UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Corso di Laurea in Scienze Statistiche ed Economiche

Tesi di Laurea Triennale

"Gli Expected Goals": modello di statistical learning per il caso del "Futbol Club Barcelona"

Relatore
prof. Riccardo Borgoni

Laureando
Filippo Maria Masi

Settembre 2020

Ringraziamenti

Se penso al mio percorso di studi, ci sono alcune persone senza le quali il mio cammino sarebbe stato rovinoso e in salita, desidero quindi esprimere gratitudine a coloro che ne hanno preso parte.

In primo luogo ringrazio il professor Borgoni che mi ha seguito con costanza durante l'elaborato e ha supportato la mia idea di argomento della tesi.

Un pensiero enorme va ai miei genitori, mamma e papà, che si sono sempre presi cura di me unendosi al mio fianco nell'affrontare ogni mia preoccupazione o ostacolo. Grazie alla loro energia e affetto mi sono sempre sentito al sicuro con loro, e sono cresciuto in un nido d'amore. Ringrazio i miei due fratelloni, Elena e Francesco, per la loro simpatia e affettuosità che mi hanno sempre dimostrato come loro fratello più piccolo, e i miei nonni, Giordano e Antonietta, che mi hanno sempre coccolato e fatto sentire il loro amore.

Al di fuori della sfera famigliare, vorrei ringraziare tutti i miei compagni dell'università coi quali ho condiviso i momenti di studio più profondi durante questi tre anni. Un grazie particolare a Guglielmo detto *Google* e a Silvia che non saranno con me durante il percorso Magistrale e che, con la loro sicurezza e abilità, mi hanno dato forza nello studiare insieme, oltre a essersi dimostrati amici preziosi; a Claudio (detto *Stradio*) e Lorenzo (detto *Lollo*), compagni di studio ma soprattutto amici fidati e simpatici che hanno reso più leggeri periodi di studio intensi della Triennale. Inoltre ringrazio il mio affezionato amico di sempre Giacomo detto *Jack* e la sua energia contagiosa che ha contribuito a farmi scaturire l'idea dell'argomento della tesi e insieme a lui gli altri amici *storici* del gruppo *Biancaneve e sette nani*. Infine vorrei ringraziare la ragazza che amo, Chiara, che, oltre ad avermi effettivamente aiutato nella stesura dell'elaborato, ha trascorso insieme a me momenti indelebili e che con la forza, serenità e dolcezza che la distinguono mi ha offerto un appoggio instancabile giorno dopo giorno in questo cammino.

Indice

Capitolo 1

Introduzione

La Statistica è una disciplina che permette di prevedere e misurare, attraverso l'utilizzo di metodi e tecniche matematiche, fenomeni di natura quantitativa e qualitativa. Inoltre consente a chi riesce a padroneggiarla di prendere decisioni in condizione di incertezza, quindi anche in situazioni in cui il quadro generale non è chiaro e ben definito, ma può dipendere da fattori esterni e soggettivi.

Il fenomeno che tratterà questo elaborato, ad esempio, sebbene venga studiato e misurato con metodi quantitativi e statistici, ha una forte natura qualitativa e appartiene ad un mondo che sembra essere ben discostato dal rigore e oggettività della statistica matematica: la disciplina Sportiva e in particolare il gioco del Calcio. Grazie alla stima degli *Expected Goals*, la scienza entra quindi all'interno del rettangolo verde.

1.1 Expected Goals - cosa sono

Gli *Expected Goals* mirano a rappresentare il potenziale offensivo prodotto da una squadra di calcio in una determinata partita o le occasioni da goal potenzialmente avute da un singolo giocatore [5].

Il numero di xG (come vengono abbreviati) indica i goal che ci si sarebbe aspettati che quella squadra o quel giocatore segnasse, ovvero quelle occasioni che si traducono in rete con una probabilità elevata condizionatamente alla situazione di gioco circostante. Perciò gli *Expected Goals* si propongono di creare metriche avanzate che ben correlano con la probabilità di ottenere risultati vincenti e che di conseguenza sono in grado non solo di valutare la prestazione di un giocatore nelle stagioni passate, ma anche di predire il livello di prestazioni future[5].

Non è presente un solo modello per il calcolo di queste statistiche avanzate e quindi il risultato può variare a seconda dell'algoritmo che le calcola.

Durante questo elaborato l'obbiettivo è quello di mostrare metodi di calcolo e il confronto tra essi a partire da un campione ottenuto aggregando statistiche

semplici che molto spesso hanno una natura unicamente descrittiva: i tiri verso la porta. Partendo dall'Analisi empirica descrittiva del DataBase di conclusioni verso la porta, verranno impiegati più metodi di *statistical learning* quali *Regressioni Logistiche*, *Alberi Decisionali* e *Random Forest* per ottenere modelli che offrono la possibilità di decretare per una qualsiasi conclusione a rete se essa è definibile *Goal atteso* o meno.

Ogni modello stimato verrà inoltre analizzato, le sue prestazioni predittive saranno testate confrontando stime ottenute con valori empirici e calcolando indicatori utili. Esso sarà eventualmente modificato con lo scopo ultimo di ottenere le performance migliori di stime e previsioni e i risultati ottenuti verranno interpretati.

Verranno infine illustrate le discrepanze tra i diversi metodi, mostrati quali sono i benefici e gli svantaggi dovuti all'impiego di ognuno di essi, ed esposte le situazioni più adatte nell'uso di un metodo piuttosto che un altro.

Capitolo 2

Contestualizzazione

In questo capitolo verrà eseguita un'illustrazione del contesto specifico dove agiranno le operazioni di *statistical learning* per il calcolo degli *Expected Goals*. Verrà presentato il campione di partenza, i suoi riferimenti temporali e spaziali, ed esposte le variabili che lo compongono.

2.1 "Barcelona Futbol Club"

La squadra di calcio considerata nell'elaborato è il Barcellona. Per questo motivo i modelli che verranno dapprima ricercati e poi trovati, si riferiscono esclusivamente alla squadra catalana. Le figure che maggiormente possono attingere ai risultati ottenuti da questo tipo di ricerca statistica sono per esempio il manager della squadra, che intende ottimizzare la capacità offensiva e illustrare all'allenatore dove migliorare, in particolare evidenziando quali azioni risultino più vincenti in media [5]. Allo stesso tempo il dirigente o l'allenatore di una squadra rivale che può voler studiare i propri avversari e capire il loro potenziale offensivo: come si sviluppa e dove maggiormente va in goal. Infine un'altra figura che si può ritenere interessata a questi studi è l'allibratore di un'agenzia di scommesse, interessato a misurare le quote dei goal della squadra *blaugrana*, in modo tale che siano in linea con le passate probabilità campionarie di rete e vantaggiose per l'agenzia. Durante l'interpretazione dei risultati ottenuti perciò sarà ragionevole mettersi nei panni di queste figure per trarre le conclusioni che meglio fanno al loro caso.

2.1.1 L'efficacia delle statistiche `tiri verso la porta`

Le osservazioni da cui partirà l'analisi statistica sono i tiri verso la porta. L'ammontare dei tiri dipende dalla capacità di avere un efficiente potenziale offensivo, ma non è sempre sinonimo di squadra vincente; nel corso di una partita di calcio, è importante infatti anche la fase difensiva, ma nella gran parte delle ultime stagioni del

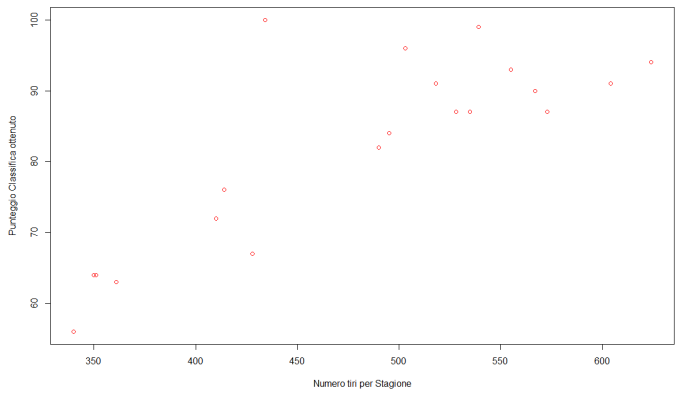

Figura 2.1: Tiri-Punteggi

Barcellona nel campionato spagnolo, *La Liga*, le conclusioni a rete si sono rivelate essere una statistica ben relazionata con la capacità di fare punti e successi sportivi. Tale relazione viene testimoniata dal grafico 2.1 tra i tiri effettuati per stagione e i punteggi per stagione.

La correlazione che c'è tra i due vettori di informazioni è evidente nella figura poiché nella gran parte dei casi le stagioni con maggiori conclusioni a rete coincidono con un elevato punteggio. Essa può essere calcolata come:

$$cor_{\underline{x},\underline{y}} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2}} = 0.8217194$$

dove

$$\underline{x} = \texttt{vettore del numero dei tiri dal 99 al 2019}$$
$$\underline{y} = \texttt{vettore del numero dei punti in classifica dal 99 al 2019}.$$

I dati sono stati ottenuti dal sito *Who Scored* [19]. Il Coefficiente di correlazione $cor_{\underline{x},\underline{y}} \in [-1,1]$ è prossimo all'uno e quindi elevato, come ipotizzato.

L'orizzonte temporale in questo caso è allargato rispetto a quello che verrà considerato poi nel campione studiato, per avere un numero più ampio di osservazioni: le stagioni vanno dal 1999/00 al 2018/19.

2.2 La costruzione del campione : conclusioni a rete del Barcellona

Attraverso il package di Rstudio [1] che mette a disposizione il sito StatsBomb [13] si è ottenuto un database contenente tutti gli eventi accaduti durante ogni partita del campionato spagnolo *La Liga* per stagione sportiva della squadra Barcellona.

Dopo aver isolato i tiri verso la porta effettuati da parte della squadra e dopo averli aggregati, sono state concatenate le unità statistiche dalla stagione 2006/07 fino al 2018/19, fino a ottenere 6726 osservazioni campionarie (righe del DataSet).

Il modello considerato ha come obbiettivo quello di prevedere statistiche future e si basa sulle serie storiche dei tiri verso la porta avvenuti in passato, della stessa squadra.

2.3 Le variabili

Per ogni unità statistica del campione sono state selezionate le variabili che potessero interessare per il calcolo degli xG: tutte quelle che descrivono la situazione di gioco nella quale è avvenuto il tiro verso la porta, da un punto di vista fisico e reale. Ad Esempio: Distanza del giocatore che calcia dalla porta, Densità dei giocatori presenti attorno a chi calcia, Parte del Corpo con cui è stata colpita la palla (vedi paragrafo 2.3.1), e infine è stato ottenuto il campione un cui stralcio è riportato nella figura 2.2.

Le variabili fin qui scelte costituiscono solo una prima scrematura delle colonne del `DataSet`. Nel seguito dell'analisi il campione subirà altre modifiche: alcune variabili saranno ulteriormente scartate in quanto risultate di scarsa rilevanza durante i procedimenti della costruzione dei modelli di calcolo.

	location.x	location.y	DistSGK	AngleToGoal	AngleToKeeper	shot.deflected	DistToGoal
1	89.5	40.3	28.900173	90.56355	97.12502	0	30.501475
2	112.6	35.1	9.861541	56.48898	146.30993	0	8.875246
3	101.9	53.5	21.349005	126.71765	98.13010	0	22.580080
4	96.3	45.8	22.777401	103.75148	86.63354	0	24.399385
5	107.3	43.8	11.985408	106.65785	140.19443	0	13.256319
6	98.6	52.1	23.281967	119.48471	122.47119	0	24.583938
7	113.9	37.6	6.280127	68.52321	141.84277	0	6.555151
8	109.8	52.1	11.321661	139.86994	156.64444	0	15.825612
9	104.0	22.6	22.447494	42.59979	55.00798	0	23.638105
10	89.7	27.6	31.123785	67.74361	49.76364	0	32.739120

Figura 2.2: Parte del primo campione

Le variabili spaziali hanno come riferimento un piano cartesiano in cui il vertice a Nord-Ovest del rettangolo di gioco è l'origine dei due assi:
Ascisse = lato lungo del campo
Ordinate = lato corto del campo.
Le coordinate cartesiane sono illustrate nella figura 2.3 e hanno come unità di misura di riferimento il metro.

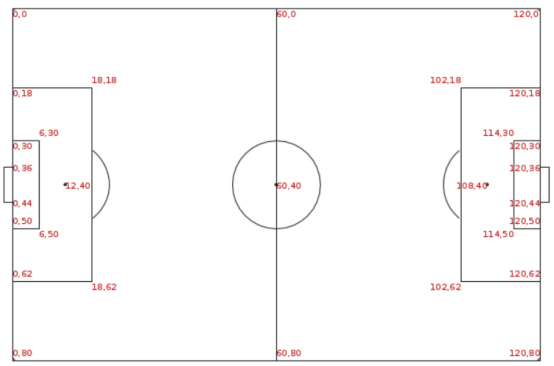

Figura 2.3: Coordinate di riferimento

Non tutte le variabili di interesse derivano direttamente dai dati di StatsBomb [13], ma alcune, in maniera indiretta, sono state create con algoritmi di calcolo dalle stesse informazioni come si può notare dal codice in figura 2.4. Molte di queste si sono verificate effettivamente poi fondamentali per la misurazione degli xG.

```r
shotinfoss <- function(dataframe){
  ##Calculate a distance from shot and distance from center of frame variable
  dataframe <- dataframe %>%
    arrange(match_id, index)

shots <- dataframe %>%
    filter(type.name == "Shot") %>%
    mutate(location.x = ifelse(location.x == 120 & location.y == 40, 119.66666, location.x)) %>%
    mutate(location.x.GK = ifelse(location.x.GK == 120 & location.y.GK == 40, 119.88888, location.x.GK)) %>%
    mutate(DistToGoal = sqrt((location.x - 120)^2 + (location.y - 40)^2),
           DistToKeeper = sqrt((location.x.GK - 120)^2 + (location.y.GK - 40)^2)) %>%
    mutate(AngleToGoal = ifelse(location.y <= 40, asin((120-location.x)/DistToGoal),
                                (pi/2) + acos((120-location.x)/DistToGoal))) %>%
    mutate(AngleToKeeper = ifelse(location.y.GK <= 40, asin((120-location.x.GK)/DistToKeeper),
                                  (pi/2) + acos((120-location.x.GK)/DistToKeeper))) %>%
    mutate(AngleToGoal = AngleToGoal*180/pi) %>%
    mutate(AngleToKeeper = AngleToKeeper*180/pi) %>%
    mutate(AngleDeviation = abs(AngleToGoal-AngleToKeeper)) %>%
    #mutate(duration = ifelse(duration <= 0.1, 0.1, duration)) %>%
    mutate(avevelocity = sqrt((shot.end_location.x - location.x)^2 +
                              (shot.end_location.y - location.y)^2)/duration) %>%
    mutate(DistSGK = sqrt((location.x - location.x.GK)^2 + (location.y - location.y.GK)^2))
  others <- dataframe %>%
    filter(type.name != "Shot")

  dataframe <- bind_rows(shots, others) %>%
    arrange(match_id, index)

  return(dataframe)
}
```

Figura 2.4: Algoritmo in RStudio

2.3.1 Spiegazione delle variabili

Nel prossimo capitolo verranno messe in evidenza le caratteristiche e le proprietà statistiche delle variabile che poi sono risultate essere di maggiore interesse attraverso un'analisi statistica descrittiva; in questa sezione invece viene solo illustrata una spiegazione dell'informazione offerta da ogni variabile ottenuta dal sito[15], che sarà necessaria per l'interpretazione finale degli esiti riscossi.

- `Location.x` = Luogo del campo nell'asse X (vedi figura 2.3)
- `Location.y` = Luogo del campo nell'asse X (vedi figura 2.3)
- `DistSGK` = Distanza in metri dal giocatore che calcia al Portiere
- `AngleToGoal` = Angolo in gradi rispetto al centro della porta
- `AngleToKeeper` = Angolo in gradi rispetto al portiere
- `Shot.deflected` = Varibile Dummy (1 = Tiro deviato, 0 = Tiro non deviato)
- `DistToGoal` = Distanza in metri dal punto di partenza del tiro, alla linea di porta
- `Shot.body_part.id` = Variabile Dummy che indica la parte del corpo che colpisce la palla
  - 37 = "Testa"

  - 38 = “Piede sinistro”
  - 70 = “Altro” (i.e ginocchio,petto etc)
  - 40 = “Piede destro”

- `Density.incone` = Densità di uomini nel cono che ha come 3 vertici i due pali e il punto di partenza del tiro
- `Under_pressure` = Variabile Dummy (1 = Calciatore in quel momento sotto pressione, 0 = Viceversa)
- `DefendersBehindBall` = Numero di difensori che si trovano tra la palla e la porta al momento del tiro
- `AttackersBehindBall` = Numero di attaccanti che si trovano tra la palla e la porta al momento del tiro
- `DefendersInCone` = Numero di difensori che si trovano nel cono di tiro
- `Duration` = Durata del tiro in secondi
- `Distance.ToD1.360` = Distanza dal primo difensore più vicino a 360° da chi calcia
- `Distance.ToD2.360` = Distanza dal secondo difensore più vicino a 360° da chi calcia
- `ElapsedTime` = Tempo di esecuzione del tiro in millisecondi
- `Distance.ToD1` = Distanza dal primo difensore più vicino davanti a chi calcia
- `Distance.ToD2` = Distanza dal secondo difensore più vicino davanti a chi calcia
- `shot.technique.id` = Variabile Dummy:
  - 89 = “Colpo di tacco”
  - 90 = “Colpo di testa in tuffo”
  - 91 = “Tiro al volo dopo un rimbalzo”
  - 92 = “Pallonetto"
  - 93 = “Normale”
  - 94 = "Sforbiciata"
  - 95 = “Tiro al volo (senza rimbalzo)”
- `shot.type.id` = Variabile Dummy:

  - 61 = “Calcio d’angolo”
  - 62 = “Punizione indiretta”
  - 65 = “Punizione diretta”
  - 87 = “Azione di gioco”
  - 88 = “Calcio di rigore”

- `shot.end.location_x` = posizione di arrivo del tiro lungo l’asse delle ascisse (vedi figura 2.3)
- `shot.end.location_y` = posizione di arrivo del tiro lungo l’asse delle ordinate (vedi figura 2.3)
- `shot.end.location_z` = altezza da terra del punto di arrivo del tiro (in porta: da 0 a 2.67 metri)

Capitolo 3

Pulizia del campione

Affinché i dati campionari possano essere studiati e analizzati, è necessaria una fase precedente di ripulitura del *DataSet*. Quest'operazione è fondamentale per evitare errori significativi negli stadi successivi dell'elaborato. In particolare la pulizia del campione deve portare alla eliminazione di variabili poco inerenti all'obbiettivo della ricerca statistica e allo stesso tempo risolvere le problematiche dei *missing values*.

3.1 Cancellazione di alcune colonne

In un primo momento, costruito il DataBase, il numero delle variabili presenti era molto elevato. Le informazioni erano state raccolte con un fine illustrativo, o probabilmente con innumerevoli altri scopi diversi dal nostro proposito di calcolare per ogni azione di gioco se essa fosse un `Expected Goal`. E' stato doveroso perciò un lavoro di scrematura accurata. L'aggiunta di intere colonne poco significative per l'analisi di statistiche prefissate, avrebbe portato a una quantità ingente di informazioni ridondanti tra i pattern e ogni azione successiva sarebbe risultata così più macchinosa e meno semplificata.

Nello specifico sono state cassate le colonne che comunicavano quale giocatore avesse compiuto l'azione, la partita in cui è avvenuta, la squadra che affrontava il Barcellona e ogni singola informazione tipica degli eventi di gioco che non fossero le conclusioni verso la porta.

3.2 Gli NA

Le osservazioni del campione erano piuttosto ordinate fin da subito, solamente alcune risultavano `"NA"`, ovvero *Not available* perciò dati mancanti.

La scelta della procedura da attuare in questi casi dipende da diversi fattori:

- Il numero di `NA` presenti,
- La percentuale di `NA` sulla specifica colonna rispetto al totale dell'informazione disponibile,
- L'incidenza che ha la variabile contenente il dato mancante sui modelli stimati.

A seconda della situazione in cui ci si trova e dei fattori che caratterizzano il campione, le tecniche per superare questo problema, spesso presente nelle indagini statistiche, sono di tre tipi:

1. Metodi di eliminazione
2. Metodi di imputazione singola
3. Algoritmi specifici di imputazione.

Nel primo caso non vengono considerate tutte quelle unità statistiche con almeno un dato mancante oppure l'intera variabile avente una percentuale mediamente elevata di *NA* rispetto ai totali; e quindi vengono eliminate. I vantaggi di questo metodo sono l'immediatezza e la semplicità poiché le analisi statistiche possono essere applicate senza modifiche ad un DataSet di dimensione ridotta, ma completo. Lo svantaggio principale è invece la perdita di informazioni: se il campione completo ha dimensioni ridotte, questo potrebbe portare a risultati di stima distorti e dunque a conclusioni errate.

L'imputazione singola è un metodo statistico che cerca di eliminare i valori mancanti all'interno di un DataSet, sostituendo gli `NA` con dei valori ammissibili per la variabile considerata. Il vantaggio principale è la non perdita di informazioni che permette di mantenere un dimensionamento del DataSet elevato e quindi una precisione maggiore nel calcolo di stime. Lo svantaggio invece consiste nel fatto che i dati imputati vengono poi considerati come realmente osservati e trattati come tali nello studio; ciò comporta una riduzione della variabilità e una distorsione delle informazioni tanto grande quanto la lontananza dai dati reali.

La portata di quest'ultimo fenomeno dipende da una corretta definizione del criterio di imputazione. Vengono riportati di seguito i valori di imputazione più utilizzati da sostituire alla cella contenente `NA` [10]:

- Media: la media dei valori osservati della variabile a cui appartiene l'`NA`.
- Campionamento aleatorio: un valore estratto in modo casuale da quelli disponibili per la variabile.
- Regressione: risultato di una regressione tra le variabili osservate della stessa unità.

- Probabilità condizionata Bayesiana: metodo utilizzato in prevalenza per le variabili Dummies. Si tratta di un'imputazione pesata dalla probabilità che un'osservazione abbia uno dei due valori della Dummy, condizionata alle informazioni della stessa unità statistica per le altre variabili disponibili.

- A discrezione dell'Analista: metodo usato nei casi di un'osservazione particolare in cui si imputa in base alle altre situazioni estreme presenti.

3.2.1 Il caso in esame

Nella matrice dei dati di riferimento erano inizialmente presenti 2199 celle `NA` su un totale di 6726 righe · 23 colonne.

Muovendoci seguendo le procedure elencate sopra, si notava in maniera immediata come delle 2199, 1682 erano facenti parte della variabile `shot.end.location_z` (vedi sezione 12.3.1) mentre le rimanenti sparse sulle altre.

Poiché l'obbiettivo finale dell'elaborato è riuscire a prevedere se, data una osservazione, in base alle sue caratteristiche, risulti Goal o viceversa, non è di interesse la posizione esatta finale del tiro, ma solamente se si è tramutato in rete. Risulta pertanto lecito eliminare dal Database la colonna relativa a tale variabile e con essa anche `shot.end.location_x` e `shot.end.location_y` riferenti lo stesso tipo di informazione ma su assi cartesiani differenti.

Per quanto riguarda le rimanenti, esaminando le altre celle delle stesse osservazioni con 1 o più *Not Available*, si è notato che di esse gran parte erano conclusioni effettuate durante un calcio di rigore. Infatti 470 delle celle mancanti avevano come variabile `Shot.type.id = 88` che, come si può notare dal paragrafo 2.3.1, testimonia il *penalty kick*. Perciò alcune rilevazioni del campione non erano state effettuate nei calci di rigore.

In una raccolta di unità statistiche volta a prevedere la probabilità che un tiro qualsiasi vada a segno, è bene trascurare casi estremi come questi perché porterebbero a distorsioni nelle previsioni in maniera sicuramente più ottimistica. Infatti i rigori sono di per sé già *Expected Goals*; prendendo le stesse osservazioni di calci di rigore, la probabilità che siano andate a segno è

$$\frac{R_g}{N_r} = 0.8745,$$

dove

$$R_g = \text{Rigori andati a segno del campione osservato}$$
$$N_r = \text{Numero totale rigori del campione osservato.}$$

Osservando il tasso di rigori andati a segno che ben si discosta dal tasso di reti di una conclusione qualsiasi, illustrato nel capitolo successivo, e considerato che l'obiettivo dell'analisi statistica è quello di osservare e prevedere una conclusione a

rete qualsiasi, allora è ragionevole omettere tutte le osservazioni relative a calci di rigore poiché sono situazioni di gioco diverse non assimilabili agli altri tiri e quindi non trattabili nella stesso modo.

Sussistono infine ancora 47 celle della matrice con dati mancanti: 32 di queste si trovavano disposte tutte sulla stessa variabile `Distance.ToD2.360` che specifica la distanza a 360° di chi effettua il tiro dal secondo difensore più vicino (vedi il paragrafo 2.3.1), mentre le restanti 15 celle erano localizzate solo in tre righe del DataSet; per ogni riga le `NA` giacevano sulle stesse cinque variabili.

In quest'ultimo caso specifico è stato ragionevole eliminare le 3 osservazioni contenenti cinque celle *missing* perché costituenti una percentuale di informazione *persa* minima ($\frac{3}{6735}$) e quindi poco significativa ai fini ultimi dell'analisi.

Al contrario per le `NA` nella stessa variabile, volendo evitare la cassazione della colonna completa o di ogni riga (che non registravano *missing values* in coincidenza con altre variabili), è stato opportuno operare con un'imputazione. Poiché la percentuale di dati mancanti comunque era estremamente ridotta ($\frac{32}{6732 \cdot 23}$), si è reputato lecito proseguire con un'imputazione *libera* (ovvero senza fare ricorso ad algoritmi specifici) e senza commettere un marcato errore di `Bias`, anche se quanto imputato non fosse stato in realtà di ottima precisione. Per ogni osservazione con cella mancante si è potuto rilevare che si trattava di conclusioni a rete in cui la `Density` era prossima allo 0. Tale variabile misura la densità nei $30m^2$ attorno al punto dal quale partiva la conclusione a rete, calcolata come rapporto tra i giocatori presenti rispetto all'area di campo. Perciò si possono considerare quelle conclusioni come inerenti a situazioni di gioco in cui l'attaccante aveva *campo libero* a 360 gradi se non per un unico difensore più vicino (la cui distanza viene però descritta nella variabile `Distance.ToD1.360`). Un'adeguata tecnica sarebbe quindi stata imputare una distanza elevata che superasse il massimo della stessa variabile, ad esempio:

$$D_i = Max_i + 1 \; metri$$

dove

D_i = Dato mancante da imputare della colonna i-esima
Max_i = Massimo valore della variabile i-esima.

Capitolo 4

Analisi descrittiva delle variabili

Ottenuto un campione completo è opportuno osservare come le variabili si distribuiscono e per ognuna di esse determinare i valori delle funzioni statistiche descrittive, ai fini di una loro conoscenza più approfondita.

Durante questo capito verrà messo in luce uno studio delle variabili più significative per la costruzione dei modelli di calcolo degli xG, di natura illustrativa.

4.1 La variabile `Goal`

L'informazione che più interessa la ricerca statistica è se un tiro sia risultato effettivamente una rete o viceversa. Per questo motivo è utile crearsi la variabile Dummy `GOAL` che tramuta l'informazione in quantitativa, attribuendo il valore 1 all'unità statistica che si è rivelata *vincente* e il valore 0 alle altre.

Calcolata la media di questo vettore informazioni essa coincide il tasso di reti segnate dal Barcellona nel campione, ovvero la percentuale di conclusioni andate a segno rispetto alle conclusioni totali:

$$\frac{N_g}{N} = 0.156$$

dove

$$N_g = \text{numero di goal}$$
$$N = \text{numero totale tiri nel campione.}$$

La Dummy specificata sarà di fondamentale importanza sia in fase di costruzione dei modelli che nella fase di previsione poiché essa, relazionata alle previsioni messe in atto da un modello, dà misura di quali delle osservazioni sono state catalogate nei goal correttamente e quali no. Essa quindi rappresenta la variabile che detiene i risultati empirici che il modello, oggetto di interesse della composizione, intende

stimare. Nella figura viene mostrato un esempio di una parte del DataSet con l'aggiunta della Variabile `Goal`.

distance.ToD2.360	milliseconds	ElapsedTime	Goal
3.860052	385	62.385	1
1.897367	949	1212.949	0
12.000417	364	1314.364	0
2.302173	391	2028.391	0
5.100980	350	2887.350	0
6.100820	610	2911.232	0
6.946222	298	2948.920	1
6.407027	974	3549.596	0
10.049876	63	4202.685	0
1.166190	45	4285.667	0
3.590265	611	5844.233	0
10.945775	905	534.905	0

Figura 4.1: Esempio della colonna Goal all'interno del campione

4.2 Indici statistici per le variabili con maggiore rilievo

Sono state selezionate otto delle variabili che risulteranno le più rilevanti per il calcolo del modello, sulla base degli esiti ottenuti nel capitolo successivamente presentato, relativo alla fase di costruzione. Si è ricavata la matrice, mostrata in figura 4.2, che per ognuna di esse (righe) espone le proprietà statistiche (colonne).

	Min.	X1st.Qu.	Median	Mean	X3rd.Qu.	Max.	rSquare	varianza	Odds_Ratio
DistSGK	0.360	8.12	14.13	15.551	22.49	63.67551	0.08277	81.27074	NA
duration	0.001	0.35	0.69	0.754	1.05	6.15065	0.00349	0.26890	NA
distance.ToD1.360	0.200	1.53	2.46	3.383	4.18	16.00562	0.00062	7.16147	NA
shot.deflected	0.000	0.00	0.00	0.011	0.00	1.00000	NA	0.01074	6.7564978
DistToGoal	0.700	12.16	17.62	18.680	24.66	81.44354	0.07721	70.81017	NA
density.incone	0.000	0.00	0.13	0.248	0.37	5.00000	0.00768	0.14028	NA
DefendersBehindBall	0.000	1.00	4.00	4.247	7.00	10.00000	0.05470	9.75529	NA
under_pressure	0.000	0.00	0.00	0.152	0.00	1.00000	NA	0.12919	0.7605049

Figura 4.2: Analisi descrittiva delle variabili

L'informazione per ogni singola variabile descrive:

- Il minimo valore che acquisisce la variabile nel campione

- Il primo quantile che per definizione è la minima modalità della variabile, tale che la somma delle frequenze relative fino ad essa (inclusa) sia almeno $\frac{1}{4} = 0.25$ e che la somma delle frequenze relative successive a quella modalità sia al massimo $\frac{3}{4} = 0.75$.
- La mediana che per definizione è la minima modalità tale che la somma delle frequenze relative fino ad essa (inclusa) sia almeno $\frac{1}{4} = 0.5$
- La Media campionaria dei valori della variabile che per definizione è :

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

dove

$$x_i = \text{le N osservazioni della variabile, } \forall i \in 1...N$$

- Il terzo quantile che per definizione è la minima modalità tale che la somma delle frequenze relative fino ad essa (inclusa) sia almeno $\frac{3}{4} = 0.75$ e che la somma delle frequenze relative successive a quella modalità sia al massimo $\frac{1}{4} = 0.25$
- Il massimo valore che assume la variabile in questione
- La Varianza campionaria dei valori della variabile che rappresenta una misura della variabilità e lo scostamento di essa dalla sua media. Per definizione è data dalla seguente formula :

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Infine le ultime due misurazioni esposte mirano a valutare la relazione tra la variabile in questione e la variabile `Goal`; ovvero cercano di descrivere come sono collegate l'informazione disponibile all'interno delle otto variabili e l'effetto che ha sull'esito di un tiro in porta, *Goal* o *Non Goal*.

- Il *Coefficiente di determinazione* tra la variabile di interesse e la colonna `Goal`. Esso è un indice che rappresenta la relazione tra le due variabili. Viene misurato come il quadrato dell'indice di correlazione parziale ρ già visto in precedenza nella sezione 2.1.1 ed è quindi sempre non negativo ($R^2(\underline{y}, \underline{x}) \geq 0,\ \forall \underline{x}, \underline{y}$) [4]. Esso è calcolato quindi nella seguente maniera:

$$R^2 = \left[\frac{cov(x,y)}{\sigma_x \sigma_y}\right]^2 = \frac{\left[\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})\right]^2}{\sum_{i=1}^{N}(x_i - \overline{x})^2 \sum_{i=1}^{N}(y_i - \overline{y})^2}$$

dove

$$\underline{x} = \text{Variabile di interesse nella descrizione}$$
$$\underline{y} = \text{Variabile risposta}(\textit{Goal}).$$

- Per le variabili di tipo Binario (o *Dummy*), cioè quelle che assumono solo due modalità opposte (ad esempio 0 o 1) a seconda che sia soddisfatta o meno una data condizione, diversamente dalle altre, non viene mostrato l'indice di determinazione. Per esse viene esibito l'*Odds Ratio* (o *OR*) anche chiamato *rapporto di probabilità*. Esso rappresenta il grado di correlazione tra due fattori di tipo binario, attraverso il rapporto tra gli *Odds*. Il calcolo di quest'ultimi prevede il confronto tra le frequenze in cui la variabile risposta `Y=Goal` assume una modalità (0 o 1), dato il valore della variabile binaria di interesse `X` [14]. Infatti:

$$OR = Odd_1/Odd_0 = \frac{Freq(Y=1|X=1)/Freq(Y=1|X=0)}{Freq(Y=0|X=1)/Freq(Y=0|X=0)}.$$

Capitolo 5

Analisi dei modelli di statistical learning per il calcolo degli *Expected Goals*

La statistica inferenziale è una branca della statistica che si occupa di analizzare i dati ottenuti da un campione della popolazione per stimare un fenomeno statistico sull'intera popolazione di riferimento. L'analisi che verrà esposta in questo capitolo perciò mira a formare il modello: un metodo di calcolo delle statistiche avanzate `Expected Goals`, utilizzando i dati del campione, definito e studiato da un punto di vista descrittivo nei capitoli precedenti che ha l'obbiettivo di stimare al meglio il fenomeno per ogni unità statistica della popolazione di tiri verso la porta da parte del Barcellona.

Verranno utilizzati diversi metodi di *statistical learning*.

Durante l'identificazione del modello è opportuno procedere a due fasi: costruzione del modello seguita dalla sperimentazione della prestazione predittiva. Sulla base dell'accuratezza nelle previsioni si baserà infine il confronto finale tra i metodi.

Per fare ciò, si procede con un approccio di *Cross Validation* a due *fold* per la quale vengono divise le unità statistiche del DataSet in due parti: `Training Set`, parte nella quale viene stimato il modello e, se necessario, parte sulla quale vengono fatte le modifiche per migliorarne la performance; `Test Set` contenente tutte le osservazioni sulla base delle quali viene testato il modello e calcolati la capacità predittiva e l'errore di stima. Questa tecnica statistica è utilizzabile perché in presenza di una buona numerosità del campione osservato.

Per operare seguendo questo criterio le righe della matrice dei dati sono state suddivise: il 60% di esse è entrato a fare parte del `Training Set` (che conta perciò 4036 osservazioni) mentre la restate parte costituisce il `Test Set`. Per riscuotere causalità nella divisione del campione, si è optato per un'estrazione random delle osservazioni senza reinserimento da immettere nel `Training Set` e le rimanenti

sono state introdotte nel `Test Set`. Per ogni tipo di tecnica di Statistical Learning che verrà esposta nelle prossime Sezioni, l'approccio sarà di questo tipo.

5.1 Albero di Classificazione

La prima metodologia adottata è l'algoritmo `CART` (*Classification and Regression Trees*). Nel nostro caso specifico verranno usati gli *Alberi di classificazione* o *Alberi decisionali* poiché la variabile da predire è di tipo categorico.

5.1.1 Come funziona un Albero Decisionale

Da un punto di vista teorico date $\underline{Y}$ variabile risposta di tipo categoriale e $y \in C = \{0,1,2,..,J\}$ modalità che essa può assumere, dove $y = j$ indica che la Y assume il valore della j-esima modalità, e dati dei vettori $\underline{x_1}, ..., \underline{x_d}$ di variabili esplicative tali che $\underline{x_i} \in D \subset \mathbb{R}^k \; \forall i \in \{1, ..., d\}$ con $k \in \mathbb{N}$, lo scopo dell'analisi è quello di costruire un `classificatore`, ossia una regola per cui a partire da un vettore di variabili esplicative $\underline{x}$ si possa associare un'etichetta $j \in C$ alla variabile d'interesse[11]. Quindi si definisce un classificatore come una funzione d($\cdot$), tale che $d : D \to C$, quindi

$$\forall \underline{x} \in D, \exists j \in C : d(\underline{x}) = j.$$

Gli algoritmi di classificazione operano partizionando D in sottoinsiemi allo scopo di classificare le unità in gruppi il più omogenei possibile al loro interno e quanto più differenziati tra di loro. La partizione è di tipo binario: ad ogni passo l'insieme di partenza viene suddiviso in due sottoinsiemi propri mediante un'operazione di *split*[17].

E' dunque possibile associare ai modelli *Classification Trees* una rappresentazione grafica a forma di Albero 5.1.

Un Albero può essere visto come un insieme finito di elementi detti nodi e il nodo iniziale da cui si diramano i successivi viene detto radice. L'insieme dei nodi, ad esclusione del nodo radice può essere suddiviso in h insiemi distinti, $S_1, ..., S_h$, detti *sottoalberi*; un nodo viene chiamato *padre* rispetto ai nodi che esso genera, mentre viene denominato *figlio* rispetto al nodo da cui discende. I nodi terminali, ovvero senza figli, sono denominati *foglie*.

Ad ogni nodo h viene associato un sottoinsieme di D, in particolare al nodo radice si associa D intero e ai due figli di un nodo, i due insiemi in cui questo è stato bipartito.

I passi fondamentali per la costruzione di un *Albero di Classificazione* sono:

- la selezione degli *splits*,
- la decisione su quando dichiarare un nodo terminale o continuare a dividerlo,

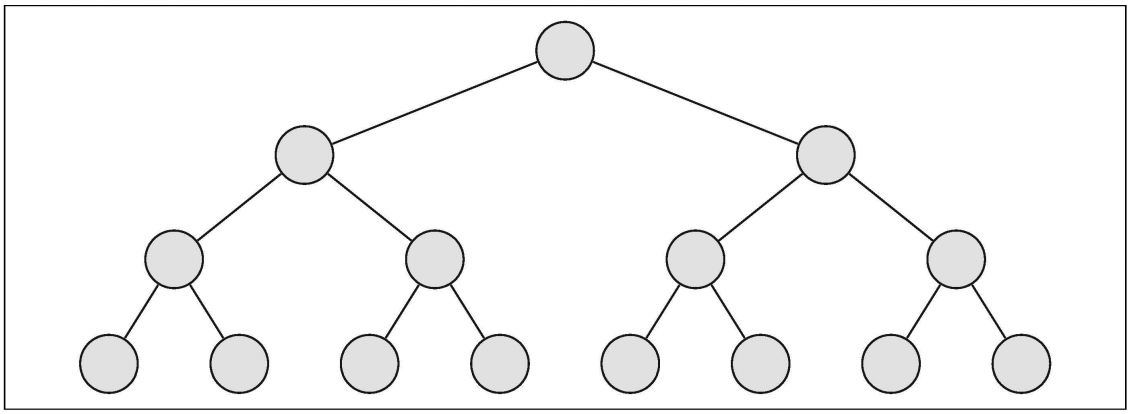

Figura 5.1: Struttura algoritmo ad Albero

- l'assegnazione ad ogni nodo terminale di una modalità per Y. [11]

Partendo dal nodo radice h_1 si effettua una successione di *splits*; in ogni nodo si sceglie quello che rende più omogenei i dati all'interno dei due nodi discendenti. L'operazione di *split* consiste nel individuare un valore soglia per una delle variabili sulla base del quale si suddividono le osservazioni.

Per tale operazione in primo luogo si deve definire la probabilità condizionata $P(j|h)$ che $y = j$ dato $\underline{x} \in h$ come:

$$P(j|h) = \frac{n_j(h)}{n_h}.$$

dove:

$$n_j(h) = \texttt{numero di casi con } \underline{x_i} \in h \texttt{ e } y = j$$
$$n(h) = \texttt{numero di casi con } \underline{x_i} \in h.$$

Da essa infatti dipende la misura di *Impurità* per il nodo h$i(h)$. Questo indice è definito come funzione delle probabilità condizionate $f(P(1|h), P(2|h), ..., P(J|h))$ con le seguenti caratteristiche:

- è massima in $(\frac{1}{j}, \frac{1}{j}, ..., \frac{1}{j})$;
- $f(1,0,...,0) = f(0,1,...,0) = ... = f(0,0,...,1) = 0$
- è simmetrica.

L'indice d'impurità più noto è quello di Gini, dato da

$$i(h) = \sum_{j \neq i} p(j|h)p(i|h) = \left[\sum_j p(j|h)\right]^2 - \sum_j p(j|h)^2 = 1 - \sum_j p(j|h)^2.$$

E' fondamentale il calcolo della misura di impurità perché il decremento in impurità è alla base dell'assegnazione degli *splits*. Per ogni nodo h lo *split* s genererà i figli h_l ed h_r, con p_l proporzione dei casi di h che va in h_l e la restante proporzione p_r in h_r, a seconda che $\underline{x_i} \in h$ sia in h_l o viceversa. In ogni nodo h si sceglie quello *split* $s*$ che massimizza il *decremento in impurità* [11], cioè :

$$s : max \{\Delta i(s,h) = i(h) - p_l i(h_L) - p_r i(h_r)\}$$

Una volta terminata la costruzione dell'Albero, vengono assegnate le modalità ai nodi terminali. Considerato l'Albero H, l'assegnazione di una modalità $j \in C \; \forall h \in H$ terminale è effettuata mediante una funzione $j(h)$. Precisamente, se $p(j|h) = max_i[p(i|h)]$, allora $j(h) = j$, ossia all'h-esimo nodo si assegna l'etichetta della variabile y più presente nel nodo considerato; se il massimo è raggiunto da più modalità, si sceglie arbitrariamente una fra queste.

5.1.2 I risultati ottenuti dall'Albero

Applicando il *Classification Tree* al campione delle conclusioni a rete del Barcellona, vengono utilizzate tutte le colonne del `DataSet` del Barcellona come componenti dei predittori $\underline{x_1}, ..., \underline{x_d}$ e viene fissata come variabile target $\underline{Y}$ la colonna *Goal*, ottenendo perciò come insieme C delle modalità che essa può assumere l'insieme {0,1}.

I risultati riscossi sono contenuti nell'albero della figura 5.2.

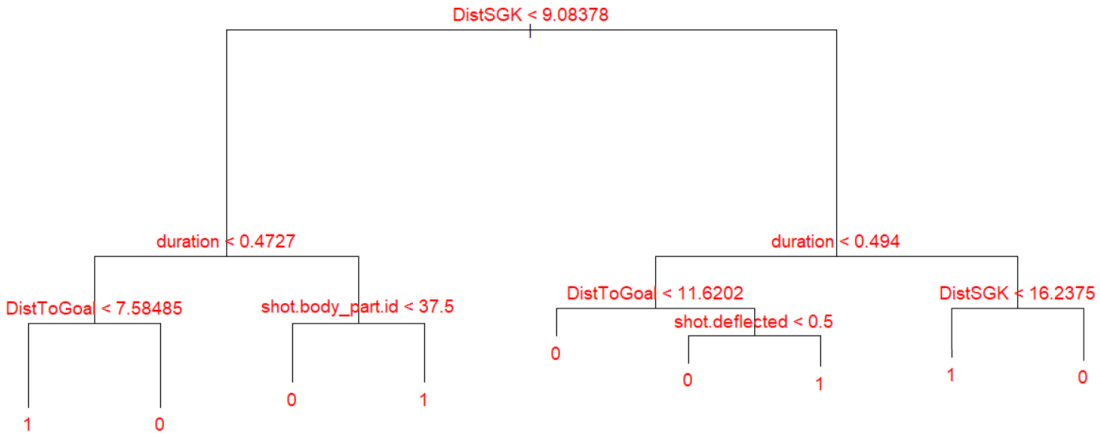

Figura 5.2: Albero Decisionale Barcellona

L'Albero ha 9 nodi terminali (o nodi foglia).

Cinque variabili sono entrate a far parte degli split dell'Albero (`DistSGK`, `duration`, `DistToGoal`, `shot.bodyPart.id`, `shot.deflected`). Esse sono quindi quelle per le quali c'è una rilevante relazione con la probabilità di segnare, ovvero per cui si è osservato nel campione che ad una loro modifica varia significativamente la probabilità di andare a segno.

Per interpretare la figura si procede come segue: data un'osservazione, ovvero un vettore di variabili esplicative $\underline{x}$, si segue il percorso dell'Albero dall'alto verso il basso. Ad ogni nodo non terminale se il vettore presenta la caratteristica della variabile indicata in esso minore del valore soglia indicato, allora si procederà andando verso la diramazione alla sinistra (il nodo figlio di sinistra), viceversa se maggiore o uguale si procede seguendo il ramo di destra. Proseguendo in questo modo lungo tutta la diramazione si giunge ai nodi foglia: in essi vi è la previsione della modalità che per quel vettore informazioni specifico assume la variabile `Goal`:

- 1 se la conclusione viene stimata come *goal*
- 0 viceversa.

5.1.3 Valutazione dell'accuratezza nelle previsioni dell'Albero

Come preannunciato, prima di poter ritenere un modello collaudato si necessita una fase in cui viene testata la capacità di predire da parte del modello. Vengono analizzate le stime messe in atto dal modello per le osservazioni del `Test Set` e vengono confrontate con i risultati campionari della variabile risposta `Goal` delle relative unità statistiche.

	0 campionari	1 campionari
0 stimati	2102	210
1 stimati	180	198

Figura 5.3: Confronto tra stime e valori reali nel Test Set

Si forma perciò una tabella doppia entrata come mostrato nella figura 5.3. Da questa semplice matrice si può ricavare quanti `0` o `1 campionari` vengono stimati correttamente dal modello e quanti invece no. Con questi dati è possibile esaminare la performance del modello analizzando:

- la sua *Capacità di predizione*, ovvero il rapporto tra il numero delle osservazioni ben stimate sulle osservazioni totali del Test Set:

$$C_p = \frac{n(0_{stim.} \cap 0_{camp.}) + n(1_{stim.} \cap 1_{camp.})}{n_{tot}(TestSet)} = \frac{2102 + 198}{2690} = 0.855$$

- *l'Errore di stima Test Set*, rapporto tra il numero delle osservazioni erroneamente stimate sul totale delle osservazioni del Test Set:

$$E_s = 1 - Cp = \frac{n(0_{stim.} \cap 1_{camp.}) + n(1_{stim.} \cap 0_{camp.})}{n_{tot}(TestSet)} = \frac{210 + 180}{2690} = 0.145$$

- la sua *Specificità*, nonché la capacità del modello di identificare correttamente le osservazioni che non sono soggette al fenomeno (che sono risultate `NON GOAL`). Se un test ha un'ottima specificità allora è basso il rischio di falsi positivi [12] (osservazioni che sono state stimate come *reti* ma che invece non lo erano):

$$Sp = 1 - \frac{n(1_{stim} \cap 0_{camp.})}{n(0_{campTot.})} = 1 - \frac{180}{2102 + 180} = 0.921$$

- la sua *Sensibilità*, la capacità del modello di identificare correttamente le osservazioni che sono soggette al fenomeno (che sono risultate `GOAL`). Se un test ha un'ottima sensibilità allora è basso il rischio di falsi negativi[12] (osservazioni che sono state stimate come *non reti* ma che invece lo erano):

$$Ss = 1 - \frac{n(0_{stim} \cap 1_{camp.})}{n(1_{campTot.})} = 1 - \frac{210}{210 + 198} = 0.485$$

Alla luce di questi risultati, si sottolinea che il modello sembra predire in maniera soddisfacente, presenta un'ottima specificità ma una bassa sensibilità.

Per studiare le prestazioni del modello in maniera più approfondita e quindi detenere risultati più robusti, è ragionevole sviluppare gli stessi calcoli degli indici statistici sulla base di diverse divisioni del campione in `Training Set` e `Test Set`. Adoperando quindi 4 diverse estrazioni fortuite senza reinserimento dal campione otteniamo 4 diversi Campioni `Training` e `Test`.

Con questo metodo sono stati ottenuti gli esiti illustrati nella tabella 5.1.

Indicatori performance predittiva				
Estrazioni	Specificità	Sensibilità	Errore di stima	Capacità predittiva
1	0.921	0.485	0.145	0.855
2	0.926	0.503	0.139	0.861
3	0.944	0.378	0.139	0.86
4	0.914	0.415	0.166	0.834

Tabella 5.1: Performance predittiva albero

5.1.4 Commento ai risultati ottenuti dell'albero di classificazione

La capacità del modello di predire, è risultata in tutti e quattro le estrazioni, di buon livello. Infatti in media nell'85% delle osservazioni dei diversi *Test Set* le stime effettuate sono in linea con i risultati empirici ed è da considerare, durante la

visione dei risultati, il fatto che il risultato di una conclusione a rete è qualcosa di molto aleatorio e quindi non facile da stimare; basti pensare a quante situazioni di gioco nelle partite, si sono tradotte in goal in maniera del tutto inaspettata, perché assolutamente soggettive.

Si deve notare come, dai risultati conseguiti, il modello riesce ad identificare in maniera più precisa le osservazioni che sono rivelate non goal (`GOAL`= 0 rispetto a quelle che sono vincenti. Infatti il valore della *Sensibilità* è significativamente minore della *Specificità* in tutte e quattro i sorteggi del *Test Set*, e inoltre essa presenta un intervallo massimo di oscillazione nei quattro esiti ottenuti, esteso e maggiore delle altre misurazioni:

$$Sens_{max} - Sens_{min} = 0.504 - 0.378 = 0.126$$

I risultati oscillano su quattro diversi campioni test del 12.6% circa. Questo fenomeno lascia intendere che gli esiti in termini di *Sensibilità* sono più variabili rispetto agli altri indici e quindi meno precisi.

5.2 Random forest e Bagging

Altri metodi sviluppati per l'estrapolazione del modello di calcolo degli xG sono il *Bagging* e il *Random Forest*. Queste tecniche di statistical learning utilizzano gli alberi decisionali come base per costituire modelli predittivi più potenti e robusti. Gli alberi di classificazione e regressione discussi nella sezione 5.1 soffrono di variabilità elevata. Ciò significa che dividendo i dati del `Training Set` in due parti forfettariamente e sviluppando un albero decisionale per ciascuna metà, i risultati ottenuti potrebbero essere abbastanza diversi. Al contrario, una procedura con bassa varianza produrrà risultati simili se applicata ripetutamente a set di dati distinti.

Il Bagging e il Random Forest sono procedure che permettono di ridurre la varianza di un metodo di Statistical Learning. Dato un insieme di n osservazioni indipendenti $Z_1, ..., Z_n$ ciascuna con varianza σ^2, la varianza della media $\overline{Z}$ delle osservazioni è data da $\frac{\sigma^2}{n}$. In altre parole, servirsi della media di una serie di osservazioni comporta una riduzione della varianza [17]. Dunque un modo naturale per ridurre la variabilità e quindi aumentare la precisione della previsione è quello di considerare molteplici sottocampioni del campione totale eseguendo un *Bootstrap*, ovvero prelevando casualmente ripetuti *Training Set* per poi costruire un modello di previsione separato per ognuno di essi e fare una media delle previsioni risultanti.

5.2.1 Il metodo Bootstrap

Entrambi le tecniche, come anticipato, si servono del metodo di *Bootstrap*. Esso è un potente strumento statistico di cui ci si avvale per quantificare l'incertezza associata a un determinato stimatore o metodo di *statistical learning*.

Supponendo di avere un campione casuale di n osservazioni, $\underline{x} = (x_1, x_2, ..., x_n)$, da una non nota distribuzione F e di voler calcolare una stima $\hat{\theta} = s(\underline{x})$, la tecnica permette quindi di stimare lo *Standard Error* (SE) di $\hat{\theta}$ [16].

Per eseguire questa stima secondo il metodo, è necessario definire il *Bootstrap sample* ovvero il *campione Bootstrap*. Esso, qualificato come $\underline{x}' = (x'_1, x'_2, .., x'_n)$, consiste in un campione casuale n-dimensionale estratto con reinserimento da $\underline{x}$. L'insieme *Bootstrap* $\underline{x}'$ contiene quindi i membri del DataSet originale $\underline{x}$, alcuni apparendo zero volte, altri due o una ecc. .

In corrispondenza del *campione Bootstrap* si può valutare $\hat{\theta}' = s(\underline{x}')$ e di conseguenza ottenere la stima *Bootstrap* di $SE_F(\hat{\theta})$: una stima *plug-in* utilizzando la distribuzione empirica $\hat{F}$ al posto della sconosciuta F indicata con

$$SE_{\hat{F}}(\hat{\theta}').$$

Tuttavia è molto inverosimile che esisti una funzione o una formula chiara per la stima di θ (eccetto ad esempio per la stima della media); per questo motivo è più opportuno procedere implementando l' *algoritmo Bootstrap* che offre una buona approssimazione di $SE_{\hat{F}}(\hat{\theta}')$.

Esso lavora selezionando K campioni *Bootstrap* $\underline{x}'_1, \underline{x}'_2, ..., \underline{x}'_K$ ognuno dei quali costituito da n valori attinti da $\underline{x}$. Calcola per ognuno di esso il valore:

$$\hat{\theta}(k) = s(\underline{x}'_k) \quad k = 1,2, ..., K \ .$$

Infine stima $SE_F(\hat{\theta})$ attraverso la *Standard Deviation* campionaria delle B stime:

$$S\hat{E}_K = \left\{ \sum_{k=1}^{K} \frac{\left[\theta(\hat{k})' - \sum_{k=1}^{K} \hat{\theta}(k)'\right]^2}{K-1} \right\}^{\frac{1}{2}} .$$

5.2.2 Foreste di Alberi di Classificazione

I metodi di *Bagging* e di *Random Forest* si costruiscono quindi sviluppando K *Alberi Decisionali* con l'algoritmo `CART` per i K campioni *Bootstrap*. Se i modelli ad albero intendono stimare una variabile di tipo categoriale (come poi accadrà nel caso dell'elaborato dove la risposta è `Goal`, variabile *Dummy*), non si procede come ultimo passo con una media, ma con un voto di maggioranza ottenendo comunque lo stesso effetto risultante sulla varianza: per una data osservazione $\underline{x}$ viene registrata la modalità della risposta $\underline{Y}$ prevista da ogni albero e infine viene scelta la previsione che maggiormente si manifesta negli K alberi.

Viene a formarsi perciò una *foresta* di Alberi Decisionali, che quindi non è esprimibile con un grafico rappresentativo come diversamente avviene per i *Classification Trees* (dove vi risultava un diagramma facilmente interpretabile), ma che ha lo stesso scopo e la stessa tipologia di risultato. Per questo motivo i metodi di *Random*

Forest e *Bagging*, nonostante nella maggioranza dei casi presentino miglioramenti in quanto a performance predittiva per la minor varianza, comportano come svantaggio la difficoltà nell'interpretazione e la poca chiarezza su quali siano le variabili più rilevanti per la procedura.

Sebbene la raccolta di una *foresta* di alberi sia molto più difficile da interpretare di un singolo albero, si può ottenere un riepilogo generale dell'importanza di ciascun predittore usando l'indice di *Gini*. La rilevanza di una variabile può essere valutata attraverso la diminuzione media (tra i K alberi della foresta) del suddetto indice. Questo, come ricordato nella sezione 5.1.1, è una misura di impurità e risulta diminuire (o in generale variare) in corrispondenza del nodo che *splitta* tale variabile.

5.2.3 Differenza tra bagging e random forest

Quanto spiegato nella sezione 5.2.2 sopra esposta, vale per entrambi gli algoritmi di *statistical learning* impiegati nell'elaborato. La differenza tra i due metodi applicati al campione delle conclusioni della squadra catalana è che la tecnica di *Random Forest* prevede che, per ogni *Training Set* estratto dal campione, vengano utilizzati per la costruzione del relativo *Classification Tree* solo m delle p variabili predittori (con $m < p$) sorteggiate in maniera casuale; differentemente dal *Bagging* che per ogni albero impone l'utilizzo del numero totale p di predittori [17].

Il *Random Forest* rispetto al *Bagging* supera il problema della correlazione tra gli alberi. Infatti, supponendo che vi sia una variabile esplicativa con caratteristica di forte predizione, negli algoritmi dei *Bagging Classification trees* il forte predittore sarà presente per ogni albero sviluppato nei diversi sottocampioni *training*, mentre in un *Random Forest* non è detto che esso faccia parte delle m su p variabili sorteggiate in ogni albero. In questo modo nel primo caso avremo una forte correlazione tra i risultati degli alberi, mentre con il secondo metodo essi saranno meno relazionati perché aventi diversi predittori e rivelando perciò una maggiore informazione [17].

5.2.4 Risultati ottenuti dal Bagging

Viene dunque dapprima adoperato l'algoritmo *Bagging tree* ai dati nel campione delle conclusioni verso la porta della squadra Barcellona per ottenere il calcolo degli *Expected Goals* utilizzando tutte le colonne del `DataSet` come variabili esplicative e la variabile `GOAL` come risposta $\underline{Y}$ da predire.

La rilevanza delle singole esplicative che sono entrate nell'algoritmo, è data dalla tabella in figura 5.2.

In esso sono riportate due misure di importanza. La prima, `Mean Decrease Accuracy` (*MDA*), si basa sulla diminuzione media dell'accuratezza nelle previsioni

Importanza Variabili		
Variabili	Mean Decrease Accuracy	Mean Decrease Gini
underPressure	2.83	3.32
shot.type.id	2.56	0.31
shot.bodyPart.id	35.8	40.16
shot.technique.id	5.62	14.46
shot.deflected	25.03	14.48
DistSGK	61.94	163.12
duration	71.15	126.99
DistToGoal	36.99	102.2
DistToKeeper	35.4	66.68
AngleToGoal	13.22	45.23
AngleToKeeper	12.22	47.81
AngleDeviation	15.63	62.62
density	15.16	30.66
density.incone	13.29	17.83
distance.ToD1	18.37	43.26
distance.ToD2	13.33	37.00
AttackersBehindBall	15.74	11.77
DefendersBehindBall	12.11	11.57
DefendersInCone	8.13	6.93
InCone.GK	6.96	4.93
DefArea	6.42	55.20
distance.ToD1.360	21.42	39.17
distance.ToD2.360	13.12	39.07
milliseconds	0.91	51.66
ElapsedTime	-1.33	45.97

Tabella 5.2: Indici importanza variabili

delle osservazioni *out-of-bag* (*OOB*) (alcune osservazioni che all'interno del Training Set vengono tralasciate durante la stima del modello), quando la variabile viene esclusa dal modello [17]. In questo caso viene definita la accuratezza come l'indice di *Capacità predittiva* (C_p), definito nella sezione 5.1.3. La seconda, come spiegato in precedenza, calcola la diminuzione totale dell'impurità del nodo che risulta dalle divisioni su quella variabile (mediando tra tutti gli alberi). Come misura di impurità è stata utilizzata come nella costruzione dell'albero decisionale l'indice di *Gini* illustrato nella sezione 5.1.1.

Più alti sono i valori, maggiore è il rilievo della variabile.

E' utile raffigurare le dimensioni di rilevanza in un grafico *importanza-variabile*. Quest'ultimo risulta facilmente interpretabile e rende chiaro quali variabili si rivelano più importanti secondo i due diversi indici.

Ad esempio, secondo entrambe le misure utilizzate risultano di fondamentale rilevanza per il calcolo del *Goal atteso* la durata del tiro in secondi (`duration`) e la distanza di chi effettua la conclusione a rete dal portiere (`DistSGK`), poiché come ben chiaro nella raffigurazione 5.4 presentano valori elevati per le due misure e sono le prime variabili esposte in ordine decrescente di importanza.

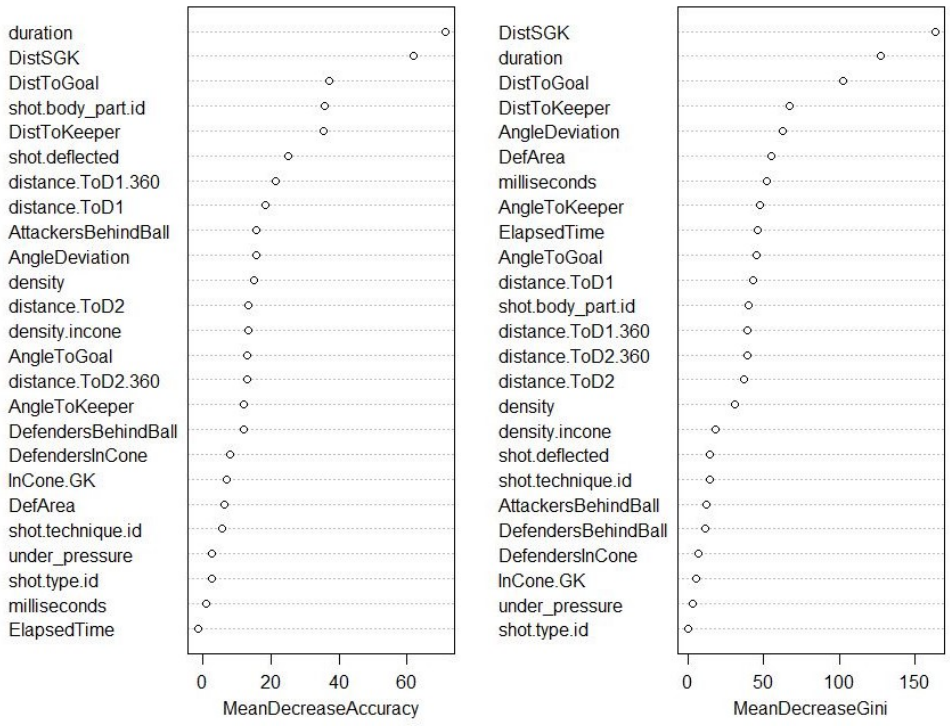

Figura 5.4: Grafico Importanza Variabili

5.2.5 Valutazione dell'accuratezza nelle previsioni del Bagging

	0 campionari	1 campionari
0 stimati	2182	249
1 stimati	100	159

Figura 5.5: Stime Bagging - Valori reali

Il modello estrapolato necessita di essere testato in termini di performance predittiva. Utilizzando quindi le osservazioni del *Test Set* vengono stimate le relative modalità `Goal` (1) e `Non Goal` (0).

Si è sviluppata, come in precedenza con gli *alberi CART* nella sezione 5.1.3, la tabella a doppia entrata per confrontare le stime delle unità statistiche nel *Test Set* e i valori campionari osservati nello stesso sottocampione per la variabile risposta.

La tabella di riferimento è quella riportata a fianco nella figura 5.5.

Viene dimensionata la performance attraverso il calcolo di:

- la sua *Capacità di predizione*:

$$C_p = \frac{n(0_{stim.} \cap 0_{camp.}) + n(1_{stim.} \cap 1_{camp.})}{n_{tot}(TestSet)} =$$

$$= \frac{2182 + 159}{2690} = 0.87$$

- l'*Errore di stima Test Set*:

$$E_s = 1 - C_p = \frac{n(0_{stim.} \cap 1_{camp.}) + n(1_{stim.} \cap 0_{camp.})}{n_{tot}(TestSet)} =$$

$$= \frac{249 + 100}{2690} = 0.13$$

- la sua *Specificità*:

$$Sp = 1 - \frac{n(1_{stim} \cap 0_{camp.})}{n(0_{campTot.})} = 1 - \frac{100}{2182 + 100} = 0.956$$

- la sua *Sensibilità*:

$$Ss = 1 - \frac{n(0_{stim} \cap 1_{camp.})}{n(1_{campTot.})} = 1 - \frac{249}{249 + 159} = 0.39.$$

Per un maggiore approfondimento dei risultati, si studiano questi indici nelle quattro diverse divisioni del DataSet in *Training* e in *Test*, già adoperate in precedenza con l'albero decisionale nella sezione 5.1.3, in modo da ottenere una gamma valori per questi indici predittivi. Si ottengono gli esiti esposti nella figura 5.3.

Indicatori performance predittiva				
Estrazioni	Specificità	Sensibilità	Errore di stima	Capacità predittiva
1	0.958	0.392	0.131	0.869
2	0.963	0.377	0.126	0.873
3	0.957	0.408	0.125	0.875
4	0.968	0.306	0.138	0.861

Tabella 5.3: Performance predittiva Bagging

5.2.6 Commento agli esiti del Bagging

La performance predittiva si è delineata di ottimo livello e in maniera più efficace del modello ad *Albero di Classificazione* studiato nella sezione precedente 5.1.2. Infatti in media in ogni diversa estrazione considerata più dell'86% delle le stime effettuate coincidono correttamente con i risultati empirici nonostante l'aleatorietà delle osservazioni prese in considerazione, nella popolazione dei tiri verso la porta del Barcellona, di stampo poco oggettivo.

Inoltre anche in questo caso la differenza tra *Sensibilità* e *Specificità* è notevole: la prima risulta molto meno significativa della seconda, a testimonianza di una capacità predittiva più efficace per le osservazioni che non presentano il fenomeno (che non si sono tradotte in rete). Infatti il valore *Specificità* risulta estremamente di rilievo e superiore anche della stessa misura rilevata dal modello ad albero. Diversamente l'indice di *Sensibilità* è peggiorato, a conferma di una differenza tra le due stime, piuttosto ampia.

5.2.7 Risultati ottenuti dal Random Forest

Così come per il *Bagging*, allo stesso modo viene eseguito il modello di *Random Forest* al campione osservato dei tiri del Barcellona. Inoltre in questo caso è necessario decidere il numero m dei predittori che saranno considerati per gli N alberi all'interno della foresta, estratte in modo causale per ognuno di essi; per esso viene preso la radice quadrata del numero delle variabili presenti

$$m = \sqrt{p} = \sqrt{25} = 5.$$

Per facilitare un futuro confronto, anche in questo caso si è proceduto rilevando l'importanza di ciascuna variabile predittore con le misure di *Mean Decrease accuracy* e di *Mean Decrease Gini*, ottenendo quindi i seguenti risultati esposti nella tabella 5.4.

I Risultati ottenuti per le variabili, affinché si goda di una veloce lettura, sono stati anche in questo caso sviluppati sul grafico *importanza-variabili* esibito in figura 5.6.

Nella raffigurazione 5.6 vengono disposte le variabili in ordine decrescente per importanza. Come si può notare le due misure di rilevanza non sono identiche e non sempre correlate. Infatti alcune variabili presentano un valore del `MDG` molto discostato dal proprio `MDA`, perché effettivamente questi indici nonostante misurino la stessa caratteristica e spesso uno implica l'altro, hanno significati diversi. Il *Mean Decrease Accuracy*, trattandosi di accuratezza per le previsioni degli *OOB* (vedi sezione 5.2.4), misura la rilevanza della variabile in termini di capacità di eseguire

Importanza variabili		
Variabili	Mean Decrease Accuracy	Mean Decrease Gini
underPressure	1.83	4.24
shot.type.id	3.44	0.56
shot.bodyPart.id	19.27	20.07
shot.technique.id	5.88	12.60
shot.deflected	15.39	11.33
DistSGK	37.20	108.08
duration	53.67	112.08
DistToGoal	33.31	104.05
DistToKeeper	31.08	70.38
AngleToGoal	17.14	49.52
AngleToKeeper	16.25	51.87
AngleDeviation	19.48	60.55
density	19.11	48.02
density.incone	15.33	22.63
distance.ToD1	17.18	52.96
distance.ToD2	16.42	50.75
AttackersBehindBall	14.18	17.73
DefendersBehindBall	15.44	26.96
DefendersInCone	8.78	10.36
InCone.GK	14.04	14.12
DefArea	5.51	49.77
distance.ToD1.360	22.23	48.76
distance.ToD2.360	14.55	44.41
milliseconds	-1.19	46.36
ElapsedTime	-1.06	44.90

Tabella 5.4: Importanza variabili nel Random Forest

una performance predittiva ottima. Il secondo si concentra invece più sulla purezza che una variabile riesce ad apportare, attraverso la propria presenza nel modello.

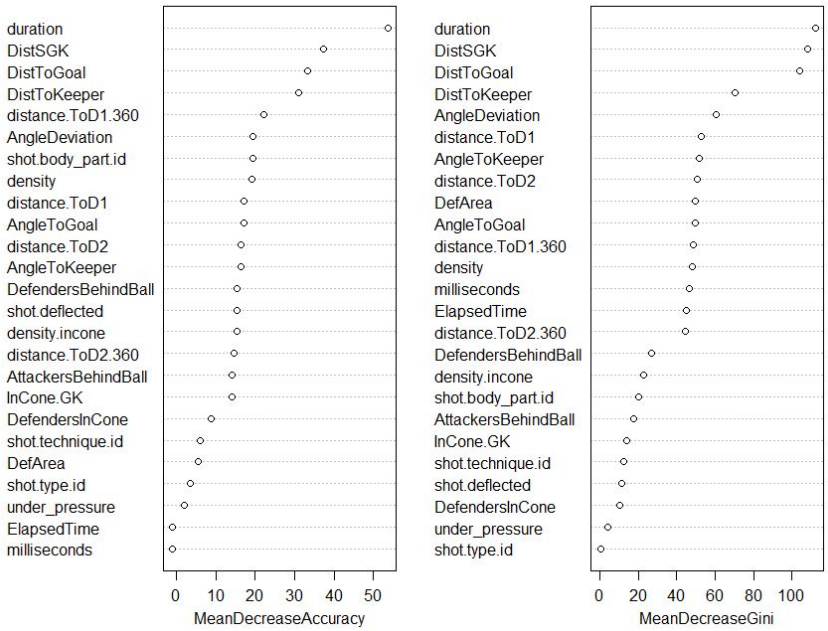

Figura 5.6: Grafico Importanza Variabili Random Forest

5.2.8 Valutazione dell'accuratezza nelle previsioni del Random Forest

	0 campionari	1 campionari
0 stimati	2229	284
1 stimati	53	124

Figura 5.7: Stime Random Forest - Valori reali

La prestazione predittiva del modello *RF* illustrato, è stata in seguito testata. Come per le due tecniche statistiche sopra esposte, vengono predette le modalità della variabile d'interesse per le osservazioni del *Test Set*, lasciate da parte durante la formazione del modello. In seguito, costruendo la tabella con le stime predette del Test set e le relative modalità campionarie, si possono determinare gli indici di performance.

A partire dagli esiti vengono calcolati :

- la sua *Capacità di predizione*:

$$C_p = \frac{n(0_{stim.} \cap 0_{camp.}) + n(1_{stim.} \cap 1_{camp.})}{n_{tot}(TestSet)} = \frac{2229 + 124}{2690} = 0.875$$

- l'*Errore di stima Test Set*:

$$E_s = 1 - C_p = \frac{n(0_{stim.} \cap 1_{camp.}) + n(1_{stim.} \cap 0_{camp.})}{n_{tot}(TestSet)} = \frac{284 + 53}{2690} = 0.125$$

- la sua *Specificità*:

$$Sp = 1 - \frac{n(1_{stim} \cap 0_{camp.})}{n(0_{campTot.})} = 1 - \frac{53}{2229 + 53} = 0.977$$

- la sua *Sensibilità*:

$$Ss = 1 - \frac{n(0_{stim} \cap 1_{camp.})}{n(1_{campTot.})} = 1 - \frac{284}{284 + 124} = 0.30$$

In seguito sviluppando questi indici anche sulle quattro diverse divisioni del Campione osservato in *Test* e *Train*, già adoperate per i precedenti metodi, si ottengono i risultati alla tabella 5.5.

Indicatori performance predittiva				
Estrazioni	Specificità	Sensibilità	Errore di stima	Capacità predittiva
1	0.974	0.304	0.129	0.871
2	0.98	0.275	0.128	0.872
3	0.973	0.33	0.119	0.881
4	0.979	0.225	0.143	0.857

Tabella 5.5: Performance predittiva Random Forest

5.2.9 Commento ai risultati del Random Forest

La prestazione predittiva generale del modello è di eccellente livello e supera sia il modello ad *Albero di Classificazione* che quello modello ottenuto dall'algoritmo *Bagging*. In media, nei quattro diversi sorteggi campionari dei *Test Set* considerati, le stime correttamente identificate superano l'87%, a testimonianza di un livello di predizione molto elevato.

Tuttavia la ottima capacità di predire è ben diversificata tra le stime effettuate delle osservazioni andate a rete e quelle non vincenti. Infatti il divario tra i due valori di *Sensibilità* e *Specificità* in tutte le diverse estrazioni prese in valutazione, è molto spiccata: la prima risulta poco convincente, con un minimo addirittura del 22%; al contrario la seconda risulta persino prossima al massimo valore conseguibile,1. Il *gap* tra le grandezze è decisamente più ampio del modello ad albero, e più considerevole anche del modello *Bagging* che presentava la stessa caratteristica.

5.3 Modelli di Regressione Logistica

Il *Modello di Regressione Logistica* o *Binaria* viene utilizzato quando si è interessati a studiare o analizzare la relazione causale tra una variabile dipendente dicotomica e una o più variabili indipendenti [9].

La Regressione Logistica è un modello statistico che presenta un approccio diverso da quelli esaminati nelle sezioni precedenti di tipo *non parametrico*. Esso infatti, data la variabile target risposta $\underline{Y}$ di tipo categoriale che assume le due modalità {0,1}, ha come obbiettivo quello di stimare la probabilità che un determinato vettore di variabili esplicative $\underline{x} = (x_1, x_2, ..., x_p)$ appartenga ad una delle due categorie, piuttosto che stimare direttamente la modalità della risposta. Attraverso un modello di regressione multiplo di tipo additivo, si necessita quindi calcolare la probabilità $P(Y = 1 \mid \underline{x})$:

$$P(Y = 1 \mid \underline{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon$$

dove β_i è il coefficiente di regressione associato alla variabile i-esima che indica l'apporto che essa fornisce alla probabilità che si vuole estrapolare, $\forall i \in \{1, ..., p\}$, con $p \in \mathbb{N}$ numero delle variabili. Mentre ϵ rappresenta l'errore aleatorio.

5.3.1 Come si sviluppa un Modello Logistico

La regressione ha quindi lo scopo di stimare una probabilità e quindi come tale essa è una quantità compresa nell'intervallo [0,1], ovvero

$$0 \leq P(Y = 1 \mid \underline{x}) \leq 1.$$

Per estrapolare la probabilità, essa viene eguagliata con la somma tra le x_i pesate con i relativi coefficienti β_i che, come le prime, si muovono su un campo diverso, più ampio, ovvero :

$$x_i \in \mathbb{R}, \beta_i \in \mathbb{R}, \forall i \in \{1, ..., p\}\,.$$

Quindi ci si espone al rischio che quanto ottenuto nel lato destro dell'uguaglianza non si trovi tra 0 e 1, cioè nell'insieme di appartenenza della probabilità da estrapolare.

Per questo motivo la risposta $P(Y = 1 \mid \underline{x})$ viene legata al lato destro dell'uguaglianza attraverso una funzione, la funzione *logistica* $g(\cdot)$, di tipo non lineare, che applicata alla risposta probabilistica trasformi essa in un intervallo in $\mathbb{R}$ [8] e quindi:

$$g : [0,1] \longrightarrow \mathbb{R}.$$

Nella regressione logistica, viene spesso adoperata la funzione `logiT(x)` che viene definita in questo modo:

$$logiT : [0,1] \longrightarrow \mathbb{R},$$

$$logiT(x) = log(\frac{x}{1-x}), \forall x \in [0,1].$$

Il modello quindi assume infine la seguente forma:
Chiamando $P(Y = 1 \mid \underline{x}) = \rho$

$$log(\frac{\rho}{1-\rho}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon$$

dove $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$ è chiamato μ.

Applicando la funzione esponenziale da ambo i lati si ottiene l'uguaglianza :

$$\frac{\rho}{1-\rho} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+....+\beta_p x_p}.$$

La quantità alla sinistra dell'uguale è denominata *Odd*. Esso indica il rapporto tra la probabilità di un evento (ρ) e la probabilità che non accada ($1 - \rho$). Gli *Odds* possono presentare qualsiasi valore da 0 a ∞, sono tradizionalmente usati nelle corse di cavalli, e rappresentano le quote per le scommesse, poiché si riferiscono in modo più naturale alla corretta strategia di puntata [17]. Infatti questa misura rappresenta la `ragione di scommessa`, ovvero quanto si è disposto a scommettere per vincere 1 se l'evento si verifica ($Y = 1$) e perdere se esso non si verifica [6].

Isolando la probabilità ρ si ottiene, dopo qualche semplice manipolazione, il seguente modello :

$$\rho = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+....+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+....+\beta_p x_p}}.$$

Una funzione logistica ha come equazione la seguente:

$$f(x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}},$$

dove

$$x \in \mathbb{R}$$
$$\alpha, \beta = \text{coefficienti} \in \mathbb{R}.$$

Si può notare che essa produrrà sempre una curva a forma di S con immagine che detiene estremi massimi in 0 e 1 così come rappresentato alla figura 5.8, dove sulle ascisse vi risiedono i valori della variabile esplicativa campionaria (x) e sulle ordinate la probabilità che la funzione stima.

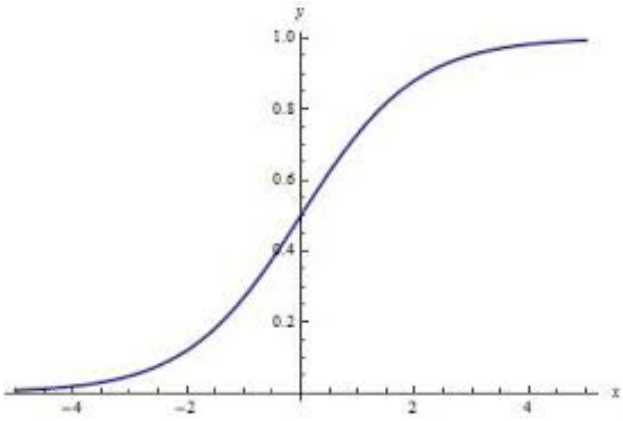

Figura 5.8: Funzione logistica

5.3.2 Stima dei coefficienti di regressione

I coefficienti β_i del modello sono ignoti e vengono stimati sulla base dei dati che si ha a disposizione. Il procedimento che si adopera è denominato *maximum likelihood*, ovvero metodo di *massima verosimiglianza*.

L'intuizione alla base dell'utilizzo di questo metodo in un modello di regressione logistica è la seguente: vengono ricercate le stime $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$ dei rispettivi coefficienti $\beta_0, \beta_1, ..., \beta_p$, tali che la probabilità stimata $\hat{\rho}$ per ogni osservazione $\underline{x}$ corrisponda il più vicino possibile a quanto osservato nel campione. In altre parole si ricercano quelle stime tali che inserendole nel modello sopra esposto riescano a produrre delle probabilità vicine allo zero per tutte quelle osservazioni che detengono una variabile risposta $y_i = 0$ e viceversa per le altre [17]. Questa teoria si sviluppa attraverso la funzione matematica chiamata *likelihood function*, ossia *funzione di verosimiglianza*

$$L(\underline{\beta}) = \prod_{i:y_i=1} P(Y_i = 1|\underline{x_i}) \prod_{j:y_j=0} (1 - P(Y_j = 1|\underline{x_j})).$$

Le stime dei coefficienti coincidono con quei β_i tali che la funzione è massima in quel punto.

5.4 Modello di Regressione con le variabili dell'albero

Tornando al caso del *Futbol Club Barcelona*, viene sviluppata una Regressione Binaria come modello di estrapolazione degli `xG`, per prevedere la probabilità che una conclusione con tutte le sue caratteristiche, descritte nelle colonne del campione di riferimento 2.3.1, risulti rete o meno. Ovvero la probabilità che dato un vettore di variabili esplicative $\underline{x}$, la colonna `Goal` risulti uguale a 1:

$$P(Goal = 1|\underline{x})$$

Se essa risulta maggiore di una certa soglia s prefissata, allora l'osservazione $\underline{x}$ che rappresenta il tiro in porta, può dirsi *Expected Goal* secondo il modello.

Per applicare la regressione sopra analizzata al DataSet delle conclusioni a rete, è opportuno, fin dal principio, sviluppare una scrematura delle variabili di interesse del campione che entreranno a far parte del modello. Infatti detenendo 25 colonne nel DataSet, una volta poste tutte nella regressione, sussisterebbe una quantità troppo elevata di variabili esplicative all'interno, che porterebbe ad una poca robustezza del modello finale e una difficile interpretazione dell'apporto che ognuna variabile dà alla risposta.

Come opportuna selezione, ci si avvale delle variabili che sono entrate negli splits dell'Albero di Classificazione sviluppato nella sezione 5.1.2. Infatti come già anticipato nella stessa sezione, esse risultano quelle che detengono una rilevante relazione con la probabilità di segnare, ovvero per cui si è osservato nel campione che ad una loro modifica varia significativamente la probabilità di andare a segno. Il modello sviluppato quindi avrà come variabili esplicative le seguenti cinque: `DistSGK`, `duration`, `DistToGoal`, `shot.bodyPart.id`, `shot.deflected`. Verranno utilizzate tutte le osservazioni del campione *Training Set* e trascurate quelle del *Test Set* che serviranno solo nella seconda fase per valutare il metodo.

Dopo aver ottenuto le stime dei β_i, $\forall i \in [1:5]$, con il metodo di massima verosimiglianza, si consegue il seguente modello stimato per l'osservazione i-esima $\underline{x}_i$:

$$logiT(\rho_i) = log(\frac{\rho_i}{1+\rho_i}) =$$

$$= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p =$$

$$= -3.313 - 0.099 \cdot DistSGK_i + 0.918 \cdot Duration_i - 0.0545 \cdot DistToGoal_i +$$

$$+0.077 \cdot shot.bodPart.id_i + 2.468 \cdot shot.deflected_i$$

dove

$$\rho_i = P(Goal = 1|\underline{x}_i).$$

Le stime dei coefficienti associati alle variabili, sono stati arrotondati alla terza cifra decimale.

5.4.1 Verifica della bontà del modello

Il modello è stato stimato e necessita, prima di essere testato in termini di performance predittiva, una fase di verifica della significatività dei suoi coefficienti e verifica della bontà di adattamento ai dati. In questa fase si compie una ricerca

di eventuali miglioramenti che possono portare ad una robustezza migliore e come ultima conseguenza ad una migliore capacità di previsione. L'obbiettivo finale dell'analisi inferenziale infatti rimane quello di riuscire a costruire un modello che preveda al meglio la probabilità che una conclusione a rete si traduca in goal o meno, e quindi ogni test, modifica e verifica che si compiono in questa sessione viene sviluppato con questo unico intento che non sempre coincide con fini, ad esempio, di adattamenti perfetti ai dati o di facili interpretazioni.

Verifica della significatività dei coefficienti

Come prima analisi è doveroso soffermarsi sulle stime dei coefficienti di regressione, e verificarne la significatività singolarmente per ognuno di essi. Qualora uno non si rivelasse tale, sarebbe opportuno eliminarlo dal modello perché poco rilevante ai fini di quanto si vuole stimare. Si procede quindi con un esercizio di *Verifica di ipotesi* dove come ipotesi nulla si pone il coefficiente uguale a zero

$$H_0 : \beta_i = 0 \qquad H_1 : \beta_i \neq 0$$

Si necessita una *Statistica Test* per risolvere il *Test d'ipotesi*. Essa è per definizione una funzione dei dati campionari che, sotto ipotesi nulla, si distribuisce come una distribuzione nota e che ha l'obbiettivo di fornire un valore di confronto per risolvere il test di ipotesi. Infatti se il valore di essa che si ottiene dall'inserimento dei dati empirici, denominato *Z-value*, si troverà all'interno della *Regione Critica*, allora l'ipotesi nulla sarà rifiutata[3].

La *Regione Critica* è un sottoinsieme dello *Spazio Campionario*, che si costruisce fissando un α, livello di significatività del test ed *Errore del primo tipo* che si ha quando viene rifiutata l'H_0 vera, abbastanza piccolo (ad esempio: 0.01 o 0.05). Essa coincide con la parte dello spazio campionario, tale che la statistica test Z sia minore di un valore Z_1 e maggiore di un Z_2, da fissare in base alla distribuzione della statistica. Per costruire la regione critica e la statistica test è bene notare che il vettore degli stimatori di massima verosimiglianza $(\hat{B}_0, \hat{B}_1, ..., \hat{B}_p)$ dei coefficienti di regressione $(\beta_0, \beta_1, ..., \beta_p)$ tende in distribuzione come una Normale p-variata, con media il vettore dei beta e varianza la matrice delle varianze e covarianze $\Sigma_{\underline{\hat{B}}}$:

$$\underline{\hat{B}} \xrightarrow{d} N_p\left(\underline{\beta}, \Sigma_{\underline{\hat{B}}}\right)$$

dove $\Sigma_{\underline{\hat{B}}} = I[\underline{\beta}]$ ovvero l' *informazione attesa di Fisher* data dalla seguente formula:

$$I[\underline{\beta}] = E\left[\frac{-\partial^2\, l(\underline{\beta})}{\partial\beta_j \delta\beta_i}\right]$$

con

$$i = (1, ..., p) \; e \; j = (1, ..., p)$$

$l()$ = logaritmo della *funzione di massima verosimiglianza*: *log-Verosimiglianza*

Viene individuata la seguente Statistica Test per il j-esimo coefficiente che, sotto ipotesi nulla, si distribuisce come una Normale univariata Standard:

$$Z = \frac{\hat{\beta}_j - \beta_j}{SE_{B_j}} \overset{H_0}{=} \frac{\hat{\beta}_j}{SE_{B_j}} \to N(0,1)$$

dove SE_{B_j} è lo *Standard Error* dello stimatore B_j del coefficiente β_j, conseguito eseguendo la radice quadrata del valore che si trova alla riga e colonna j della matrice varianze e covarianze stimata.

La Regione Critica rilevata in questo test è quindi definita da:

$$C = \{\underline{x} : H_0 \text{ è } respinta\} = \left\{\underline{x} : Z < Z_{\frac{\alpha}{2}}\right\} \cup \left\{\underline{x} : Z > Z_{1-\frac{\alpha}{2}}\right\}$$

dove

$Z_{\frac{\alpha}{2}}$ = quantile della distribuzione normale standard di ordine $\frac{\alpha}{2}$
$Z_{1-\frac{\alpha}{2}}$ = quantile della distribuzione normale standard di ordine $1-\frac{\alpha}{2}$

Per ogni coefficiente, inserendo i dati campionari si ottengono i seguenti esiti, mostrati in tabella 5.6, distribuiti in questa matrice.

Test d'ipotesi di significatività				
Variabili	Stime	Standard-Error	Z-value	P-value
Intercept	-3.31	0.92	-3.58	0.000343
DistSGK	-0.10	0.01	-6.53	6.74e-11
DistToGoal	-0.05	0.02	-3.40	0.000669
Shot.bodyPart.id	0.08	0.24	3.22	0.001273
Shot.deflected	2.47	0.35	6.96	3.36e-12

Tabella 5.6: Significatività coefficienti

Nelle colonne della figura vengono mostrate:

- le stime dei coefficienti: $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_5$,
- Gli `Standard error` per ogni stima,
- Gli `Z-value` che rappresentano i valori campionari della statistica test rilevata per ogni stima,
- Il `P-value`: probabilità di osservare un valore della statistica test Z più estremo del modulo del valore osservato $Z - value$, data l'ipotesi nulla vera [3].

$$Pvalue = P(|Z| > |Zvalue| \ |H_0) =$$

$$= 2 \cdot P(Z > |Zvalue| \ |H_0) = 2[1 - P(Z \leq |Zvalue| \ |H_0)]$$

Nelle casistiche in cui il valore del *P-value* associato al coefficiente è minore del livello di significatività del test, α, rifiutare l'ipotesi nulla è appropriato.

Fissando un errore di primo tipo pari a $\alpha = 0.01$, ogni coefficiente regressivo del modello detiene valori dei *P-value* minori.

Per questo motivo è opportuno rifiutare l'H_0 che ipotizzava la non significatività dei coefficienti e mantenere perciò tutte le variabili nel modello.

Verifica della devianza

Una seconda operazione che il modello necessita di subire è una verifica della grandezza della *Devianza* che esso presenta.

La *Devianza* è una misura che rappresenta l'allontanamento del modello corrente dal modello *saturo* ovvero quello ideale che presenta un *fit* perfetto per ogni istanza osservata [8]. Questa grandezza si calcola facendo la differenza tra il valore massimo del logaritmo naturale della funzione di verosimiglianza (la *log-Verosimiglianza*) del modello corrente e quello del modello *saturo* e ha la caratteristica di distribuirsi approssimativamente come una *Chi quadro* con $n - p - 1$ gradi di libertà[6] (dove n indica il numero di unità statistiche nel campione utilizzato) :

$$D = -2\left[l(\underline{\hat{\beta}}, \underline{Y}) - l(\underline{\hat{\beta}_s}, \underline{Y_s})\right] \xrightarrow{d} \chi^2_{n-p-1}$$

Affinché il valore rilevato sia adeguato e quindi non vi sia un troppo alto allontanamento tra i due modelli, la devianza del modello non deve superare i gradi di libertà della *Chi quadro* a cui essa tende in distribuzione, in quanto essi costituiscono il valore atteso della distribuzione.

La devianza risultante risulta minore dei gradi di libertà del modello:

$$Df = n - p - 1 = 4036 - 5 - 1 = 4030$$

$$D = -2\left[l(\underline{\hat{\beta}}, \underline{Y}) - l(\underline{\hat{\beta}_s}, \underline{Y_s})\right] = 2957.3$$

$$D < Df.$$

Per questo motivo la verifica eseguita senza problematiche e l'allontanamento non troppo elevato.

5.4.2 Valutazione dell'accuratezza della Regressione Logistica

In seguito alla fase di verifica e studio della bontà del modello, giunge il momento di testare la sua qualità nel predire.

RBL_Forecast	0	1
Goal	28	27
NonGoal	2254	381

Figura 5.9: Stime Regressione Logistica Albero - Valori reali

Vengono stimate le probabilità relative alle osservazioni messe da parte, all'interno del campione *Test*, e ognuna di esse vengono classificate come `GOAL` o `Non GOAL` in base al loro valore se superiore o meno la soglia $s = 0.5$. Perciò tutte le probabilità stimate maggiori di s vengono classificate come vincenti e viceversa. In questo modo è possibile costruire la tabella a doppia entrata illustrata in figura 5.9, tra valori stimati e valori campionari per ogni unità statistica del *Test Set*, così come in precedenza eseguito per i tre diversi modelli ad algoritmo ad albero specificati nelle sezioni precedenti.

Successivamente si eseguono il calcolo degli indici predittivi, ottenendo:

- la sua *Capacità di predizione*:

$$C_p = \frac{n(0_{stim.} \cap 0_{camp.}) + n(1_{stim.} \cap 1_{camp.})}{n_{tot}(TestSet)} = \frac{2254 + 27}{2690} = 0.848$$

- l'*Errore di stima Test Set*:

$$E_s = 1 - C_p = \frac{n(0_{stim.} \cap 1_{camp.}) + n(1_{stim.} \cap 0_{camp.})}{n_{tot}(TestSet)} = \frac{284 + 53}{2690} = 0.152$$

- la sua *Specificità*:

$$Sp = 1 - \frac{n(1_{stim} \cap 0_{camp.})}{n(0_{campTot.})} = 1 - \frac{28}{2229 + 28} = 0.987$$

- la sua *Sensibilità*:

$$Ss = 1 - \frac{n(0_{stim} \cap 1_{camp.})}{n(1_{campTot.})} = 1 - \frac{381}{381 + 27} = 0.07$$

5.4.3 Commento ai risultati della Regressione Logistica con Variabili dell'albero

Il modello predice in modo piuttosto convincente, con un errore solo nel 15,2% delle situazioni, ma tuttavia peggiore rispetto ai primi tre modelli studiati *non parametrici*.

Il metodo inoltre presenta una scarsissima *Sensibilità*, testimoniata dal suo indice del 7%, che si contrappone alla *Specificità* eccellente quasi del 99%. Esso perciò ha presenta una facilità particolare a classificare le osservazioni che non si sono rivelate rete, ma una forte debolezza nello stimare le altre osservazioni.

5.5 Modello di Regressione a partire dall'algoritmo Step-Wise

Il modello costruito inserendo le variabili dell'albero di classificazione si può considerare "svantaggiato" rispetto agli altri modelli sviluppati in quanto utilizza un numero di variabili minore e detiene quelle variabili scelte in base a un altro metodo distinto. Infatti nel *CART* e nei *Bagging* e *Random Forest* venivano rese disponibili tutte le istanze delle variabili del campione che poi erano scremate solo durante la procedura, in base ai criteri propri degli stessi algoritmi.

Si necessita quindi di trovare un modello di regressione che superi questo discapito. A questo proposito è utile servirsi dell'algoritmo *Forward-Step-Wise*.

Esso è un metodo di selezione delle variabili che faranno parte di un modello di regressione. Nasce dalla necessità di selezionare un insieme "ottimo" tra un numero elevato di variabili esplicative per la costruzione di un modello efficiente[17]. Qualora si debba estrapolare un modello regressivo da un insieme limitato di predittori, l'operazione è gestibile con consueti metodi manuali. Quando invece, come nel caso in esame, le variabili sono troppe, se non si preferisce fare una scelta soggettiva, è ragionevole ricorrere a metodi di selezione efficienti, come appunto la *Step-Wise* perché diversamente, inserendo ogni predittore nel modello, si andrebbe incontro a problemi di elevata variabilità delle stime dei coefficienti regressivi o di *Overfitting* che di conseguenza produrrebbe distorsioni nelle previsioni.

5.5.1 Il procedimento del Forward-Step-Wise

La procedura comincia a partire da un modello di regressione che non contiene alcuna variabile e in seguito lavora aggiungendo predittori al modello. In particolare, ad ogni passaggio la variabile che dà un maggiore miglioramento addizionale al *fit* del modello viene man mano inserita.

Più formalmente, i passaggi si susseguono come segue [17]:

- 1. Sia M_0 il modello nullo, che non contiene predittori.
- 2. Per $k = 0, ..., p-1$:
  - (a) Considera tutti i $p-k$ modelli detti M_k che man mano hanno un predittore in aggiunta
  - (b) Sceglie il miglior modello in termini di minor Devianza, tra i $p-k$ modelli M_k e lo chiama M_{k+1}
- 3. Seleziona un singolo modello tra $M_0, ..., M_p$ considerato il migliore in termini di

– *AIC*: *Akaike's information criterion*[6] metodo per la valutazione e il confronto tra modelli statistici, nel quale si preferiscono quelli con il valore più basso, calcolato come:

$$AIC = -2(l_M - (p+1)).$$

dove l_M =logaritmo naturale del valore massimizzato della funzione di verosimiglianza.

– Errore di previsione e la capacità predittiva.

5.5.2 Il Modello derivante dello Step-Wise

Viene dunque sviluppato l'algoritmo descritto per formare un nuovo modello a partire dalle venticinque variabili del campione delle conclusioni del Barcellona. Come esito finale del metodo statistico, si consegue un modello contenente sedici variabili selezionate in base alla procedura illustrata nella sezione precedente.

Prima di considerare il modello come conclusivo e collaudato, è necessario verificare per ogni variabile la significatività delle stime dei coefficienti di regressione delle variabili raccolte.

In effetti esse, durante la tecnica dello *Step-Wise*, sono state selezionate sulla base dell'apporto che hanno dato alla misura dell'*AIC* e della capacità di predire, ma la loro scelta non ha un sostegno per quanto riguarda la significatività all'interno di un modello di regressione logistica. Per questo motivo, è opportuno sviluppare il *Test d'ipotesi* che nella sezione 5.4.1 viene ampiamente definito, con $H_0 : \beta_i = 0 \ \ \forall i$.

Per ogni coefficiente perciò vengono raccolti i relativi valori campionari delle *Statistiche Test* (gli *Z-value*) e i *P-value* che descrivono la probabilità che essi appartengano o meno alla *Regione Critica*. Essa è descritta come segue:

$$C = \{\underline{x} : H_0 \text{ è } respinta\} = \left\{\underline{x} : Z < Z_{\frac{\alpha}{2}}\right\} \cup \left\{\underline{x} : Z > Z_{1-\frac{\alpha}{2}}\right\}.$$

dove $Z_{\frac{\alpha}{2}}$ = quantile della distribuzione normale standard di ordine $\frac{\alpha}{2}$
$Z_{1-\frac{\alpha}{2}}$ = quantile della distribuzione normale standard di ordine $1-\frac{\alpha}{2}$
α = livello di significatività e *errore del primo tipo*

Si ottengono dunque i risultati esposti nella tabella 5.7:

Dopo aver fissato il livello di significatività $\alpha = 0.05$, dalla tabella si evince che il *P-value*, illustrato nella quarta colonna, per alcune delle stime dei coefficienti non è inferiore ad α. La regola decisionale del test sancisce per giunta che, per quelle stime, l'ipotesi nulla è da non rifiutare e quindi essi si considerano `non significativi` a livello 0.05.
E' bene dunque eliminare dal modello le variabili relative ai coefficienti che presentano questa problematica: `density`, `shot.bodyPart.id`, `DefArea`, `AngleToKeeper`, `milliseconds`, `shot.technique.id`.

Test d'ipotesi di significatività				
Variabili	Stime	Standard-Error	Z-value	P-value
Intercept	15.09	4.55	3.32	0.000910
DistSGK	-0.8	0.01	-5.73	1.0e-08
Duration	1.03	0.08	13.04	<2e-09
distance.ToD1.360	0.14	0.02	5.84	5.1e-09
Shot.deflected	2.92	0.28	10.41	<2e-16
DistToGoal	-0.23	0.02	-11.78	<2e-16
location.X	-0.16	0.017	-9.50	<2e-16
density	-0.003	0.13	-0.20	0.839611
density.incone	-0.55	0.15	-3.7	0.000215
shot.type.id	-0.03	0.01	-3.58	0.000346
DefendersBehindBall	-0.11	0.03	-3.63	0.000287
UnderPressure	-0.28	0.12	-2.39	0.016909
Shot.bodyPart.id	0.03	0.017	1.92	0.054984
DefArea	7.34e-4	3.91e-4	-1.876	0.060699
AngleToKeeper	1.55e-3	8.15e-4	1.90	0.056957
milliseconds	-2.01e-4	1.30e-4	-1.55	0.120286
shot.technique.id	0.064	0.04	1.51	0.13239

Tabella 5.7: Significatività coefficienti Step-Wise

Si prosegue perciò la costruzione del modello con i dieci predittori rimasti.

Si ottiene la seguente stima del *Modello di Regressione Binaria* per la i-esima osservazione $\underline{x_i}$:

$$logit(\rho_i) = log(\frac{\rho_i}{1+\rho_i}) =$$

$$= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p =$$

$$= 23.521 - 0.094 \cdot DistSGK_i + 0.931 \cdot Duration_i + 0.12639 \cdot Distance.ToD1.360_i +$$

$$2.750 \cdot shot.deflected_i - 0.22276 \cdot DistToGoal_i - 0.166 \cdot location.x_i$$

$$-0.54939 \cdot density.incone_i - 0.04031 \cdot shot.type.id_i$$

$$-0.10202 \cdot DefendersBehindBall_i - 0.32577 \cdot underPressure_i.$$

5.5.3 Verifica della Bontà del Modello

E' opportuno anche per questo modello eseguire uno studio della bontà in termini di adattamento ai dati campionari, per assicurare la massima prestazione predittiva.

Verifica della Devianza

Si compie il controllo della *Devianza* ottenuta che misura l'allontanamento del modello corrente dal *saturo* (modello ideale, che ha un adattamento perfetto per ogni osservazione del *DataSet*), allo stesso modo di quanto effettuato con la regressione *logistica* analizzata nella sezione 5.4.1.

Si calcola il valore della *Devianza* del modello attraverso la differenza tra i valori della *log-Verosimiglianza* dei due modelli.

$$D = -2\left[l(\underline{\hat{\beta}}, \underline{Y}) - l(\underline{\hat{\beta}_s}, \underline{Y_s})\right] = 2794.27$$

Il confronto da eseguire avviene tra quest'ultimo valore e i grandi di libertà (Df) della distribuzione *Chi quadro* a cui essa tende.

L'allontanamento tra modello saturo e corrente non risulta troppo vasto in quanto la *Devianza* non supera i Df come si evince da quanto segue:

$$Df = n - p - 1 = 4036 - 10 - 1 = 4025$$

$$D = 2794.27$$

$$D < Df$$

.

Il *fit* sembra dunque di buon livello.

Verifica della linearità

Un'ulteriore verifica necessaria per assicurare la bontà del modello ottenuto è l'accertamento della *linearità* nel $logit(\rho)$ del modello la cui equazione viene ricordata:

$$logit(\rho) = log(\frac{\rho}{1-\rho}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon$$

dove $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$ è chiamato μ

Questo test decreta se il modello stimato, affinché abbia un adattamento ai dati di miglior livello, debba perdere la sua *linearità* passando ad una forma *polinomiale* di grado superiore al primo[8].

Per compiere la verifica si inseriscono i valori di μ all'interno del modello di regressione elevandoli alla potenza seconda, allo stesso modo di una semplice variabile; successivamente si compie il controllo della significatività del coefficiente di regressione associato ad esso: se significativo, allora il modello necessita una forma polinomiale, contrariamente è confermata la linearità di esso.

Eseguendo la verifica dell'ipotesi nulla: $H_0 : \beta_{\mu^2} = 0$ Si ottengono i seguenti esiti nella tabella 5.8.

Si evince una perfetta linearità del modello: il *P-value* è elevato e maggiore di ogni livello di significatività consueto. Non è dunque da rifiutare l'ipotesi nulla secondo la *regola decisionale* e di conseguenza il modello non necessita una modifica.

Test significatività variabile aggiunta				
Variabile	Stima	Standard-Error	Z-value	P-value
μ	-2.36	1.02	-2.31	0.5

Tabella 5.8: Test per linearità

5.5.4 Analisi della Performance predittiva del modello ottenuto dallo Step-Wise

STEP_Forecast	0	1
Goal	46	100
NonGoal	2236	308

Figura 5.10: Stime Regressione SW - Valori reali

Dopo aver sviluppato gli accertamenti circa la significatività e il *fit* del modello stimato, è doveroso testare la prestazione delle predizioni che esso riesce a conseguire. Ci si concentra perciò sulle unità statistiche della parte del campione *Test*. Per ognuna di esse il modello stima le relative probabilità $\rho = P(Goal = 1|\underline{x})$; successivamente viene assegnata una modalità per la variabile risposta tale che se $\rho > 0.5$ essa è definita *Expected Goal* e viceversa.

Infine si confrontano le modalità stimate e i valori rilevati campionari delle osservazioni del *Test Set* in una matrice raffigurata nella immagine 5.10.

Con i dati nella tabella è possibile svolgere gli indici predittivi, già definiti, che testano la capacità di predire del modello. Vengono calcolati:

- la sua *Capacità di predizione*:

$$C_p = \frac{n(0_{stim.} \cap 0_{camp.}) + n(1_{stim.} \cap 1_{camp.})}{n_{tot}(TestSet)} =$$

$$= \frac{2236 + 100}{2690} = 0.868$$

- l'*Errore di stima Test Set*:

$$E_s = 1 - C_p = \frac{n(0_{stim.} \cap 1_{camp.}) + n(1_{stim.} \cap 0_{camp.})}{n_{tot}(TestSet)} =$$

$$= \frac{308 + 46}{2690} = 0.132$$

- la sua *Specificità*:

$$Sp = 1 - \frac{n(1_{stim} \cap 0_{camp.})}{n(0_{campTot.})} =$$

$$= 1 - \frac{46}{2236 + 46} = 0.98$$

- la sua *Sensibilità*:

$$Ss = 1 - \frac{n(0_{stim} \cap 1_{camp.})}{n(1_{campTot.})} =$$

$$= 1 - \frac{308}{308 + 100} = 0.25$$

Commento agli esiti

Il modello detiene un'ottima attitudine a predire che supera di netto la relativa capacità predittiva della regressione con variabile ad albero.

La sua *Sensibilità* è tuttavia non di buon livello, ma è comunque in miglioramento rispetto alla corrispondente *Regressione Logistica* precedentemente stimata e la *Specificità* è ottima con un valore prossimo al massimo. La predizione in tutte e quattro le sue misure ricorda spesso i risultati ottenuti nel modello della foresta *Random*.

Capitolo 6

Il Confronto e l'interpretazione dei modelli

I modelli sono stati stimati e i risultati raccolti possono essere frutto di decisioni in condizioni di incertezze o di studi nel settore calcistico. Tuttavia un ultimo passo è fondamentale per poter procedere adeguatamente. E' necessario svolgere un'interpretazione dei dati disponibili e mostrare un utilizzo corretto ed efficiente di essi in base alla situazione particolare. Infatti il dato grezzo raccolto e analizzato nei modelli e tecniche statistiche, senza un'interpretazione corretta, non può essere applicato per alcuna scelta.

Come ultimo passo dunque è necessario eseguire un'interpretazione finale dei modelli stimati volta a permettere di capire quali dei cinque sono più adeguati, e quindi preferibili in base a diversi contesti di utilizzo e in aggiunta una ricapitolazione conclusiva.

Per poter selezionare il metodo più adeguato nel contesto specifico, è opportuno fornire un confronto tra i diversi modelli, indagando differenze, vantaggi e svantaggi di ognuno di essi.

6.1 Confronto tra modelli *parametrici* e *non parametrici*

Due tipologie distinte di modelli sono stati stimati nel corso della ricerca statistica, e in seguito studiati e analizzati. Tecniche che non possiedono un'equazione *parametrica*, in particolare *Albero di classificazione*, *Bagging* e *Random Forest*, e quelle chiamate *parametriche*, delle quali sono stati studiati i metodi di *regressione logistica*.

Intercorrono molteplici differenze tra i due metodi. Esse sono presenti durante il procedimento di costruzione, ma anche nei risultati finali e nell'interpretazione di essi. E' utile osservare le distinzioni per poter capire i vantaggi e svantaggi.

In un primo luogo i modelli non parametrici *Bagging* e *Random Forest* forniscono direttamente la modalità finale della specifica osservazione, classificandola quindi direttamente in *Expected Goals* o meno, mentre usando i metodi di regressione stimati e gli *alberi* si giunge in una prima fase alla probabilità di rete della osservazione, che poi è catalogata come *Goal atteso* se maggiore di una certa soglia.

Questa distinzione può rivelarsi un beneficio in una situazione in cui è richiesto ottenere oltre al calcolo degli *Expected goals* anche un valore probabilistico associato. La probabilità è un valore semplice da interpretare anche per un individuo senza competenze altamente specifiche. Inoltre una *Regressione Logistica* fornisce il valore del *Odd* per ogni unità statistica, che come già anticipato nella sezione 5.3.1, rappresenta la *ragione di scommessa* ovvero quanto si è disposto a scommettere per vincere 1 se l'evento si verifica (*Goal*) e perdere se esso non si verifica.

Questa misura può destare l'interesse di una ditta di scommesse, che per esempio deve quotare l'evento di un goal e si basa sugli xG delle conclusioni passate.

Tuttavia questa differenza può risultare favorevole anche per i metodi non parametrici poiché saltando la fase di ottenimento della probabilità, presentano caratteristiche di grande immediatezza nell'individuazione degli xG; seppur è necessario un contesto di disponibilità tecnologica di software statistici avanzati per implementare l'algoritmo. Ad esempio in una situazione ipotetica in cui, terminata una partita di calcio, durante un programma televisivo nel quale si commentano le azioni dell'incontro, si vuole analizzare nel dettaglio il match individuando i goal *attesi* delle due squadre, ma si ha un tempo ristretto per il calcolo, è di maggior rapidità un utilizzo di tecniche algoritmiche quali ad esempio il *Random Forest*.

In aggiunta un'importante considerazione da svolgere nell'analisi dei pro e i contro tra le due tipologie di metodi statistici è la facilità di interpretazione che riesce ad offrire un modello di regressione, come i due stimati nel capitolo quinto, rispetto agli altri tre metodi. Infatti il modello *logistico* offre una chiara equazione leggibile e interpretabile in maniera poco complessa, e soprattutto mette in luce l'apporto che ogni variabile fornisce al modello attraverso le stime dei coefficienti di regressione. Considerato un coefficiente β_i, esso costituisce l'aumento o la diminuzione media nel *logit* della probabilità di andare a segno, a parità delle altre variabile del modello, nel caso in cui il valore della variabile a cui esso è associato, la i-esima, si modifichi di 1.

Il peso delle singole variabili non è ottenibile in un *Albero Decisionale* e costituisce perciò un difetto della tecnica. Inoltre come le altre due tecniche non parametriche non possiede la stessa semplicità di interpretazione della regressione. Da essi difatti si possono estrapolare solo i valori dell'importanza dei singoli predittori mediante le tecniche di *Mean Decrease Accuracy* e della diminuzione media dell'indice di *Gini*, che costituiscono delle misure complesse da interpretare e da

calcolare per un individuo che non presenta elevate competenze specifiche.

6.2 Confronto in termini di facilità di implementazione

E' doverosa un'attenta valutazione in aggiunta delle differenze tra la complessità nell'impiego di un modello piuttosto di altri.

Innanzitutto si osserva facilmente la differenza tra il numero delle variabili da imputare nei diversi modelli. Questa caratteristica definisce la complessità dei singoli modelli e del loro utilizzo per prevedere la modalità di un unità statistica.

In relazione a ciò, è preferibile sicuramente tra i modelli ad algoritmo, l'*Albero di Classificazione* che presenta un diagramma facilmente interpretabile e utilizzabile per prevedere un tiro a rete sulla base di solamente cinque variabili. In particolare per impiegare questa tecnica è necessario rilevare la distanza del giocatore che effettua la conclusione dal portiere (`DistSGK`), la durata del tiro(`duration`), la distanza dalla porta (`DistToGoal`), la parte del corpo con la quale è stato effettuato (`Shot.bodyPart`) e infine se esso è stato deviato o meno (`shot.deflected`)

Diversamente i modelli di *Bagging* e di *Random Forest* presentano una maggiore complessità. Per ottenere delle stime previsive, come già specificato nella sezione precedente, sono necessari macchinari tecnologici con software statistici avanzati. Inoltre per la prevedere se risulta xG un tiro, sono necessarie un insieme molto ampio di informazioni.

Per quanto riguarda i modelli di regressione, ci si può servire di essi anche manualmente risolvendo l'equazione del modello. Inoltre la tecnica che richiede il minor numero di dati è la prima regressione illustrata che al suo interno dispone le variabili dell'*Albero Decisionale*.

Tuttavia anche nell'altro modello logistico studiato non sono necessari tutti i dati che invece risultano obbligatori per l'utilizzo dei metodi di algoritmo a *foresta*.

6.3 Confronto tra le variabili di rilievo selezionate

Le variabili che risultano fondamentali nelle previsioni degli xG sono molteplici. Alcune sono ritenute tali da ogni modello analizzato, mentre altre hanno una forte rilevanza solo in uno specifico.

La distanza del giocatore dalla porta e dal portiere (rispettivamente `DistToGoal` e `DistSGK`) insieme alla durata del tiro (`duration`) costituiscono delle informazioni fondamentali da rilevare se si deve prevedere un goal *atteso*. Infatti esse sono presenti in ogni modello stimato, inoltre nelle regressioni possiedono i coefficienti regressivi con maggior peso significativo, come si evince dall'esiguità del *P-value*

del test d'ipotesi di non significatività mostrato per entrambi i modelli di regressioni analizzati (figure 5.6 e 5.7), e infine presentano l'importanza più marcata nei modelli di *Bagging* e *Random Forest*, come si nota nel grafico 5.6 e in 5.4.

Al contrario, tra le variabili che offrono un apporto marcato solamente in alcuni dei modelli selezionati, ad esempio si segnala la colonna relativa all'informazione che specifica se chi effettua il tiro si trovi in una situazione di pressione o meno (`Under Pressure`). Essa risulta fondamentale solamente nel modello di regressione *Step-Wise*.

E' quindi utile sottolineare che in alcune circostanze è opportuno scegliere il modello corretto anche in base a cosa facilmente si riesce a rilevare, o alle informazioni già disponibili nel momento dell'impiego del metodo.

6.4 Confronto tra le performance predittive

Qualora fosse disponibile qualsiasi tipologia di informazione delle variabili del campione adoperato, è opportuno preferire il modello con una maggiore capacità predittiva.

E' giusto sottolineare che ogni tecnica adoperata riesce a garantire una percentuale di corretta previsione almeno dell'84/85% delle osservazioni del *Test Set*, garantendo un errore predittivo mai superiore al 16%.

Tuttavia una gerarchia in termini di performance è osservabile dai risultati ottenuti. Per un richiamo dei risultati ottenuti viene esposta la tabella riassuntiva 6.1 contenente i valori conseguiti degli indicatori dell'accuratezza della performance predittiva di tutti i modelli.

Modelli stimati					
Indicatori	Albero dec.	Bagging	Random Forest	Regres. Albero	Regres. Step-Wise
Specificità	0.921	0.958	0.977	0.976	0.98
Sensibilità	0.485	0.304	0.392	0.07	0.25
Errore di stima	0.145	0.131	0.125	0.152	0.132
Cap. predittiva	0.855	0.869	0.875	0.848	0.868

Tabella 6.1: Tabella riassuntiva performance predittiva modelli

Indagando la tabella, il modello stimato che garantisce le stime più precise è il *Random Forest*. Il metodo infatti detiene un errore di stima estremamente modesto in media circa del 12.5%.

In situazioni in cui è più produttiva una stima corretta delle osservazioni che non presentano il fenomeno, ovvero che non si sono tradotte in rete, è opportuno considerare un modello che presenti un'alta specificità ovvero che detenga un rischio

basso di *falsi positivi*. Dunque è preferibile il modello a Regressione *logistica Step-Wise* che dispone di una *Specificità* del 98% circa.

Ad esempio, considerando le vesti di un manager di una squadra avversaria del Barcellona che vuole compiere un'analisi degli xG del team *blaugrana* in una partita passata, per migliorare la fase difensiva e conferire consigli all'allenatore, è di grande utilità servirsi di questo modello perché riesce a ridurre al minimo i casi nei quali un'osservazione è ritenuta erroneamente Goal *atteso* e quindi non si rischia di sviluppare studi inutili su situazioni errate.

Viceversa, qualora si presentasse un contesto calcistico nel quale è preferibile detenere un rischio di falsi negativi minimale, è più adeguato il modello ad *Albero Decisionale* in quanto, mantenendo una *specificità* ottima, offre l'indice di *Sensibilità* più alto seppur comunque non ottimale del 48.5%.

Capitolo 7

Conclusione

Le tecniche statistiche permettono di ricavare dati, informazioni e misurazioni. Ciò che offre la materia è di poter effettuare scelte in condizioni di incertezza, sulla base di dati oggettivi e così le tecniche messe in luce nei capitoli possono risultare di grande utilità in un settore di stampo *qualitativo*, poco razionale, come il mondo dello sport ed in particolare il gioco del calcio.

Durante l'elaborato cinque modelli sono stati stimati utilizzando quattro tecniche di *statistical learning* quali *Regressioni Logistiche*, *Alberi Decisionali*, e metodi di *Bagging* e *Random Forest*. Ogni metodo impiegato, considerata una conclusione a rete e le sue caratteristiche, ha la facoltà di prevedere l'eventuale successo (`GOAL`) o insuccesso (`NON GOAL`). Nel primo caso il tiro verso la porta preso in considerazione è perciò definibile *Expected Goal* o goal *atteso*, mentre nel secondo caso non è considerabile tale.

Ogni metodo stimato è stato studiato, sviluppato e analizzato. Le stime prodotte nelle osservazioni del campione *Test Set* sono state confrontate con i valori empirici per testare la capacità del modello di predire. I risultati sono stati esposti e interpretati.

Per eseguire ciò, in primo luogo è stato studiato il campione raccolto. Si è mostrato il contesto specifico empirico nel quale le osservazioni rilevate hanno vissuto: la squadra ispezionata (i blaugrana del Barcellona), la situazione temporale (le annate dal 2006 al 2019) e spaziale (le partite del campionato spagnolo *La Liga*). In seguito ogni informazione contenuta nelle variabili osservate che sono risultate di elevato rilievo per la ricerca degli xG è stata specificata, e per ogni colonna del DataSet è stata fornita una descrizione statistica sulla base di classiche proprietà quali la *media*, i *quantili*, ecc. e sulla base della relazione con la variabile `GOAL`.

Si è notato che le informazioni che maggiormente hanno importanza per l'investigazione degli *Expected Goals* sono la distanza del giocatore che effettua il tiro dal portiere e dalla porta, la durata del tiro, quanti difensori sono presenti al momento dell'impatto col pallone, a 360° e tra il punto di tiro e la porta. In alcuni metodi acquisisce molto rilievo anche l'angolatura del tiro rispetto al portiere (come nel

Random Forest) e in altri se il tiro è stato deviato o con quale parte del corpo è stato effettuato.

E' stato ottenuto il grafico ad albero dell'algoritmo *CART*, con il quale anche manualmente, date le informazioni di un tiro a rete delle cinque variabili impiegate, scalando i rami verso il basso, si può ottiene la stima `GOAL`= 1 o `Non GOAL`= 0 e di conseguenza se la conclusione è considerabile xG.

In aggiunta sono stati stimati due modelli di regressione *logistica*, che hanno i vantaggi di possedere un'equazione illustrativa che mette in mostra le variabili necessarie e il valore dell'apporto che ognuna fornisce alla probabilità attraverso i coefficienti regressivi associati. Inoltre si è notato come il modello ha la facoltà di fornire per prima la probabilità che un evento sia xG e la misura dell' *Odd*.

Per gli algoritmi a *foresta*, *Bagging* e *Random Forest*, invece si è notato come non è possibile detenere una raffigurazione ne un'equazione descrittiva, ma se si ha a disposizione software statistici avanzati e un buon numero di informazioni per conclusione a rete, godono di un'ottima capacità predittiva e di una buona immediatezza nel fornire le stime.

Le performance predittive di ogni tecnica sono considerate di buon livello (gli errori di stima non superano mai il 16%) ma un confronto è stato necessario e sviluppato al capitolo 6.

La statistica applicata alla disciplina sportiva è una dottrina che riscuote molto interesse ma che non è stata molto sperimentata. Tuttavia il calcolo degli *Expected Goals* può risultare di interesse per ogni studioso del settore.

Esso può rivelarsi importante con l'obbiettivo di analizzare le partite di calcio per fornire consigli che apportano miglioramenti alla strategia di gioco di una squadra. Ad esempio mediante queste statistiche si possono intravedere in quale zona del campo o in quale situazione favorevole il Barcellona ha sviluppato al meglio il proprio potenziale offensivo, e quindi si offre il consiglio di replicare la tattica che si è rivelata più efficace; inoltre ad esempio anche gli avversari della squadra che dispongono di queste statistiche possono preparare il match contro i catalani osservando le mosse e le manovre offensive che si sono rivelate più efficienti degli avversari.

Allo stesso modo gli *Expected Goals* possono risultare utili con il fine di sviluppare uno studio e un'esposizione di stampo divulgativo, come ad esempio articoli di giornali o dvd e libri sulla squadra, oppure per confrontare il *team* con altre rivali di altre epoche storiche.

Si nota quindi come i metodi di *statistical learning* mostrati, nonostante adoperino tecniche matematiche oggettive, forniscono informazioni utili nel mondo del calcio.

Bibliografia

[1] J.J. Allaire. *RStudio*. 2009. URL: https://rstudio.com/.

[2] Christoph Biermann. *Football Hackers: The Science and Art of a Data Revolution*. A cura di Blink publishing. 2019.

[3] Minozzo M. Cicchitelli G. D'Urso P. *Statistica: Principi e Metodi*. A cura di Pearson. Terza edizione. 2018.

[4] G. Cicchitelli. *Statistica principi e metodi*. A cura di Pearson. II. 2012.

[5] Manusia Daniele. «l'Ultimo Uomo». In: (2016). A cura di Emanuele Atturo. URL: http://www.ultimouomo.com/.

[6] Barnett A.G. Dobson A.J. *An introduction to Generalized Linear Models*. A cura di CRC Press. 2008.

[7] A. Agresti C. Franklin. *Statistica: l'arte e la scienza d'imparare dai dati*. A cura di Pearson Education Italia. mylab. 2016.

[8] Hilbe J.M. Hardin J.W. *Generalized Linear Models and Extensions*. A cura di Stata Press. 2007.

[9] K.P.Murphy. *Machine Learning: A Probabilistic Perspective*. A cura di The MIT press. 2012.

[10] Roderick J.A. Little e Donald B. Rubin. *Statistical analysis with missing data*. A cura di John Wiley & Sons. 2014.

[11] Antonio Manno. *cart.pdf*. 2016. URL: http://www.adbarno.it/rep/biblio/cart.pdf.

[12] Rajul Parikh. «Understanding and using sensitivity, specificity and predictive values». In: (2008).

[13] T. Knutson C. Randall. *StatsBomb*. A cura di James Kick. 2016. URL: https://statsbomb.com/.

[14] Tamas Rudas. *Odds Ratios in the Analysis of Contingency Tables*. A cura di SAGE pubblications Inc. 1997.

[15] *StatsBomb Open Data Specification*. URL: https://OpenDataSpecification.com/.

[16] B. Efron R.J. Tibshirani. *An Introduction to the Bootstrap*. A cura di SPRINGER SCIENCE + BUSINESS MEDIA. 1993.

[17] G. James D. Witten T. Hastie R. Tibshirani. *An Introduction to Statistical Learning*. A cura di Springer. StatsBomb, 2017.

[18] L. Molteni G. Troilo. *Ricerche di marketing. Metodologie e tecniche per le decisioni strategiche e operative di marketing*. A cura di EGEA. Reference. 2012.

[19] *Who Scored*. 2004. URL: https://whoscored.com/.