Big Data Analysis on NBA & NCAA

Colleges' impact on professionistic games

Marco Faretra, Gabriele Marini

3 luglio 2017

RIASSUNTO

Lo scopo di questo progetto è quello di produrre un ranking dei college americani sulla base delle performance dei primi anni di carriera dei giocatori professionistici. Allo scopo si sono analizzate tutte le statistiche dei giocatori NBA dai primi anni '50 fino ad oggi e si è semplificato il gioco fino ad individuare 6 categorie di giocatori, non mutuamente esclusive. Ad ogni giocatore viene assegnato un punteggio a seconda della particolare categoria che si sta analizzando, il punteggio di un college è la somma di tutti i punteggi dei giocatori provenienti da quel college, tenendo conto della particolare categoria analizzata. Lo scopo dell'analisi è quello di comprendere come alcuni college possano puntare su particolari aspetti del gioco, scelta che si dovrebbe riflettere nelle statistiche dei primi anni di carriera di un giocatore professionistico.

1 INTRODUZIONE

La lega di pallacanestro professionistica americana, meglio conosciuta come NBA, è stata una delle prime realtà a fruire dei dati dei suoi giocatori allo scopo di perfezionare questi ultimi e rendere le franchigie partecipanti sempre più competitive. Questo flusso di dati permette oggigiorno di avere a disposizione una quantità di dati immensa, delle tipologie più disparate, dalle statistiche dei singoli giocatori o delle singole partite, fino ad arrivare alle statistiche più dettagliate play-by-play [1].

Quello che non tutti sanno è che esiste un mondo dietro a quello della lega professionistica, altrettanto vasto, ovvero quello dei college e della NCAA. Prima di rendersi disponibili per il draft i giocatori tendono a svolgere qualche anno di preparazione in uno dei molti college americani. Questi, oltre che a fornire borse di studio ai giocatori, permettono loro di affinare gli aspetti legati al gioco, oltre che a dare loro una discreta visibilità agli occhi degli scout NBA.

Per questo progetto è di interesse proprio la correlazione tra gli insegnamenti del college e l'effettiva applicazione di questi nella sfera professionistica. Il nostro scopo è quello di analizzare, utilizzando tecniche Big Data, le statistiche di tutti i giocatori NBA dai primi anni '50 fino ad oggi, al fine di stilare una classifica dei college americani che hanno avuto nella loro storia almeno un giocatore che è riuscito a fare il salto di categoria nella lega professionistica.

A tale scopo si sono individuate 6 categorie di giocatori: tiratori (divisa a sua volta in tiratori da 2 e tiratori da 3), rimbalzisti, all-around, +/- guys, difensori, attaccanti. Per ognuna di queste categorie si è calcolato una score per ogni giocatore sulla base di alcune particolari statistiche tra le

[1] https://www.bigdataball.com/nba-historical-playbyplay-dataset

molte disponibili[2]. Ogni giocatore contribuisce, per la particolare categoria scelta, al punteggio totale del college di appartenenza. In questo modo college che curano di più una particolare caratteristica avranno un punteggio più alto se analizziamo la categoria legata a quella caratteristica. Una volta calcolato lo score per ogni categoria per ogni college è possibile stilare un ranking dei college sulla base delle categorie.

2 SETTING

I dati sulle statistiche di tutta la storia NBA non vengono forniti attraverso REST API. Quindi per effettuare le nostre analisi, abbiamo effettuato delle operazioni di scraping da basketball reference [3], un motore che contiene tutte le statistiche NBA dal 1946 ad oggi. Il nostro processo di estrazione dati, si è suddiviso in diversi passaggi, per prima cosa abbiamo estratto gli identificativi di tutti i giocatori con associato il college di appartenenza e gli anni in cui hanno giocato in NBA. Successivamente abbiamo scaricato le pagine html che contengono le statistiche (misurate per partita) per ogni stagione di ogni giocatore, per poi infine abbiamo estratto attraverso degli xpath i dati che avremmo utilizzato per calcolare le statistiche.

3 APPROCCIO

Per i dati estratti abbiamo pensato di utilizzare un database document-store ed un key-value store per salvare i nostri dati.
Il database orientato ai documenti utilizzato per persistere i dati completi, che contiene ogni giocatore, con tutte le stagioni da lui giocate e tutte le partite.
Mentre il key-value store utilizzato per prendere dati che vengono acceduti spesso in modo abbastanza veloce, ad esempio ci siamo resi conto che poteva tornarci utile calcolarci media e varianza per ogni stagione, per il calcolo dello score, e risulta comodo avere un key-value store che mantenga questi dati, poiché verranno acceduti frequentemente.
Per il calcolo dello score, vengono effettuate delle verifiche riguardo delle soglie specificate, e vengono prese in considerazione delle percentuali da applicare alla particolare statistica di interesse. Ad esempio per gli attaccanti abbiamo fissato come soglie che i tiri provati e i minuti giocati siano maggiori della media, e abbiamo dato un peso del 30% alla percentuale effettiva di successo al tiro (sia tiri da due che tiri da tre), ed un 70% ai punti totalizzati.
Per ogni categoria vengono quindi definite le soglie (campi che siano maggiori della media stagionale) e le percentuali sulle statistiche di interesse. Per alcune categorie, lo score non era del tutto pulito, quindi abbiamo deciso di utilizzare anche una formula di bonus/malus da poter applicare, ad esempio nella categoria dei difensori, oltre ai valori utilizzati per calcolare lo score, abbiamo pensato di tener conto anche dei falli personali (caratteristica che caratterizza un difensore), quindi assegniamo anche qui un bonus dove viene specificato la caratterstica di interesse (in questo esempio i falli personali), ed il peso. Questo bonus/malus andrà a sottrarsi allo score originale, di modo che se viene un numero negativo, allora con la sottrazione andrà a sommarsi e quindi ad assegnare un bonus, altrimenti andrà ad assegnare un malus abbassando lo score complessivo.

[2]per i tiratori, ad esempio, è di fondamentale importanza la % del tiro, mentre le statistiche relative ai rimbalzi sono state ignorate per questa categoria

[3]http://www.basketball-reference.com/

4 SOLUZIONE TECNOLOGICA

Come soluzione tecnologica per il nostro approccio abbiamo pensato di utilizzare MongoDB come document-store e Redis come key-value store, perché ci sono sembrate le tecnologie più famose e stabili per le rispettive famiglie di database.

5 RISULTATI

6 CONCLUSIONI

7 SVILUPPI FUTURI