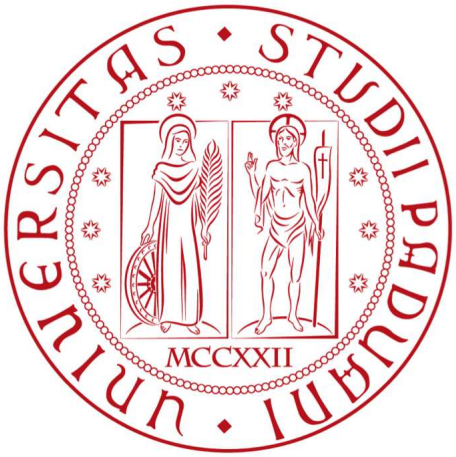
Università degli studi di Padova

Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in Statistica per l'Economia e l'Impresa



Relazione Finale

La previsione dei risultati delle partite di calcio mediante modelli statistici

Relatore Prof. Luigi Grossi
Dipartimento di Scienze Statistiche

Laureando Nicola Pregnolato
Matricola n.2012096

Anno Accademico 2022/23

INDICE

1 LA STATISTICA E LO SPORT

1.1 La statistica oggi

La statistica assieme ai modelli matematici sta subendo un'importante evoluzione che sta facendo progredire il mondo in modo esponenziale soprattutto negli ultimi decenni.

Dall'introduzione dell'intelligenza artificiale con l'esplosione di ChatGPT passando alle statistiche descrittive arrivando ai modelli di previsione. Questi ultimi oggigiorno vengono applicati in moltissimi settori come ad esempio quello economico (la previsione del PIL), quello finanziario (la previsione dei tassi d'interesse), quello sociale (la previsione delle tendenze demografiche), quello sanitario (la previsione dell'andamento dei contagi) e via dicendo.

1.2 L'evoluzione dell'intreccio tra la statistica e lo sport

Tra i vari settori in cui viene applicata la statistica è presente anche quello sportivo: i modelli di previsione nello sport nascono a partire dalla metà del XX secolo, più precisamente attorno agli anni '50 e '60 quando il matematico statunitense Bill James sviluppa la formula dell'Aspettativa di Pitagora per il baseball. Questo modello utilizza il rapporto tra punti fatti e subiti per calcolare la probabilità di vittoria o sconfitta di ogni determinata squadra del campionato.

Un momento significativo nell'intreccio tra statistica e sport è rappresentato dall'introduzione del Moneyball. L'approccio Moneyball si basa sull'utilizzo di misure statistiche più sofisticate per valutare i giocatori senza fare affidamento

alle opinioni personali dei talent scout. Questa tecnica nasce nei primi anni 90' quando Billy Beane (dirigente sportivo degli Oakland Athletics della Major League Baseball), assieme al suo team, si pone l'obiettivo di individuare giocatori sottovalutati che hanno buone prestazioni così da poterli acquistare ad un costo inferiore rispetto ai grandi nomi del baseball. La filosofia del Moneyball ,resa famosa anche grazie all'omonimo film uscito nel 2011, cambiò significativamente il mondo dello sport, tant'è che ancora oggi questo metodo viene utilizzato in qualsiasi disciplina sportiva di gruppo.

Infatti proprio in questo periodo è emerso come il Milan, una delle squadre italiane più famose al mondo, stia utilizzando l'analisi dei dati e statistiche avanzate per assumere decisioni più accurate inerenti la formazione della nuova squadra.

1.3 I dati nel calcio

Come accennato in precedenza, il progresso portato dalla statistica coniugata all'intelligenza artificiale e all'innovazione tecnologica sta influenzando l'intero pianeta. Risulta quindi complicato pensare che l'ambiente calcistico non risenta anch'esso della capacità di analizzare e interpretare il dato.

Infatti come vedremo i dati sono diventati uno degli strumenti essenziali al supporto di direttori sportivi, allenatori e giocatori.

1.3.1 Come vengono utilizzati i dati nel calcio?

Le applicazioni principali dell'analisi dei dati nel calcio vengono suddivise in quattro aree differenti:

1. Scouting: come visto in precedenza abbiamo citato il Moneyball. L'analisi dei dati permette di valutare le prestazioni dei giocatori, di individuare

nuovi talenti, di confrontare tra di loro giocatori e individuare caratteristiche simili e di ridurre il rischio di investimento. In sintesi è un aiuto non indifferente ai talent scout e ai dirigenti sportivi in cerca di nuovi giocatori su cui investire.

2. Valutazione delle prestazioni in allenamento: essere presenti alla partita nella miglior forma possibile ad oggi è diventato cruciale nel mondo del calcio. Grazie alle numerose innovazioni tecnologiche come le pettorine GPS, gli smartwatch e i droni si possono rilevare dati atletici molto precisi per ogni giocatore giorno per giorno. Questo aiuta sia gli allenatori che, assieme ai preparatori atletici, riescono a capire quale calciatore risulta essere nella migliore forma atletica ma anche i calciatori a tenere sotto controllo i propri valori di salute e individuare subito la natura di un possibile problema.

3. Analisi della squadra avversaria: l'analisi tecnica e tattica della squadra avversaria che si andrà a sfidare nella partita successiva può aiutare gli allenatori a sviluppare la migliore tattica in base ai punti di forza e di debolezza individuati.

4. Analisi del mercato delle scommesse: con la nascita del dato nello sport sono nate anche le scommesse e i cosiddetti *bookmakers* (agenzie di scommesse). Le statistiche delle squadre e dei giocatori sono gli elementi che influenzano maggiormente questo mercato poichè le varie quote vengono calcolate in base alle probabilità che un determinato evento si verifichi o meno. Vedremo in seguito con l'applicazione dei dati come queste probabilità potrebbero venire calcolate.

1.4 Obiettivo della tesi

L'obiettivo di questa tesi è sviluppare un modello di previsione abbastanza affidabile per i risultati delle partite di calcio, individuando quali potrebbero essere i coefficienti che più influenzano il numero di gol segnati da una determinata squadra in una determinata partita.

2 MODELLI DI PREVISIONE PER I RISULTATI DELLE PARTITE DI CALCIO

In questo capitolo vengono spiegati i principali modelli già presenti in letteratura che hanno contribuito allo sviluppo del modello utilizzato al Capitolo 3, quando verrà utilizzato per l'applicazione ai dati del campionato di Serie A.

2.1 La distribuzione di Poisson

L'approccio econometrico predominante nella maggior parte dei casi è quello di modellare il numero dei gol segnati dalle due squadre che si affrontano in una partita basandosi su due distribuzioni di Poisson.

La distribuzione di Poisson è una distribuzione di probabilità discreta che viene utilizzata per determinare il numero di eventi che si verificano in un intervallo di tempo e in uno spazio specifico, dato un tasso medio di occorrenza. Questa distribuzione è caratterizzata da due parametri principali:

1. il tasso medio λ (lambda): rappresenta il numero medio ma anche la varianza degli eventi che si verificano.
2. La variabile casuale X: rappresenta il numero effettivo di eventi che si verificano in tale intervallo o spazio.

La funzione di probabilità è data dalla seguente formula:

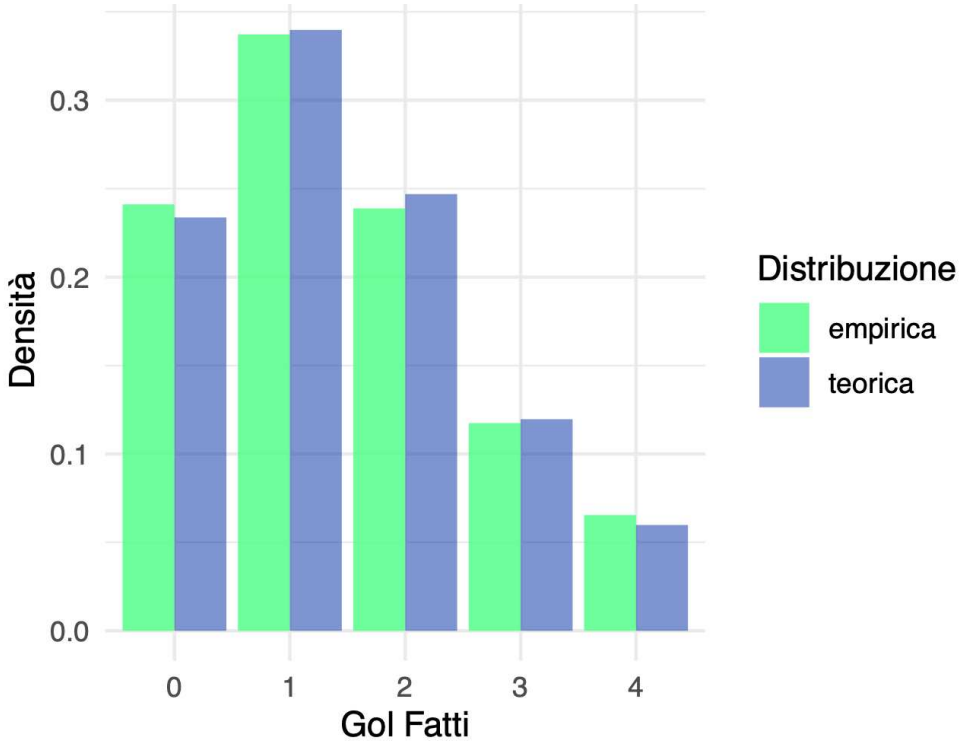
$$P(x = k) = \frac{e^{-\lambda} \cdot \lambda^k}{k!} \quad x \in N \, e \, \lambda > 0 \tag{1}$$

Una particolarità della Poisson è che il valore atteso e la varianza si equivalgono

infatti:

$$E[x] = V[x] = \lambda \tag{2}$$

Sembra quindi ragionevole assumere che questa distribuzione sia la più adatta a descrivere il numero di gol fatti da ciascuna squadra in una partita.




```
X-squared = 5.6215, df = 7, p-value = 0.5846

> var(data_golfatti$golfat)
[1] 1.522692
> mean(data_golfatti$golfat)
[1] 1.453618
```

Figura 1: Distribuzione empirica e teorica dei gol segnati in Serie A dalla stagione 2018/19 alla stagione 2021/22 e output del comando Chisq.test di R.

Infatti dalla Figura 1 si può notare come la distribuzione empirica dei dati sui gol fatti da una squadra in una partita è molto simile alla distribuzione teorica di una Poisson. Due ulteriori conferme sono:

1. il p-value del test Chi-Quadro (0.5846), il quale suggerisce che non si può

rifiutare l'ipotesi che le due distribuzioni siano uguali.

2. media e varianza, che, come detto precedentemente in una distribuzione teorica di Poisson, dovrebbero essere uguali, risultano essere molto simili (media=1.522692, varianza=1.453618).

Questa assunzione distributiva, come vedremo in seguito, è la base di quasi tutti i modelli previsivi dei risultati delle partite di calcio presenti in letteratura, compreso il modello utilizzato per l'applicazione dei dati di questa tesi.

2.2 Il modello di Maher

Il modello di Maher (1982) è considerato uno dei primi modelli, se non il vero e proprio primo modello, utilizzato per la previsione dei risultati delle partite di calcio. Esso si basa sul fatto che ogni volta che una squadra è in fase di possesso ha l'opportunità di attaccare e segnare.

La probabilità p che un attacco si tramuti effettivamente in un gol è generalmente bassa, ma al contrario il numero di volte in cui una squadra ha il possesso del pallone durante una partita è molto elevato. Se assumiamo che la probabilità p sia costante e che gli attacchi siano indipendenti l'uno dall'altro, allora il numero di gol segnati dalla squadra seguirà una distribuzione binomiale. Per questo motivo, soprattutto quando il numero di attacchi è grande, l'approssimazione di Poisson risulta essere molto accurata e valida.

In altre parole, considerando che il numero di attacchi di una squadra durante una partita è molto alto e che la probabilità di segnare un gol in un singolo attacco è piccola e costante, possiamo approssimare il numero di gol segnati da una squadra come una variabile casuale di Poisson. Questo rende più semplice

e pratico il trattamento statistico dei dati relativi ai gol segnati in una partita di calcio.

2.2.1 La Double Poisson

Basandosi quindi su questa assunzione, Maher propone un modello chiamato Double Poisson: il risultato di una partita è dato dai gol segnati dalla squadra di casa e dalla squadra fuori casa. Questi vengono rappresentati da due variabili di Poisson indipendenti così definite:

$$X_{ij} \sim \text{Pois}(\alpha_i \beta_j) \qquad\qquad Y_{ij} \sim \text{Pois}(\alpha_j \beta_i), \tag{3}$$

dove Xij rappresenta il numero dei gol segnati dalla squadra i (in casa) e Yij rappresenta il numero di gol segnati dalla squadra j (fuori casa). Il parametro α esprime la forza offensiva di ogni squadra, mentre il parametro β esprime la forza difensiva della squadra avversaria. Questi due parametri sono necessari in quanto il numero di gol segnati in una partita da una squadra non dipende solamente dalle capacità offensive, bensì anche dalle capacità difensive avversarie. Infatti si può dimostrare che una squadra di forza offensiva media avrà una probabilità di segnare molto più alta contro una squadra considerata scarsa piuttosto che contro una squadra forte.

Essendo, a detta di Maher, Xij e Yij indipendenti tra loro, allora la distribuzione congiunta è data dal prodotto delle due densità marginali, vale a dire:

$$P(X_{ij} = x, Y_{ij} = y) = (\alpha_i \beta_j)^x \frac{e^{-\alpha_i \beta_j}}{x!} (\alpha_j \beta_i)^y \frac{e^{-\alpha_j \beta_i}}{y!} \tag{4}$$

$$con\ X_{ij} \perp Y_{ij},\ x = 0, 1, 2, \ldots \quad \text{e} \quad \alpha_k, \beta_k > 0.$$

Inoltre, deve valere l'equazione seguente che garantisce l'unicità dei parametri:

$$\sum_i \alpha_i = \sum_j \beta_j; \quad \sum_j \alpha_j = \sum_i \beta i. \tag{5}$$

2.3 Il Modello Dixon-Coles

Uno dei motivi che più ha stimolato lo sviluppo di nuovi modelli di previsione per le partite è stato sicuramente il mercato delle scommesse. Ciò ha condotto gli studiosi Mark J. Dixon e Stuart G. Coles ad implementare un modello più completo rispetto alla semplice Double Poisson di Maher.

2.3.1 La Bivariate Poisson

Sebbene Maher avesse capito che assumere totale indipendenza tra le due distribuzioni non fosse del tutto esatto, il suo modello non prevede una correlazione tra le due Poisson. Questo è uno dei primi punti sviluppati da Dixon e Coles. Infatti essi propongono un allontanamento dall'assunzione di indipendenza per i risultati con pochi gol (0-0, 1-0, 0-1, 1-1) analizzando le partite della Premier League e delle coppe nazionali inglesi dal 1992 al 1995.

Suggeriscono quindi un modello differente dalla Double Poisson, aggiun-

gendo alla funzione di densità la funzione τ (definita nella 7) e il parametro di correlazione ρ. Questo modello viene chiamato Bivariate Poisson e la sua funzione di densità è così definita:

$$\begin{aligned} P(X_{ij} = x, Y_{ij} = y) &= P(X_{ij} = x) \cdot P(Y_{ij} = y) \cdot \tau_{\lambda,\mu}(x, y, \rho) \\ &= \frac{\lambda_1^x \cdot e^{-\lambda_1} \cdot \lambda_2^y \cdot e^{-\lambda_2}}{x! \cdot y!}; \quad \text{con } \lambda_1 = \alpha_i \beta_j \gamma \quad , \quad \lambda_2 = \alpha_j \beta_i \end{aligned} \tag{6}$$

$$\text{e} \quad \tau_{\lambda_1,\lambda_2}(x, y, \rho) = \begin{cases} 1 - \lambda_1 \lambda_2 \rho, & \text{se } x = y = 0, \\ 1 + \lambda_1 \rho, & \text{se } x = 0, y = 1, \\ 1 + \lambda_2 \rho, & \text{se } x = 1, y = 0, \\ 1 - \rho, & \text{se } x = y = 1, \\ 1, & \text{altrimenti.} \end{cases} \tag{7}$$

dove γ rappresenta l'*home effect*, ovvero il vantaggio che possiede la squadra che gioca in casa giustificato da molteplici fattori come la presenza di più tifosi, l'adattamento al campo di gioco e la mancanza di stanchezza dovuta ad una possibile trasferta lontana. Si può notare che γ non è indicizzato, questo perchè viene considerato costante sia per semplicità sia perchè provando a non assumerlo costante le previsioni non migliorano. A conferma di ciò, la Figura 2 mostra come il parametro non cambi significativamente nel tempo.

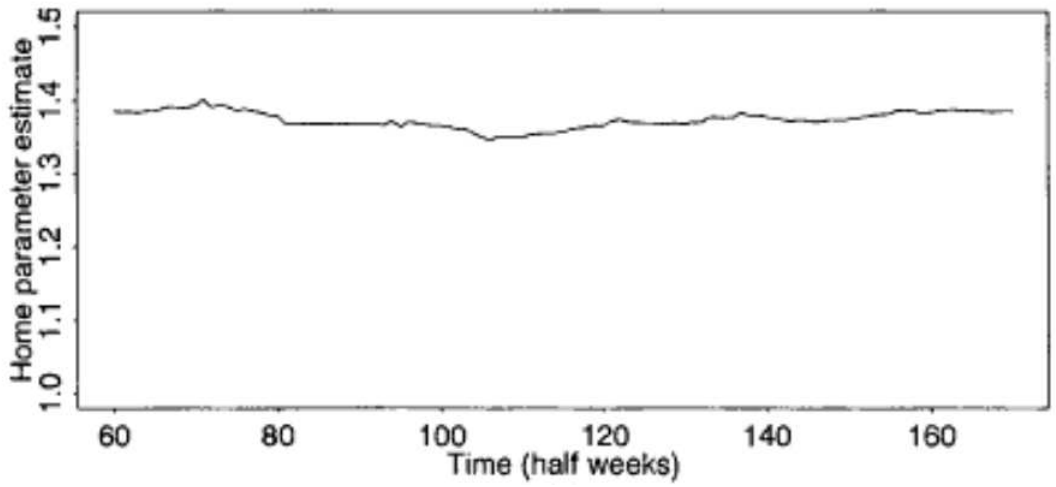


Figura 2: serie storica del parametro γ che rappresenta l'*home effect* delle squadre della Premier League dal 1992 al 1995, output del comando ts.plot()

La funzione τ invece, con $max(-\frac{1}{\lambda_1}, -\frac{1}{\lambda_2}) \leq \rho \leq min(\frac{1}{\lambda_1\lambda_2}, 1)$, rappresenta il parametro di dipendenza tra le due distribuzioni di Poisson. In termini calcistici si può pensare come al fatto che quando una squadra (qualsiasi siano le sue capacità) sta perdendo con un gol di scarto, sarà più propensa ad attaccare rispetto a quanto attaccava sullo 0-0. Si noti che, se $\rho = 0$, allora la funzione di densità della Bivariate Poisson torna ad essere la funzione di densità della Double Poisson sviluppata da Maher.

L'ultimo aggiustamento introdotto in questo modello riguarda il fatto che le performance recenti dovrebbero avere un peso maggiore rispetto alle partite degli anni precedenti. Infatti, si può facilmente pensare che una squadra che proviene da una serie di cinque partite consecutive in calo di forma (ad esempio Vittoria-Vittoria-Sconfitta-Sconfitta-Sconfitta) sarà più indicata a perdere contro una squadra in uno stato di forma ascendente (ad esempio Sconfitta-Sconfitta-Sconfitta-Vittoria-Vittoria) sebbene il numero di vittorie e sconfitte siano le stesse.

Questo tipo di fenomeno è ancora più rilevante ad inizio ed a metà campi-

onato, quando le finestre di calciomercato permettono alle varie squadre di acquistare o cedere giocatori consentendo così rispettivamente di potenziarsi o indebolirsi. Per questo i due studiosi decidono di aggiungere un'altra funzione $\phi(t)$ che permette loro di far variare la funzione di verosimiglianza nel tempo. Quest'ultima così definita:

$$\phi(t) = e^{(-\xi_t)}. \tag{8}$$

abilita quindi di pesare le capacità delle squadre al variare del tempo determinato dal parametro ξ. Quest'ultimo può essere visto come un indice di variabilità delle prestazioni delle squadre e del campionato: un valore prossimo a zero indica che i dati provenienti dalle stagioni passate sono molto affidabili, mentre un valore più alto (ad esempio $\xi = 0.0015$ relativo ai dati della Premier League dal 2012 al 2016) indica che esiste un'alta variabilità delle prestazioni, confermata anche dalle classifiche finali che mostrano un diverso piazzamento ogni anno.

In sintesi, il modello Dixon-Coles apporta le seguenti modifiche rispetto al modello di base:

- aggiunta del parametro γ che rappresenta il vantaggio del giocare in casa,
- aggiunta della funzione τ che identifica una dipendenza in caso di risultati bassi (0-0, 1-0, 0-1, 1-1),
- le capacità offensive e difensive (α_k e β_k con $k = i, j$) non sono un dato costante preso ad inizio anno come prevedeva il modello di Maher, bensì una valutazione delle performance recenti.

2.4 Il Modello di Karlis-Ntzoufras

Questo modello risulta essere molto simile al modello Dixon-Coles, con una diversa precisazione per quanto riguarda l'analisi della correlazione nella Bivariate Poisson. Infatti, Karlis e Ntzoufras estendono l'assunzione della dipendenza non solo per risultati bassi, ma in generale per qualsiasi risultato. Questo perchè secondo loro la probabilità di un pareggio in una Bivariate Poisson è maggiore rispetto ad una Double Poisson, anche nel caso in cui λ_3, che rappresenta la covarianza tra λ_1 e λ_2, sia relativamente piccolo.

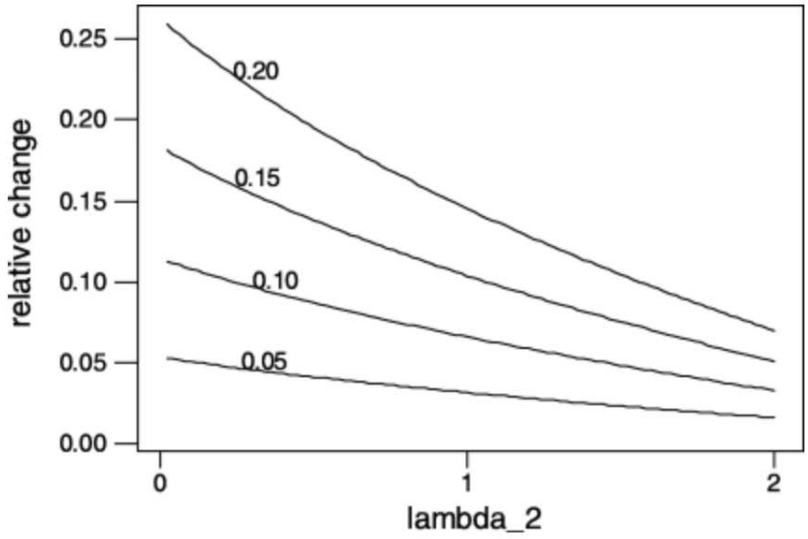


Figura 3: Variazione relativa della probabilità di pareggio quando le due squadre hanno media marginale $\lambda_1 = 1$ e λ_2 che varia tra 0.1 e 2. Le linee differenti rappresentano i vari valori di λ_3 che rappresenta la covarianza tra λ_1 e λ_2. Immagine trovata dall'articolo originale sul modello Karlis-Ntzoufras

La Figura 3 mostra il variare della probabilità di pareggi al variare della covarianza λ_3. Quindi, per estendere la probabilità di pareggio, Karlis e Ntzoufras

decidono di sviluppare un modello *diagonal inflated* così specificato:

$$P_D(x,y) = \begin{cases} (1-p)BP(x,y|\lambda_1,\lambda_3,\lambda_3) & \text{se } x \neq y, \\ (1-p)BP(x,y|\lambda_1,\lambda_2,\lambda_3) + pD(x,\theta) & \text{se } x = y. \end{cases} \tag{9}$$

dove $D(x,\theta)$ è una distribuzione discreta tra una Poisson, Geometrica o Bernoulli di parametro θ e p, il rappresenta il peso che si vuole dare alla diagonale della matrice dei risultati possibili (tutti i pareggi possibili).

Riassumendo, questo modello prevede l'aggiunta di una funzione per estendere la probabilità di pareggio non solo per risultati bassi, ma per qualsiasi risultato.

2.5 Il Modello Koopman-Lit

Uno dei modelli più recenti è stato sviluppato da Siem Jan Koopman e Rutger Lit, i quali aggiungono alcuni aggiustamenti ai modelli già conosciuti in precedenza. Anche secondo loro la distribuzione Poisson Bivariata è quella che più si adatta a rappresentare i risultati delle partite di calcio. Poi, come nel modello Dixon-Coles, propongono l'aggiustamento alla Double Poisson aggiungendo la funzione τ definita però in questo modo:

$$\text{e} \quad \tau_{\lambda,\mu}(x,y,\rho) = \begin{cases} 1-\lambda_1\lambda_2\rho, & \text{se } x=y=0, \\ 1+\lambda_1\rho, & \text{se } x=0, y=1, \\ 1+\lambda_2\rho, & \text{se } x=1, y=0, \\ 1-\frac{\rho}{1+\frac{\lambda_3}{\lambda_1\lambda_2}}, & \text{se } x=y=1, \\ 1, & \text{altrimenti.} \end{cases} \tag{10}$$

2.5.1 Introduzione di parametri di regressione dinamici

L'aggiustamento più significativo è l'inserimento di parametri di regressione (α_k e β_k) non più costanti fissati ad inizio anno, bensì dinamici. Per far sì quindi che i parametri λ_1 e λ_2 dipendano dal tempo è necessario considerare i parametri di regressione come delle serie storiche che vengono modellate come dei processi autoregressivi con previsione ad un passo (AR(1)).

I parametri vengono definiti in questo modo:

$$\lambda_{1,t} = e^{(\gamma+\alpha_{i,t}+\beta_{j,t})}; \quad \lambda_{2,t} = e^{(\alpha_{j,t}+\beta_{i,t})}$$

$$\alpha_{i,t} = \mu_{\alpha_i} + \phi_{\alpha_i}\alpha_{i,t-1} + \epsilon_{\alpha_i,t}; \quad \beta_{i,t} = \mu_{\beta_i} + \phi_{\beta_i}\beta_{i,t-1} + \epsilon_{\beta_i,t} \tag{11}$$

$$\epsilon_{k_i,t} \sim WN(0,\sigma^2_{ki}), \quad \text{con } k = \alpha,\beta \quad \text{e } \sigma^2_{ki} > 0$$

dove μ_{α_i} e μ_{β_i} sono costanti ignote, ϕ_{α_i} e ϕ_{β_i} sono i coefficienti che caratterizzano la parte autoregressiva e $\epsilon_{\alpha_i,t}$ e $\epsilon_{\beta_i,t}$ sono gli errori indipendenti e identicamente distribuiti con media zero e varianza costante (*White Noise*).

Questa scelta è stata fatta perchè è molto piu ragionevole che le capacità di una squadra non siano sempre le stesse per un anno intero (o metà campionato nel caso in cui si volessero ricalibrare i parametri durante le sessioni di calciomercato invernale oltre che quello estivo). Infatti è chiaro che la forma della squadra non è costante: possono infortunarsi o rimanere squalificati giocatori importanti, oppure una squadra può arrivare più stanca rispetto all'avversaria per il maggior numero di partite disputate (ad esempio a causa di un turno infrasettimanale di coppe europee).

Per questo motivo, il modello Koopman-Lit risulta uno dei più completi per fare previsioni sui risultati delle partite di calcio e fornisce quindi, assieme ai modelli citati precedentemente, un grosso spunto sul modello sviluppato per l'applicazione dei dati al Capitolo 3.

3 APPLICAZIONE DEL MODELLO SUI DATI DELLA SERIE A

Per mostrare empiricamente dei risultati sul modello creato viene applicato il modello specificato in seguito sui dati riguardanti la Serie A per gli anni 2018/19, 2019/2020, 2020/2021 e 2021/22.

3.1 Raccolta dei dati

Per analizzare i dati e testare il modello più adatto, sono stati scaricati quattro dataset (uno per ogni campionato elencati in precedenza) contenenti tutte le informazioni di ogni partita. I dataset sono stati scaricati da *Kaggle*, una piattaforma gratuita online dove si possono trovare informazioni utili e dati già raccolti per svariati argomenti.

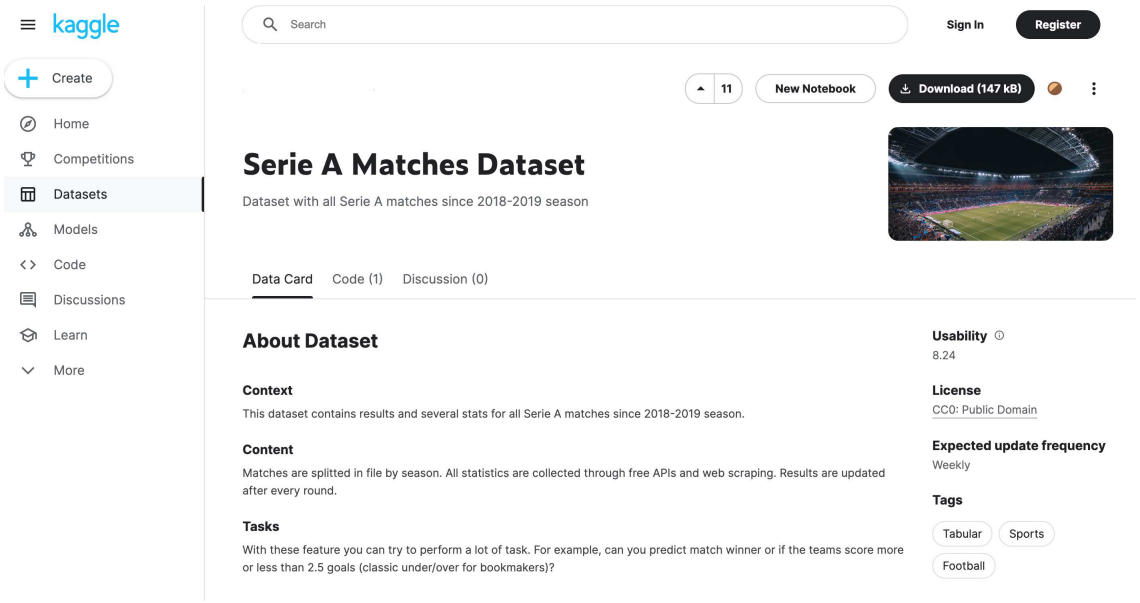


Figura 4: Pagina web della piattaforma Kaggle, in alto a destra il tasto *download* e nel mezzo alcune informazioni utili sulle variabili definite nei dataset.

Le tabelle sono state importate poi in formato .csv su R, dove è stata fatta tutta l'analisi dati.

round	match_id	home_team_id	home_team	away_team_id	away_team	home_goals	away_goals	home_pred	draw_pred
1	28993	491	Chievo	496	Juventus	2	3	0%	50%
1	28995	487	Lazio	492	Napoli	1	2	0%	50%
1	28999	503	Torino	497	AS Roma	0	1	0%	50%
1	28998	488	Sassuolo	505	Inter	1	0	50%	50%
1	28992	500	Bologna	493	Spal	0	1	33%	33%
1	28994	511	Empoli	490	Cagliari	2	0	0%	50%
1	28996	523	Parma	494	Udinese	2	2	0%	50%
1	28991	499	Atalanta	512	Frosinone	4	0	50%	50%
2	29005	496	Juventus	487	Lazio	2	0	45%	45%
2	29006	492	Napoli	489	AC Milan	3	2	45%	45%
2	29008	493	Spal	523	Parma	1	0	45%	45%
2	29009	494	Udinese	498	Sampdoria	1	0	35%	35%
2	29000	490	Cagliari	488	Sassuolo	2	2	10%	45%
2	29001	502	Fiorentina	491	Chievo	6	1	45%	45%
2	29002	512	Frosinone	500	Bologna	0	0	10%	45%
2	29003	495	Genoa	511	Empoli	2	1	10%	45%
2	29004	505	Inter	503	Torino	2	2	10%	45%
2	29007	497	AS Roma	499	Atalanta	3	3	10%	45%
3	29010	489	AC Milan	497	AS Roma	2	1	10%	45%
3	29012	500	Bologna	505	Inter	0	3	10%	45%

Figura 5: Esempio di un dataset scaricato: sulle colonne le variabili d'interesse (come *home team, home goals,...*) e sulle righe ogni unità statistica (ogni partita).

3.2 Il Modello Utilizzato

Come detto nelle sezioni precedenti, tutte le informazioni apprese nei modelli trovati in letteratura sono servite per creare il modello più efficiente in termini previsivi.

3.2.1 Adattamento dei gol alla distribuzione di Poisson

Come si è notato la distribuzione di Poisson è l'assunzione base su cui si fondano la maggior parte dei modelli presenti in letteratura.

Infatti, si è visto come nella Figura 1 l'assunzione sia chiaramente corretta: il *p-value* = 0.5846 suggerisce che l'ipotesi che i gol segnati si distribuiscano come una Poisson (output della funzione *chisq.test()*) non si debba rifiutare.

Per mostrare che questa assunzione valga anche per le singole squadre si può notare la Figura 6.

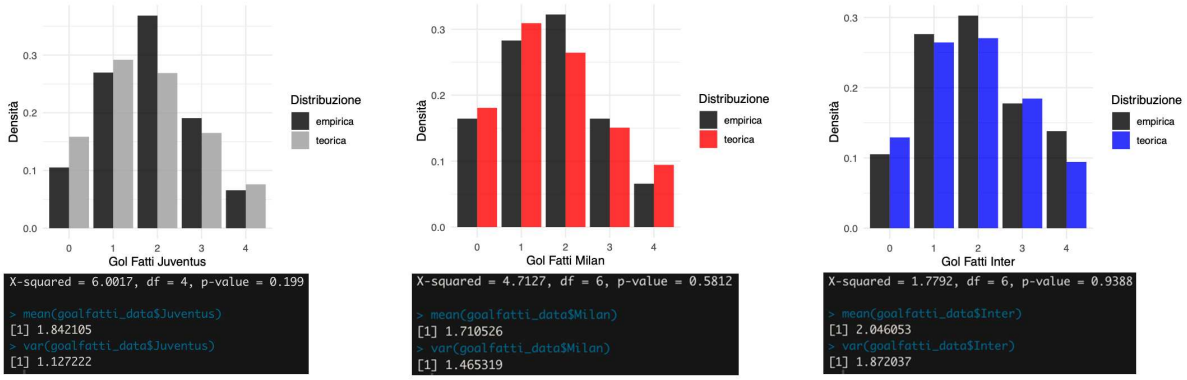


Figura 6: Distribuzione empirica e teorica (distribuzione di Poisson) dei gol fatti dalle tre squadre più tifate in Italia.

E' da notare che nei grafici le classi sono cinque poichè per un grafico visivamente accettabile si sono dovute raggruppare le classi che prevedono quattro o più gol all'interno della medesima classe. Invece, per quanto riguarda il test (output R sotto ogni grafico), non c'è stato nessun raggruppamento di classi per far sì che fosse il più corretto possibile. Infatti, nel test per il Milan e per l'Inter, i gradi di libertà *df* sono sei poichè ci sono stati casi in cui le due squadre hanno segnato anche più di quattro gol (ad esempio Torino-Milan terminata 0-7 nell'anno 2020/21).

3.2.2 Spiegazione del modello

Il modello utilizzato per l'applicazione è simile al modello Koopman-Lit. La diiferenza è che nel modello già presente in letteratura i parametri $\lambda_{1,t}$ e $\lambda_{2,t}$ vengono trattati come serie storiche modellate come dei processi autoregressivi con previsione ad un passo, mentre in quello sviluppato qui i parametri vengono determinati attraverso due modelli di regressione di Poisson.

Quindi:

$$log(\lambda_{i,t}) = \phi_{0it} + \phi_{1it}\alpha_{it} + \phi_{2it}\beta_{jt} \tag{12}$$

$$log(\lambda_{j,t}) = \phi_{0jt} + \phi_{1jt}\alpha_{jt} + \phi_{2jt}\beta_{it} \tag{13}$$

dove α_{kt} e β_{kt} con $k = i, j$ sono i parametri che rappresentano rispettivamente la forza offensiva e difensiva delle due squadre. Da notare che dipendono tutti dal tempo, in quanto sono parametri dinamici e non fissati ad inizio campionato. I parametri ϕ invece sono i paramentri di regressione del modello. Per ottenere le stime dei parametri $\lambda_{1,t}$ e $\lambda_{2,t}$ è necessario applicare l'esponenziale in questo modo:

$$\lambda_{i,t} = \gamma e^{\phi_{0it} + \phi_{1it}\alpha_{it} + \phi_{2it}\beta_{jt}} \tag{14}$$

$$\lambda_{j,t} = e^{\phi_{0jt} + \phi_{1jt}\alpha_{jt} + \phi_{2jt}\beta_{it}}. \tag{15}$$

Si ricorda che γ è l'*home effect* poichè come detto nel par. 2.3.1 la squadra che gioca in casa ha un vantaggio determinato da vari aspetti già spiegati.

3.2.3 I parametri α e β

Come detto in precedenza per i parametri α (misura della capacità offensiva) e β (misura della capacità difensiva avversaria) non sono costanti bensì dipendono dal tempo t. Per quanto riguarda il parametro α, è stato fatto un confronto per capire quale metodo per calcolare il parametro fosse più adatto comparando i due indici U di Theil.

L'Indice U di Theil fornisce una misura di quanto le previsioni fatte con il modello scelto sono più efficaci rispetto al *modello naive* (modello semplice che assegna al valore previsto al tempo $t+1$ il valore osservato al tempo t). L'indice

può assumere valori che vanno da 0 a $+\infty$ e si calcola in questo modo:

$$U = \frac{\sqrt{EQM_{modello}}}{\sqrt{EQM_{modellonaive}}} \tag{16}$$

I valori vanno così interpretati:

$$\begin{aligned} U = 0 &\quad \text{previsione perfetta} \\ U < 1 &\quad \text{previsione modello} > \text{previsione modello naive} \\ U = 1 &\quad \text{soglia di discriminazione, il modello è inutile} \\ U > 1 &\quad \text{il modello è inaccetabile} \end{aligned} \tag{17}$$

Per calcolare α è stato prima utilizzato il lisciamento esponenziale di Holt con previsione ad un passo e poi la media dei gol fatti nelle cinque partite precedenti a quella disputata. Gli output di R dei due indici per ciascuna squadra sono i seguenti:

```
    Atalanta      Bologna     Cagliari       Empoli   Fiorentina        Genoa
   0.7707510    0.7383531    0.8341794    0.7363988    0.8769924    0.7850612
Hellas Verona        Inter     Juventus        Lazio        Milan       Napoli
   0.7162372    0.7914965    0.7148493    0.7521091    0.7894272    0.7785174
        Roma   Salernitana    Sampdoria     Sassuolo       Spezia       Torino
   0.7489611    0.9972907    0.8105462    0.7961502    0.7665525    0.7542158
     Udinese      Venezia
   0.7407489    0.8350985
```

Figura 7: Indice U di Theil utilizzando il lisciamento esponenziale di Holt ad un passo per le previsioni dei risultati nei campionati 2018/19, 2019/2020, 2020/2021, 2021/22.

```
    Atalanta      Bologna     Cagliari       Empoli   Fiorentina        Genoa
   0.8153881    0.7276460    0.8339485    0.7412499    0.7326877    0.7724388
Hellas Verona        Inter     Juventus        Lazio        Milan       Napoli
   0.7128747    0.7846054    0.7461594    0.7625316    0.7855453    0.7881156
        Roma   Salernitana    Sampdoria     Sassuolo       Spezia       Torino
   0.7544960    0.8183333    0.7378647    0.7941740    0.7164658    0.7836412
     Udinese      Venezia
   0.7737175    0.7807251
```

Figura 8: Indice U di Theil utilizzando la media dei gol fatti le 5 partite precedenti per le previsioni dei risultati nei campionati 2018/19, 2019/2020, 2020/2021, 2021/22.

Come si può osservare, gli indici risultano tutti < 1, il che significa che i due metodi sono entrambi più efficaci rispetto all'utilizzo di un semplice *modello naive*. Dunque, per scegliere quale tra il primo e il secondo metodo utilizzato è il più adatto, si calcola la media, per i due metodi, degli indici U di Theil di ogni squadra: 0.7867 utilizzando Holt, 0.7681 utilizzando la media dei gol delle ultime cinque partite. Pertanto, dato che dalla 17 si capisce che più l'indice si avvicina a 0 più le previsioni sono efficaci, la scelta è quella di usare la media dei gol delle ultime cinque partite come parametro della forza offensiva (α).

Per quanto riguarda il parametro β, la decisione è quella di utilizzare la media dei gol subiti dalla squadra avversaria in tutte le partite precedenti. Anche in questo caso viene calcolato l'Indice U di Theil per dare una misura di quanto possono essere efficaci le previsioni, mostrando i seguenti risultati:

Atalanta	Bologna	Cagliari	Empoli	Fiorentina	Genoa
0.7913835	0.6807984	0.7670334	0.6578337	0.7122926	0.7622392
Hellas Verona	Inter	Juventus	Lazio	Milan	Napoli
0.6708360	0.7668687	0.7389116	0.6953986	0.7771372	0.7564956
Roma	Salernitana	Sampdoria	Sassuolo	Spezia	Torino
0.6756150	0.8527200	0.6792638	0.7371050	0.6798797	0.6900842
Udinese	Venezia				
0.7015736	0.8000048				

Figura 9: Indice U di Theil utilizzando la media dei gol subiti in tutte le partite precedenti dalla squadra avversaria per le previsioni dei risultati nei campionati 2018/19, 2019/2020, 2020/2021, 2021/22.

In sintesi α sarà rappresentato dalla media dei gol fatti nelle ultime 5 partite antecedenti alla partita di cui bisogna fare la previsione, mentre β sarà la media dei gol subiti dalla squadra affrontata in tutte le partite precedenti alla partita analizzata.

3.2.4 Il parametro γ: l'*home effect*

Come mostrato al par. 2.3.1 alla Figura 2, il parametro γ non varia di molto nel tempo,al contrario di α e β, ma rimane perlopiù costante. Per questo motivo non è stato indicizzato, ma è stato preso un valore unico per ogni singola previsione, dividendo la somma totale dei gol fatti dalle squadre in casa alla somma totale dei gol fatti dalle squadre fuori casa. Il risultato è $\gamma = \frac{2369}{2050} = 1.1556$, il che si può interpretare come il vantaggio che la squadra ha di giocare in casa, appunto *home effect*.

3.2.5 La funzione τ e il parametro p di dipendenza

Come visto nel par. 2.3.1 uno dei fattori più importanti sviluppati da Dixon e Coles è il parametro p di dipendenza. Difatti, i risultati con un numero limitato di gol (0-0, 1-0, 0-1, 1-1) hanno una maggiore probabilità di verificarsi rispetto ai risultati con un numero elevato di gol. Questa dipendenza, introdotta nel modello Dixon-Coles, è stata poi sviluppata, seppur diversamente, in tutti gli altri modelli. Per questo motivo, viene inserita anche in questo modello e viene definita come nel par. 2.3.1, quindi:

$$\text{e} \quad \tau_{\lambda_1,\lambda_2}(x,y,\rho) = \begin{cases} 1-\lambda_1\lambda_2\rho, & \text{se } x=y=0, \\ 1+\lambda_1\rho, & \text{se } x=0, y=1, \\ 1+\lambda_2\rho, & \text{se } x=1, y=0, \\ 1-\rho, & \text{se } x=y=1, \\ 1, & \text{altrimenti.} \end{cases} \tag{18}$$

con $max(-\frac{1}{\lambda_1}, -\frac{1}{\lambda_2}) \leq \rho \leq min(\frac{1}{\lambda_1 \lambda_2}, 1)$ e con funzione di densità che diventa:

$$P(X_{ij} = x, Y_{ij} = y) = P(X_{ij} = x) \cdot P(Y_{ij} = y) \cdot \tau_{\lambda,\mu}(x, y, \rho). \tag{19}$$

Verrà quindi applicato il modello con p differenti per analizzare quale possa essere il miglior valore da utilizzare (in seguito si vedrà alla Figura 11).

3.3 Risultati Ottenuti

Il modello definito in tutto il par. 3.2, è stato utilizzato per analizzare i risultati e fare previsioni *out-of-sample*, cioè previsioni su dati non utilizzati per calcolare i parametri del modello. Questo aspetto risulta molto importante poichè quando si cerca di prevedere un risultato i dati di quella determinata partita, e soprattutto delle partite successive, non sono ancora disponibili. Per questo sono stati creati modelli per ogni giornata di campionato e per ogni squadra di quella determinata giornata.

E' stato analizzato il campionato di Serie A 2021/2022, usando i dati disponibili dal campionato di Serie A del 2018/2019 fino alla partita precedente a quella di cui si è voluto fare la previsione (come se si dovesse fare realmente). Per le previsioni sono state prese in considerazione le giornate 16ª, 21ª e 30ª.

Non sono state scelte giornate ad inizio campionato poichè non si hanno dati abbastanza affidabili, dato che, come spiegato in precedenza, durante il calciomercato estivo le squadre possono cambiare la loro identità e il loro stile di gioco, migliorando o peggiorando le qualità di squadra. Inoltre Salernitana e Venezia sono squadre neopromosse e per questo motivo non si hanno molti dati a disposizione (confermato anche dagli indici U di Theil elevati).

Dopo aver creato il modello e aver trovato il parametro λ per ogni squadra,

viene costruita la matrice dei risultati possibili, da cui poi viene estratto il risultato esatto con più probabilità di successo. In seguito un esempio di come è costruita la matrice appena spiegata (Figura 10):

	0	1	2	3	4	5	6
0	0.046776006	0.073978190	0.0546962751	0.0277171914	0.0105342096	3.202910e-03	8.115329e-04
1	0.075467886	0.099105286	0.0837022322	0.0424158828	0.0161206015	4.901443e-03	1.241897e-03
2	0.055423108	0.084256537	0.0640451627	0.0324547153	0.0123347553	3.750363e-03	9.502431e-04
3	0.028271504	0.042979528	0.0326696414	0.0165552536	0.0062919979	1.913072e-03	4.847220e-04
4	0.010816038	0.016442995	0.0124986661	0.0063336657	0.0024071761	7.318980e-04	1.854437e-04
5	0.003310377	0.005032575	0.0038253656	0.0019384938	0.0007367449	2.240061e-04	5.675725e-05
6	0.000844317	0.001283566	0.0009756656	0.0004944159	0.0001879080	5.713312e-05	1.447603e-05

Figura 10: Esempio di matrice dei risultati possibili. La partita analizzata è Bologna-Fiorentina del 5/12/2021.

Poi, per calcolare le probabilità di vittoria della squadra di casa e fuori casa, vengono sommate rispettivamente le probabilità presenti nelle celle del triangolo superiore e del triangolo inferiore, mentre per la probabilità di pareggio vengono sommate le probabilità presenti nelle celle diagonali.

Alla Figura 11 si mostra la tabella contenente i primi risultati di previsione del modello.

Partita 16esima giornata Anno 2021/22	P = 0.05		p = 0.10		p = 0.15		p = 0.20		Osservato	
	Esito	Risultato	Esito	Risultato	Esito	Risultato	Esito	Risultato	Esito	Risultato
Empoli-Udinese	1 (0.5334)	1-1 (0.1014)	1 (0.5340)	2-1 (0.0979)	1 (0.5346)	2-1 (0.0979)	1 (0.5352)	2-1 (0.0979)	1	3-1
Cagliari-Torino	1 (0.3920)	1-1 (0.1211)	1 (0.3930)	1-1 (0.1147)	1 (0.3940)	1-0 (0.1086)	1 (0.3950)	1-0 (0.1096)	X	1-1
Bologna-Fiorentina	1 (0.3853)	1-1 (0.0936)	1 (0.3838)	1-1 (0.0991)	1 (0.3853)	1-1 (0.0936)	1 (0.3868)	1-1 (0.0881)	2	2-3
Juventus-Genoa	1 (0.6541)	2-0 (0.1026)	1 (0.6539)	2-0 (0.1026)	1 (0.6537)	2-0 (0.1026)	1 (0.6535)	2-0 (0.1026)	1	2-0
Milan-Salernitana	1 (0.8670)	2-0 (0.2465)	1 (0.8311)	2-0 (0.2465)	1 (0.7952)	2-0 (0.2465)	1 (0.7594)	2-0 (0.2465)	1	2-0
Napoli-Atalanta	1 (0.4184)	2-1 (0.0739)	1 (0.4197)	2-1 (0.0739)	1 (0.4209)	2-1 (0.0739)	1 (0.4221)	2-1 (0.0739)	2	2-3
Roma-Inter	2 (0.4595)	1-2 (0.0884)	2 (0.4604)	1-2 (0.0884)	2 (0.4613)	1-2 (0.0884)	2 (0.4622)	1-2 (0.0884)	2	0-3
Sampdoria-Lazio	2 (0.4366)	1-2 (0.0876)	2 (0.4376)	1-2 (0.0876)	2 (0.4385)	1-2 (0.0876)	2 (0.4394)	1-2 (0.0876)	2	1-3
Spezia-Sassuolo	2 (0.4637)	1-1 (0.0955)	2 (0.4645)	1-2 (0.0915)	2 (0.4652)	1-2 (0.0915)	2 (0.4660)	1-2 (0.0915)	X	2-2
Venezia-Hellas Verona	2 (0.4060)	1-1 (0.1161)	2 (0.4065)	1-1 (0.1101)	2 (0.4069)	1-1 (0.1039)	2 (0.4074)	0-1 (0.0980)	2	3-4

Figura 11: Tabella che rappresenta il confronto tra risultati previsti dal modello (più colonne per ogni parametro p) e i risultati osservati nei campionati 2018/19, 2019/2020, 2020/2021, 2021/22.

La tabella è composta da cinque colonne e dieci righe (una riga per ogni partita). Nelle prime quattro colonne ci sono esito ("1" indica la vittoria della squadra di casa, "X" indica pareggio e "2" indica la vittoria della squadra ospite) e risultato esatto per ogni parametro p (0.05, 0.10, 0.15, 0.20). Nella quinta colonna, invece, sono presenti esiti e risultati realmente osservati. Le celle colorate di verde indicano che l'esito o il risultato sono stati previsti in modo esatto.

Analizzando la tabella si nota inoltre che, all'aumentare del parametro p, la probabilità dell'esito previsto aumenta nel caso in cui la probabilità sia relati-

vamente bassa (< 0.6), mentre diminuisce nel caso in cui sia molto alta (≥ 0.6, ad esempio per le partite Juventus-Genoa e Milan-Salernitana). Le colonne con più esiti e risultati esatti azzeccati sono la prima e la seconda, con ben 6/10 esiti e 3/10 risultati esatti. Per questo nelle prossime tabelle verranno analizzate le partite utilizzando il valore 0.10 per il parametro p, il quale sembra un giusto compromesso tra un valore troppo basso (sebbene abbia lo stesso numero di previsioni esatte) e un valore troppo alto (meno previsioni esatte).

Nelle due facciate seguenti sono presenti le due tabelle che si riferiscono all'analisi della 21ª e della 30ª giornata del campionato di Serie A 2021/22.

Partita 21esima giornata Anno 2021/22	p = 0.10		Osservato	
	Esito	Risultato	Esito	Risultato
Cagliari-Bologna	1 (0.3853)	1-1 (0.1056)	1	2-1
Empoli-Sassuolo	2 (0.4770)	1-2 (0.0829)	2	1-5
Genoa-Spezia	1 (0.4866)	1-1 (0.0999)	2	0-1
Hellas Verona-Salernitana	1 (0.7775)	2-0 (0.1986)	2	1-2
Inter-Lazio	1 (0.6057)	2-1 (0.0899)	1	2-1
Napoli-Sampdoria	1 (0.6341)	2-1 (0.0925)	X	2-2
Roma-Juventus	2 (0.4828)	1-2 (0.0877)	2	3-4
Torino-Fiorentina	2 (0.4080)	1-1 (0.1069)	1	4-0
Udinese-Atalanta	2 (0.6763)	0-2 (0.1028)	2	2-6
Venezia-Milan	2 (0.6940)	0-3 (0.0981)	2	0-3

Figura 12: Tabella che rappresenta il confronto tra risultati previsti dal modello (con parametro $p = 0.10$) e i risultati osservati per la 21esima giornata di campionato di Serie A 2021/22.

Anche in questo caso le previsioni sono buone: 7/10 esiti e 2/10 risultati esatti.

Partita 30esima giornata Anno 21/22	p = 0.10		Osservato	
	Esito	Risultato	Esito	Risultato
Bologna-Atalanta	2 (0.6239)	1-2 (0.0881)	2	0-1
Cagliari-Milan	2 (0.5019)	0-1 (0.1214)	2	0-1
Empoli-Hellas Verona	2 (0.4597)	1-2 (0.0891)	X	1-1
Genoa-Torino	2 (0.4878)	0-1 (0.1112)	1	1-0
Inter-Fiorentina	1 (0.5741)	2-1 (0.9091)	X	1-1
Juventus-Salernitana	2 (0.8126)	2-0 (0.1782)	1	2-0
Napoli-Udinese	2 (0.5223)	2-0 (0.1069)	1	2-1
Roma-Lazio	1 (0.4392)	2-1 (0.0831)	1	3-0
Sassuolo-Spezia	1 (0.5551)	2-1 (0.0965)	1	4-1
Venezia-Sampdoria	2 (0.5124)	0-2 (0.0981)	2	0-2

Figura 13: Tabella che rappresenta il confronto tra risultati previsti dal modello (con parametro $p = 0.10$) e i risultati osservati per la 30esima giornata di campionato di Serie A 2021/22.

Anche in quest'ultima analisi le previsioni sono abbastanza precise: 7/10 esiti e 2/10 risultati esatti.

Conclusioni

Il gioco del calcio continua ad essere forse lo sport dove i risultati delle partite rimangono ancora molto imprevedibili. Capita non così raramente infatti che una squadra molto svantaggiata con poche capacità riesca a vincere contro un avversario dalle doti elevate. Il motivo è che in questo sport dominano moltissime variabili (come la forma fisica dei giocatori, gli episodi-partita, i calci di rigore o le giocate individuali che possono decidere le sorti di una gara) che difficilmente si riescono a riassumere con un semplice modello.

L'obiettivo di questa tesi però è proprio questo: cercare di fare previsioni più accurate possibili attraverso un modello parsimonioso che tenga conto solo di alcune di queste variabili.

Considerazioni sulle previsioni

Si nota che le previsioni ottenute a partire dal modello utilizzato per l'applicazione dei dati sono abbastanza buone, con un'accuratezza di circa il 66,67% per quanto riguarda gli esiti ("1", "X", "2") e di circa il 23,33% per i risultati esatti. Ciò significa che il modello è riuscito a fornire un'indicazione su quanti gol una squadra è in grado di fare in una partita, data la squadra avversaria.

Nonostante ciò, il modello dispone di alcuni limiti. Il primo tra tutti è la sottostima di pareggi, problema già affrontato dalla maggior parte degli studiosi citati al Capitolo 2. Infatti si può notare come questo problema sia in parte risolto dai risultati esatti (da notare nella Figura 11, Cagliari-Torino con risultato pronosticato 1-1 e osservato 1-1), ma persiste nel caso delle probabilità degli esiti: si può notare come non venga mai pronosticato l'esito "X". Il motivo è che

la somma delle probabilità presenti nelle celle diagonali sarà sempre minore di almeno uno tra l'esito "1" o "2" perchè le celle sulle diagonale sono solamente 7 (0-0, 1-1, ..., 6-6), mentre le celle del triangolo inferiore e superiore sono 21 ciascuna. Così, anche se la partita dovesse essere molto equilibrata, la probabilità di pareggio è sempre più bassa della probabilità che almeno una delle due squadre vinca. Questo aspetto viene confermato anche dal mercato delle scommesse: ci sarà sempre un esito tra "1" e "2" che avrà una quota più bassa della quota rappresentata dall'esito "X".

Un altro limite del modello è il fatto che, come ribadito in precedenza, il calcio è uno sport dove domina l'imprevedibilità, quindi molti risultati inaspettati spesso non coincidono con le previsioni del modello, il quale, invece, favorisce la maggior parte delle volte la squadra con più capacità.

Considerazioni Finali

L'obiettivo della tesi è stato in parte raggiunto attraverso la creazione di un modello per la previsione delle partite, il quale ha ottenuto una buona percentuale di risultati previsti uguali a quelli osservati. Si ricorda che il modello utilizzato è stato applicato per previsioni *out-of-sample*, quindi è applicabile anche in un contesto reale. Infatti potrebbe essere un dato utile sia per le squadre, sia per le agenzie di scommesse per calcolare le quote attraverso le varie probabilità.

E' necessario riconoscere il progresso portato dall'utilizzo del dato in qualsiasi tipo di azienda, comprese quelle che operano nel settore sportivo, in particolare in quello calcistico attorno al quale ruotano milioni di fatturato.

Da sempre nel calcio ha dominato l'esperienza personale e le opinioni dei più esperti, ma ora con l'introduzione dell'analisi del dato le informazioni possono essere molto più precise e molto più utili sia per la squadra e il suo staff, sia per

chi investe il proprio denaro in questo settore.

I dati e le statistiche, a differenza delle opinioni soggettive, sono informazioni inconfutabili che permettono di analizzare e confrontare squadre e giocatori ad un livello più oggettivo rispetto a prima.

Naturalmente i numeri non bastano, ma c'è il costante bisogno di chi questi dati li sappia interpretare: per un'analisi moderna del calcio, analisi del dato ed esperti del settore hanno il dovere di unire le proprie idee e permettere a questo sport di continuare ad evolversi.

BIBLIOGRAFIA

Maher, M. J. (1982). Modelling association football scores. Statistica Neerlandica, 36(3), 109–118.

Dixon, M. J., & Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. Applied Statistics, 46(2), 265–280.

Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. Journal of the Royal Statistical Society, Series D, 52(3), 381–393.

Koopman, S. J., & Lit, R. (2015). A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League. Journal of the Royal Statistical Society, Series A, 178(1), 167–186.

Giovanni Angelini, Luca De Angelis (2017). PARX model for football match predictions.

A. P. Rotshtein, M. Posner & A. B. Rakityanskaya(2005).Football Predictions Based on a Fuzzy Model with Genetic and Neural Tuning.

SITOGRAFIA

Per i dati utilizzati al Capitolo 3,
https://www.kaggle.com/datasets/giovannicarlozzi/serie-a-matches-dataset

Steve Round(2023). La statistica applicata al calcio.
https://www.studiamo.it/pages/la-statistica-applicata-al-calcio/

https://www.intelligenzaartificialeitalia.net/post/come-vengono-utilizzati-i-dati-nel-calcio: :text=I%20dati%20sono%20usati%20come,fase%20di%20scouting%20pi%C3%B9%20tradizionale.