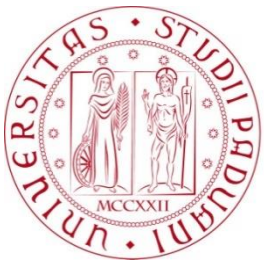
Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in

STATISTICA PER L'ECONOMIA E L'IMPRESA



RELAZIONE FINALE

PREVISIONE DELL'ESITO DI UNA PARTITA DI BASKET NBA

Relatore: Prof. Matteo Grigoletto
Dipartimento di Scienze Statistiche

Laureando: Raffaele William D'Agostino
Matricola: 1148195

Anno Accademico 2019/2020

Indice

Capitolo 1

Introduzione

Da sempre, per i tifosi e gli appassionati, lo sport è semplicemente questione di passione, emozioni e drammaticità, ovvero quelle componenti astratte che lo rendono un'esperienza piacevole e spettacolare. Se proviamo, invece, a spostare l'attenzione sulla prospettiva dei principali "attori" dello sport (atleti, allenatori, staff tecnici, società, ecc.), possiamo facilmente notare come quest'ultimo sia sinonimo di competizione. E come in tutti gli ambiti altamente competitivi, l'obiettivo principale degli sportivi è quello di ottenere un vantaggio competitivo sui concorrenti, che permetta loro di collezionare successi con costanza. A questo proposito, uno dei principali mezzi ai quali si ricorre con una sempre più crescente insistenza è sicuramente la statistica e l'attuazione di strategie basate sull'analisi dei dati raccolti dai diretti interessati.
Tra gli sport maggiormente influenzati dall'avvento delle "*Analytics*", ossia le analisi dei dati relativi alle principali variabili che determinano l'esito di un incontro sportivo, c'è sicuramente il basket, sport di cui ci occuperemo in quest'elaborato.

1.1 La statistica nel basket

Nello sport del basket, la statistica ha da sempre ricoperto un ruolo di una certa rilevanza, seppur nei primi decenni della sua esistenza i dati raccolti fossero piuttosto basilari. Nel caso specifico della NBA (acronimo di "*National Basketball Association*"), ovvero la più prestigiosa lega cestistica americana, per decenni l'utilizzo della statistica è stato limitato alla raccolta di dati rappresentanti i principali avvenimenti di una partita, anche chiamati dati da "*box score*". Tra questi troviamo il conteggio di alcune delle azioni più rilevanti di una partita di basket,

quali i punti segnati, gli assist forniti ai compagni di squadra, le palle perse, i falli commessi ed altre informazioni di questo tipo. E' però a cavallo tra gli anni '90 e i primi anni 2000 che la statistica comincia ad assumere un ruolo sempre più importante nel gioco del basket, grazie alla continua espansione relativa ai dati raccolti ed alla combinazione degli stessi volta ad evidenziare i principali fattori determinanti dell'esito di una partita. A questo proposito, un'importante rivoluzione è data da Oliver (2004), successivamente rivisitata in chiave accademica da Kubatko et al. (2007), con l'introduzione dei "*Four Factors*", ovvero le quattro variabili dal maggiore impatto sull'esito di un incontro. In seguito, molteplici studi basati sulle intuizioni sopracitate si sono susseguiti, quali Teramoto e Cross (2010), in cui i *Four Factors* vengono utilizzati per studiare la differenza presente tra le variabili più statisticamente significative di partite di *regular season* e *playoff*, e Štrumbelj e Vračar (2012), incentrato sulla simulazione e previsione dell'esito di partite di basket NBA mediante l'utilizzo dei *Four Factors* ed un modello di Markov. Ad approfondire, poi, la natura dinamica di quest'ultimo troviamo, tra gli altri, Štrumbelj et al. (2016) e Manner (2016).
Il presente elaborato si pone l'obiettivo di individuare, per ciascuna squadra, le variabili che influenzano maggiormente l'esito di una partita, così da poter basare l'applicazione dei vari modelli previsivi su tali variabili specifiche, invece che sui più generici *Four Factors*.

1.2 Le principali variabili di una partita di basket

Nel seguente paragrafo verranno presentati gli eventi più comuni e di maggior rilievo di un incontro cestistico. Ciascuna variabile è associata ad una precisa azione di gioco, la quale influisce inevitabilmente sull'andamento della partita permettendoci, quindi, di quantificare l'impatto di ciascuna di esse su quest'ultima. Per questo elaborato verranno utilizzate principalmente le seguenti variabili, ciascuna rilevata per una singola squadra in un singolo incontro:

- **PTS** n. punti messi a segno dalla squadra
- **FGA** n. tiri dal campo tentati

- **FGM** n. tiri dal campo segnati
- **FG%** percentuale di tiri dal campo segnati (FGM/FGA)
- **FG3A** n. tiri da 3 punti tentati
- **FG3M** n. tiri da 3 punti segnati
- **FG3%** percentuale di tiri da 3 punti segnati (FG3M/FG3A)
- **FTA** n. tiri liberi tentati
- **FTM** n. tiri liberi segnati
- **FT%** percentuale di tiri liberi segnati (FTM/FTA)
- **AST** n. di assist (ossia passaggi direttamente seguiti da un FGM)
- **BLK** n. di tiri avversari stoppati
- **STL** n. di palle rubate agli avversari
- **TOV** n. di palle perse
- **OREB** n. di rimbalzi offensivi (ossia palle recuperate dopo un FGA della propria squadra)
- **DREB** n. di rimbalzi difensivi (ossia palle recuperate dopo un FGA della squadra avversaria)

Tali variabili, fatta eccezione per **FG%**, **FG3%** e **FT%** che sono ovviamente combinazioni di altri caratteri, possono essere visti come le rilevazioni dei principali eventi che si osservano durante una partita di basket, e quindi come dati grezzi. Quest'ultimi possono essere sia utilizzati direttamente per analizzare l'andamento di un incontro, sia combinati tra loro per creare variabili capaci di fornire informazioni più utili per la medesima analisi. Ne sono un esempio i tre caratteri percentuali sopra riportati, i quali forniscono una misura adimensionale dell'efficienza di tiro, e i precedentemente citati *Four Factors*, che altro non sono che 4 combinazioni delle variabili di cui sopra, volte ad enfatizzare i concetti di efficienza di tiro, frequenza delle palle perse, frequenza dei tiri liberi e capacità di ottenere rimbalzi offensivi.

1.3 Il Dataset

Per il presente elaborato sono state raccolte le principali statistiche relative a ciascuna partita di *regular season* della NBA svoltasi nelle stagioni che vanno dalla 2013-14 alla 2018-19. Tali dati sono stati ricavati dal sito https://stats.nba.com/ , ovvero il portale online ufficiale della NBA contenente tutte le statistiche rilevate durante gli incontri della lega. La numerosità campionaria del dataset che andremo ad utilizzare è di 14.760, che corrisponde al doppio del numero di partite giocate nel periodo di riferimento (in quanto in ciascuna partita si incontrano due squadre ed abbiamo bisogno delle statistiche di entrambe). Le variabili che andremo ad analizzare (ed eventualmente ad inserire nel nostro modello) sono quelle riportate nel precedente paragrafo. Inoltre, per ciascuna unità statistica, insieme alle variabili relative alla squadra di riferimento sono riportate anche le medesime variabili relative alla squadra avversaria, indicate col suffisso "_opp". In questo modo potremo usufruire di una più precisa rappresentazione dell'impatto della difesa (intesa come il limitamento delle potenzialità offensive della squadra avversaria) sull'esito di un incontro.

Capitolo 2

Individuazione delle variabili di maggiore impatto

Nel seguente capitolo andremo a cercare di individuare le variabili che riteniamo possano avere un'incidenza fondamentale sull'andamento di una partita, provando a specificare un modello statistico che si adatti in maniera corretta e parsimoniosa ai dati di cui disponiamo, e che produca risultati soddisfacenti in termini di studio e previsione della nostra variabile di interesse, in questo caso l'esito della partita.

2.1 Analisi esplorativa dei dati

Il nostro primo compito propedeutico al perseguimento dell'obiettivo è quello di conoscere ed approfondire le proprietà che caratterizzano le variabili che abbiamo osservato. Un primo accorgimento è quello di verificare l'eventuale presenza di variabili ridondanti e/o poco utili allo studio che andremo a svolgere di seguito. Una prima esclusione, dettata esclusivamente dalla logica, è quella delle variabili relative ai punti segnati rispettivamente dalla squadra oggetto dell'unità statistica e da quella avversaria. Tale esclusione è giustificata dal fatto che conoscere la quantità di punti messi a segno dalle due squadre di un incontro equivale a conoscere l'esito dello stesso, per cui sarebbe quindi inutile ed insensato includere nel nostro modello una variabile sostanzialmente equivalente a quella che intendiamo studiare.

Veniamo ora alla ricerca di eventuali variabili ridondanti, che comportano una situazione problematica detta *multicollinearità*. La multicollinearità (perfetta) è una condizione che si verifica qualora una variabile indipendente sia una funzione lineare esatta di una o più altre variabili indipendenti. L'immediata conseguenza è che un eventuale modello di regressione basato su predittori linearmente dipendenti non è identificato. Nel caso in cui disponiamo di variabili esplicative linearmente indipendenti, ma vicine alla dipendenza lineare (ossia una variabile indipendente è *approssimativamente* una funzione lineare di una o più altre variabili esplicative), ci troviamo in una situazione di multicollinearità (non perfetta). In questo caso il modello di regressione esiste e tutte le sue assunzioni sono formalmente soddisfatte, ma diventa pressoché impossibile ottenere stime accurate dei coefficienti dei regressori in quanto, a causa di un'elevata varianza degli stimatori, piccole variazioni nei dati possono comportare variazioni notevoli in suddette stime (Grigoletto et al., 2017). Uno dei metodi che possiamo utilizzare per ovviare a tale problema consiste nell'accertarci di non disporre di caratteri altamente correlati e, nel caso in cui fossero presenti, valutare con attenzione e criterio la loro esclusione dallo studio.
Utilizziamo, quindi, l'indice di correlazione, che viene così calcolato:

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \ ,$$

dove $\sigma_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$ è la covarianza tra le variabili esplicative X e Y, $\sigma_X^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$ e $\sigma_Y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}$ sono rispettivamente le varianze di X e Y, e l'indice di correlazione ρ è un valore compreso tra -1 ed 1 che osserva la relazione lineare presente tra le due variabili.

Il seguente grafico (*Figura 2.1*) mostra il grado di correlazione per ciascuna coppia di variabili.

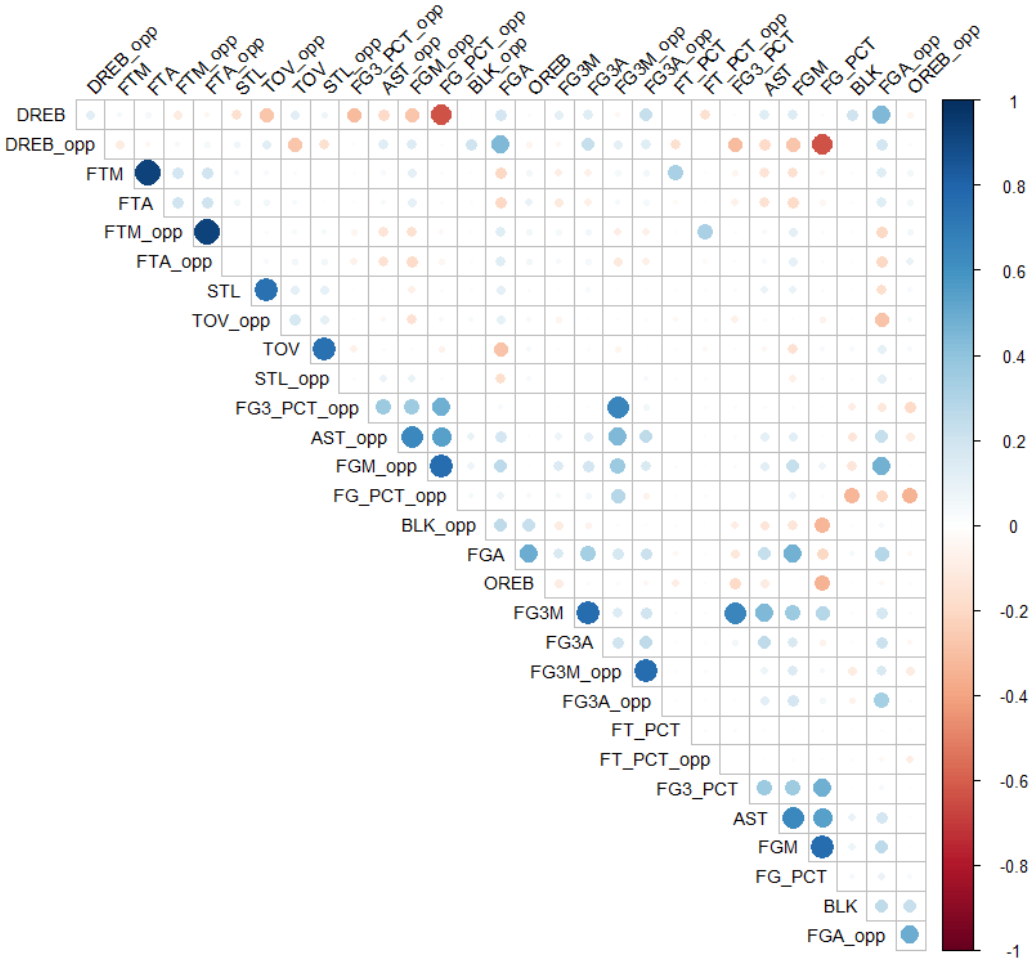


Figura 2.1: Matrice di correlazione delle variabili. Il colore ed diametro dei cerchi indicano il livello di correlazione presente tra le due variabili considerate

Com'è possibile vedere dal grafico, il nostro dataset presenta molteplici variabili altamente correlate. È quindi ora necessario analizzare più attentamente le relazioni presenti tra le variabili maggiormente correlate, così da poter decidere accuratamente se e quali di esse riteniamo opportuno escludere dal modello che proporremo in seguito. La prime relazioni possibilmente problematiche che andremo ad esplorare sono quelle relative alle statistiche di tiro. Come abbiamo infatti già spiegato nel primo capitolo, le variabili FG% , FG3% e FT% altro non sono che il rapporto tra i tiri messi a segno e quelli tentati per le rispettive tipologie di tiro. È quindi evidente che ci sia una forte ridondanza tra i caratteri in questione, e questo ci costringe a doverne escludere alcuni. I risultati della matrice di correlazione, insieme ai diagarammi di dispersione riportati nella *Figura 2.2*, ci indicano una dipendenza generalmente forte per le variabili di tipo "tiri segnati" sia

con i rispettivi tiri tentati (in quanto si possono segnare solo tiri effettivamente tentati), sia con le relative percentuali al tiro (in quanto l'efficienza di tiro è direttamente collegata al successo dello stesso).

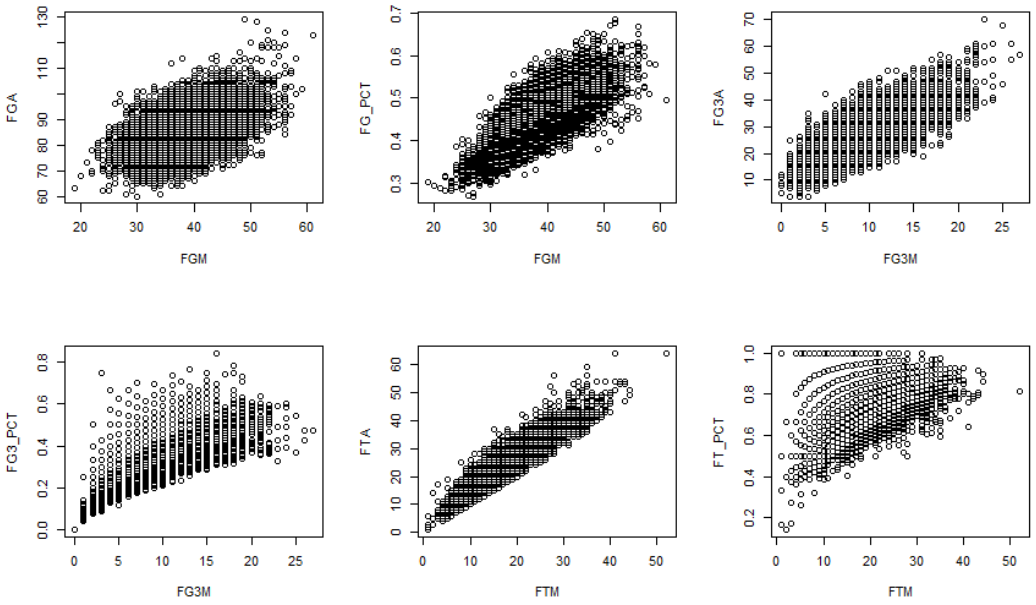


Figura 2.2: Diagrammi di dispersione rispettivamente tra
FGM-FGA, FGM-FG%, FG3M-FG3A, FG3M-FG3%, FTM-FTA, FTM-FT%

Ragionevolmente, decidiamo quindi di escludere le variabili FGM, FG3M e FTM (e le controparti della squadra avversaria) dal nostro studio. Una simile relazione è presente tra i caratteri STL e TOV_opp (e ugualmente tra STL_opp e TOV) (*Figura 2.3)*, in quanto ogni palla rubata corrisponde ad una palla persa della squadra avversaria. Inoltre, una palla persa può avvenire anche senza che venga registrata una palla rubata per la squadra avversaria. Per questo motivo, considerando che TOV e TOV_opp contengono maggiore informazione, decidiamo di escludere le variabili STL e STL_opp.

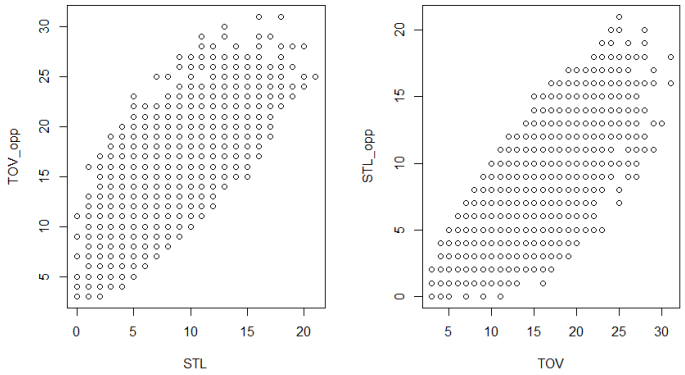


Figura 2.3: Diagrammi di dispersione rispettivamente tra STL-TOV_opp e TOV-STL_opp

Restano da valutare le relazioni tra DREB e FG%_opp e tra AST e FG% . Il grafico in *Figura 2.4* sottolinea una forte correlazione per entrambe le coppie di variabili, ma è doveroso chiedersi le motivazioni di tali relazioni prima di valutare eventuali esclusioni dal modello. La dipendenza lineare osservata nella prima coppia di variabili è frutto della natura del gioco, che prevede la possibilità di ottenere un rimbalzo difensivo solo in seguito ad un tiro sbagliato dagli avversari, per cui, pur essendo correlate, le due variabili osservano due aspetti dell'incontro sostanzialmente diversi e ,dunque, decidiamo di non escludere nessuna delle due. Un discorso analogo si può fare per la relazione tra AST e FG% , motivo per il quale optiamo per la loro inclusione, almeno per il momento, nel modello che andremo ora a vedere.

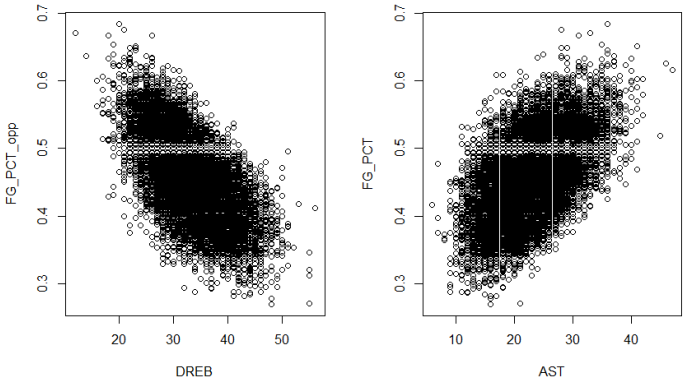


Figura 2.4: Diagrammi di dispersione rispettivamente tra DREB-FG_PCT_opp e AST-FG_PCT

2.2 Modelli di Regressione

2.2.1 Il modello di Regressione Lineare

Molto spesso, nello studio di un determinato fenomeno, un aspetto a cui si è particolarmente interessati è la relazione tra una specifica variabile (chiamata dipendente o risposta) ed una o più altre variabili (indipendenti o esplicative). Alla base di tutti i possibili modelli potenzialmente utili a rispondere a tale quesito abbiamo quello di *Regressione Lineare*. Se chiamiamo Y la variabile risposta e $(x_1, \dots, x_p)$ le variabili esplicative, e disponiamo di n osservazioni del fenomeno d'interesse, possiamo specificare il modello di regressione lineare attraverso i seguenti assunti:

- $Y_i = \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \epsilon_i$, per la generica osservazione $i \in \{1, \dots, n\}$,

  da cui ricaviamo la forma matriciale $Y = X\beta + \epsilon$ per le n osservazioni:

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} ,$$

  dove ϵ indica l'*errore*, β è il vettore dei coefficienti di regressione (o parametri), e X è una matrice di costanti note $n \times p$, con rango pieno $p < n$.

- $\epsilon \sim N_n(0, \sigma^2 I_n)$ con $\sigma^2 > 0$, I_n matrice identità
- I vettori $x_j \in \mathbb{R}^n$, $j = 1, \dots, p$, sono linearmente indipendenti
- $E(Y_i) = \mu_i = \beta^T x_i = \beta_1 x_{i1} + \dots + \beta_p x_{ip}$,

  dove $\mu_i = \beta^T x_i$ è detto **predittore lineare.**

Dagli assunti di cui sopra ricaviamo che

$$Y \sim N_n(X\beta, \sigma^2 I_n).$$

La regressione lineare è uno strumento importantissimo della statistica, in quanto ci permette di fare inferenza sull'entità delle diverse relazioni presenti tra le variabili che intendiamo studiare. Il modello ha infatti posto le fondamenta per lo studio di tali relazioni, ed è stato successivamente esteso attraverso la specificazione di nuovi e più complessi modelli.

2.2.2 I modelli lineari generalizzati

Il modello di regressione lineare, seppur molto basilare, è estremamente diffuso per via della sua semplicità di implementazione e per la sua capacità di risolvere una vasta gamma di problemi di inferenza. Il modello presenta, però, alcuni limiti. Uno dei limiti principali è quello dovuto alla condizione imposta dallo stesso per la distribuzione della variabile risposta. La regressione lineare semplice, infatti, è correttamente specificata solo quando Y è normalmente distribuita. Questo implica una sostanziale impossibilità di applicazione del modello per tutti i casi in cui Y segue una distribuzione diversa dalla Normale (ad esempio Poisson, Binomiale, Gamma, ecc.). Una soluzione a tale problema è data dall'implementazione di *modelli lineari generalizzati*. Quest'ultimi sono caratterizzati dalla seguente specificazione:

- Si assume per la variabile risposta una distribuzione della forma

$$f(y_i; \theta_i, v) = exp\left\{\frac{y_i\theta_i - c(\theta_i)}{v} + b(y_i, v)\right\}, \qquad i = 1, \dots, n \ ,$$

  con $b(\cdot)$ e $c(\cdot)$ funzioni note, $\theta_i \in \mathbb{R}$ e $v > 0$.
  Quando v è noto tale classe viene chiamata *famiglia esponenziale* e θ_i , $i = 1, \dots, n$, sono i *parametri naturali*

- $g\big(E(Y_i)\big) = \beta^T x_i$,

  dove $g(\cdot)$ è una funzione invertibile e derivabile.

La funzione $g(\cdot)$ viene chiamata *funzione di legame* ed il suo compito è quello di spiegare il legame presente tra il valore atteso della variabile risposta e le variabili esplicative (tramite il predittore lineare).
Un'applicazione particolarmente interessante dei modelli lineari generalizzati è data da quei casi in cui la variabile dipendente segue una distribuzione discreta. Per tali distribuzioni il metodo utilizzato per fare inferenza sui parametri è quello della massima verosimiglianza. Quando $v = 1$ possiamo esprimere la funzione di probabilità di Y nella seguente forma:

$$p(y;\theta) = exp\{y\theta - c(\theta) + b(y)\},$$

con $b(\cdot)$ e $c(\cdot)$ funzioni note e $\theta \in \mathbb{R}$, da cui ricaviamo (Casella e Berger, 2002)

$$E(Y) = \mu = c'(\theta) \qquad \text{e} \qquad V(Y) = c''(\theta).$$

Essendo $V(Y) = c''(\theta)$, abbiamo che $c''(\theta) > 0$. Questo comporta che $c'(\theta)$ sia strettamente crescente per θ, e quindi invertibile. Per cui possiamo ora riparametrizzare la famiglia di distribuzioni per la media μ invece che per θ, tramite la funzione

$$\theta(\mu) = c'^{-1}(\mu),$$

dalla quale ricaviamo anche la *funzione di varianza*

$$V(\mu) = V(Y) = c''(\theta) = c''(\theta(\mu)).$$

Se consideriamo il vettore di osservazioni indipendenti $y = (y_1, \dots, y_n)$, dove ogni y_i proviene dalla famiglia esponenziale sopra specificata con parametri θ_i e media $\mu_i = c'(\theta_i),\ i = 1, \dots, n$, la funzione di log-verosimiglianza per $\theta = (\theta_1, \dots, \theta_n)$, basata su y, è così calcolata

$$\ell(\theta) = \sum_{i=1}^{n} (y_i\theta_i - c(\theta_i)).$$

Possiamo ora definire la struttura dei parametri θ_i imposta da un modello di regressione con variabile dipendente discreta con:

$$g(\mu_i) = g\big(c'(\theta_i)\big) = \beta^T x_i \,.$$

In questo modo rendiamo facilmente visibile il legame tra il vettore dei coefficienti β, θ_i e μ_i. Dalle definizioni precedenti possiamo scrivere la funzione di log-verosimiglianza per β come

$$\ell(\beta) = \sum_{i=1}^{n} \Big(y_i \theta(\mu_i) - c\big(\theta(\mu_i)\big)\Big) = \sum_{i=1}^{n} \big(y_i \theta\big(g^{-1}(\beta^T x_i)\big) - c(\theta(g^{-1}(\beta^T x_i))\big) \,.$$

Tra le plausibili funzioni di legame $g(\cdot)$ ve ne è una solitamente privilegiata, ossia quella per cui $g(\cdot) = \theta(\cdot)$, che implica

$$g(\mu_i) = \theta(\mu_i) = \theta_i = \beta^T x_i \;, \qquad i = 1, \dots, n \,.$$

La funzione $g(\cdot) = \theta(\cdot)$ viene chiamata **funzione di legame canonica**, e come abbiamo mostrato è caratterizzata dal fatto il parametro naturale θ_i coincide col predittore lineare $\beta^T x_i$. Per quanto riguarda invece l'inferenza sui parametri β, otteniamo le stime risolvendo il sistema di p equazioni di verosimiglianza ottenute ponendo la funzione di score uguale a zero, ossia

$$\ell'(\beta) = \frac{\partial \ell(\beta)}{\partial \beta} = 0 \,.$$

Generalmente la soluzione di tale sistema non è ottenibile in forma esplicita e si ricorre, dunque, all'utilizzo di algoritmi iterativi (per esempio Newton-Raphson) per il calcolo numerico delle stime.

Attraverso la specificazione di un modello lineare generalizzato siamo, quindi, capaci di estendere l'utilizzo della regressione a dati che presentano variabili dipendenti che seguono diversi tipi di distribuzione. Nel prossimo paragrafo

andremo ad esplorare il modello di regressione più diffuso per lo studio di variabili dipendenti distribuite come una Binomiale.

2.2.3 Il modello di Regressione Logistica

Nei precedenti paragrafi abbiamo effettuato delle operazioni preliminari di analisi esplorativa dei dati e di esclusione delle variabili che riteniamo possano essere le più problematiche, per via della loro ridondanza, nella buona riuscita del nostro studio. Adesso ci troviamo di fronte ad una scelta cruciale per l'analisi che intendiamo portare a termine. È in questo momento, infatti, che dobbiamo scegliere un modello statistico adatto alle nostre esigenze, e per fare questo dobbiamo conoscere bene ciò che vogliamo studiare. Nel nostro caso l'obiettivo è quello di studiare la relazione presente tra una variabile dipendente (l'esito della partita) e un set di variabili indipendenti costituito dai restanti caratteri considerati. Inoltre, sappiamo che la variabile dipendente è dicotomica, in quanto una partita ammette solamente due tipi di esito: vittoria o sconfitta. Tutte queste informazioni ci portano a pensare che il modello di *regressione logistica* possa essere adeguato alle nostre esigenze.

Il modello di regressione logistica è un modello statistico appartenente alla classe dei modelli lineari generalizzati, introdotti nel precedente paragrafo. Nello specifico, la regressione logistica si presta a risolvere problemi di stima dell'impatto di ciascuna variabile esplicativa su una variabile risposta bernoulliana, ovvero una variabile che permette due soli esiti.

Se assumiamo che le osservazioni della variabile risposta $y_1, \dots, y_n$ siano realizzazioni di variabili casuali indipendenti di Bernoulli, abbiamo che $E(Y_i) = \pi_i$ per $i = 1, \dots, n$, dove π_i è la probabilità di successo (nel nostro caso di vittoria della partita) per la variabile casuale Y_i. Se consideriamo il vettore delle variabili indipendenti $(x_1, \dots, x_p)$ e specifichiamo la funzione di legame canonica del nostro modello lineare generalizzato come $g(\pi) = \log(\frac{\pi}{1-\pi})$, otteniamo il modello di regressione logistica così specificato:

$$g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} \ , \qquad i = 1, \dots, n \ ,$$

da cui ricaviamo anche che

$$\pi_i = \frac{e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}}} .$$

Una volta specificato il modello, ci poniamo l'obiettivo di fare inferenza sui coefficienti $\beta_0, \beta_1, \ldots, \beta_p$, i quali descrivono le relazioni presenti tra le corrispondenti variabili indipendenti e la probabilità di successo della variabile dipendente. Il metodo più diffuso per la stima di suddetti parametri è quello della massima verosimiglianza. Quest'ultimo, nel caso della regressione logistica, viene solitamente applicato tramite l'utilizzo di algoritmi numerici, per via della natura della funzione di verosomiglianza del modello che risulta difficile da elaborare analiticamente. Nel presente elaborato ci affideremo al software statistico R per risolvere tutti gli aspetti computazionali del caso, permettendoci di ottenere risultati con la maggior precisione possibile.

2.3 Applicazione del modello

Nei precedenti paragrafi abbiamo definito le variabili esplicative ed il modello statistico che riteniamo essere più adeguato a rispondere al nostro quesito iniziale, ossia quello di stabilire i caratteri più rilevanti nella determinazione dell'esito di una partita. Formalizzando quindi il nostro modello, consideriamo l'esito della partita come variabile risposta dicotomica e le restanti variabili come regressori del nostro modello di regressione logistica. Ricordiamo che il nostro obiettivo è quello di trovare le variabili maggiormente influenti sul risultato di una partita specificatamente a ciascuna squadra. Come primo esperimento, però, adattiamo il modello ai dati indipendentemente dalla squadra, in modo da ottenere un primo risultato "universale" da poter poi comparare con i risultati storicamente più diffusi apportati dai *Four Factors*. Un ultimo accorgimento prima di procedere con l'applicazione del modello è quello di dividere il nostro dataset in due subset, così da poter adattare la regressione ai dati di un singolo subset ed in seguito sfruttare i

dati presenti nell'altro per valutare l'accuratezza previsiva del modello al di fuori del campione su cui quest'ultimo è stato adattato. Decidiamo di utilizzare un subset contenente le stagioni dalla 2013-14 alla 2017-18 per adattare il modello, mentre la stagione più recente (2018-19) verrà utilizzata per testarne l'accuratezza previsiva. A questo punto siamo pronti per procedere all'applicazione del nostro primo modello, contenente tutte le variabili esplicative a nostra disposizione:

predictor	estimate	std.error	z value	p.value
FGA	0.98	0.05	19.80	≈ 0
FG_PCT	214.08	8.46	25.30	≈ 0
FG3A	0.41	0.02	20.78	≈ 0
FG3_PCT	28.02	1.34	20.88	≈ 0
FTA	0.90	0.04	24.01	≈ 0
FT_PCT	26.27	1.22	21.51	≈ 0
OREB	0.12	0.04	3.04	2e-03
DREB	0.07	0.03	2.31	0.02
AST	0.04	0.02	1.96	0.05
BLK	0.01	0.03	0.36	0.72
TOV	-0.19	0.04	-5.01	5e-07
FGA_opp	-0.98	0.05	-19.80	≈ 0
FG_PCT_opp	-214.08	8.46	-25.30	≈ 0
FG3A_opp	-0.41	0.02	-20.78	≈ 0
FG3_PCT_opp	-28.02	1.34	-20.88	≈ 0
FTA_opp	-0.90	0.04	-24.01	≈ 0
FT_PCT_opp	-26.27	1.22	-21.51	≈ 0
OREB_opp	-0.12	0.04	-3.04	2e-03
DREB_opp	-0.07	0.03	-2.31	0.02
AST_opp	-0.04	0.02	-1.96	0.05
BLK_opp	-0.01	0.03	-0.36	0.72
TOV_opp	0.19	0.04	5.01	5e-07

Tabella 2.1: Tabella dei coefficienti del primo modello

Come possiamo vedere dalla tabella ci sono alcuni coefficienti a cui è associato un p-value maggiore di 0.05, il che ci indica che le corrispondenti variabili non sembrano essere statisticamente significative per quanto concerne la determinazione dell'esito della partita. Dobbiamo ora stabilire se e quali delle quattro variabili incriminate (AST, BLK, AST_opp, BLK_opp) possiamo escludere senza apportare cambiamenti significativi al modello. Nel seguente paragrafo andremo ad introdurre gli strumenti che ci permetteranno di compiere tale scelta in maniera oculata.

2.3.1. Bontà di adattamento: Devianza Residua e Accuratezza

Un metodo che andremo ad utilizzare per comparare i diversi modelli e la loro bontà di adattamento ai dati è quello dell'Analisi della varianza (ANOVA), e nello specifico andremo ad effettuare dei test di verifica d'ipotesi basati sulla differenza tra le devianze residue di due modelli. Siamo interessati a decidere quale modello sia preferibile tra due modelli che prevedono l'utilizzo di un numero diverso di variabili esplicative, dobbiamo quindi verificare la seguente ipotesi:

$$\begin{cases} H_0 : modello\ con\ p_0\ parametri \\ H_1 : modello\ con\ p_1\ parametri \end{cases},$$

dove $p_0 < p_1$.

Se consideriamo un modello saturo, ossia un modello in cui il numero di parametri *p* è uguale al numero di unità statistiche *n* del dataset, possiamo ricavare la *Devianza Residua* del modello che intendiamo utilizzare tramite il seguente test di rapporto di log-verosimiglianza:

$$Devianza\ Residua_{M_c} = 2\left(\widehat{\ell(M_s)} - \widehat{\ell(M_c)}\right),$$

dove M_s ed M_c sono rispettivamente il modello saturo e quello corrente (ossia quello che intendiamo adattare ai dati).

Per cui, tornando al nostro problema di verifica d'ipotesi, si può dimostrare che la differenza tra le devianze residue dei modelli con rispettivamente p_1 e p_0 parametri ignoti è equivalente a

$$TRV = 2\left(\widehat{\ell(H_1)} - \widehat{\ell(H_0)}\right) \ \dot{\sim}\ \chi^2_{(p-p_0)}\ , \qquad sotto\ H_0\ .$$

Questo ci permette di utilizzare la differenza tra le devianze residue dei due modelli come statistica test utile a verificare la nostra ipotesi ad un determinato livello di significatività approssimato. Un'interpretazione pratica del risultato di tale test è

quella di stabilire se il modello più complesso è significativamente migliore (ovvero si adatta meglio ai dati) di quello con meno variabili esplicative.
Proviamo ora a comparare il nostro modello con il modello nullo, ossia il modello contenente solamente l'intercetta (e quindi nessuna variabile indipendente).

Resid..Df	Resid..Dev	df	Deviance	p.value
12299	17051.42			
12278	1184.01	21	15867.41	≈ 0

Tabella 2.2: Tabella ANOVA comparativa tra modello nullo ed il nostro modello

Com'era facilmente intuibile, il test basato sulla differenza tra devianze residue mostra una forte evidenza contro l'ipotesi nulla, portandoci a preferire nettamente il nostro modello a discapito del modello nullo.
Un altro degli aspetti che maggiormente ci interessano relativamente alla bontà di adattamento del modello è vedere la sua capacità nel prevedere il risultato di una partita dati i valori delle variabili indipendenti. Per fare ciò andremo ad utilizzare i dati relativi alla stagione 2018-19 che, ricordiamo, non sono stati presi in considerazione in fase di adattamento del modello. Il procedimento consiste nell' utilizzare il vettore dei coefficienti $(\widehat{\beta_1}, \widehat{\beta_2}, \dots, \widehat{\beta_p})$, stimati in precedenza dal modello, per ottenere una stima della probabilità di vincere una partita dati i valori dei regressori. Tale stima viene ottenuta grazie alla formula:

$$\widehat{\pi_i} = \frac{e^{\widehat{\beta_1} x_{i1} + \dots + \widehat{\beta_p} x_{ip}}}{1 + e^{\widehat{\beta_1} x_{i1} + \dots + \widehat{\beta_p} x_{ip}}} \quad , \quad i = 1, \dots, n \ .$$

Una volta ottenuta $\widehat{\pi_i}$ dobbiamo decidere in che modo stabilire se la probabilità stimata è a favore del successo o meno. Uno strumento utile a tale scopo è la curva ROC, ovvero un grafico che associa ad ogni possibile valore di soglia per $\widehat{\pi_i}$ la rispettiva coppia di valori formata dai veri positivi (ovvero i successi previsti correttamente) ed i falsi positivi (i successi non previsti correttamente). Nel nostro caso, però, preferiamo fissare una sola soglia, sia per semplicità, che per il fatto che utilizzare tutte le soglie implica che la valutazione sia basata anche su soglie molto

piccole o molto grandi, che sono poco realistiche. Optiamo, dunque, per la scelta di $\widehat{\pi_i} = 0.5$ come valore di soglia, per cui decidiamo di assegnare una vittoria per ogni unità statistica la cui probabilità di successo stimata è maggiore o uguale a 0.5; viceversa assegniamo una sconfitta in caso contrario. Arriviamo ora al punto in cui compariamo i risultati che abbiamo ottenuto utilizzando le stime fornite dal modello con quelli effettivamente osservati nel dataset. È facile capire come un modello abbia maggiori capacità previsive quante più sono le osservazioni della variabile risposta che vengono previste correttamente. Un indice che ci aiuta a comprendere la capacità previsiva di un modello è quello dell'*accuratezza*, ossia il rapporto tra il numero delle previsioni corrette e la numerosità campionaria del subset considerato. Nel caso del modello da noi adottato otteniamo i seguenti risultati:

	OSSERVATI	
PREVISIONE	0	1
0	1202	28
1	28	1202
		1230

Tabella 2.3: Tabella di contingenza per le previsioni del primo modello.

Alla tabella sopra riportata si associa un livello di accuratezza di 0.9772, il che significa che il nostro modello ha previsto correttamente il 97.72% degli esiti delle partite della stagione 2018-19.

2.3.2. Riduzione della complessità del modello

A primo impatto ci verrebbe da dire che quest'ultimo è un ottimo risultato, e tutto sommato lo è sicuramente. Una domanda che però vogliamo porci è se sia possibile ottenere risultati di valore simile utilizzando meno variabili esplicative. Questo porterebbe una serie di vantaggi tra cui quelli di una diminuzione della probabilità di incorrere in problemi di ridondanza delle variabili o di "overfitting", in cui si corre il rischio di studiare anche la componente casuale di errore naturalmente presente nel fenomeno osservato. L'esclusione di alcuni regressori non viene, però, senza i suoi svantaggi. Il più importante, dal nostro punto di vista, è quello di una

possibile (ma non per forza presente) riduzione dell'accuratezza previsiva, per cui è fondamentale trovare il giusto compromesso per mantenere una capacità previsiva elevata col minor numero di variabili esplicative possibile.
Proviamo ora, quindi, ad escludere le variabili che abbiamo visto in precedenza avere p-value maggiore o uguale a 0,05, verificando allo stesso tempo che il nuovo modello non sia significativamente diverso da quello iniziale. Otteniamo il seguente modello:

predictor	estimate	std.error	z value	p.value
FGA	0.99	0.05	20.3	≈ 0
FG_PCT	216.80	8.44	25.7	≈ 0
FG3A	0.42	0.02	21.4	≈ 0
FG3_PCT	28.19	1.33	21.1	≈ 0
FTA	0.91	0.04	24.1	≈ 0
FT_PCT	26.28	1.22	21.6	≈ 0
OREB	0.12	0.04	3.0	2e-03
DREB	0.07	0.03	2.3	0.02
TOV	-0.19	0.04	-5.0	5e-07
FGA_opp	-0.99	0.05	-20.3	≈ 0
FG_PCT_opp	-216.80	8.44	-25.7	≈ 0
FG3A_opp	-0.42	0.02	-21.4	≈ 0
FG3_PCT_opp	-28.19	1.33	-21.1	≈ 0
FTA_opp	-0.91	0.04	-24.1	≈ 0
FT_PCT_opp	-26.28	1.22	-21.6	≈ 0
OREB_opp	-0.12	0.04	-3.0	2e-03
DREB_opp	-0.07	0.03	-2.3	0.02
TOV_opp	0.19	0.04	5.0	5e-07

Tabella 2.4: Tabella dei coefficienti del modello senza AST, AST_opp, BLK, BLK_opp

Vediamo subito che il modello non presenta coefficienti non significativi ad un livello di significatività di 0,05 (DREB e DREB_opp lo sono ad un livello di significatività 0,01). Per cui disponiamo ora di un modello in cui tutti regressori sembrano essere statisticamente significativi (per lo meno ad un livello dello 0.05). Dobbiamo ora verificare che il modello abbia una bontà di adattamento adeguata. Come prima verifica compariamo i due modelli visti finora (Tabella 2.5).

Resid..Df	Resid..Dev	df	Deviance	p.value
12282	1190.60			
12278	1184.01	4	6.59	0.16

Tabella 2.5: Tabella ANOVA comparativa tra
il primo ed il secondo modello.

Il valore del p-value ci indica che la differenza tra le devianze residue dei due modelli non è statisticamente significativa, ed è quindi preferibile utilizzare il modello meno complesso. Inoltre, otteniamo un livello di accuratezza *out of sample* (d'ora in poi intenderemo sempre quella calcolata fuori dal campione) del 97.64%, che è pressoché identico a quello del primo modello, confermando la preferenza per il secondo modello. Possiamo continuare utilizzando questo approccio, ossia rimuovendo mano a mano le variabili meno significative, ed arrivare a raggiungere un modello che soddisfi le nostre esigenze sia previsive che di semplicità di implementazione. Un altro metodo può essere quello di procedere "al contrario", ovvero partire dal modello nullo ed aggiungere mano a mano nuovi regressori fino a raggiungere il modello preferibile. In questo caso è opportuno avere delle conoscenze preliminari relative al fenomeno che si sta studiando, così da poter scegliere quali variabili inserire nel modello seguendo una logica dettata dalle sue caratteristiche. Per esempio, nel nostro caso, se proviamo ad attare un modello contenente le sole percentuali di tiro dal campo per le due squadre, otteniamo i seguenti risultati:

predictor	estimate	std.error	z value	p.value
FG_PCT	31.06	0.56	55.75	≈ 0
FG_PCT_opp	-31.06	0.56	-55.75	≈ 0

Tabella 2.6: Tabella dei coefficienti del modello con le sole % di tiro dal campo

Resid..Df	Resid..Dev	df	Deviance	p.value
12299	17051.42			
12298	10270.35	1	6781.07	≈ 0

Tabella 2.7: Tabella ANOVA comparativa tra il modello nullo
e quello con le % di tiro dal campo.

Possiamo quindi vedere che entrambi i regressori sono significativi e che il modello è decisamente migliore del modello nullo. Ma il dato veramente interessante è quello dell'accuratezza: considerando come unici regressori le percentuali di tiro dal campo delle due squadre otteniamo un'accuratezza del 78.70%. Tale risultato ci indica che l'efficienza di tiro è di gran lunga il carattere con maggiore impatto sull'esito di una partita, ed un modello contenente solo tale informazione può già vantare una discreta capacità previsiva. A questo punto possiamo utilizzare questo modello come modello di partenza ed aggiungere eventuali altre variabili verificando i risultati ottenuti, fino a raggiungere un modello ottimale. Seguendo le intuizioni dettate dalla natura dello sport e i risultati dei vari modelli plausibili, arriviamo a formulare il seguente modello, che riteniamo ottimale per le nostre esigenze.

predictor	estimate	std.error	z value	p.value
FG_PCT	66.49	1.48	45.03	≈ 0
FG_PCT_opp	-66.49	1.48	-45.03	≈ 0
FG3_PCT	10.76	0.47	22.72	≈ 0
FG3_PCT_opp	-10.76	0.47	-22.72	≈ 0
FTA	0.16	0.01	26.80	≈ 0
FTA_opp	-0.16	0.01	-26.80	≈ 0
TOV	-0.38	0.01	-31.20	≈ 0
TOV_opp	0.38	0.01	31.20	≈ 0
OREB	0.35	0.01	29.21	≈ 0
OREB_opp	-0.35	0.01	-29.21	≈ 0

Tabella 2.8: Tabella dei coefficienti del modello ottimale.

Vediamo subito che tutti i coefficienti sono ampiamente significativi e calcolando l'accuratezza otteniamo una precisione previsiva del 90.65%, che è un ottimo risultato considerando che questo modello include solo 10 variabili indipendenti (ben 12 in meno del modello iniziale, per una perdita di circa il 7% dell'accuratezza). Ciò che balza all'occhio di tale modello sono sicuramente le variabili scelte: osserviamo, infatti, che il modello segue perfettamente i principi esposti dai *Four Factors*, in quanto i caratteri scelti enfatizzano l'importanza delle medesime quattro aree della natura dello sport (efficienza di tiro, frequenza delle palle perse, frequenza dei tiri liberi e capacità di ottenere rimbalzi offensivi).

2.3.3. Analisi delle singole squadre

Possiamo, quindi, dire che a livello "universale" i *Four Factors* forniscono un'ottima sintesi delle variabili di maggiore rilevanza nella determinazione dell'esito di un incontro. Vogliamo, però, verificare se tale risultato vale anche per le singole squadre o se in determinati casi è preferibile utilizzare variabili esplicative differenti. Per prima cosa effettuiamo una rapida analisi esplorativa dei dati relativi alle singole squadre. Con l'aiuto del seguente grafico (*Figura 2.5*), andiamo a distinguere le squadre migliori e peggiori delle stagioni oggetto del nostro studio.

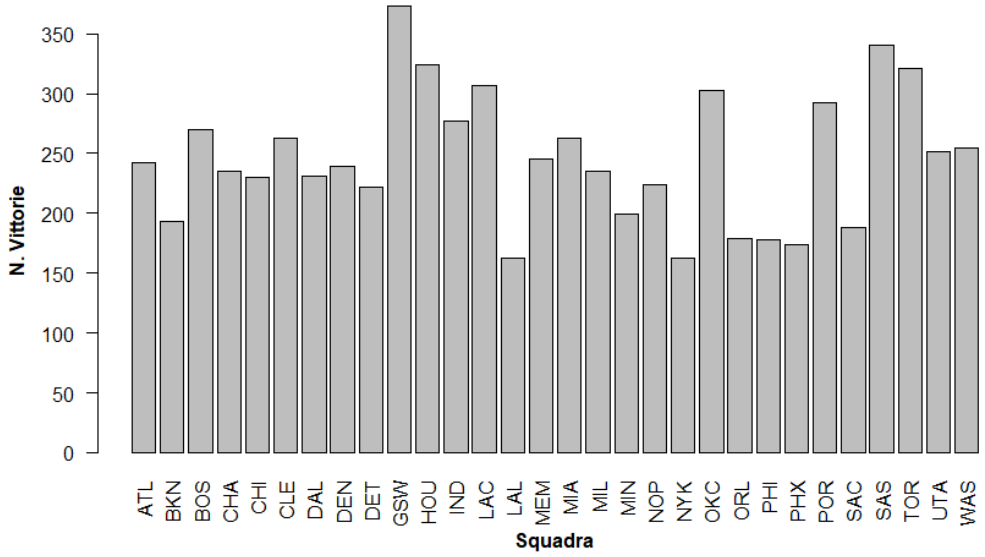


Figura 2.5: Bar-chart del numero di vittorie per ciascuna squadra

Dal grafico possiamo subito vedere che i Golden State Warriors sono stati indubbiamente la squadra più vincente delle ultime sei stagioni, vincendo 373 partite su 492 giocate. All'opposto, osserviamo come i Los Angeles Lakers ed i New York Knicks siano state le peggiori, in quanto entrambe hanno vinto solamente 163 incontri. Inoltre, i dati ci indicano che nel complesso delle passate sei stagioni le squadre della lega hanno vinto in media 246 partite, con una deviazione standard

di circa 55 vittorie. Quest'analisi preliminare ci aiuta a comprendere come, in un determinato periodo di tempo, le situazioni delle differenti squadre possono essere drasticamente diverse tra loro. Per questo motivo è logico pensare di doverle trattare come tali: è infatti altamente improbabile che i problemi e le esigenze affrontati dalle migliori squadre della lega corrispondano con quelli delle peggiori, sia per quanto ne concerne la dimensione, che per la loro natura. Riteniamo, dunque, necessario ed opportuno estendere il nostro studio anche alle singole squadre, ponendoci l'obiettivo di migliorare il modello visto in precedenza proponendo modelli che si adattino alle specifiche esigenze di una determinata squadra. In seguito prenderemo in esame alcune delle 30 squadre della lega, ma ricordiamo che il processo può essere ripetuto in qualsiasi momento per ogni squadra.
Prima di iniziare, ci poniamo il limite massimo di 10 variabili esplicative (tante quante quelle del modello ottimale) per ogni modello proposto. Iniziamo applicando il nostro modello ottimale alle due squadre peggiori della lega. Partiamo con i New York Knicks:

predictor	estimate	std.error	z value	p.value
FG_PCT	103.31	15.88	6.51	≈ 0
FG_PCT_opp	-101.40	15.73	-6.45	≈ 0
FG3_PCT	17.51	3.77	4.64	≈ 0
FG3_PCT_opp	-24.24	4.96	-4.89	≈ 0
FTA	0.22	0.05	4.59	≈ 0
FTA_opp	-0.16	0.04	-3.96	≈ 0
TOV	-0.49	0.10	-5.09	≈ 0
TOV_opp	0.42	0.10	4.38	≈ 0
OREB	0.57	0.12	4.66	≈ 0
OREB_opp	-0.59	0.12	-4.95	≈ 0

Tabella 2.9: Tabella dei coefficienti del modello ottimale universale per i New York Knicks.

Osserviamo che tutti i regressori sono statisticamente significativi a qualsiasi livello di significatività ragionevole e otteniamo un livello di accuratezza previsiva del 95,12%. Entrambi i risultati sono eccellenti e, in questo caso specifico, tale modello rimane il migliore, in quanto nessuna combinazione alternativa delle variabili esplicative ha fornito risultati migliori. Proviamo ora con i Los Angeles Lakers:

predictor	estimate	std.error	z value	p.value
FG_PCT	61.49	8.40	7.32	≈ 0
FG_PCT_opp	-64.46	8.13	-7.93	≈ 0
FG3_PCT	11.58	2.64	4.38	1e-05
FG3_PCT_opp	-12.26	2.65	-4.63	3e-06
FTA	0.08	0.03	2.61	9e-03
FTA_opp	-0.15	0.03	-4.77	1e-06
TOV	-0.38	0.07	-5.24	2e-07
TOV_opp	0.53	0.09	6.10	≈ 0
OREB	0.33	0.07	4.88	1e-06
OREB_opp	-0.27	0.06	-4.19	2e-05

Tabella 2.10: Coefficienti del modello ottimale universale per i Los Angeles Lakers.

Nel caso dei Los Angeles Lakers otteniamo un'accuratezza dell'89,02%. Inoltre, possiamo vedere che la variabile FTA sia sensibilmente meno significativa delle altre, per cui possiamo provare ad escluderla per far spazio ad un altro regressore potenzialmente più utile alla nostra causa. Scambiando la variabile FTA con FG3A_opp otteniamo il seguente modello:

predictor	estimate	std.error	z value	p.value
FG_PCT	67.75	9.03	7.51	≈ 0
FG_PCT_opp	-65.70	8.36	-7.86	≈ 0
FG3_PCT	10.29	2.57	4.00	6e-05
FG3_PCT_opp	-11.92	2.69	-4.43	9e-06
FG3A_opp	-0.06	0.03	-2.19	0.03
FTA_opp	-0.15	0.03	-4.77	1e-06
TOV	-0.37	0.07	-5.15	≈ 0
TOV_opp	0.54	0.09	6.28	≈ 0
OREB	0.39	0.07	5.47	≈ 0
OREB_opp	-0.22	0.06	-3.93	8e-05

Tabella 2.11: Coefficienti del modello ottimale specifico per i Los Angeles Lakers.

Seppur possiamo osservare che il nuovo parametro sia meno significativo del precedente, il presente modello vanta un'accuratezza previsiva del 92,68%, quasi il 4% in più del modello precedente. Questo, unito alle ridotte differenze tra le devianze residue e le significatività dei parametri dei due modelli, ci porta a preferire il modello creato su misura per i Lakers. Abbiamo quindi il primo caso in

cui i risultati universali forniti dai *Four Factors* risultano in qualche modo inferiori a quelli ottenuti attraverso un modello creato appositamente per le esigenze della specifica squadra. Procediamo ora analizzando la squadra più vincente delle ultime 6 stagioni, i Golden State Warriors.

predictor	estimate	std.error	z value	p.value
FG_PCT	74.83	10.75	6.96	≈ 0
FG_PCT_opp	-70.19	11.14	-6.30	≈ 0
FG3_PCT	11.57	3.53	3.28	1e-03
FG3_PCT_opp	-10.50	3.03	-3.46	5e-04
FTA	0.17	0.04	4.22	2e-05
FTA_opp	-0.18	0.04	-4.41	1e-05
TOV	-0.50	0.09	-5.63	≈ 0
TOV_opp	0.43	0.09	4.84	1e-06
OREB	0.40	0.08	4.84	1e-06
OREB_opp	-0.41	0.08	-4.89	1e-06

Tabella 2.12: Coefficienti del modello ottimale universale per gli Warriors.

A tali dati si abbina un livello di accuratezza dell'86,59%. Provando ad ottenere un modello che fornisca un'accuratezza migliore, senza perdere la significatività dei parametri, troviamo il seguente:

predictor	estimate	std.error	z value	p.value
FG_PCT	60.72	8.90	6.82	≈ 0
FG_PCT_opp	-66.33	10.02	-6.62	≈ 0
FG3_PCT	9.42	3.13	3.01	2e-03
FG3_PCT_opp	-9.98	2.77	-3.60	3e-04
FT_PCT	4.69	1.99	2.35	0.02
FTA	0.10	0.04	2.87	4e-03
TOV	-0.48	0.08	-5.79	≈ 0
TOV_opp	0.36	0.08	4.65	3e-06
OREB	0.33	0.07	4.52	6e-06
OREB_opp	-0.38	0.08	-5.08	≈ 0

Tabella 2.13: Coefficienti del modello ottimale specifico per gli Warriors.

Osserviamo che tutti i coefficienti sono significativi ad un livello di significatività dello 0.05, ed otteniamo un livello di accuratezza dell'90,24%, superiore a quello del modello universale.
Come ultimo esempio andiamo ad analizzare una squadra che nelle sei stagioni di riferimento ha ottenuto un numero di risultati positivi vicino alla media della lega.

Scegliamo i Memphis Grizzlies, in quanto con le loro 245 vittorie sono la squadra che più si avvicina alla media di 246. Adattiamo il modello ottimale universale:

predictor	estimate	std.error	z value	p.value
FG_PCT	85.86	11.72	7.33	≈ 0
FG_PCT_opp	-87.57	11.97	-7.32	≈ 0
FG3_PCT	12.84	2.59	4.95	≈ 0
FG3_PCT_opp	-12.05	3.21	-3.76	1e-04
FTA	0.21	0.04	5.42	≈ 0
FTA_opp	-0.22	0.04	-5.69	≈ 0
TOV	-0.42	0.09	-4.92	≈ 0
TOV_opp	0.43	0.08	5.68	≈ 0
OREB	0.37	0.08	4.81	1e-06
OREB_opp	-0.36	0.08	-4.40	1e-05

Tabella 2.14: Coefficienti del modello ottimale universale per i Grizzlies.

Ricaviamo un'accuratezza dell'85,37%. Utilizzando il solito procedimento troviamo il seguente modello:

predictor	estimate	std.error	z value	p.value
FG_PCT	60.98	7.43	8.21	≈ 0
FG_PCT_opp	-66.40	7.74	-8.58	≈ 0
FG3_PCT	8.14	1.94	4.20	2e-05
FG3_PCT_opp	-10.38	2.58	-4.03	5e-05
FT_PCT	6.92	1.62	4.26	2e-05
FG3A_opp	-0.09	0.03	-2.83	4e-03
TOV	-0.38	0.06	-5.95	≈ 0
TOV_opp	0.34	0.06	5.83	≈ 0
OREB	0.28	0.06	4.91	≈ 0
OREB_opp	-0.25	0.06	-4.05	5e-05

Tabella 2.15: Coefficienti del modello ottimale specifico per i Grizzlies.

Notiamo come anche in questo modello i coefficienti siano tutti significativi, ed osserviamo un miglioramento nell'accuratezza previsiva con un valore dell'89,02%. Anche in questo caso, quindi, preferiamo il modello specifico.
Ricapitolando, dunque, abbiamo osservato i seguenti risultati relativamente all'accuratezza (*Tabella 2.16* a pagina successiva), senza alterare drasticamente la significatività dei parametri e la devianze residue dei modelli.

Squadra/Modello	M.Universale	M.GSW	M.MEM	M.LAL
Tutte le squadre	90,65%	87,85%	86,83%	88,05%
GSW	86,59%	90,24%	86,59%	87,80%
MEM	85,37%	79,27%	89,02%	84,15%
LAL	89,02%	85,37%	84,15%	92,68%
NYK	95,12%	92,68%	90,24%	93,90%

Tabella 2.16: Accuratezza dei modelli trattati per le diverse squadre

Dai risultati ottenuti vediamo subito che per tre delle quattro squadre analizzate è possibile implementare un modello più efficiente di quello basato sui *Four Factors*. Osserviamo inoltre come tali modelli risultino essere i più efficienti solo se applicati ai dati relativi alle squadre su cui ciascuno di essi è stato progettato. Questo conferma la nostra teoria secondo la quale la personalizzazione dei modelli, basata sui dati osservati per ciascuna squadra, fornisca risultati migliori e maggiormente specifici di un singolo modello standard, a parità di numerosità delle variabili esplicative. Un'ulteriore osservazione che ricaviamo dai risultati ottenuti riguarda il grado di prevedibilità delle singole squadre. È facile notare come ciascuna squadra differisca (in certi casi anche in maniera importante) nel valore di accuratezza previsiva calcolato tramite l'utilizzo del rispettivo modello ottimale. Questo può essere dovuto a diversi fattori: diversi livelli di variabilità presenti nei casi specifici, variabili non incluse nei modelli che possono impattare l'accuratezza più o meno significativamente a seconda dei casi, o altri possibili fattori. Per le squadre da noi analizzate osserviamo come gli esiti delle partite dei New York Knicks siano quasi completamente prevedibili mediante l'utilizzo dei *Four Factors*, mentre i risultati dei Memphis Grizzlies risultano più difficili da prevedere nonostante l'impiego di un modello creato appositamente per migliorarne l'accuratezza previsiva (89,02% contro i 95,12% dei New York Knicks).

Capitolo 3

Conclusioni

Il presente elaborato ci ha fornito un'introduzione e diversi spunti di riflessione riguardanti il vasto mondo della statistica nel basket, cercando di trovare metodi alternativi che potessero migliorare l'efficienza di quelli ampiamente studiati nella letteratura ed applicati nello studio di dati reali. Le teorie sviluppate nel corso dello studio dettagliato in questo documento possono essere seguite per le più disparate applicazioni, a seconda dell'interesse delle parti coinvolte. I modelli proposti, inoltre, possono essere utilizzati come base per eventuali studi estensivi e/o per l'implementazione di modelli più complessi, che possano ad andare ad integrare i risultati da noi esposti per ottenere una visione ancor più accurata o specifica dello svolgimento di una partita di NBA.

3.1. Utilizzi delle teorie proposte

I risultati che abbiamo ottenuto nel precedente capitolo trovano, tra le altre, un'immediata applicazione pratica nella preparazione tecnica e tattica che precede ogni incontro. Lo staff tecnico di ogni singola squadra potrà infatti consultare i risultati ottenuti dall'applicazione di un modello, appositamente creato, per decidere su quali aspetti del gioco lavorare con maggiore attenzione, oltre che per avere una misura, sostenuta dai dati, dell'impatto degli stessi sull'esito delle partite. Se prendiamo come esempi le squadre che abbiamo analizzato nel capitolo precedente, possiamo realisticamente ipotizzare che allenatore e staff tecnico dei Los Angeles Lakers possano decidere di dare priorità e porre maggiore enfasi sull'allenamento dei fondamentali tecnici volti a limitare la frequenza di tiri da tre

punti tentati dagli avversari, invece che concentrarsi maggiormente su quelli volti ad aumentare la frequenza dei tiri liberi tentati, in quanto i dati suggeriscono che i primi siano più significativi nel prevedere l'esito di un incontro rispetto ai secondi. Analogamente gli Warriors potrebbero preferire l'allenamento dell'efficienza di tiro dalla lunetta, invece di quello incentrato sulla limitazione della frequenza dei tiri liberi tentati dagli avversari, e così via per tutte le squadre della lega. Un altro tipo di utilizzo dei modelli, utile alle squadre, può essere quello di consultare i risultati ottenuti per determinare un profilo specifico delle caratteristiche ottimali da cercare in un potenziale nuovo acquisto. In questo modo le società potranno seguire l'evidenza statistica portata dai dati per mettere a fuoco le aree di gioco in cui le proprie squadre presentano lacune, così da poter in seguito prendere scelte oculate relativamente al reclutamento di giocatori capaci di colmare suddette lacune.

Per quanto riguarda gli enti esterni alla lega, possiamo ragionevolmente pensare ad applicazioni delle teorie mostrate per usi giornalistici, informativi, accademici, o anche come materiale utilizzabile per cercare di migliorare l'efficienza dei rendimenti di scommesse sportive incentrate sul basket NBA.

3.2. Miglioramenti e considerazioni finali

Le teorie ed i modelli presentati in questo elaborato sono sicuramente basilari, e pertanto possono essere migliorate o estese tramite svariati metodi o approcci. Nei nostri modelli, infatti, ci siamo limitati ad introdurre le statistiche "base" di una partita di basket, ovvero quelle più classiche e immediate. Dalla seconda metà degli anni 2000, la NBA ha iniziato a fornire statistiche molto più complesse per descrivere l'andamento degli incontri, di cui alcune consistenti in combinazioni di quelle classiche, capaci di sintetizzare specifiche aree di gioco, ed altre basate sulla rilevazione di particolari eventi precedentemente non rilevati. Un possibile approccio volto a migliorare i nostri modelli potrebbe, dunque, essere quello di integrare tali statistiche avanzate, che, insieme a quelle base, possono fornire risultati più accurati. Un ulteriore possibilità è quella di utilizzare modelli statistici diversi dalla regressione logistica. Si potrebbe infatti pensare di applicare qualsiasi

altro modello di classificazione (per esempio Support Vector Machines, k-Nearest Neighbours, ecc.) o di applicare un semplice modello di regressione lineare dove la variabile risposta è la differenza tra i punti segnati rispettivamente dalla squadra di riferimento e quella avversaria, così da ottenere informazioni riguardanti gli eventi che portano una partita ad essere più o meno combattuta (e quindi l'esito più o meno incerto). Infine, un altro utilizzo dei nostri modelli può essere quello di porli alla base di studi più complessi. Un esempio è dato dallo studio di Štrumbelj et al. (2016), in cui viene proposto un modello Markoviano intento a simulare l'andamento di una partita di basket utilizzando dati storici. Tra i fattori che influiscono nella simulazione troviamo un set di variabili volto a sintetizzare le caratteristiche delle due squadre coinvolte. Le variabili utilizzate in tale elaborato sono quelle basate sui *Four Factors*. È ragionevole, quindi, pensare di poter sostituire quest'ultime con i caratteri suggeriti dai nostri modelli per le singole squadre, così da ottenere risultati più precisi e accurati.

Concludiamo, perciò, il presente elaborato avendone illustrato alcuni dei numerosi vantaggi e possibili utilizzi da esso apportati, siano questi ultimi di natura accademica, informativa, lavorativa o di qualsiasi altro genere.

Bibliografia

- Casella G. e Berger R.L. (2002) “Statistical Inference (Second Edition)” Duxbury
- Grigoletto M. , Ventura L. e Pauli F. (2017) “Modello Lineare: Teoria e Applicazioni con R” G Giappichelli Editore
- Hosmer D.W. Jr, Lemeshow S. e Sturdivant R.X. (2013) “Applied Logistic Regression, Third Edition” Wiley
- Kubatko J. , Oliver D. , Pelton K. e Rosenbaum D.T. (2007) "A Starting Point for Analyzing Basketball Statistics" Journal of Quantitative Analysis in Sports: Vol. 3: Iss. 3, Article 1
- Manner H. (2016) “Modeling and forecasting the outcomes of NBA basketball games” Journal of Quantitative Analysis in Sports 2016; 12(1): 31–41
- Oliver, D. (2004) “Basketball on Paper” Brassey’s, Washington, DC
- Štrumbelj E. e Vračar P. (2012) “Simulating a Basketball Match with a Homogeneous Markov Model and Forecasting the Outcome” International Journal of Forecasting 28, 532–542
- Štrumbelj E. , Vračar P. e Kononenko I. (2016) “Modeling basketball play-by-play data” Expert Systems With Applications 44 (2016) 58–66
- Teramoto M. e Cross C.L. (2010) "Relative Importance of Performance Factors in Winning NBA Games in Regular Season versus Playoffs" Journal of Quantitative Analysis in Sports: Vol. 6: Iss. 3, Article 2
- Materiale didattico di "Modelli Statistici 1". A.A. 2017-2018. A cura della Prof.ssa Ventura L.

Sitografia

- https://www.basketball-reference.com/
- https://cran.r-project.org/
- https://stackoverflow.com/
- https://stats.nba.com/

Appendice

Codice R

```
# Importiamo il dataset contenente i nostri dati, eliminiamo la superflua colonna di indice e svolgiamo una preliminare analisi del dataset

df = read.csv('dataset_2.csv', sep=';', dec=',', header = TRUE)
df = df[,-1]
head(df)
summary(df)

# Rimuoviamo le colonne contenenti variabili che non ci interessano
df = df[,-c(3,4,6,22,23,39,40)]

# Effettuiamo un'analisi della correlazione tra le variabili del dataset

# Matrice di correlazione
crl=cor(df[,-c(1,2,3)])
crl

# Grafico livello di correlazione
library(corrplot)
corrplot(crl, type = "upper", order = "hclust",
      tl.col = "black", tl.srt = 45, diag = FALSE)

# Genero i diagrammi di dispersione per le variabili correlate che possono causare problemi

# Scatter plots per le statistiche di tiro
par(mfrow=c(2,3))
plot(df[,c('FGM','FGA')])
plot(df[,c('FGM','FG_PCT')])
plot(df[,c('FG3M','FG3A')])
plot(df[,c('FG3M','FG3_PCT')])
plot(df[,c('FTM','FTA')])
plot(df[,c('FTM','FT_PCT')])

# Scatter plots per palle rubate/perse dagli avversari
par(mfrow=c(1,2))
plot(df[,c('STL','TOV_opp')])
plot(df[,c('TOV','STL_opp')])

# Scatter plots per DREB e % di tiro avversaria e per assist e % di tiro
par(mfrow=c(1,2))
plot(df[,c('DREB','FG_PCT_opp')])
plot(df[,c('AST','FG_PCT')])
```

```
# Escludo le Variabili altamente correlate e divido il dataset in train e test
df = df[,-c(4,7,10,16,19,22,25,31)]
df_train = df[(df$SEASON_ID<22018),]
df_test = df[(df$SEASON_ID==22018),]

# Adatto il modello contenente tutte le variabili ai dati train
fit_all = glm(data = df_train[,-c(1,2)], WL~.-1 , family = 'binomial')
summary(fit_all)

# Test Anova modello nullo
null_fit = glm(data = df_train[,-c(1,2)], WL~ 1 , family = 'binomial')
anova(null_fit,fit_all, test="Chi")

# Previsione degli esiti delle partite della stagione 2018-19
pred_values = ifelse((predict(fit_all,newdata=df_test[,-c(1,2)], type=
'response'))>=0.5,1,0)
y_2018=ifelse(df_test$WL=='W',1,0)
acc_all= sum(ifelse(pred_values==y_2018,1,0))/nrow(df_test)

# Adatto il modello senza AST e BLK e lo comparo al primo modello
fit_AB = glm(data = df_train[,-c(1,2)], WL~. -1-AST-AST_opp-BLK-BLK_opp , family =
'binomial')
summary(fit_AB)
anova(fit_AB,fit_all, test="Chi")
pred_values = ifelse((predict(fit_AB,newdata=df_test[,-c(1,2)], type=
'response'))>=0.5,1,0)
acc_AB= sum(ifelse(pred_values==y_2018,1,0))/nrow(df_test)

# Adatto il modello di sole % al tiro
fit_pct = glm(data = df_train[,-c(1,2)], WL~ -1+FG_PCT+FG_PCT_opp, family =
'binomial')
summary(fit_pct)
anova(null_fit,fit_pct, test="Chi")
pred_values = ifelse((predict(fit_pct,newdata=df_test[,-c(1,2)], type=
'response'))>=0.5,1,0)
acc_pct= sum(ifelse(pred_values==y_2018,1,0))/nrow(df_test)

# modello ottimale
fit_best = glm(data = df_train[,-c(1,2)], WL~ -
1+FG_PCT+FG_PCT_opp+FG3_PCT+FG3_PCT_opp+FTA+FTA_opp+TOV+TOV_opp+O
REB+OREB_opp , family = 'binomial')
summary(fit_best)
pred_values = ifelse((predict(fit_best,newdata=df_test[,-c(1,2)], type=
'response'))>=0.5,1,0)
acc_best= sum(ifelse(pred_values==y_2018,1,0))/nrow(df_test)

# analisi esplorativa dati squadre
count=as.data.frame(table(df[df$WL=='W','TEAM_ABBREVIATION']))
barplot(count$Freq,names.arg = count$Var1, las=2, ylab = 'N. Vittorie',
xlab='Squadra', font.lab=2)
summary(count$Freq)
sd(count$Freq)
```

```
# Modelli per NYK e LAL
nyk_df_train = df_train[(df_train$TEAM_ABBREVIATION=='NYK'),]
nyk_df_test = df_test[(df_test$TEAM_ABBREVIATION=='NYK'),]

nyk_fit_best = glm(data = nyk_df_train[,-c(1,2)], WL~ -
1+FG_PCT+FG_PCT_opp+FG3_PCT+FG3_PCT_opp+FTA+FTA_opp+TOV+TOV_opp+O
REB+OREB_opp , family = 'binomial')
summary(nyk_fit_best)

pred_values = ifelse((predict(nyk_fit_best,newdata=nyk_df_test[,-c(1,2)], type=
'response'))>=0.5,1,0)

y_2018=ifelse(nyk_df_test$WL=='W',1,0)
acc_best= sum(ifelse(pred_values==y_2018,1,0))/nrow(nyk_df_test)

lal_df_train = df_train[(df_train$TEAM_ABBREVIATION=='LAL'),]
lal_df_test = df_test[(df_test$TEAM_ABBREVIATION=='LAL'),]

lal_fit_best1 = glm(data = lal_df_train[,-c(1,2)], WL~ -
1+FG_PCT+FG_PCT_opp+FG3_PCT+FG3_PCT_opp+FTA+FTA_opp+TOV+TOV_opp+O
REB+OREB_opp , family = 'binomial')
summary(lal_fit_best1)

pred_values = ifelse((predict(lal_fit_best1,newdata=lal_df_test[,-c(1,2)], type=
'response'))>=0.5,1,0)

y_2018=ifelse(lal_df_test$WL=='W',1,0)
acc_best= sum(ifelse(pred_values==y_2018,1,0))/nrow(lal_df_test)

lal_fit_best2 = glm(data = lal_df_train[,-c(1,2)], WL~ -
1+FG_PCT+FG_PCT_opp+FG3_PCT+FG3_PCT_opp+FG3A_opp+FTA_opp+TOV+TOV_o
pp+OREB+OREB_opp , family = 'binomial')
summary(lal_fit_best2)

pred_values = ifelse((predict(lal_fit_best2,newdata=lal_df_test[,-c(1,2)], type=
'response'))>=0.5,1,0)
acc_best= sum(ifelse(pred_values==y_2018,1,0))/nrow(lal_df_test)

# Modelli per GSW
gsw_df_train = df_train[(df_train$TEAM_ABBREVIATION=='GSW'),]
gsw_df_test = df_test[(df_test$TEAM_ABBREVIATION=='GSW'),]

gsw_fit_best = glm(data = gsw_df_train[,-c(1,2)], WL~ -
1+FG_PCT+FG_PCT_opp+FG3_PCT+FG3_PCT_opp+FTA+FTA_opp+TOV+TOV_opp+O
REB+OREB_opp , family = 'binomial')
summary(gsw_fit_best)

y_2018=ifelse(gsw_df_test$WL=='W',1,0)
pred_values = ifelse((predict(gsw_fit_best,newdata=gsw_df_test[,-c(1,2)], type=
'response'))>=0.5,1,0)
acc_best= sum(ifelse(pred_values==y_2018,1,0))/nrow(gsw_df_test)
```

```
gsw_fit_best2 = glm(data = gsw_df_train[,-c(1,2)], WL~ -
1+FG_PCT+FG_PCT_opp+FG3_PCT+FG3_PCT_opp+FT_PCT+FTA+TOV+TOV_opp+OR
EB+OREB_opp , family = 'binomial')
summary(gsw_fit_best)

pred_values = ifelse((predict(gsw_fit_best2,newdata=gsw_df_test[,-c(1,2)], type=
'response'))>=0.5,1,0)
acc_best= sum(ifelse(pred_values==y_2018,1,0))/nrow(gsw_df_test)

# Modelli per MEM
mem_df_train = df_train[(df_train$TEAM_ABBREVIATION=='MEM'),]
mem_df_test = df_test[(df_test$TEAM_ABBREVIATION=='MEM'),]

mem_fit_best = glm(data = mem_df_train[,-c(1,2)], WL~ -
1+FG_PCT+FG_PCT_opp+FG3_PCT+FG3_PCT_opp+FTA+FTA_opp+TOV+TOV_opp+O
REB+OREB_opp , family = 'binomial')
summary(mem_fit_best)

y_2018=ifelse(mem_df_test$WL=='W',1,0)
pred_values = ifelse((predict(mem_fit_best,newdata=mem_df_test[,-c(1,2)], type=
'response'))>=0.5,1,0)
acc_best= sum(ifelse(pred_values==y_2018,1,0))/nrow(mem_df_test)

mem_fit_best2 = glm(data = mem_df_train[,-c(1,2)], WL~ -
1+FG_PCT+FG_PCT_opp+FG3_PCT+FG3_PCT_opp+FT_PCT+FG3A_opp+TOV+TOV_op
p+OREB+OREB_opp , family = 'binomial')
summary(mem_fit_best2)

pred_values = ifelse((predict(mem_fit_best2,newdata=mem_df_test[,-c(1,2)], type=
'response'))>=0.5,1,0)
acc_best= sum(ifelse(pred_values==y_2018,1,0))/nrow(mem_df_test)
```