UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento di
Ingegneria Informatica

Corso di laurea in
Ingegneria Informatica

Classe n. 21-270

"Expected Goals Model: Applicazione del modello logistico per la stima della probabilità di segnare nella Lega Serie A di Calcio"

Candidato:
Andrea Vaerini
Matricola n.
1067398

Relatore:
Chiar.mo Prof. Rodolfo Metulini

Anno Accademico
2023/2024

INDICE

1. Abstract

Nell'era digitale in cui siamo immersi, l'analisi dei dati sta rivoluzionando il mondo dello sport, e il calcio non fa eccezione. L'intento di questo studio è focalizzarsi sull'applicazione del modello degli Expected Goals (xG) come metodologia innovativa per valutare qualitativamente i tiri dei giocatori nelle partite di calcio della stagione 22-23 della Serie A italiana.

L'obiettivo principale di questa ricerca è quello di sviluppare un modello di analisi basato sugli xG per fornire una valutazione più oggettiva ed accurata delle prestazioni calcistiche. Per raggiungere questo obiettivo, sono stati raccolti dati sulle partite, includendo informazioni dettagliate su ogni tentativo di tiro.

Il modello è stato applicato a un vasto campione di partite di club della lega italiana nell'ultima stagione, con l'obiettivo di valutare la sua affidabilità e precisione. I risultati ottenuti sono stati confrontati e analizzati con l'intento di dimostrare l'utilità degli xG nell'offrire una visione più completa e approfondita delle prestazioni delle squadre e dei calciatori.

Inoltre, in questa tesi viene data importanza all'interpretazione dei risultati ottenuti, mettendo in evidenza i possibili utilizzi pratici degli xG nel contesto delle decisioni tattiche, degli acquisti e delle strategie di gioco. Le implicazioni di questa ricerca potrebbero influenzare le scelte degli allenatori e dei dirigenti sportivi, migliorando la competitività delle squadre e, di conseguenza, l'esperienza dei tifosi.

Infine, vengono anche discusse alcune delle sfide e limitazioni legate all'utilizzo degli xG nel calcio. Questo studio si pone come punto di partenza per ulteriori ricerche nel campo della soccer analytics, promuovendo una maggiore integrazione tra l'Ingegneria Informatica in campo Data Science e il mondo del calcio.

2. Stato dell'arte

Negli ultimi anni, il modello xG ha guadagnato crescente popolarità ed è sempre più utilizzato nel mondo del calcio come indicatore per valutare le prestazioni di finalizzazione dei calciatori e la forza offensiva delle squadre durante una partita. Per questo motivo, alcune ricerche e siti web hanno affrontato questo argomento: per esempio, Rathke (2017) [2] e Umami et al. (2021) [3] hanno analizzato i tiri, concentrandosi solamente sulla distanza e l'angolazione rispetto alla porta, mentre Fairchild et al. (2018) [4] hanno condotto un'analisi spaziale dei tiri nella Major League Soccer, utilizzando una regressione logistica. Un altro studio recente (Ruan et al., 2022 [5]) ha cercato di quantificare l'efficacia degli stili di gioco difensivi nella Chinese Football Super League mediante l'uso degli xG.

La principale lacuna è rappresentata dal fatto che, attualmente, i modelli xG si basano solamente su dati degli eventi e non tengono conto delle caratteristiche dei calciatori. Con l'intento di apportare una notevole innovazione, Cefis e Carpita [6] attraverso il loro studio, hanno puntato a migliorare il modello xG introducendo degli indicatori compositi basati sulle prestazioni dei calciatori e ottenuti attraverso un modello di equazioni strutturali a minimi quadrati parziali (PLS-SEM), al fine di considerare le caratteristiche degli attaccanti e dei portieri. Per raggiungere questo obiettivo, ha effettuato una fusione di dati provenienti da diverse fonti (per esempio, Understat [1] per i dati sugli eventi, e Sofifa [7] per la costruzione degli indicatori delle prestazioni dei calciatori). Il dataset finale è composto da un campione di ben 600 tiri, con 23 variabili (come ad esempio 1 risultato binario e 22 regressori), relativi a 50 partite ufficiali di calcio della stagione 2019/2020 della Serie A italiana.

Nella figura 1 dal documento di Cefis Carpita [6], possiamo notare un'applicazione del modello PLS-SEM in ambito soccer analytics: il modello xG.

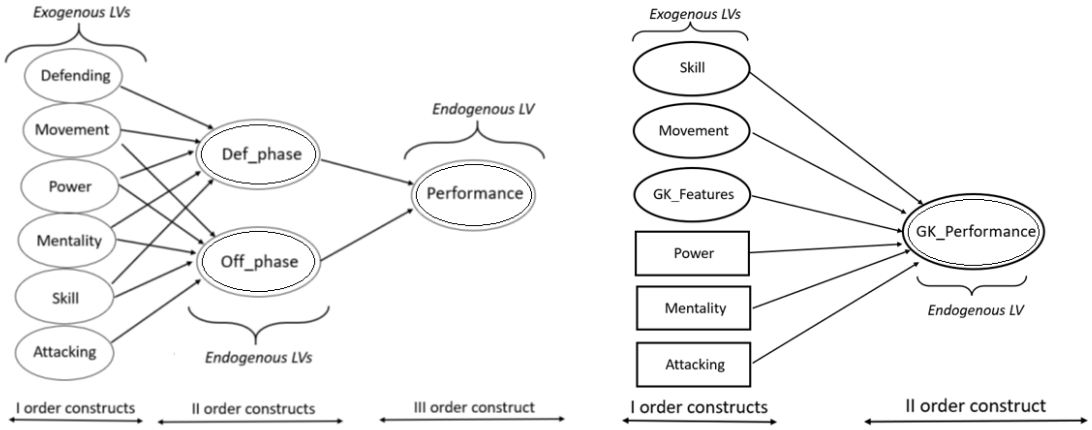

Fig.1

I movimenti dei portieri e dei giocatori utilizzati per creare gli indicatori mediante l'approccio PLS-SEM.

I principali studi riguardanti gli expected goals trovano il loro testo di riferimento nel documento redatto da Anzel e Bauer nel 2021 "A Goal Scoring Probability Model for Shots" [8]. In questo studio, i due analisti tedeschi decisero per la prima volta di basare le loro valutazioni sui tiri dei giocatori, introducendo così il modello degli xG. Gli studi precedenti si basavano, appunto, unicamente sui goal segnati dove il gol è l'evento che si verifica con la probabilità minore in una partita. Inoltre, concentrandosi solo su dati di tipo quantitativo relativi ai gol e alle partite, non venivano tenuti in considerazione variabili come la zona del campo del tiro, le caratteristiche del marcatore, del portiere, la situazione di gioco, l'attitudine difensiva e offensiva di una squadra, l'allenatore, le tattiche, lo stato di forma dei giocatori e così via. Maggiore è la quantità di futures inserite nel modello, maggiore sarà la precisione restituita.

3. Introduzione

Nell'era moderna del calcio, l'utilizzo dei dati e dell'analisi statistica ha rivoluzionato il modo in cui questo sport viene compreso, interpretato e valutato. Le tecnologie avanzate e l'ingegneria informatica hanno aperto nuove prospettive per una visione più approfondita e oggettiva delle prestazioni sportive, sia a livello individuale che a livello di squadra. In questo contesto, l'impiego degli Expected Goals (xG) si è affermato come una delle metodologie più innovative e promettenti per valutare l'efficacia delle azioni offensive nel calcio.

"Lo scopo del calcio è segnare gol", questo è un concetto che da sempre caratterizza il gioco. Tuttavia, valutare l'efficacia di una squadra o di un giocatore basandosi solo sul numero di gol segnati può risultare ingannevole e poco esaustivo. Infatti, i gol sono spesso il risultato di un intricato insieme di variabili, tra cui la precisione dei tiri tenendo quindi conto della posizione del portiere avversario e della distanza dalla porta, grazie all'utilizzo delle coordinate cartesiane. Di conseguenza, una squadra potrebbe segnare un numero considerevole di gol, nonostante la qualità delle sue occasioni sia piuttosto bassa.

L'obiettivo principale del modello xG è assegnare un valore compreso tra zero e uno per ciascun tiro, che rappresenta la probabilità che un tiro che si trasformi in un gol. Il gol è un evento che si verifica raramente. È proprio in questo contesto che gli Expected Goals entrano in gioco. Gli xG rappresentano un indicatore statistico che quantifica la probabilità di un tiro di concludersi con un gol, sulla base delle caratteristiche dell'azione stessa. In altre parole, gli xG tengono conto di variabili come la distanza dalla porta, l'angolazione del tiro, la parte del corpo utilizzata per il tiro e la pressione degli avversari.

Questa metodologia offre un'analisi più dettagliata ed equilibrata delle prestazioni offensive, permettendo di valutare in modo più accurato l'efficacia delle azioni in fase di attacco.

Nell'ambito della Soccer Analytics, gli xG hanno suscitato un crescente interesse da parte di analisti, allenatori e addetti ai lavori. La loro capacità di svelare aspetti nascosti delle prestazioni calcistiche ha aperto nuove prospettive per migliorare l'efficacia delle strategie di gioco e per ottimizzare le decisioni tattiche. Come afferma Luke Bornn, direttore del Dipartimento di Analisi dei Dati del Sacramento Kings e precedentemente del calcio professionistico, "Gli xG sono uno strumento potente per valutare le prestazioni e i modelli di gioco di una squadra. Forniscono una chiara indicazione delle abilità offensive e della capacità di creare opportunità da gol".

In questo studio, si propone l'applicazione degli Expected Goals in relazione alle varie partite di club in Serie A, nell'ultima stagione 2022-23. Per raggiungere l'obiettivo prefissato, questa ricerca si articolerà in diverse fasi fondamentali:

Raccolta e Preparazione dei Dati: Verranno raccolti dati relativi alla scorsa stagione comprendenti informazioni sulle 380 (38 turni da 10 partite) partite con maggior enfasi sui tiri effettuati. L'analisi comprenderà la quantità di tiri, con particolare attenzione a quelli convertiti a rete, alcune variabili, tra cui la distanza del giocatore dalla porta, il tipo di situazione che ha portato al tiro (azione, calcio piazzato, corner…) e la modalità (tiro effettuato con il piede destro o sinistro oppure con la testa). Questi dati saranno fondamentali per la costruzione del modello predittivo.

Sviluppo del Modello xG: verrà sviluppato un modello predittivo in grado di calcolare gli xG al fine di renderlo universale e facilmente implementabile per altri campionati e stagioni. Questo modello sarà poi applicato ai dati raccolti per valutare l'efficacia delle squadre e dei giocatori nella stagione scelta, riguardo alla creazione di opportunità da gol.

Implicazioni e Applicazioni Pratiche: Verranno discussi gli utilizzi pratici degli xG nel contesto delle decisioni tattiche e strategiche. Inoltre, esploreremo come queste informazioni possano influenzare le scelte degli allenatori e dei dirigenti sportivi, migliorando la competitività delle squadre e l'esperienza dei tifosi.

Sfide e Limitazioni: Infine, analizzeremo le sfide e le limitazioni legate all'utilizzo degli xG nel calcio, evidenziando le aree in cui ulteriori ricerche potrebbero essere utili per sviluppare ulteriormente questa metodologia.

In conclusione, questa tesi si propone di apportare un contributo significativo all'ambito delle Soccer Analytics, dimostrando il valore e l'utilità degli Expected Goals nel contesto del campionato di Serie A TIM 2022-23. Gli xG rappresentano uno strumento potente per l'Ingegneria Informatica applicata allo sport, offrendo un approccio innovativo e promettente per valutare le prestazioni delle squadre e dei calciatori.

In sintesi, questo studio si propone di esplorare e analizzare dati calcistici estremamente ricchi e dettagliati in un periodo di tempo limitato, al fine di offrire una visione approfondita delle prestazioni delle squadre e dei giocatori.

La speranza è che i risultati ottenuti possano essere di grande interesse e utilità per tutti coloro che seguono il calcio e che desiderano comprendere meglio le dinamiche di questo affascinante sport.

4. Dati

Nel corso di questo studio, si propone di analizzare un vasto archivio di informazioni calcistiche reperibili online in particolare su Understat[1], una fonte affidabile che fornisce una ricca raccolta di dati relativi alle principali leghe europee di calcio. All'interno di questo archivio, troviamo una miriade di informazioni riguardanti le partite e i dati sulle varie situazioni che si verificano negli incontri. Questa raccolta di dati è una miniera d'oro per gli appassionati di calcio e gli analisti, in quanto ci offre una panoramica completa e dettagliata delle prestazioni delle squadre e dei giocatori nelle principali leghe europe.

La nostra analisi si concentra su dati che coprono un intervallo temporale ristretto e limitato a una sola stagione, la precedente 2022-23. Questo ci consente di osservare la tendenza in merito ai tiri effettuati e convertiti a rete in un periodo recente e, per questo, relativamente carente in merito ad analisi statistiche aggiornate. Ci aspettiamo che questa vasta serie di dati ci permetta di individuare pattern e relazioni interessanti non ancora rilevati dagli analisti e che potrebbero essere di grande utilità per tifosi, allenatori e studi futuri.

L'analisi su concentrerà sul campionato italiano principale: la Serie A. Ogni lega presenta le proprie caratteristiche e peculiarità, e la nostra analisi cercherà di scoprire quali fattori sono in grado di creare dei pattern e influenzare le prestazioni delle squadre.

Il dataset contiene i tiri realizzati nella stagione 2022-23 del campionato italiano di Serie A e conta un totale di ben 944 gol in 380 partite, di cui 38 turni da 10 partite ciascuno.

5. Modello utilizzato e sua applicazione

5.1 Modello scelto

Nel calcio l'obiettivo è segnare ed essendo il Goal un evento raro [2], il miglior modello per confrontare dati sulla probabilità di segnare è il "Modello Logistico" con parametri stimati mediante massima likelihood [9]. Il modello logistico è un modello statistico utilizzato per analizzare dati binari, dove la variabile dipendente può assumere solo due valori possibili, quali "successo" o "fallimento", "positivo" o "negativo". Questo tipo di modello è particolarmente comune nell'analisi della regressione logistica, dove si cerca di modellare la probabilità di un evento che si verifica rispetto a uno che non si verifica. Inoltre, si rivela adatto in questo caso, grazie alla sua facile interpretazione per quanto riguarda i regressori. Ciò permette di introdurre nuovi indici per migliorare la stima nel xG model.

I coefficienti di regressione sono stimati con procedura "maximum likelihood". Il modello Expected Goals è stato valutato utilizzando standard di classificazione. [10]

$$xG = P(Goal|\mathbf{X}) = \frac{e^{\mathbf{X}\hat{\beta}}}{1 + e^{\mathbf{X}\hat{\beta}}} \tag{1}$$

$$logL(\hat{\beta}) = \sum_{i=1}^{n} y_i log(xG_i) + \sum_{i=1}^{n} (1 - y_i) log(1 - xG_i) \tag{2}$$

L'espressione numero (1) rappresenta la formula della funzione logistica (o sigmoide) utilizzata nei modelli di regressione logistica. Nello specifico:

- xG: è la variabile dipendente o la probabilità condizionata di un evento "goal" dato un insieme di predittori (X).

- P (Goal | X): è la probabilità condizionata di un goal dato l'insieme di predittori (X).
- Numeratore: è l'esponenziale di (X) moltiplicato per il vettore dei coefficienti $\hat{\beta}$, è il predittore lineare.
- Denominatore: è la somma tra 1 e l'esponenziale di (X) moltiplicato per $\hat{\beta}$.
- Numeratore/Denominatore: è la formula della funzione logistica, la quale trasforma l'output dell'espressione al numeratore in un intervallo compreso tra 0 e 1, rendendolo interpretabile come una probabilità.

Una probabilità non può essere modellata attraverso un modello di regressione lineare, poiché non è garantito, così facendo, di ottenere delle stime, per tale probabilità, comprese tra 0 e 1. Questo è il motivo della scelta di usare il modello logistico. La funzione logistica è spesso utilizzata per modellare la probabilità di appartenenza a una determinata classe (in questo caso, la classe "goal") in funzione di un insieme di variabili predittive (X) e dei loro pesi associati $\hat{\beta}$.

L'espressione numero (2) rappresenta la funzione di verosimiglianza logaritmica utilizzata in modelli di regressione logistica. Nel dettaglio:

- $\log L(\hat{\beta})$: è la funzione di verosimiglianza logaritmica, che misura quanto i dati osservati sono probabili dati i parametri stimati $\hat{\beta}$ del modello.
- La prima sommatoria: somma i logaritmi delle probabilità predette xG_i associati agli eventi osservati y_i, ovvero misura quanto bene il modello si adatta ai dati osservati per gli eventi positivi.
- La seconda sommatoria: somma i logaritmi delle probabilità complementari per gli eventi negativi $(1-y_i)$, misura quanto bene il modello si adatta ai dati osservati per gli eventi negativi.

In una regressione logistica, l'obiettivo è massimizzare la funzione di verosimiglianza logaritmica, il che significa trovare i valori dei coefficienti $\hat{\beta}$, che rendono i dati osservati più verosimili. Questo processo è spesso effettuato utilizzando metodi di ottimizzazione numerica. La funzione di verosimiglianza logaritmica è centrale nelle stime e nelle valutazioni di modelli di regressione. [6]

Nel contesto dell'analisi statistica condotta mediante il codice in R, l'utilizzo della stepwise regression ha giocato un ruolo significativo nel processo di selezione delle variabili esplicative da includere nel modello finale. La stepwise regression è una tecnica di selezione delle variabili che consente di determinare quali predittori includere nel modello sulla base della loro capacità predittiva e del loro contributo alla varianza.

Nel codice ho applicato la stepwise regression al modello completo inizialmente costruito, che includeva tutte le variabili esplicative disponibili. Utilizzando la funzione `step`, ho esplorato iterativamente tutte le possibili combinazioni di variabili esplicative, aggiungendo o rimuovendo variabili in base ai criteri definiti. Nello specifico, ho specificato la direzione come "both", indicando che la selezione delle variabili potrebbe includere sia l'aggiunta che la rimozione.

La stepwise regression ha quindi valutato la significatività delle variabili esplicative presenti nel modello completo, utilizzando un criterio di selezione basato sul p-value, che rappresenta la probabilità di osservare un determinato risultato quando l'ipotesi nulla è vera. In particolare, la stepwise regression ha cercato di ottimizzare il modello selezionando le variabili che contribuiscono in modo significativo alla predizione della variabile dipendente, ovvero la probabilità di segnare un goal.

Attraverso questo processo, sono state selezionate solo le variabili esplicative più rilevanti e significative per la predizione del goal, riducendo così la complessità del modello e migliorandone l'interpretabilità. L'obiettivo finale era quello di ottenere un modello logistico più accurato, in grado di fornire previsioni affidabili in base alle differenti condizioni di gioco.

5.2 Regressori utilizzati

Le variabili esplicative utilizzate nel modello per calcolare gli Expected Goals (xG) comprendono:

-Variabili indipendenti:

1. home_away:

- indica se la squadra che effettua il tiro è in casa o in trasferta.

2. situation:

- rappresenta il tipo di situazione di gioco in cui si verifica il tiro, ad esempio "Open Play" (situazione con palla in movimento) o "Set Piece" (palla inattiva).

3. shotType:

- identifica il tipo di tiro effettuato dal giocatore, che può essere "Left Foot" (tiro con il piede sinistro), "Right Foot" (tiro con il piede destro), "Headed" (colpo di testa), o altri.

4. lastAction_aggregated:

- rappresenta l'azione precedente al tiro, aggregata in categorie rilevanti. Ad esempio, può essere "Standard" per azioni di gioco comuni o "Altro" per azioni di gioco meno frequenti.

5. distanza_porta:

- la distanza del punto di tiro dalla porta, calcolata utilizzando la formula del teorema di Pitagora, considerando la posizione del tiro rispetto al punto centrale della porta.

-Variabili aggiuntive:

6. home_team:

- indica la squadra di casa coinvolta nella partita in cui si è verificato il tiro.

7. away_team:

- rappresenta la squadra ospite, coinvolta nella partita in cui il tiro è stato registrato.

5.3 Applicazione del modello

Il modello logistico è stato utilizzato per stimare le probabilità di segnare un gol (Expected Goals) in base alle variabili esplicative sopra elencate. Il processo di raccolta dei dati includeva l'acquisizione delle informazioni sui tiri effettuati durante le partite di calcio, l'elaborazione dei dati per calcolare la distanza dalla porta e l'aggregazione delle azioni precedenti al tiro. Il modello è stato poi addestrato utilizzando il dataset risultante.

Sono state applicate alcune modifiche al dataset, tra cui: le categorie con una frequenza inferiore a 160 osservazioni sono state accorpate in una categoria "Altro" per la variabile `lastAction_aggregated`, inoltre è stata utilizzata la procedura stepwise per selezionare le variabili rilevanti per il modello.
La scelta del cut-off di 160 per la frequenza delle categorie di `lastAction_aggregated` è stata motivata dalla necessità di garantire un numero sufficiente di osservazioni per ciascuna categoria e di evitare il sovrapprendimento del modello. Mentre con la procedura stepwise, le variabili che non hanno un effetto significativo sulla probabilità di segnare un gol vengono eliminate dal modello.
Le modifiche apportate hanno permesso di ottenere un modello più robusto e interpretabile, in grado di guidare analisi più accurate sulle prestazioni delle squadre e dei giocatori. Le variabili rimaste nel modello selezionato hanno dimostrato di avere un impatto significativo sulla probabilità di segnare un goal.

Il prossimo passo è l'analisi dei coefficienti beta, i quali forniscono informazioni sull'effetto delle variabili esplicative sulla probabilità di segnare un gol. Un coefficiente positivo indica un aumento della probabilità di segnare un gol quando il valore della variabile aumenta, mentre un coefficiente negativo indica una diminuzione della probabilità.

5.4 Analisi dei risultati

Sono stati raccolti un totale di 9457 tiri che si sono concretizzati in un Goal in 944 casi per la stagione 2022-23 di Serie A. Possiamo quindi constatare che in media ogni 10 tiri c'è stato un goal, con una percentuale realizzativa calcolata sul totale di circa 9.98 %.

Il tiro con il parametro di xG maggiore si è verificato il 04/06/2023 nel match Ac Milan contro Verona, con un valore pari a 0.979 ed è stato convertito a rete con tiro di destro dell'attaccante del Milan Rafael Leao, il quale ha siglato il 3 a 1 nel match.
Il tiro con il parametro di xG maggiore non convertitosi a rete è capitato sul piede destro dello sfortunato Tyronne Ebuehi, difensore dell'empoli, il quale ha completamente mancato la porta, nel match perso per 2 a 0 all'Olimpico contro la Roma.
Un altro dato molto curioso riguarda il tiro convertitosi a rete con il minor xG, primato che per la stagione 22-23 di Serie A appartiene al centrocampista offensivo dell'Atalanta, Teun Koopeminers. Il valore registrato è stato di 0.00817, ciò non ha comunque impedito all'olandese di siglare con il piede preferito, il sinistro, il goal del definitivo 5 a 2 a discapito del Monza, con un pallonetto da centrocampo, a seguito di un ottimo recupero palla.

Ponendo, invece, il focus sulle variabili esplicative utilizzate, i risultati ottenuti sono stati raccolti e riassunti nella tabella che segue. Possiamo vedere i valori dei differenti coefficienti per quanto riguarda il modello completo e il modello selezionato. Infine, troviamo anche il P-value associato a ciascun coefficiente.

Variabili	Modello completo	Modello selezionato	P Value
distanza_porta	-0.17065	-0.16995	< 2e-16
shotTypeRightFoot	1.02521	1.01133	< 2e-16
shotTypeLefttFoot	1.17178	1.15141	< 2e-16
shotTypeOtherBodyParts	0.76560	0.67844	0.0816
situationFromCorner	-1.64203	-1.65235	1.97e-08
situationSetPiece	-1.56454	-1.57388	1.39e-06
situationOpenPlay	-1.07533	-1.06687	7.73e-05
situationPenalty	1.77052	1.73465	1.35e-06
home_awayh	0.01809	/	/

Fig.2

Tabella valori coefficienti da R

Analizzando la variabile distanza_porta nel modello selezionato e in quello completo notiamo che un coefficiente negativo per la variabile "distanza dalla porta" indica che all'aumentare della distanza dalla porta, le probabilità di segnare un goal tendono a diminuire. In questo caso, il coefficiente ha valore -0.17 nel modello completo e -0.169 nel modello selezionato, ciò suggerisce che per ogni unità di aumento della distanza dalla porta, ovvero ogni metro aggiuntivo, le probabilità di segnare un goal diminuiscono di circa lo 0.17%. Questo risultato è intuitivo e conferma quanto ci si potrebbe aspettare nel calcio: tiri effettuati da distanze maggiori hanno una minore probabilità di convertirsi in rete, rispetto a tiri effettuati da distanze più vicine. È opportuno fare queste considerazioni avendo un campione di tiri relativamente elevato; infatti, se ci riferissimo a una singola partita, il relativo coefficiente potrebbe avere un valore controintuitivo.

Nel modello selezionato, abbiamo osservato che il coefficiente per i tiri effettuati con il piede sinistro è pari a 1.15. Ciò suggerisce che i tiri effettuati con il piede sinistro hanno una maggiore probabilità di convertire in goal rispetto ai tiri effettuati con altre parti del corpo. Questo risultato è in linea con le aspettative, poiché i calciatori solitamente sviluppano una maggiore precisione e potenza con il loro piede dominante. D'altra parte, il coefficiente per i tiri effettuati con il piede

destro è di poco inferiore, pari a 1.01. Anche se i tiri con il piede destro hanno comunque una probabilità più alta di convertire in gol rispetto a quelli effettuati con altre parti del corpo, l'effetto è meno pronunciato rispetto ai tiri effettuati con il piede sinistro.
Nel modello completo, i coefficienti per i tiri con entrambi i piedi rimangono coerenti con quelli del modello selezionato, sebbene con valori leggermente diversi. Il coefficiente per i tiri effettuati con il piede sinistro è leggermente più alto, pari a 1.17, confermando che questi tiri hanno la più alta probabilità di successo. Allo stesso modo, il coefficiente per i tiri con il piede destro è 1.02, indicando che questi tiri hanno una probabilità più alta di segnare rispetto ai tiri con altre parti del corpo, ma l'effetto è meno forte rispetto ai tiri con il piede sinistro. Infine, il coefficiente per i tiri effettuati con altre parti del corpo è 0.76 nel modello completo. Anche se questi tiri hanno una probabilità inferiore di segnare rispetto ai tiri con i piedi, l'effetto è meno marcato rispetto al modello selezionato. In generale, i risultati confermano l'importanza del piede dominante nel calcio.

Riguardo alla significatività dei dati, il p-value dei vari coefficienti è di gran lunga inferiore a 0.05 e prossimo allo zero, ciò indica la significatività dei dati raccolti e analizzati per i coefficienti che indicano la modalità con la quale avviene il tiro. Tuttavia, la variabile che indica i goal realizzati con altre parte del corpo, “ShotTypeOtherBodyPart” presenta un p-value di 0.0816. Questo indica che c'è una buona probabilità che l'effetto osservato di questa variabile sul risultato dipenda dal caso, piuttosto che rappresentare un effetto reale e significativo. In generale, un p-value superiore a 0.05 indica che non ci sono prove statisticamente significative per respingere l'ipotesi nulla, la quale afferma che non vi è alcun effetto della variabile sul risultato.

Invece nel confronto mancini destrorsi, è interessante notare il vantaggio, seppur ridotto, dei tiratori mancini rispetto ai giocatori di piede destro. Questa differenza è già stata ampiamente discussa da alcuni studi in altri sport, tra cui il tennis, il ping pong e il baseball. [11][12] Le conclusioni sono state che, a quanto pare, essendo la prevalenza di atleti a dominanza destra, nell’eventualità di confronti

sportivi con giocatori mancini, i secondi trovino impreparati i primi. Questa difficoltà ad adattarsi dei destrorsi ad un tipo di gioco differente poiché praticato da una minoranza di atleti, porta a prestazioni di livello mediamente inferiore dei giocatori a dominanza destra. A livello calcistico gli studi che hanno provato a dimostrare questa sottile differenza sono ridotti. Un esempio significativo è uno studio del 2018 appartenente alla Facoltà di scienze sportive di Istanbul in Turchia. Due ricercatori hanno infatti raccolto dati sulle prestazioni calcistiche di 61 giocatori under 15 di cui 31 destri e 30 mancini, di differenti pesi e altezze, provenienti da altrettanti club di Istanbul. [13] I risultati hanno evidenziato, dopo diverse prove calcistiche, un leggero vantaggio nell'abilità di tiro e nel dribbling in favore dei mancini, nonostante i destri evidenzino un piccolo vantaggio nell'esecuzione dei lanci lunghi.

Nell'analisi dei coefficienti relativi alle varie situazioni di gioco che influenzano la probabilità di segnare un goal, emerge un quadro interessante che riflette alcune dinamiche comuni nel calcio moderno. Innanzitutto, osserviamo che i goal segnati dopo un calcio d'angolo, coefficiente di -1.65, e le situazioni con palla in movimento, coefficiente di -1.06, tendono ad avere una probabilità inferiore di convertirsi a rete rispetto ad altre situazioni di gioco.

Questo risultato è in linea con la ricerca condotta da alcuni analisti nel 2018 sui calci d'angolo nel calcio. [14] Con il loro studio hanno analizzato alcuni fattori che intervengono sul calcolo della probabilità del goal da situazione di calcio d'angolo, come per esempio la quantità di uomini offensivi portati in area: maggiore è il numero di attaccanti e maggiore sarà la probabilità di segnare.

Inoltre, è interessante notare che il valore dei coefficienti mette in risalto come le squadre difensive, in un campionato come la Serie A, in cui lo studio della tattica rappresenta un tassello importante nella preparazione delle partite, spesso aumentino la loro organizzazione e concentrazione durante queste situazioni da calcio piazzato, riducendo così le opportunità di concludere a rete per le squadre in attacco.

D'altra parte, i tiri segnati da calcio di rigore, il cui valore è di 1.73, mostrano una probabilità significativamente più alta di segnare rispetto ad altre situazioni di

gioco. Tutto ciò è in linea con le aspettative, infatti i calci di rigore offrono un'opportunità più favorevole per fare goal, in quanto il tiro è eseguito senza la pressione difensiva diretta e da una posizione favorevole.
Infine, i tiri segnati da situazioni di palla inattiva, come i calci d'angolo o i calci da fermo, coefficiente di -1.57 presentano una probabilità inferiore di convertirsi a rete rispetto ad altre situazioni di gioco. Questo risultato può essere attribuito alla maggiore capacità delle squadre avversarie di organizzarsi difensivamente durante queste situazioni, rendendo più difficile trovare spazi liberi e creare opportunità di segnare.

In conclusione, i risultati dei modelli confermano alcune dinamiche comuni nel calcio professionistico e sono in linea con la letteratura scientifica esistente sull'argomento. Tuttavia, è importante considerare che i modelli statistici possono offrire solo una prospettiva parziale sulle complesse dinamiche del gioco, e ulteriori ricerche qualitative e quantitative possono essere necessarie per approfondire la comprensione di tali fenomeni.

A seguire, l'analisi della variabile "home_awayh" nel modello logistico suggerisce che il suo ruolo nella previsione della probabilità di segnare un gol potrebbe non essere significativo. Questa conclusione emerge dal confronto tra il modello completo, che include tutte le variabili esplicative disponibili, e il modello selezionato, che è stato ottimizzato attraverso la stepwise selection. Nel modello completo, la variabile "home_awayh" è stata inclusa insieme ad altre variabili come parte del set di covariate utilizzate per predire la probabilità di segnare un gol. Tuttavia, nel modello selezionato, la variabile è stata esclusa, indicando che la sua presenza non ha contribuito in modo significativo alla capacità predittiva del modello.
Questo risultato può essere interpretato considerando diversi fattori. Innanzitutto, il p-value associato alla variabile nel modello completo, pari a 0.81 e di gran lunga superiore a 0.05, potrebbe essere risultato non significativo, suggerendo che non vi è sufficiente evidenza statistica per supportare l'effetto della variabile sulla probabilità di segnare un gol. Ciò potrebbe indicare che il fatto di giocare in casa

o in trasferta non influenza in modo sostanziale la capacità di una squadra di segnare. Inoltre, l'esclusione della variabile potrebbe essere stata determinata anche da considerazioni di collinearità o ridondanza con altre variabili presenti nel modello.
La stepwise selection mira a semplificare il modello mantenendo solo le variabili più informative e significative. Se l'esclusione della variabile casa trasferta ha migliorato le prestazioni predittive complessive del modello, pur mantenendo la sua interpretabilità, potrebbe essere stata considerata una scelta ragionevole.

Analizzando i valori dei coefficienti delle varie squadre di serie A in casa e in trasferta possiamo notare che per alcune vengono influenzate in modo significativo le probabilità di segnare un gol durante una partita. Nel passaggio dal modello completo al modello selezionato, sono state scartate tutte le variabili che indicavano la presenza delle squadre di casa e ospiti, tuttavia nel modello completo possiamo notare alcune squadre con una relativa importanza statistica. In particolare, le squadre di casa, Empoli, Sampdoria e Spezia, nel rispettivo stadio, sono emerse come significative a livello statistico, con p-values rispettivamente di 0.0284, 0.0272 e 0.0698. Ciò suggerisce che la presenza di queste squadre potrebbe avere un impatto sulle probabilità di segnare un gol durante una partita.

Ad esempio, considerando la squadra di casa Empoli, il coefficiente associato è stato stimato a -0.509, con uno standard error di 0.232. Questo indica che, mantenendo costanti tutte le altre variabili nel modello, la presenza di Empoli è associata a una diminuzione della log-odds di segnare un gol. Tuttavia, poiché il p-value associato è appena al di sotto del valore di tolleranza standard di 0.05, bisogna essere cauti nell'interpretare questo risultato come un'associazione significativa.

Analogamente, la presenza delle squadre di casa Sampdoria e Spezia ha mostrato coefficienti negativi, rispettivamente -0.528 e -0.422, indicando una diminuzione della log-odds di segnare un gol. Tuttavia, poiché il p-value di Sampdoria è inferiore al valore di tolleranza, possiamo considerare questa associazione più

attendibile rispetto a Spezia, la cui significatività potrebbe essere influenzata da altri fattori non inclusi nel modello, p value rispettivi di 0.027 e 0.069.

In conclusione, l'analisi suggerisce che diverse squadre di Serie A possono influenzare le probabilità di segnare un gol, in maniera differente durante una partita. Tuttavia, è importante considerare altri fattori contestuali e interpretare i risultati con cautela per comprendere appieno l'impatto delle singole squadre e dei singoli giocatori sulle prestazioni complessive.

6. Conclusioni

Dalle analisi condotte sui dati relativi alla stagione 2022-2023 della Serie A emerge un quadro ricco di informazioni e considerazioni degne di approfondimento. L'obiettivo principale di questo studio era quello di utilizzare modelli logistici per predire la probabilità di segnare un gol durante una partita di calcio, esplorando l'effetto di diversi fattori contestuali e situazionali.

Inizialmente, è stato raccolto un ampio dataset contenente informazioni dettagliate su tiri effettuati durante le partite di Serie A, includendo variabili quali la situazione di gioco: corner, azione con palla in movimento, rigore; il tipo di tiro: di sinistro, di destro o con altre parti del corpo; la distanza dalla porta e altri fattori correlati al contesto della partita.

L'applicazione di modelli logistici ha consentito di esplorare la relazione tra queste variabili e la variabile di risposta binaria rappresentante il successo, ovvero il goal, o il fallimento di un tiro. La fase di analisi ha incluso sia la costruzione di un modello completo, includendo tutte le variabili disponibili, sia l'applicazione di un approccio di selezione stepwise per identificare le variabili più rilevanti per la predizione del successo nel segnare un goal.

Durante l'analisi dei risultati ottenuti dai modelli, sono emerse diverse considerazioni degne di nota. Ad esempio, è stato osservato che la distanza dalla porta ha un impatto significativo sulle probabilità di segnare, con tiri effettuati da distanze più corte che mostrano una maggiore probabilità di successo.
Allo stesso modo, la situazione di gioco ha giocato un ruolo importante, con tiri da situazioni di "Open Play" che hanno mostrato una maggiore probabilità di successo rispetto a tiri da "Corner" o "Set Piece".
Infine, è emerso un interessante pattern riguardante i tiri effettuati con il piede sinistro rispetto a quelli effettuati con il piede destro. In particolare, i tiri con il sinistro sembrano mostrare una leggera tendenza a ottenere una maggiore

probabilità di successo rispetto a quelli con il destro. Questa osservazione può essere di particolare interesse per gli allenatori e gli analisti di squadra, che potrebbero considerare l'importanza di avere giocatori in grado di effettuare tiri accurati con entrambi i piedi durante la pianificazione delle strategie di gioco. Potrebbe risultare interessante confrontare questi dati con le stagioni passate per capire meglio se questo sia un dato ricorrente oppure che ha avuto luogo in alcune caselle temporali, dovuto alla presenza di mancini di maggior talento.

Durante l'analisi sono emerse anche alcune criticità e dubbi da tenere in considerazione. Ad esempio, l'assenza di significatività statistica per alcune variabili potrebbe indicare la presenza di fattori non inclusi nel modello che potrebbero influenzare le probabilità di segnare un gol. Inoltre, l'esclusione di alcune variabili durante la fase di selezione del modello potrebbe sollevare domande sulla loro effettiva rilevanza nel contesto del gioco, come ad esempio il dato generale riguardante i goal in casa e in trasferta. Sicuramente spostando il focus sulle singole squadre, in alcuni casi si può avere rilevanza statistica.

Nonostante queste criticità, lo studio fornisce una base solida per futuri approfondimenti e analisi. Sarebbe interessante confrontare i dati raccolti per la stagione 2022-2023 con dati storici delle stagioni precedenti per identificare tendenze e pattern nel tempo e per sviluppare modelli predittivi più accurati. Inoltre, concentrarsi su singole squadre potrebbe fornire informazioni preziose per gli addetti ai lavori nel calcio, aiutandoli nella scelta degli acquisti dei giocatori e nello sviluppo di tattiche strategiche per migliorare le prestazioni delle squadre.

In conclusione, questo studio rappresenta un punto di partenza importante per future analisi nel campo del calcio, offrendo nuove prospettive e approcci per comprendere e predire le dinamiche del gioco. La combinazione di metodologie statistiche avanzate e conoscenze approfondite del contesto sportivo potrebbe portare a risultati ancora più significativi e utili agli stakeholder del mondo del calcio per scelte ponderate e più consapevoli.

7. Elenco figure

8. Bibliografia

1. https://www.understat.com/
2. Rathke, A. (2017). An examination of expected goals and shot efficiency in soccer. Journal of Human Sport and Exercise, 12(2), 514-529
3. Umami, (2013). Implementing the expected goal (xG) Model to predict scores in soccer matches.
4. Fairchild, A. ; Pelechrinis, K. ; Kokkodis, M. (2018). Spatial analysis of shots in MLS: A model for expected goals and fractal dimensionality.
5. Ruan, L. (2022). Quantifying the Effectiveness of Defensive Playing Styles in the Chinese Football Super League.
6. Mattia Cefis, Maurizio Carpita, (2023). An original application of PLS-SEM for football analytics: the xG Model,
7. https://www.sofifa.com/
8. Gabriel, A.; Pascal, B., (2021). A Goal Scoring Probability Model for Shots Based on Synchronized Positional and Event Data in Football (Soccer).
9. James, G.; Witten, D.; Hastie, T.; Tibshirani, R.; vol.112. Springer (2013). An introduction to statistical learning.
10. Hossin, M. & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process, 5(2), 1.
11. Petro, B.; Szabo, A. (2016). The Impact of Laterality on Soccer Performance. Strength & Conditioning Journal. 38(5), 66-74
12. Loffing, F.; Hagemann, N. (2016). Performance Differences between Left- and Right Sided Athletes in One-on-One Interactive Sports. Laterality in Sports Theories and Applications.12, 249-277
13. Sinan, B.; Veysel, K., (2018). Comparing of Technical Skills of Young Football Players According to Preferred Foot,
14. Claudio, C.; Rubén, M.; Toni, A; José L.; Antonio, R., (2017). Analysis of Corner Kick Success in Elite Football

9. Appendice

Di seguito il codice utilizzato in R:

```
# Installa e carica le librerie necessarie
install.packages(c("worldfootballR", "dplyr"))
library(worldfootballR)
library(dplyr)

# Lista per memorizzare i dati di tutti i match
all_matches_list <- list()

# Costruisci la lista di URL dei match
start_match_id <- 18582
end_match_id <- 18961

match_urls <- sprintf("https://understat.com/match/%d",
start_match_id:end_match_id)

# Iterazione su ogni URL del match
for (url in match_urls) {
  match_data <- understat_match_shots(match_url = url)
  match_data$name <- url  # eventualmente un nome piu descrittivo
  all_matches_list[[url]] <- match_data
}

# Unisci i dati di tutti i match in un unico dataframe
all_matches_df <- bind_rows(all_matches_list)

# Calcola la distanza dalla porta utilizzando la formula modificata
```

```
all_matches_df$distanza_porta <- sqrt((abs(all_matches_df$X - 1) * 105)^2 +
(abs(all_matches_df$Y - 0.5) * 65)^2)
# Aggiungi la variabile aggregata per lastAction
frequenza_lastAction <- table(all_matches_df$lastAction)
categorie_selezionate <- names(frequenza_lastAction[frequenza_lastAction >=
160])
all_matches_df <- all_matches_df %>% filter(lastAction %in%
categorie_selezionate)
all_matches_df$lastAction_aggregated <- ifelse(all_matches_df$lastAction ==
"Standard", "Standard", "Altro")

# Crea la variabile di risposta goal
all_matches_df$goal <- 0
all_matches_df$goal[all_matches_df$result == "Goal"] <- 1

# Crea il modello completo con tutte le esplicative
modello_completo <- glm(goal ~ home_away + situation + shotType +
lastAction_aggregated + distanza_porta + home_team + away_team,
                        family = "binomial", data = all_matches_df)

# Applica la stepwise selection al modello completo
modello_selezionato <- step(modello_completo, direction = "both")
# Visualizza il modello completo
summary(modello_completo)

# Visualizza il modello selezionato
summary(modello_selezionato)
# Tabella di conteggio per la variabile goal
table(all_matches_df$goal)
```

Ringraziamenti

Desidero esprimere la mia sincera gratitudine a Mattia Cefis, il cui lavoro è stato fonte di ispirazione e guida preziosa durante la stesura della mia tesi sugli Expected goals. La sua ricerca ha illuminato il mio percorso e mi ha fornito preziosi spunti di riflessione.

Un sentito ringraziamento va al mio stimato relatore, il prof. Rodolfo Metulini, per la sua disponibilità, competenza e preziosi consigli durante tutto il percorso di elaborazione della tesi. La sua guida e il suo sostegno sono stati fondamentali per portare a termine questo lavoro.

Ringrazio la mia famiglia a cui voglio molto bene, che mi ha sostenuto anche troppo in ogni fase del mio percorso accademico. Un ringraziamento speciale va a mia mamma e mio papà, per il loro amore, educazione trasmessa, senso del sacrificio, impegno, sostegno e passione. Un pensiero affettuoso va anche a mia sorella Sofia, per il suo coraggio e per un futuro universitario ricco di soddisfazioni.

Un ringraziamento particolare va ai miei amati nonni di Monasterolo, Renzo e Angela e ai nonni di Torino, Rosa e Leonardo, che con il loro affetto incondizionato e la loro saggezza hanno illuminato il mio cammino fin dall'infanzia.

Desidero inoltre ringraziare la mia fidanzata Benedetta, per il suo amore, il suo sostegno, i suoi consigli e la sua comprensione durante i momenti più impegnativi di questo percorso.

Infine, un caloroso ringraziamento va a tutti i miei GreAtS aMici e al resto della mia famiglia, per il loro incoraggiamento, il loro appoggio e i bei momenti condivisi che hanno reso questo viaggio ancora più significativo. Grazie di cuore a tutti coloro che hanno contribuito al mio percorso accademico e personale.