Eleonora Baracco. Relazione Indian Premier League

I dataset da analizzare riguardano la Premier League indiana di cricket. In “matches.csv ” troviamo i dati suddivisi per partita (dal 2008 al 2017) mentre in “deliveries.csv ” troviamo i dati riguardanti ogni palla (o battuta) di ogni partita, sempre dal 2008 al 2017.

Si è deciso di sostituire i valori nulli (che rappresentavano sempre stringhe) con la stringa “ --- ”, per indicare comunque che il dato è mancante. Tutte le volte che si calcola la frequenza di valori in cui sono presenti queste stringhe, la frequenza della stringa “ --- ” indica il numero di dati mancanti per quella variabile.

IL TORNEO

Come prima analisi ho deciso di individuare gli stadi e le città in cui si sono giocate più partite nell’ intero arco di tempo del dataset, cioè dal 2008 al 2017.

Gli stadi:

```
M Chinnaswamy Stadium                                66
Eden Gardens                                         61
Feroz Shah Kotla                                     60
Wankhede Stadium                                     57
Rajiv Gandhi International Stadium, Uppal            49
MA Chidambaram Stadium, Chepauk                      48
Punjab Cricket Association Stadium, Mohali           35
Sawai Mansingh Stadium                               33
Subrata Roy Sahara Stadium                           17
Dr DY Patil Sports Academy                           17
Maharashtra Cricket Association Stadium              15
Kingsmead                                            15
SuperSport Park                                      12
Sardar Patel Stadium, Motera                         12
Brabourne Stadium                                    11
Dr. Y.S. Rajasekhara Reddy ACA-VDCA Cricket Stadium  11
Punjab Cricket Association IS Bindra Stadium, Mohali 11
Saurashtra Cricket Association Stadium               10
Himachal Pradesh Cricket Association Stadium          9
New Wanderers Stadium                                 8
JSCA International Stadium Complex                    7
Barabati Stadium                                      7
Dubai International Cricket Stadium                   7
St George's Park                                      7
Sheikh Zayed Stadium                                  7
Newlands                                              7
Shaheed Veer Narayan Singh International Stadium      6
Sharjah Cricket Stadium                               6
Holkar Cricket Stadium                                5
Nehru Stadium                                         5
Green Park                                            4
Vidarbha Cricket Association Stadium, Jamtha          3
```

```
De Beers Diamond Oval                                3
Buffalo Park                                         3
OUTsurance Oval                                      2
```

Le città:

```
Mumbai            85
Bangalore         66
Kolkata           61
Delhi             60
Hyderabad         49
Chennai           48
Chandigarh        46
Jaipur            33
Pune              32
Durban            15
Centurion         12
Ahmedabad         12
Visakhapatnam     11
Rajkot            10
Dharamsala         9
Johannesburg       8
Ranchi             7
Port Elizabeth     7
Cuttack            7
Cape Town          7
Abu Dhabi          7
---                7
Raipur             6
Sharjah            6
Indore             5
Kochi              5
Kanpur             4
East London        3
Nagpur             3
Kimberley          3
Bloemfontein       2
```

LE SQUADRE

Le squadre sono state analizzate sia nell' intero arco temporale dei dataset sia anno per anno. Si è deciso di valutare il numero di partite giocate, il numero di partite vinte e perse, il tasso di vittoria e di perdita, il numero di volte in cui si è vinto il lancio della moneta. Tramite questi valori si comprende subito che non tutte le squadre giocano lo stesso numero di partite, infatti, stiamo analizzando un torneo e non un campionato(63 partite in media per anno, mediana, invece, 60) , di conseguenza le squadre che rimango di più in gara avranno un numero più alto di vittorie, ma anche di sconfitte. Osservando i dati generali possiamo subito classificare le squadre in base a chi ha vinto più partite nel corso degli anni o chi ha il tasso di vincita più alto. Un altro modo di trovare le squadre più forti (in generale) è quello di osservare il numero di volte in cui le squadre si sono classificate in alto

negli anni. Questa classificazione, però, non risulta del tutto corretta poiché non si basa sull'effettiva classifica del torneo, anno per anno, poiché non disponibile. Si è invece contato quante volte una squadra abbia vinto più partite alla fine della stagione e in questo modo si è definito quante volte una squadra si si classificata prima, seconda, terza o quarta. Purtroppo non risulta immediato come posizionare le squadre a parimerito di vittorie, soprattutto per le posizioni successive al primo posto. Infatti per la vetta della classifica si è deciso di controllare da chi fosse stata vinta l'ultima partita di ogni stagione. Si evince dunque che l'ultima partita della stagione sia presumibilmente la finale. Inoltre dal confronto dei due metodi di classificazione possiamo assegnare con sicurezza il primo posto per ogni annata, poiché effettivamente coincidono.

Ho deciso anche di calcolare le statistiche sulle vittorie, sia in base all' annate che in base alle squadre, e di rappresentarle tramite dei box-plot. Da queste statistiche si può osservare il numero di partite necessarie per vincere in un determinato anno, oppure si possono fare previsioni su quante partite si pensi possa vincere una squadra nella stagione successiva basandosi sui dati di questi dieci anni.

I GIOCATORI

Un altro valore interessante è l'assegnazione del titolo " player of the match " ad un giocatore di una delle due squadre in gioco. Tramite questo valore si sono valutati i giocatori migliori e il loro rendimento durante gli anni. Ciò che risulta evidente è che il titolo venga assegnato pochissime volte alla maggior parte dei giocatori. Infatti, in media ad ognuno dei 202 giocatori che compaiono come player of the match, il titolo viene assegnato 0,31 volte l'anno. In ogni caso ci sono giocatori che in un anno hanno ricevuto fino a 6 volte il titolo e altri anni in cui non l'hanno proprio ricevuto. Osservando quindi l'andamento di questi giocatori negli anni, si possono individuare ad esempio la stagione d'oro di un giocatore, l'ascesa di un giocatore giovane, il tramonto di un altro e così via. Ad esempio il giocatore CH Gayle, colui che ha ricevuto più volte il titolo in assoluto e che registra anche il record di volte in cui in un anno è stato assegnato quel titolo ad un giocatore, dal 2014 non sta più avendo grandissime prestazioni.

Proprio per lasciare la possibilità di osservare i giocatori che più interessano si è scelto d'inserire un widget per permettere a chi utilizza il notebook di scegliere quanti giocatori visualizzare, inoltre, avendo usato un grafico interattivo si possono selezionare oppure deselezionare i giocatori che si vogliono analizzare.

Il fatto che alcuni giocatori ricevano pochissime volte il titolo è coerente con il fatto che ci siano squadre che giocano poche partite poiché perdono nelle fasi iniziali del

torneo. Questa caratteristica si nota anche da quanto effettivamente giochino gli atleti. Infatti osservando le distribuzioni dei giocatori per battuta osserviamo immediatamente una distribuzione di Pareto per tutti e tre i ruoli che hanno spazio nel dataset (batsman, non-striker, bowler). Calcolando le statistiche di base sul ruolo batsman, infatti, osserviamo che la media non è una valore significativo per questo tipo di distribuzioni (massimo e minimo sono lontanissimi, lo scarto quadratico medio è molto alto). Invece, la mediana è in questo caso un valore rappresentativo, poiché effettivamente va a dividere i valori a metà e quindi possiamo inferire, ad esempio, che il 50 % dei battitori ha eseguito solo 70 battute nei 10 anni analizzati.

Per quanto riguarda i battitori, si nota che possono ricoprire in ogni over o il ruolo di batsman o quello di non striker (la correlazione tra queste variabili è pari ad 1 proprio perché sono gli stessi giocatori che ricoprono quei due ruoli e di conseguenza chi gioca molto in un ruolo giocherà molto anche nell'altro).

Per i giocatori in questo ruolo si può anche controllare quante volte sono stati mandati in out dalla squadra avversaria, questo perché la colonna " player dismissed" del dataset " deliveries.csv " indica proprio i giocatori mandati in out. Si è dunque deciso di osservare il numero di out per ogni giocatore e di trovare il tasso di out per giocatore. Questo valore risulta correlato negativamente sia con il numero di battute che con il numero di out.

COMPRENDERE IL GIOCO

Per comprendere meglio il gioco ho deciso di valutare l'influenza della vittoria del lancio della moneta e la conseguente vincita della partita, valutando poi così anche la scelta effettuata (se iniziare battendo o meno).

Per fare questa valutazione ho calcolato quante volte le squadre hanno vinto sia il lancio della moneta che la partita intera inoltre calcolando anche quante volte hanno scelto di battere e quante no, anche al fine di valutare le scelte delle squadre e capire quali squadre hanno più necessità di scegliere al fine di vincere e cosa scelgono.

Ciò che scaturito dall' analisi è che in generale la vittoria del lancio della moneta non influenza particolarmente la vittoria della partita. Infatti il numero di vittorie dopo aver vinto il lancio è poco più alto del numero di vittorie senza aver vinto il lancio. Andando però nello specifico di ogni squadra e calcolando il tasso di vittoria "fortunata " si nota che la maggior parte delle squadre ha comunque un tasso maggiore o uguale al 50%. Infatti la correlazione di queste due variabili risulta positiva e lo scatter plot creato con queste variabili mostra una retta ascendente.

Valutando ancora squadra per squadra possiamo notare che la squadra con il tasso più alto di vittorie “ fortunate ”, i “ Gujarat Lions ”, in tutte le partite in cui hanno potuto scegliere hanno preso la decisione di non essere i primi a battere. Si può dunque immaginare che quella squadra sia molto forte nei lanci o che abbia bisogno di battitori migliori.

Per osservare invece i tipo di punti ho creato un dataset con il numero di palle per partita e il numero di punti fatti nella partita diviso per tipo. Ho calcolato poi le statistiche per queste variabili e ho notato che in generale le squadre fanno molti più punti “ by run ” (massimo 146) piuttosto che “ by_wicked ” (massimo 10!) .

Possiamo inoltre facilmente osservare che la maggior parte delle partite si svolge in due inning, un’infima porzione arriva giocare il terzo e il quarto inning.

Nel dataset “ deliveries.csv ” abbiamo per ogni palla lanciata in che over si sta giocando e che numero di palla (in quell’over) si sta lanciando / battendo. Incrociando queste due colonne possiamo ottenere facilmente la frequenza del numero di palla per over. Innanzitutto si evince che i gli over sono 20 per inning Ad esempio, ciò che si nota immediatamente, anche dal box plot creato sul dataset, è che solitamente negli over si lanciano circa 6 palle, con un massimo di 9, infatti la frequenza di lanci, dalla sesta palla in poi diminuisce drasticamente. Probabilmente la settima, l’ottava e la nona palla vengono lanciate solo se ci sono stati lanci invalidi. Un altro fattore è quello di notare che anche il numero di palla lanciata (se la palla lanciata è la prima, la seconda o così via) influisce sulla frequenza delle palle lanciate, infatti ,notiamo che aumentando le palle diminuisce la frequenza anche se di poco, questo è imputabile al fatto che ci possano essere degli out, che fanno terminare l’over prima della fine dei lanci. Inoltre si nota che con il proseguire degli over la frequenza di palle lanciate diminuisce.

La correlazione tra le variabili è stata calcolata usando il coefficiente di Pearson, e sono state rappresentate tramite delle heatmap (tranne la correlazione tra vittorie e vittorie “fortunate” che è stata visualizzata tramite scatter plot).

Tramite le variabili del dataset “ matches.csv ”, a cui ho aggiunto il numero di palle per partita, ho compreso che c’è una relazione negativa tra il tipo di punteggio (win by run e win by wicked), inoltre, per quanto forse meno interessante, si nota una correlazione positiva tra l’id della partita e la stagione in cui si è giocata la partita . Per quanto riguarda il set che ho utilizzato per fare una presentazione generale delle squadre, vediamo che c’è una correlazione positiva tra partite giocate e numero di vittorie e numero di perdite, ma la correlazione tra il numero di partite giocate e i

tassi di vincita e perdita risulta invece invertita (positiva del 0,38 con il tasso di vincita e viceversa).

Come ultima analisi ho fatto un piccolo focus sulla squadra più forte, i “ Mumbai Indians” e visualizzando quante volte hanno vinto con un arbitro e osservando in percentuale le vittorie in cui ha potuto scegliere il ruolo iniziale e le scelte effettuate. Questa squadra ha vinto il 49% delle volte senza aver bisogno di scegliere e quando ha potuto e ha anche vinto, ha scelto il 21% delle volte di battere e il 30% di lanciare.