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

Siamo partiti da un dataset iniziale che contiene tutte le statistiche NBA dal 1946 ad oggi.
Abbiamo poi effettuato delle operazioni di clean dei dati di questo dataset, e di salvare i dati in un database di tipo big data.

3 APPROCCIO

Per i dati estratti dal dataset, poiché si tratta di dati storici (append-only), abbiamo pensato di utilizzare un database document-store, e di utilizzare un key-value store per le statistiche di base per ogni stagione e per la profilazione.
Per il calcolo dello score, vengono specificate delle soglie da soddisfarre per considerare un elemento, questo per evitare falsi positivi e falsi negativi, e delle percentuali riguardanti il peso per la singola statistica da considerare.

4 SOLUZIONE TECNOLOGICA

Nella nostra soluzione tecnologica abbiamo pensato di utilizzare python, perché rispetto a Java risulta meno prolisso, con l'utilizzo del modulo pyspark.
Come document-store abbiamo pensato di utilizzare MongoDB, il databse più popolare della sua famiglia al momento, offre diversi connettori che permettono l'integrazione completa con pyspark, e consente un installazione molto semplice attraverso docker.
Per i stessi motivi come database key-value store, abbiamo scelto Redis.

5 RISULTATI

6 CONCLUSIONI

7 SVILUPPI FUTURI

[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