Statistica Superiore - Relazione II

L'importanza delle statistiche dei giocatori dell'NBA nel loro Draft year

Eleonora Basilico

1 Descrizione del problema

La seguente analisi è rivolta ai general manager delle squadre di basket dell'NBA. L'obiettivo è quello di capire se l'andamento del solo primo anno di gioco consenta di predire la durata della carriera di un giocatore, e se questo, quindi, sia un buon metro di giudizio per scelte riguardanti la compra-vendita o lo scambio di giocatori.

Abbiamo considerato le statistiche di 500 giocatori dell'NBA nel loro primo anno di gioco. Gli anni selezionati sono quelli che vanno dal 2002 al 2015 (per povertà di dati non sono stati presi in considerazione tutti i giocatori che hanno iniziato la loro carriera in quegli anni ma solo una parte di essi).
I giocatori sono stati divisi in due classi: appartengono alla classe 1 coloro che hanno giocato nell'NBA per almeno 5 anni (regular season), alla classe 0 coloro che hanno giocato nell'NBA per meno di 5 anni. Vogliamo capire se e con quanta precisione sia possibile predire la durata della carriera di un giocatore (più o meno di 5 anni) se ci si basa sulle sole statistiche del primo anno di gioco.

Abbiamo considerato i seguenti dati:

- GP: numero di partite giocate;
- MIN: minuti di gioco;
- PTS: Punti fatti;
- FGA: tiri dal campo tentati;
- FGP: percentuale di tiri dal campo realizzati;
- X3PA: tiri da 3 punti tentati;
- X3PP: percentuale di tiri da 3 punti realizzati;
- FTA: tiri liberi tentati;
- FTP: percentuale di tiri liberi realizzati;
- OREB: rimbalzi offensivi catturati;
- DREB: rimbalzi difensivi catturati;
- AST: assist;
- BLK: palle stoppate;
- PF: falli commessi;
- Classe: 1 se il giocatore ha giocato almeno 5 anni nell'NBA, 0 altrimenti[1].

Tutti i fattori ad eccezione di GP sono dati per partita[2]: essi sono reperibili già in questa forma sul sito ufficiale dell'NBA (link in fondo al documento). Nella tabella è indicato anche il primo anno di gioco (Draft-Year), ma il dato non è stato utilizzato perchè non utile ai fini della nostra analisi.

[1] Come ci aspettiamo la classe 1 è molto più numerosa della classe 0 (341 giocatori su 500 appartengono alla classe 1). Questo fatto andrà tenuto in considerazione nella nostra analisi.

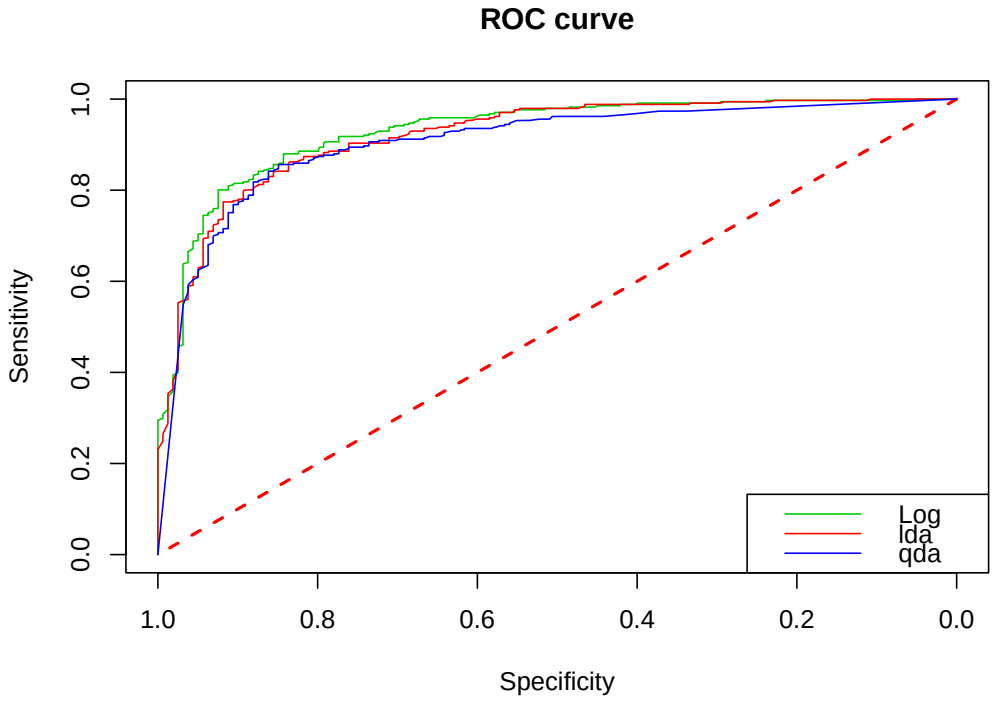
[2] Tra i fattori considerati, nonostante sia presente la percentuale di tiri andati a segno (tiri dal campo, da tre punti e liberi), è stato tenuto anche il fattore relativo al numero di tiri di ciascun tipo tentati. Tale scelta è dovuta al fatto che per alcuni giocatori la percentuale di tiri realizzati è 0, ed è quindi importante conoscere il numero di tiri tentati per distinguere coloro che sbagliano da coloro che invece non hanno fatto alcun tiro.

2 Classificazione

Al fine di predire se la durata della carriera di un giocatore nell'NBA sarà maggiore o minore di 5 anni, e lo facciamo a partire dall'andamento del suo primo anno di gioco, abbiamo implementato i seguenti modelli di classificazione: regressione logistica, analisi discriminante lineare e analisi discriminante quadratica.

2.1 Una prima valutazione dei modelli

Prima di analizzare in dettaglio i modelli usati, ne confrontiamo le prestazioni in termini di curva ROC:



I risultati sono buoni: le tre curve non sono ideali ma sono comunque molto lontane dalla bisettrice che rappresenta una predizione di tipo randomico.

Facciamo in particolare le seguenti osservazioni:

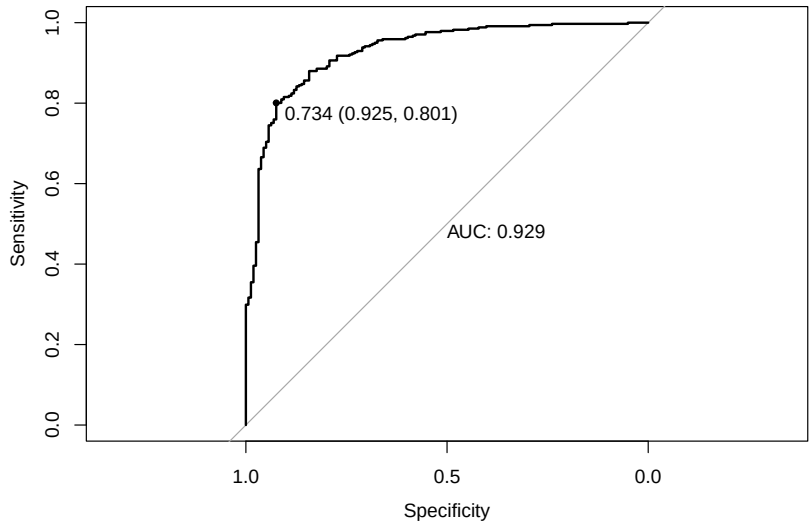
- Non ci sono differenze sostanziali tra le curve. Tuttavia, va osservato che le curve ROC relative ad analisi discriminante lineare ed analisi discriminante quadratica sono al di sotto della curva ROC relativa alla regressione logistica per quasi ogni valore del parametro soglia p. Ciò può essere sintomo del fatto che l'ipotesi di normalità dei dati sia troppo forte;
- Il modello di analisi discriminante quadratica sembra performare peggio rispetto agli altri due: la sua curva ROC è quasi sempre al di sotto di quella della regressione logistica e supera quella dell'analisi discriminante lineare solo per pochi valori del parametro p.

In conclusione, i tre calssificatori sembrano essere confrontabili: in termini di curva ROC nessuno di essi spicca di molto sugli altri. Tuttavia, può essere interessante cercare di capire sotto quali aspetti ciascuno di essi performa meglio o peggio[3].

[3] L'analisi svolta sul modello di analisi discriminante lineare non è riportata all'interno della relazione in quanto quest'ultimo ha prestazioni molto simili agli altri due modelli (è a metà tra i due). Per ulteriori dettagli si rimanda allo script.r allegato alla relazione.

2.2 Regressione logistica

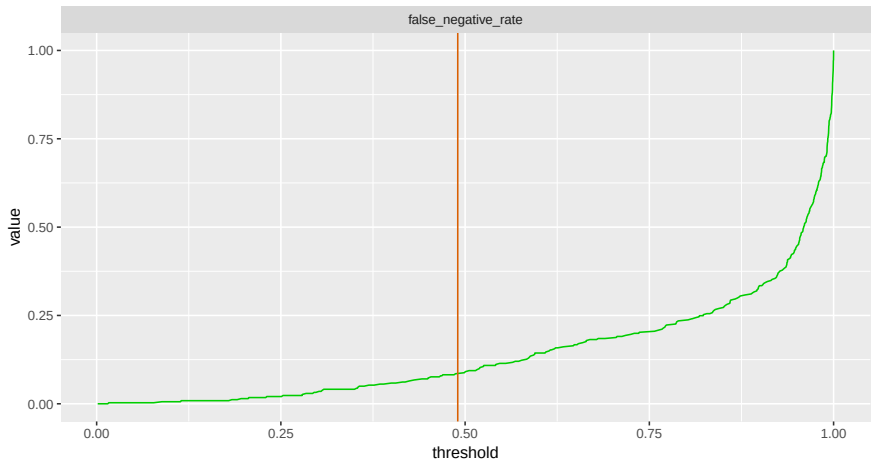
Poichè per la regressione logistica non ci sono problemi di convergenza, abbiamo mantenuto tutti i fattori precedentemente introdotti. Per una prima valutazione della performance del modello osserviamo, assieme alla curva ROC, il valore del parametro AUC:

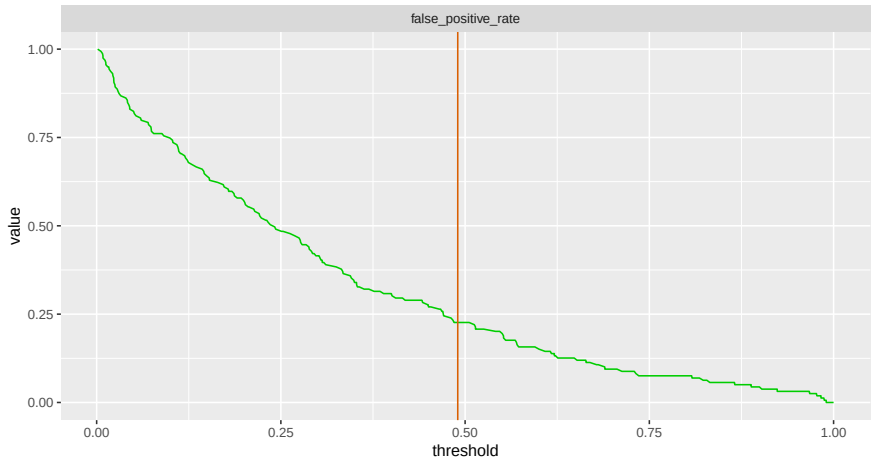


Il risultato è soddisfacente: abbiamo già detto nel precedente paragrafo che la curva ROC relativa alla regressione logistica è lontana dalla bisettrice, seppur non ideale. Inoltre, il valore del parametro AUC è molto vicino a 1.
Il grafico mostra anche che 0.734 è il valore del parametro p che massimizza la grandezza sensitivity + specificity. Tuttavia, tramite matrice di confusione si osserva che per un tale valore di p si ha un alto numero di falsi negativi e un basso numero di falsi positivi e che l'elevato valore di sensitivity + specificity è dovuto al fatto che la classe 1 sia molto più numerosa della classe 0 e non ad una maggiore accuratezza del modello.

Il nostro obiettivo è quello di limitare il più possibile il numero di errori, senza dare più importanza all'uno o all'altro tipo, e uno dei valori di p in corrispondenza dei quali si ha il minor numero di errori è 0.49. Pertanto, decidiamo di utilizzare nella nostra analisi, laddove richiesto, un valore di p pari a 0.49. Si guadagna così qualcosa in accuratezza, infatti per $p = 0.49$ si ha un'accuratezza di 0.868 contro lo 0.84 che si ha invece per $p = 0.734$.

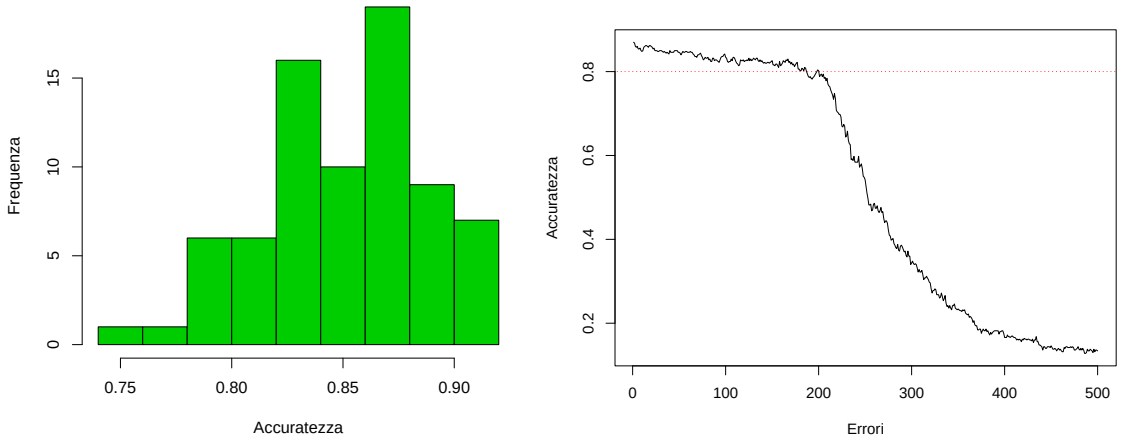
Studiamo ora in particolare l'accuratezza del nostro classificatore sui positivi e sui negativi rispettivamente. A tal fine osserviamo la percentuale di falsi negativi e di falsi positivi al variare della soglia p di accettazione:





La regressione logistica riesce a predire l'appartenenza di un giocatore alla classe 1 con maggiore precisione rispetto all'appartenenza di un giocatore alla classe 0: la percentuale di falsi negativi supera lo 0.25 solo per un valore del parametro p pari a 0.824. Va osservato, però, che questo è anche dovuto alla diversa numerosità delle due classi.
Notiamo, inoltre, che in corrispondenza di $p = 0.49$ i falsi negativi sono l' $8,8\%$, mentre i falsi positivi sono il $22,6\%$: il risultato è molto buono nel primo caso, meno buono nel secondo.

Concludiamo l'analisi della regressione logistica studiando capacità di predizione e robustezza del modello:

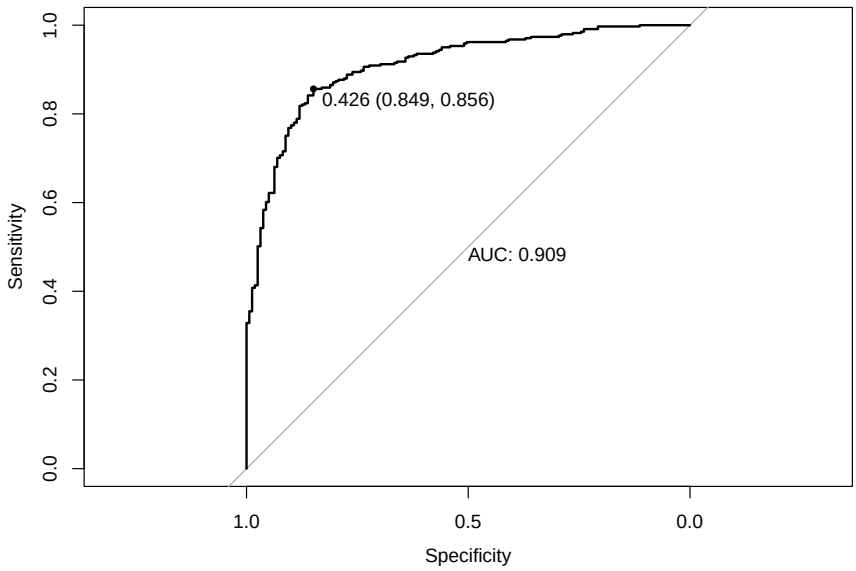


L'istogramma dell'accuratezza mostra che la capacità di predizione del modello è abbastanza soddisfacente: con una cross-validation ottentiamo una media dell'accuratezza di 0.854 ed una deviazione standard di circa 0.038. La deviazione standard è leggermente alta: notiamo che l'accuratezza scende al di sotto dell' 80% per 8 volte su 75 e sale al di sopra del 90% per 7 volte. Ciò manifesta la presenza nei dati di un certo "rumore" che andrebbe maggiormente approfondito.

Il secondo grafico si ottiene introducendo un'informazione falsa alla volta nella tabella di dati e replicando via via la regressione logistica sulla nuova tabella ottenuta, con l'obiettivo di studiare l'andamento dell'accuratezza. Più l'accuratezza rimane alta all'aumentare del numero di errori introdotti, e più il modello è robusto. Nel nostro caso, prima che l'accuratezza scenda al di sotto dell' 80% dobbiamo introdurre quasi 200 errori, pertanto possiamo dire che il nostro modello è sufficientemente robusto.

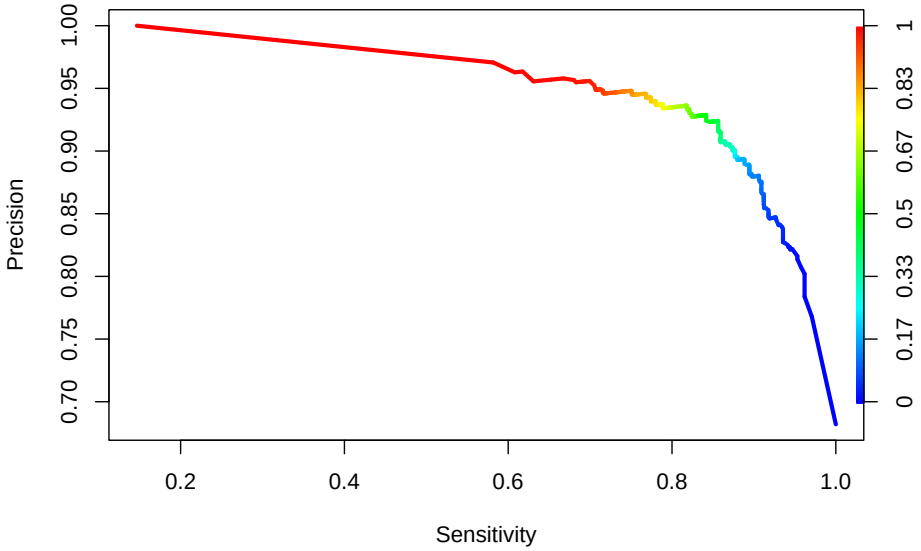
2.3 Analisi discriminante quadratica

Per una valutazione iniziale della performance del modello di analisi discriminante quadratica osserviamo, assieme alla curva ROC, il valore del parametro AUC:



Anche in questo caso il risultato è soddisfacente: sebbene la curva ROC non sia ottimale, l'area al di sotto di essa è di ben 0.909. Inoltre, il valore del parametro p in corrispondenza del quale la grandezza sensitivity + specificity è massima è 0.426: per $p = 0.426$ si ottiene l' $84,9\%$ di veri negativi e l' $85,6\%$ di veri positivi. 0.426 è anche uno dei valori di p in corrispondenza dei quali si ha il minor numero di errori (l'accuratezza è di 0.854), pertanto nell'analisi che segue manteniamo tale valore come soglia di accettazione.

Proseguiamo la nostra analisi osservando il seguente grafico:



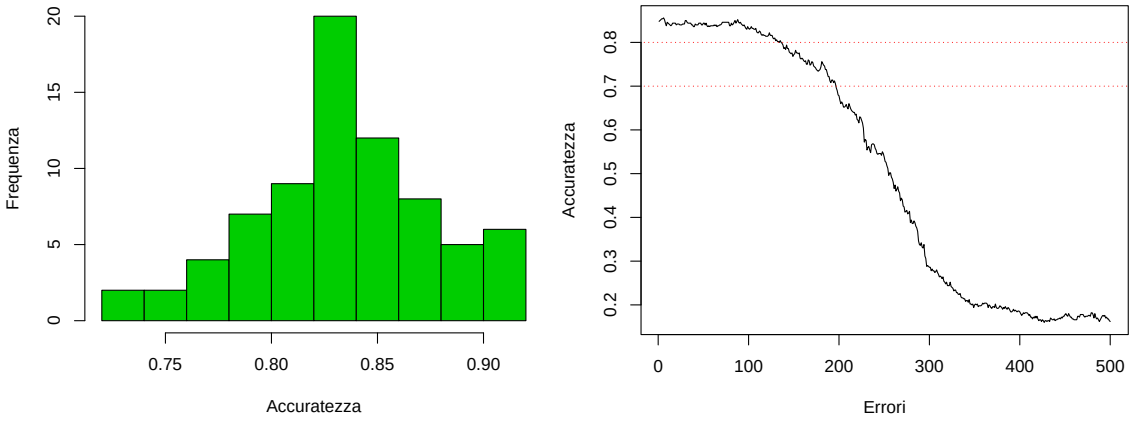
La grandezza *precision* (P) può essere vista come una misura della rilevanza del risultato. Se indichiamo con T_p il numero di veri positivi e con F_p il numero di falsi positivi, essa è definita nel seguente modo:

$$P = \frac{T_p}{T_p + F_p}.$$

La curva mostra il rapporto che c'è tra le grandezze sensitivity e precision al variare del parametro soglia p. Esse ci consentono di studiare l'accuratezza del nostro classificatore sui negativi e sui positivi rispettivamente: quanto più il valore di precision è alto tanto più è bassa la percentuale di falsi positivi; viceversa, quanto più il valore di sensitivity è alto, tanto più è bassa la percentuale di falsi negativi. Pertanto, un elevato valore dell'area al di sotto della curva sta ad indicare una buona precisione ed una buona sensibilità del modello.

Il risultato è pittosto buono: il valore dell'area al di sotto della curva sembra essere sufficientemente grande. Notiamo, inoltre, che la precisione del modello si mantiene quasi sempre al di sopra del 70%, cosa che indica un'elevata accuratezza del nostro classificatore sui negativi.

Valutiamo, infine, capacità di predizione e robustezza del modello:



La capacità di predizione è abbastanza buona ma non ottimale: con una cross-validation otteniamo una media dell'accuratezza di circa 0.835 ed una deviazione standard di 0.042. Anche in questo caso, come per la regressione logistica, la deviazione standard è leggermente alta.
Notiamo, inoltre, che occorre introdurre più di 150 errori nella tabella di dati affinchè l'accuratezza scenda al di sotto dell' 80% e più di 200 per scendere al di sotto del 70%. Possiamo ritenerci abbastanza soddisfatti.

2.4 Modelli a confronto

L'analisi svolta mostra che la performance del modello di regressione logistica è migliore, seppur di poco, di quella del modello di analisi discriminante quadratica.

Osserviamo in particolare i seguenti fatti:

- In corrispondenza del parametro p che massimizza la grandezza sensitivity + specificity la regressione logistica riesce a raggiungere un valore di poco maggiore: 1,726 contro l' 1,705 dell'analisi discriminante quadratica. Va notato, però, che il secondo modello riesce a raggiungere rispetto al primo una maggiore percentuale di veri positivi: $85,6\%$ contro l' $80,1\%$ della regressione logistica;

- La capacità di predizione del modello di regressione logistica è migliore di quella del modello di analisi discriminante quadratica: la media dell'accuratezza del primo supera quella del secondo di 0.19, mentre la deviazione standard è inferiore di 0.004.

3 Conclusioni

Abbiamo implementato tre diversi modelli di classificazione (regressione logistica, analisi discriminante lineare e analisi discriminante quadratica) con l'obiettivo di predire, a partire dall'andamento del suo primo anno di gioco, se la durata della carriera di un giocatore nell'NBA sarebbe stata maggiore o minore di 5 anni. In particolare, abbiamo analizzato in dettaglio i modelli di regressione logistica e analisi discriminante quadratica e abbiamo visto che il primo supera, seppur di poco, il secondo sotto ogni aspetto.

Possiamo ritenerci soddisfatti del risultato ottenuto: il nostro modello riesce a predire la classe di un giocatore con l' $85,4\%$ di accuratezza. Pertanto, l'andamento del primo anno di gioco è un buon metro di giudizio per scelte riguardanti la compra-vendita oppure lo scambio di giocatori, in base a quelle che sono le esigenze della squadra. É chiaro che bisognerebbe tenere conto anche di altri fattori esterni (eventuali infortuni o simili) che variano di giocatore in giocatore, ma questo esula dall'obiettivo della nostra analisi.

4 Fonti

I dati utilizzati provengono dal sito ufficiale dell'NBA: `https://www.nba.com/stats/`.

Per le statistiche dei giocatori relative all'anno 2015-2016 abbiamo utilizzato la seguente tabella:

- `https://www.nba.com/stats/players/traditional/?sort=PTS&dir=-1&Season=2015-16&SeasonType=Regular%20Season`

Per ottenere quelle relative agli altri anni (le abbiamo considerate tutte a partire dall'anno 2002-2003 fino ad arrivare all'anno 2015-2016) basta modificare l'anno nell' URL precedente (2014-15, 2013-14 e così via).

Per conoscere chi sono i giocatori che hanno iniziato nel 2015 e la durata della loro carriera abbiamo utilizzato la seguente tabella:

- `https://www.basketball-reference.com/draft/NBA_2015.html`

Per ottenere gli stessi dati relativi agli anni che vanno dal 2002 al 2014 basta sostituire l'anno nell'URL precedente[4].

[4]La seconda tabella che ho riportato non proviene dal sito ufficiale dell'NBA. Le informazioni che racchiude sono comunque reperibili sul sito ufficiale ma non in maniera compatta (bisogna andare a guardare le statistiche personali di ciascun giocatore).