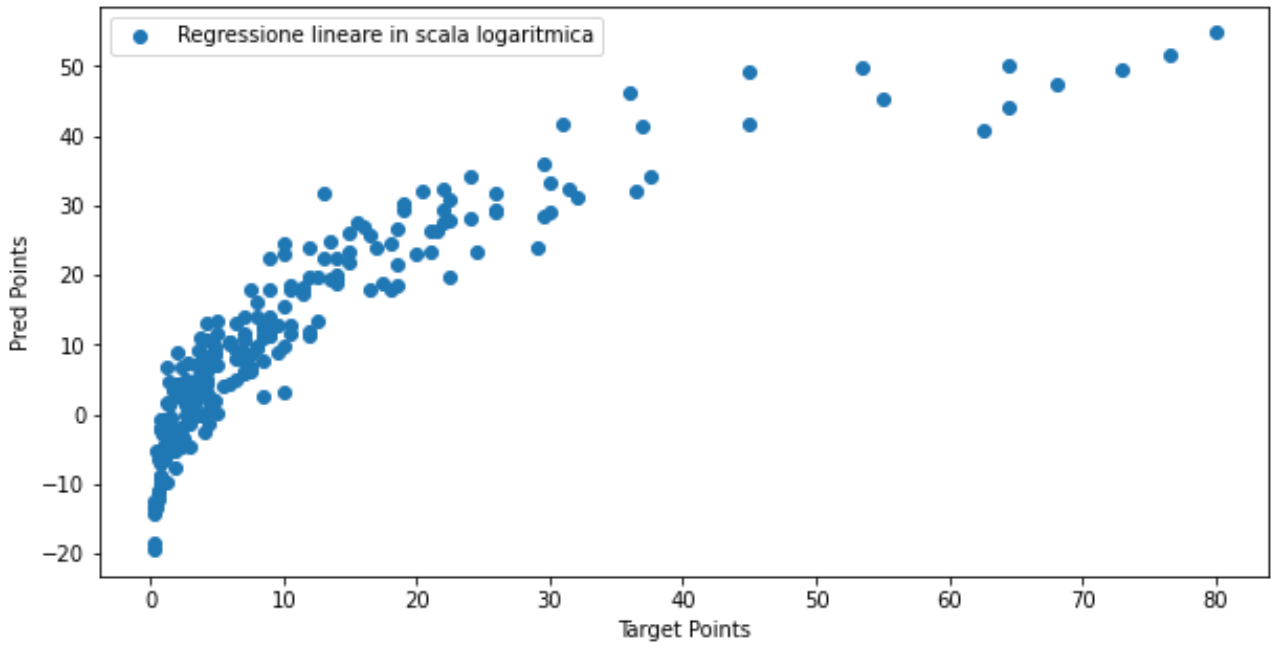
Lo *String Indexer* indicizza una colonna contenente stringhe in una colonna di indici, questo risulta particolarmente utile se usato in combinazione con il *One Hot Encoder* che assegna ad ognuna delle due categorie un valore 0 o 1 e crea una nuova colonna con questi valori. Infine, il *Vector Assembler* viene utilizzato per unire tutte le colonne contenenti le variabili presenti nel modello in una colonna che contiene un unico vettore.

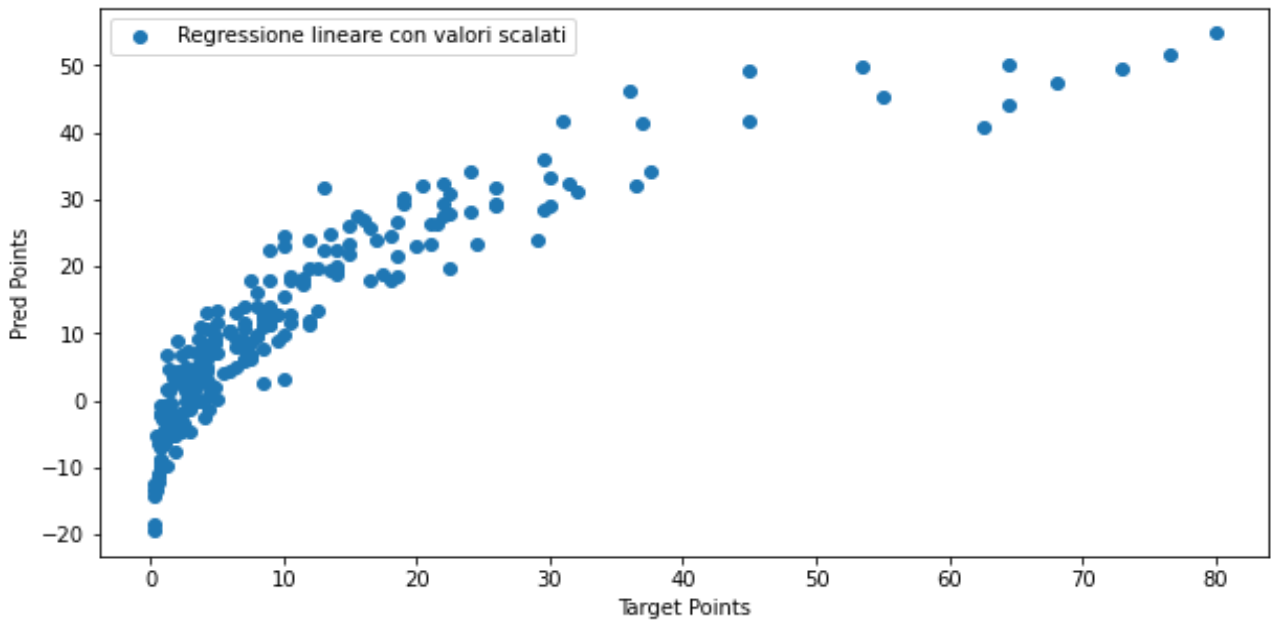
I modelli vengono fittati in due modi diversi: sulla variabile di risposta espressa in scala logaritmica e sulle feature scalate utilizzando il MinMaxScaler, un modo per normalizzare le variabili del modello in modo che tutte rientrino in un range [0,1], per fare sì che ogni variabile contribuisca in maniera uguale al fit del modello e ridurne il bias.

- # Linear Regression

Questo è il primo algoritmo che viene fittato, le due regressioni lineari che vengono implementate (scala logaritmica e con features scalate). La regressione lineare con le feature scalate performa in maniera decisamente peggiore rispetto a quello che utilizza la scala logaritmica: RMSE: 0.107183 contro 0.0973704, ed un R2 leggermente superiore: R2: 0.776803 contro 0.960327

Un'altra differenza tra esse risulta essere che l'utilizzo di features scalate porta ad avere un valore pari allo 4,5% di valori predetti negativi (chiaramente un errore), mentre la scala logaritmica registra lo 0%, valori comunque molto bassi ma che vanno tenuti in considerazione.

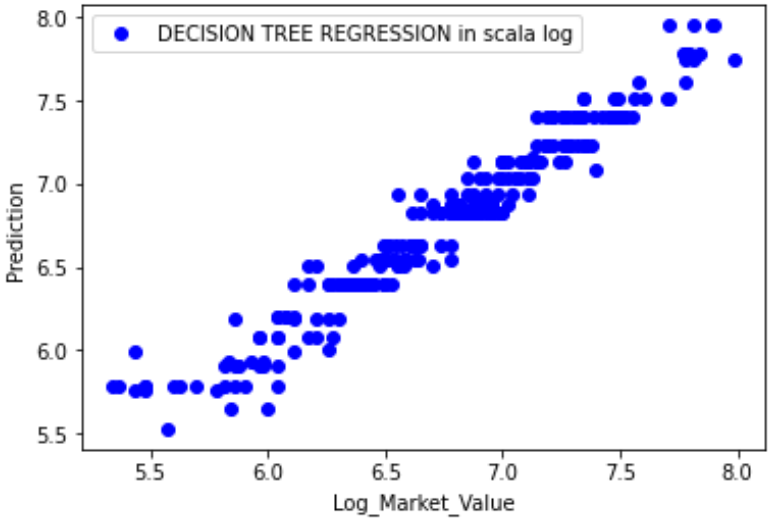




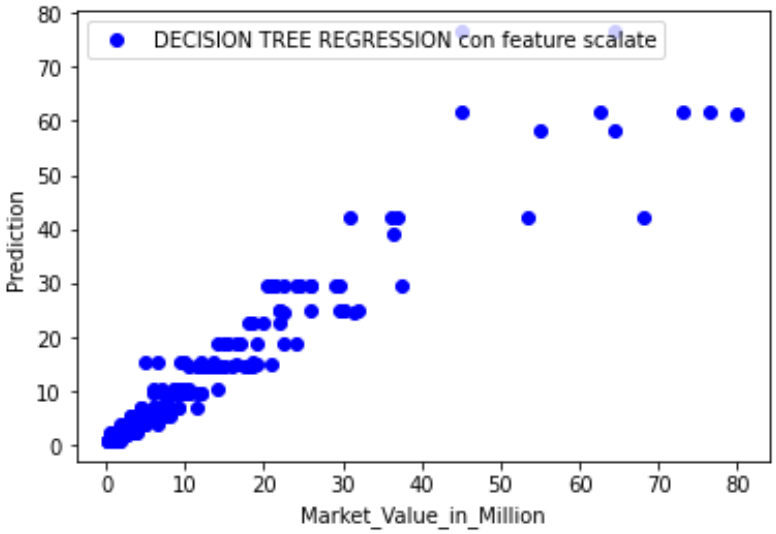
- # Decision Tree Regression

Il secondo algoritmo utilizzato è stato un Decision Tree Regression, anch'esso utilizzando sia features scalate che no, essi presentano un RMSE molto differente: 4.79799 per le features scalate, che risulta essere decisamente superiore a quello ricavato dai dati in scala logaritmica (0.125508), risulta evidente una certa variabilità, al di sopra dei valori medi.

Model Performance RMSE: 0.131702



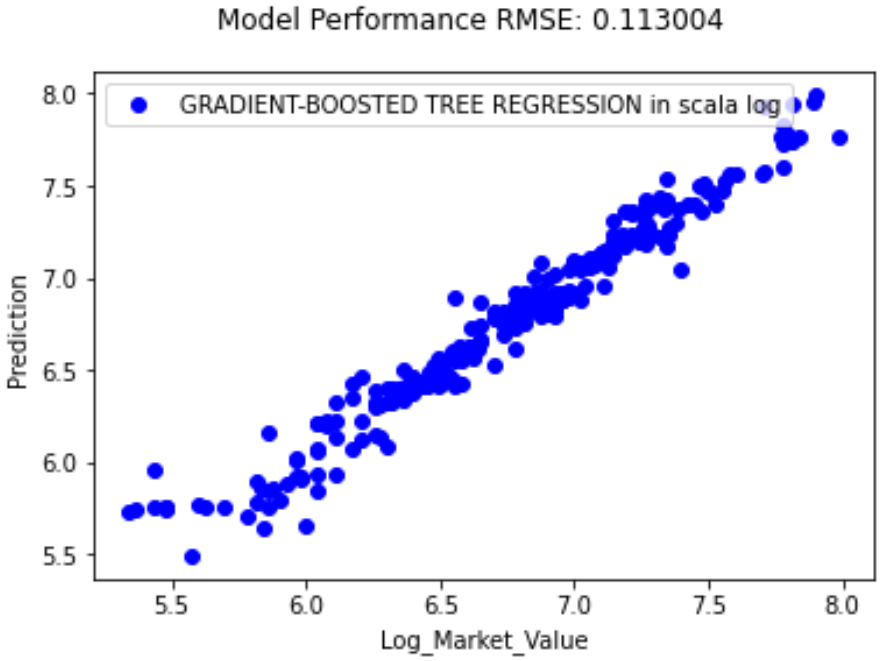
Model Performance RMSE: 4.181261

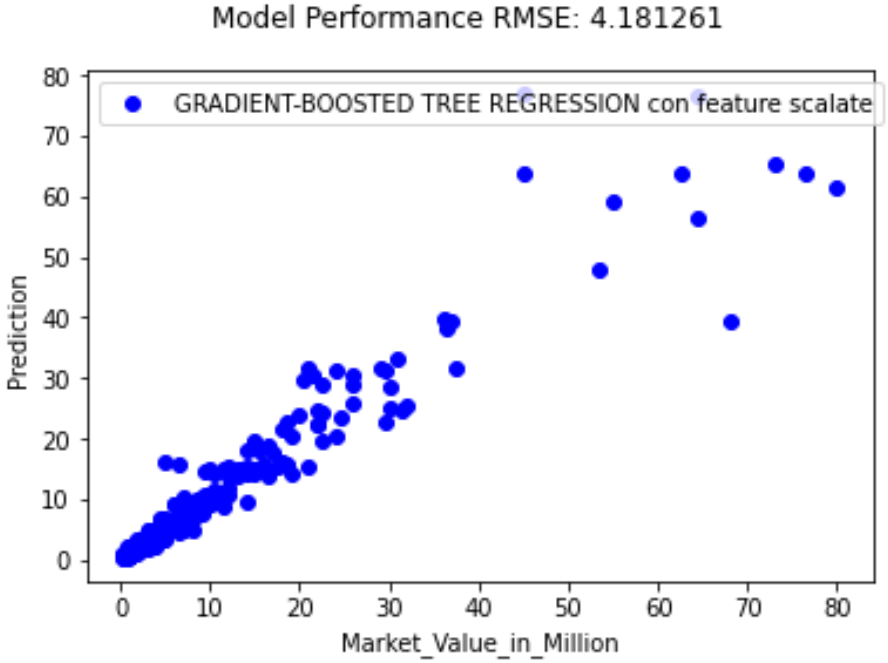


- # Gradient -Boosted Tree Regression

Il terzo ed ultimo algoritmo utilizzato è un Gradient-Boosted Tree Regression, che fornisce, generalmente, predizioni migliori rispetto al singolo albero. In questo caso si attestano ad un valore di RMSE pari a 0.112361 (scala logaritmica) e 6.99645 (con features scalate), sicuramente quest'ultimo mostra un valore peggiore rispetto al modello precedente e al Decision Tree (di cui costituisce un modello ensembled).

Anche in questo grafico (features scalate) si può notare una certa variabilità al di sopra dei 40 milioni di €,


Model Performance RMSE: 0.113004
GRADIENT-BOOSTED TREE REGRESSION in scala log
Prediction
Log_Market_Value
8.0
7.5
7.0
6.5
6.0
5.5
5.5
6.0
6.5
7.0
7.5
8.0


Model Performance RMSE: 4.181261
GRADIENT-BOOSTED TREE REGRESSION con feature scalate
Prediction
Market_Value_in_Million
80
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
80