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

Statistica per l'Economia e l'Impresa

RELAZIONE FINALE

VALUTAZIONE DI PREVISIONI PROBABILISTICHE, MOTIVATA DA UN'APPLICAZIONE AGLI ESITI DELLE PARTITE DI BASKET DELLA NBA

Relatore Prof. Matteo Grigoletto
Dipartimento di Scienze statistiche

Laureando: Camilo Ernesto Carranza Mellana
Matricola N:1220660

Anno Accademico 2021/2022

Indice

Introduzione

Uno dei principali desideri dell'uomo è quello di fare previsioni sul futuro. Le previsioni caratterizzano e riducono, ma in genere non eliminano, l'incertezza. Di conseguenza, le previsioni dovrebbero essere di natura probabilistica, sotto forma di distribuzioni di probabilità degli eventi futuri. Sono molti i diversi campi di applicazione, tra i quali: lo sport, la politica, la medicina ma, nello specifico, negli ultimi due decenni la ricerca di buone previsioni probabilistiche è diventata una forza trainante soprattutto nella meteorologia. Molte di queste previsioni vengono fatte inizialmente prima che si verifichi l'evento in questione e vengono poi continuamente aggiornate man mano che si rendono disponibili nuove informazioni. L'obiettivo di questa tesi è quello di descrivere strumenti semplici e facilmente interpretabili per valutare la taratura e l'abilità delle previsioni probabilistiche continuamente aggiornate e usare questi metodi per valutare previsioni probabilistiche pertinenti il pronostico della partita di basket NBA pubblicato su ESPN. La taratura si riferisce alla coerenza statistica tra le previsioni distributive e le osservazioni. L'abilità si riferisce alla concentrazione delle distribuzioni predittive. In termini di valutazione della taratura del modello di predizione probabilistica, gli strumenti standard sono diagrammi di affidabilità e grafici di calibrazione, in cui le frequenze dei risultati sono tracciati rispetto a intervalli di previsione. Di seguito verrà illustrato come tali curve possono essere estese nel caso di dati continuamente aggiornati e come tale superfici possono essere riassunte per mostrare in sintesi se un dato metodo è ben calibrato.

Al fine di valutare la relativa abilità di un modello di previsione continuamente aggiornato rispetto ad un altro, è stato utilizzato il metodo di Lai, Lordo, e Shen (2011) per costruire intervalli di confidenza per la differenza di perdita media misurata dal punteggio Brier (Brier 1950) tra due modelli ad un tempo dato durante il processo di aggiornamento. Per misurare la significatività statistica cumulativa delle differenze osservate in un grafico, viene sviluppato un nuovo test di significatività per le differenze di abilità aggregate nel tempo sulla base di un nuovo grande campione risultato dalla stima delle curve di differenza di perdita continua.

Con lo scopo di dimostrare questi metodi e valutare le previsioni di ESPN, sono introdotti una serie di modelli di previsione competitivi continuamente aggiornati per la stima dei risultati delle partite di basketball. Alcuni sono progettati per essere più

semplici ai fini della dimostrazione, mentre altri si basano su modelli lineari generalizzati logistici che fanno uso di informazioni sul gioco come ad esempio la differenza di punteggio. Utilizzando i metodi proposti, si osserva che il modello di ESPN è generalmente ben calibrato e presenta significativamente migliore abilità di alcuni modelli primitivi, anche se non dimostra la superiorità su modelli di regressione logistica relativamente semplici basati sul punteggio differenza e relativa forza di squadra presa da sola.

Il resto della tesi è organizzata come segue. Il primo capitolo descrive cosa si intende per previsione probabilistica, dove è utile questo tipo di previsione e come vengono valutate le previsioni probabilistiche, illustrando i concetti di taratura e abilità. Il secondo capitolo introduce i dettagli dei dati di previsione ESPN che verranno considerati e i relativi problemi riscontrati. Il terzo capitolo spiega come sono stati sviluppati alcuni modelli di previsione concorrenti che verranno utilizzati a scopo di confronto. Nel quarto capitolo si discute la valutazione della taratura e i metodi proposti per valutare l'abilità relativa dei modelli per previsioni probabilistiche continuamente aggiornate. Un confronto dettagliato delle previsioni ESPN, così come quelle dei modelli proposti applicati alle partite della NBA, è dato nel quinto capitolo.

Capitolo 1

Previsioni Probabilistiche

Le previsioni probabilistiche e, più in generale, tutte le previsioni sono onnipresenti nella moderna società, tanto che molti individui agiscono e prendono decisioni basandosi su tali previsioni durante la vita di tutti i giorni. Ad esempio, negli Stati Uniti, la probabilità di predire le precipitazioni divenne disponibile pubblicamente a partire dalla fine degli anni '60.

Nel corso del tempo il numero e la portata di previsioni probabilistiche facilmente accessibili al pubblico sono aumentati in modo costante, di pari passo con la necessità da parte dell'uomo di avere delle risposte sicure; ciò ha consentito un rapido ed efficace sviluppo di metodi scientifici precisi e sempre più sofisticati utili a tale scopo.

1.1 Cosa si intende per previsione probabilistica

Una previsione viene definita probabilistica, quando contiene un insieme di probabilità associate a tutti i possibili esiti futuri. Molte di queste previsioni sono fatte inizialmente ben prima che l'evento in questione si verifichi e vengono poi continuamente aggiornate appena nuovi dati diventano disponibili. Questo tipo di previsione è utile in tutti quegli ambiti in cui l'incertezza gioca un ruolo importante, circostanza che si verifica quasi sempre quando si tratta di sistemi complessi; pertanto le previsioni si propongono come un valido strumento di supporto alla decisione.

Le previsioni di probabilità di eventi futuri sono ampiamente utilizzate in diversi campi di applicazione. Gli oncologi prevedono abitualmente la probabilità di sopravvivenza libera da progressione di un paziente oncologico oltre un certo orizzonte temporale (Hari et al. 2009). Gli economisti forniscono previsioni sulla probabilità di una ripresa economica o di una recessione entro la fine di un anno fiscale. Le banche, per valutare i loro requisiti patrimoniali, devono prevedere periodicamente il rischio di insolvenza dei prestiti erogati. Gli ingegneri sono regolarmente chiamati a prevedere la probabilità di sopravvivenza di un sistema o di un'infrastruttura oltre i cinque o dieci anni; ciò include ponti, sistemi fognari e altre strutture. Infine, anche gli avvocati valutano la probabilità di un particolare esito processuale per decidere se andare in giudizio o patteggiare in via extragiudiziale (Fox and Birke 2002). Questo elenco non sarebbe

completo senza menzionare il campo più avanzato nelle previsioni di probabilità quotidiane, ovvero la meteorologia.
Il lavoro del meteorologo australiano W. E. Cooke (1906) è di solito citato come il primo tentativo di trattare il problema dell'incertezza nelle previsioni meteorologiche in modo esplicito, anche se tuttavia è emerso che il problema era stato sollevato più di 200 anni fa da altri individui che avevano apportato importanti innovazioni (Murphy 1998).
Negli ultimi 60 anni sono stati fatti notevoli progressi nella previsione delle probabilità di precipitazione, delle temperature e delle quantità di pioggia in termini di ampiezza e accuratezza. Murphy e Winkler (1984) forniscono una storia illuminante della transizione del *National Weather Service* degli Stati Uniti dalle previsioni non probabilistiche a quelle probabilistiche e dello sviluppo di misure di affidabilità e accuratezza per queste previsioni probabilistiche.

1.2 Valutazioni: taratura e abilità

La verifica delle previsioni è la pratica di determinare la qualità delle previsioni e rappresenta una componente essenziale di qualsiasi sistema di previsione scientifica.
La domanda naturale da porre di fronte a qualsiasi previsione probabilistica, comprese quelle che sono continuamente aggiornate è "sono queste previsioni accurate?" o, ancora, "potrebbero queste previsioni essere migliori?"
Seguendo il lavoro determinante di Murphy e Winkler (1987, 1992) sulla valutazione della qualità delle previsioni in meteorologia, valutando il metodo per produrre le previsioni probabilistiche, queste sono spesso supportate dalle diverse verifiche di misurazione della taratura (calibration) e della abilità (skill). Un modello è considerato ben tarato o calibrato se le sue previsioni sono compatibili con i risultati osservati. In altre parole, un modello che prevede un risultato con una data probabilità è ben calibrato se la frequenza relativa che il risultato si verifichi corrisponde alla previsione di probabilità nel lungo termine. Per comprendere meglio cosa sia la taratura definiamo una funzione fatta da due variabili: denotando la previsione con f e l'osservazione (di un evento, di un valore osservato o la variabile d'interesse) con x, la funzione $p(f, x)$ rappresenta la distribuzione congiunta di f e x. Questa distribuzione contiene

informazioni riguardo la previsione, riguardo il valore osservato, e riguardo la relazione tra previsione e valore osservato.

Sebbene la distribuzione congiunta contenga tutte le informazioni rilevanti alla verifica delle previsioni, le informazioni sono più accessibili quando si fattorizza la distribuzione. Prendendo in considerazione la fattorizzazione detta *calibration-refinement*, includiamo la distribuzione condizionata delle osservazioni data la previsione e la distribuzione marginale delle previsioni:

$$p(f, x) = p(x|f)p(f). \quad (1)$$

La distribuzione condizionale $p(x|f)$ indica la frequenza con cui si sono verificate osservazioni diverse quando è stata fornita una particolare previsione f.

Tenendo presente la nostra attenzione alla verifica delle previsioni, si può affermare che preferiamo le distribuzioni congiunte che assegnano alte frequenze relative alle coppie (f, x) con f uguale o vicino a x e basse frequenze relative alle coppie (f, x) con f non vicino a x.

Per esempio, nel caso in cui x rappresenti la variabile esito di una partita di basket della squadra che gioca in casa, essa assume due valori:

$$x = \begin{cases} 1, in\ caso\ di\ vittoria; \\ 0, in\ caso\ di\ sconfitta. \end{cases}$$

Assumendo che f possa assumere diversi valori tale che $f \in [0, 1]$, l'obiettivo è avere

$$p(x = 1|f) = f\ . \quad (2)$$

Se (2) è soddisfatto per tutti gli f, la previsione o il sistema predittivo è detto perfettamente calibrato.

Nel caso generale in cui esistano più di due valori per x, come nella previsione delle temperature, la distribuzione condizionata $p(x|f)$ consiste in diverse frequenze relative e non può essere rappresentata da una sola di queste frequenze. In questo caso, diciamo che la previsione è perfettamente calibrata se

$$E(x|f) = f\ . \quad (3)$$

dove $E(x|f)$ è il valore atteso di x data la previsione f.

Si ritiene che un modello abbia una maggiore abilità di un concorrente modello se le sue previsioni sono "più nitide" o "più concentrate" rispetto al suo rivale. Ad esempio, un modello di previsione che riporta sempre e solo la probabilità climatologica, come per esempio il tasso di fondo a lungo termine delle giornate piovose a New York (pari al 33.1%), non avrebbe mai alcuna variazione nelle previsioni e non sarebbe in grado di distinguere tra giorni con precipitazioni e giorni senza precipitazioni (Gneiting, Balabdaoui e Raftery 2007, Gneiting e Katzfuss 2014).

Pertanto, la taratura si riferisce alla coerenza statistica tra le previsioni distributive e le osservazioni ed è una proprietà congiunta delle previsioni e degli eventi che si verificano. L'abilità si riferisce alla concentrazione delle distribuzioni predittive ed è una proprietà esclusivamente delle previsioni.

Capitolo 2

Dati utilizzati e metodologia

2.1 Dati utilizzati

L'NBA è un campionato professionistico di pallacanestro, che viene spesso definito una delle "Quattro grandi" leghe sportive del Nord America.

L'esempio su cui si concentrerà questa tesi è la previsione dei risultati delle partite di basket della National Basketball Association (NBA). Siti web come espn.com, la pagina web principale della rete sportiva multinazionale statunitense ESPN, pubblicano e aggiornano in tempo reale, previsioni probabilistiche sulla vittoria della squadra di casa per ogni partita di NBA. Sebbene il metodo con cui ESPN produce queste previsioni sia in gran parte riservato, le previsioni probabilistiche iniziali sulla vittoria della squadra di casa sono costruite sulla base delle informazioni disponibili prima della partita, ad esempio, l'abituale vantaggio del campo di casa nell'NBA, la forza relativa della squadra, gli infortuni dei giocatori, ecc. Dopo l'inizio e l'avanzamento della partita queste previsioni vengono aggiornate con nuove informazioni come il punteggio, il tempo di gioco rimanente, il possesso palla, i falli e gli infortuni dei giocatori in partita. I dati specifici che verranno presi in considerazione sono le registrazioni play-by-play[1] e le previsioni probabilistiche in tempo reale delle partite della stagione regolare NBA scaricate da espn.com/nba (ESPN 2020).

Dal 2004, ad eccezione della stagione di lockout[2] del 2011 e delle stagioni influenzate dalla COVID-19 nel 2020 e nel 2021, la NBA è composta da 30 squadre, ognuna delle quali disputa un calendario di 82 partite nella stagione regolare. A partire dalla stagione NBA 2017-2018, ESPN Analytics ha iniziato a fornire previsioni probabilistiche in tempo reale sulla vittoria della squadra di casa per ogni partita NBA giocata; un esempio delle previsioni di una partita è mostrato nella *Figura 2.1*. I dati disponibili

[1] Registrazioni dettagliate di un evento sportivo.

[2] Periodo in cui, a causa di motivi finanziari, le squadre non potevano muoversi sul mercato, non potevano avere rapporti con i propri giocatori e soprattutto non si poteva giocare né gare ufficiali né amichevoli.

da ESPN sono molto ricchi, e comprendono informazioni in tempo reale su dettagli quali sostituzioni, falli e possesso palla. In questa sede verranno considerati solo un sottoinsieme di questi dati che include le previsioni probabilistiche in tempo reale fornite da ESPN, nonché l'evoluzione del punteggio durante la partita, per le stagioni 2017-2018 e 2018-2019. Questi dati vengono aggiornati ogni volta che si verifica un "evento" nella partita, che include principalmente cambi di punteggio, falli e cambi di possesso. Una partita tipica presenta tra i 460 e i 480 eventi.

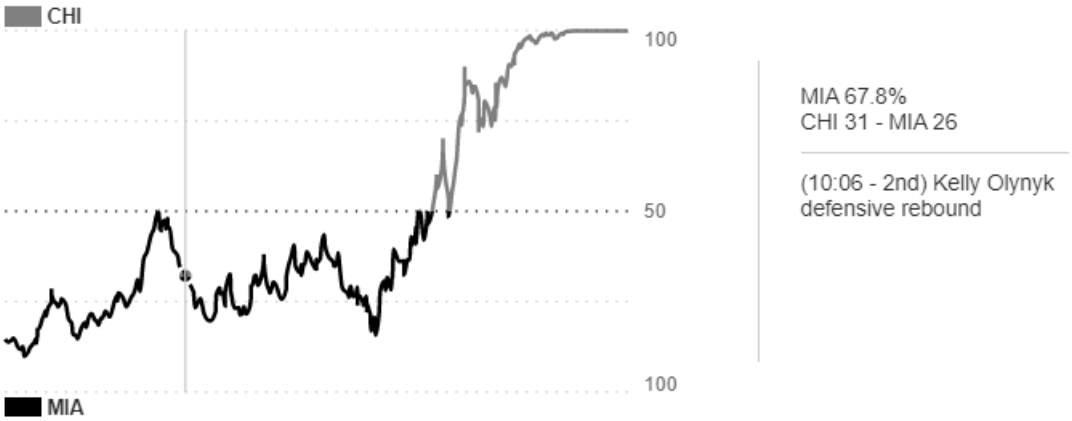

Figura 2.1 *Previsioni probabilistiche congiunte in tempo reale, in funzione del tempo di gioco, di una partita del 31 gennaio 2019 in cui i Chicago Bulls (CHI) hanno ospitato i Miami Heat (MIA). Il grafico cambia in base agli eventi successi durante la partita e il pallino nero, riportato sul grafico, rappresenta uno specifico evento della partita in cui i MIA avevano probabilità di vittoria pari al 67.8% nonostante fossero in svantaggio, come riportato a destra del grafico. (ESPN 2022)*

2.1.1 Problemi riscontrati

Sono stati esclusi una piccola parte di questi dati dall' analisi a causa di due problemi. Molto spesso si verificano più eventi nello stesso istante in una partita. Uno degli esempi principali che contribuiscono a questo fenomeno è la sostituzione di più giocatori nello stesso momento. Sebbene questi eventi vengano registrati nello stesso momento, essi si verificano nel set di dati in una sequenza ordinata. Le probabilità pubblicate da ESPN durante un evento di questo tipo sono tipicamente condizionate da questo ordine. Pertanto, l'analisi si limita alla media delle previsioni in un tale scenario per produrre una previsione probabilistica in quell'istante, sebbene ci siano altri modi per gestire questa situazione, come l'utilizzo della prima o dell'ultima previsione probabilistica tra gli eventi registrati.

Il secondo problema è dovuto alle partite che vanno ai tempi supplementari. Se due squadre sono in parità al termine dei 48 minuti di gioco regolamentare, le squadre giocheranno un periodo supplementare di 5 minuti ai tempi supplementari. Per queste partite, viene eliminato il periodo di *overtime* dall'analisi, e si considerano solo le previsioni probabilistiche fino alla fine della partita, in modo che siano comparabili con quelle che derivano da partite che non sono andate ai tempi supplementari. Le partite ai tempi supplementari rappresentano meno del 10% delle partite totali. Inoltre, un piccolo numero di dati sono stati scartati a causa di difetti evidenti o di eccessivi valori mancanti.

2.2 Metodologia

I dati rimanenti analizzati sono riassunti nella *Tabella 2.1*, in ogni stagione ci sono più di 1100 partite con un totale di oltre 350.000 registrazioni play-by-play disponibili. Di seguito verranno utilizzati i dati della stagione 2017-2018 come dati di addestramento per i modelli, per poi produrre e valutare le previsioni per la stagione 2018-2019.
Lasciando che N denoti il numero totale di partite con previsioni che si desidera valutare, quindi N = 1213 quando vengono considerate le previsioni per il 2018 e il 2019, i dati possono essere indicati come $\hat{p}_i^{ESPN}(t),\ 1 \leq i \leq N, t \in [0,1]$ che rappresentano le previsioni probabilistiche della vittoria della squadra di casa nell'i-esima partita al tempo di gioco t. Si assume che il parametro del tempo di gioco t sia normalizzato per essere compreso tra zero e uno, in modo da rappresentare la partita completa.
Queste previsioni sono disponibili solo quando si verificano gli eventi ma, dato che gli eventi sono molto densi nel corso della partita, queste previsioni vengono completate come una funzione costante a tratti impostata sull'ultima previsione di probabilità tra un evento e l'altro per produrre curve di previsione di probabilità complete sull'intervallo [0, 1], il che le rende più comparabili tra una partita e l'altra.
Si considerano anche i dati $H_i(t)$ e $A_i(t),\ 1 \leq i \leq N, t \in [0,1]$, che indicano, rispettivamente, il punteggio della squadra di casa e quello della squadra ospite nell'i-esima partita, all'istante t della partita. Durante l'analisi di seguito, si utilizza spesso la

differenza di punteggio (score difference) $ScD_i(t) = H_i(t) - A_i(t), 1 \leq i \leq N, t \in [0,1]$.

L'obiettivo dei metodi che saranno sviluppati di seguito è quello di valutare la qualità delle previsioni $\hat{p}_i^{ESPN}(t)$, $1 \leq i \leq N, t \in [0,1]$. Per sviluppare un certo numero di modelli di riferimento che vengono utilizzati a scopo di confronto di seguito verrà utilizzata la seguente notazione. Y_i indica la variabile casuale indicatrice della vittoria della squadra di casa nella partita, in modo che $Y_i = 1$ se la squadra di casa vince la partita i-esima, e $Y_i = 0$ se la squadra di casa perde la partita i-esima. L' interesse sta nel prevedere o stimare la probabilità $p_i(t)$ che la squadra di casa vinca, date le informazioni fino al momento t della partita, in modo che

$$p_i(t) = P(Y_i = 1 | tutte\ le\ informazioni\ al\ momento\ t\ nella\ partita\ i) \qquad (4)$$

$\hat{p}_i^{ESPN}(t)$ è in linea di principio una stima di $p_i(t)$.

Season	Mode	Games	Events	Max. events	Min. events	Avg. # of events
17–18	Raw	1158	530,032	606	234	457.7133
	Selected	1137	517,983	572	240	455.5699
	Processed	1137	354,749	375	173	312.0343
18–19	Raw	1229	583,443	700	124	474.7299
	Selected	1213	572,546	598	366	472.0082
	Processed	1213	396,991	385	241	327.2803

Tabella 2.1 *Riassunto dei dati ottenuti da ESPN dalla stagione regolare 2017-2018 alla stagione regolare 2018-2019. RAW rappresenta il numero totale di partite per il quale ESPN è in grado di fornire previsioni probabilistiche. SELECTED si riferisce alle partite che non contengono errori o valori mancanti. PROCESSED rappresenta i dati dopo aver calcolato la media di eventi multipli rilevati allo stesso tempo in una partita. (ESPN 2020)*

Capitolo 3

Previsione dei risultati nel basketball

3.1 Modelli di confronto

Per valutare la qualità di queste previsioni, verranno considerati una serie di modelli di riferimento in competizione tra loro, da quelli più semplici a quelli più realistici. I modelli di riferimento considerati sono per lo più modelli lineari generalizzati (GLM) per dati a risposta binaria, che sono spesso definiti modelli di regressione logistica. Utilizzando $g(\cdot)$ per indicare la funzione di legame GLM viene utilizzato il legame $logit$ (McCullagh e Nelder 1989, cap. 4). Tutti i modelli sono della forma

$$g\big(p_i(t)\big) = \beta_0(t) + \sum_{j=1}^{D} \beta_j(t) X_j(t), \tag{5}$$

dove i termini $X_j(t)$ denotano le covariate utilizzate per prevedere $p_i(t)$. Tutti i GLM sono stati adattati in modo puntuale su una griglia discreta di 721 punti temporali equidistanti t, che corrisponde a una risoluzione di 4 secondi nel gioco.

Per ogni modello vengono utilizzati i dati della stagione 2017 e 2018 come dati di addestramento, e poi si producono *rolling forecast*[3] sui dati della stagione 2018 e 2019, per confrontarle con le previsioni di ESPN.

Le covariate più complicate che sono state considerate per costruire questi modelli di riferimento sono la differenza di punteggio $ScD_i(t)$ e una misura della forza relativa delle squadre RS_i. Viene utilizzata anche la covariata *status di leader,* che codifica se la squadra di casa sta vincendo o meno al tempo t. Definita questa variabile come $LS_i(t)$, essa assume il valore 1 se $ScD_i(t) > 0$, -1 se $ScD_i(t) < 0$ e 0 se $ScD_i(t) = 0$.

Per quanto riguarda la definizione di RS_i, ci sono diversi modi per valutare la forza relativa delle squadre, tra cui l'utilizzo del sistema di Elo rating (Elo 1978), che nacque per valutare la forza relativa di un giocatore di scacchi, successivamente utilizzato

[3] Previsioni continuamente aggiornate non appena vengono aggiornati i dati.

anche per valutare la forza delle squadre di basket (Silver e Fischer-Baum 2015; Silver, Boice, e Paine 2019).

In questa sede invece, verrà utilizzata come proxy della forza relativa della squadra $RS_i = \hat{p}_i^{ESPN}(0)$, la probabilità di vittoria della squadra di casa prima della partita, come previsto da ESPN. Sono state considerate una serie di metriche alternative per definire l'RS_i, ma in generale, i risultati e le conclusioni delle analisi seguenti non cambiano in modo significativo e quindi saranno utilizzate queste quantità per evitare di introdurre nuove metriche e dati. Le descrizioni dei modelli di riferimento di base che sono stati considerati per prevedere $p_i(t)$ sono raccolte nella *Tabella 3.1* e sono elencate in ordine dal più semplice al più realistico. Si noti che i GLM con termini di intercetta sono in grado di modellare implicitamente il vantaggio della squadra di casa, che si riferisce al fenomeno per cui nella NBA la squadra di casa tende a vincere una percentuale maggiore di partite rispetto a quella della squadra in trasferta. Quindi da un modello basato sulla differenza di punteggio ma senza un termine d'intercetta, come ScDnoInt nella *Tabella 3.1*, ci si aspetta che possa essere poco calibrato, almeno all'inizio della partita. Come accennato, ogni modello GLM viene adattato considerando il tempo t di gioco. Questo permette che ad esempio, l'effetto, come determinato dai modelli, della forza relativa della squadra, del vantaggio della squadra di casa e della differenza di punteggio si possa evolvere nel corso della partita.

Model	Covariates	Description
CF	None	A constant forecast for all intra-game times of 1/2, that is, $\hat{p}_i(t) = 0.5$ for all $t \in [0, 1]$. Abbreviation is for **C**oin **F**lip.
HomeWP	None	A constant forecast for all intra-game times set to the observed home team win rate in the prior 10 NBA regular season, spanning 2008–2017. This amounts to forecasting $\hat{p}_i(t) = 0.593$ for all $t \in [0, 1]$.
PgRS	RS_i	GLM for the home team win probability in terms of **Pre**game **R**elative team **S**trength, as measured by $\hat{p}^{\text{ESPN}}(0)$.
LS	$LS_i(t)$	GLM for the home team win probability in terms of the **L**eading **S**tatus for the home team.
ScDnoInt	$ScD_i(t)$	GLM for the home team win probability in terms of the **Sc**ore **D**ifference between the home and away teams. **No Int**ercept term is included in the model.
ScD	$ScD_i(t)$	GLM for the home team win probability in terms of the **Sc**ore **D**ifference between the home and away teams.
PgRSLS	$RS_i, LS_i(t)$	GLM for the home team win probability in terms of **Pre**game **R**elative team **S**trength as measured by $\hat{p}^{\text{ESPN}}(0)$, and the **L**eading **S**tatus.
PgRSScD	$RS_i, ScD_i(t)$	GLM for the home team win probability in terms of **Pre**game **R**elative team **S**trength as measured by $\hat{p}^{\text{ESPN}}(0)$, and the **Sc**ore **D**ifference.

Tabella 3.1. *Descrizioni dei modelli di confronto utilizzati per prevedere* $\hat{p}_i(t)$*, elencati da quelli più semplici a quelli più realistici.* (Yeh et al., 2021)

Nei modelli di regressione non lineari, come in questo caso la regressione logistica, non è possibile calcolare il valore dell'R quadro. Esistono però degli indici che hanno un interpretazione simile come lo pseudo R quadro o di R quadro di McFadden (McFadden 1973).

La misura di R quadro di McFadden è definita come:

$$R^2_{McFadden} = 1 - \frac{\log(L_c)}{\log(L_{Null})} \quad (6)$$

dove L_c denota il valore della massima verosimiglianza del modello previsto e L_{Null}[4] denota il valore corrispondente per il modello nullo, con solo un'intercetta e nessuna covariata. Sebbene lo pseudo R^2 sia compreso tra 0 e 1, secondo McFadden, i valori da 0,2-0,4 indicano un eccellente adattamento al modello, pertanto i valori ottenuti vanno interpretati con attenzione, poiché questo indice non si comporta come l'R^2semplice utilizzato per la regressione lineare.

I grafici diagnostici dello pseudo R^2 rispetto l'importanza relativa delle variabili nel corso della partita del modello più realistico, che utilizza sia la differenza di punteggio e la forza relativa della squadra denominato PgRSScD, sono visualizzati nella *Figura 3.1*. Da questa figura è chiaro che le previsioni dei modelli migliorano con il progredire della partita, evidentemente perché alla fine, la variabile della differenza di punteggio determina il vincitore. L'importanza relativa della differenza di punteggio rispetto alla forza della squadra cambia inversamente con il progredire della partita; la forza relativa della squadra è la variabile esplicativa più importante all'inizio della partita, ma diventa meno importante nel proseguo della partita, man mano che la differenza di punteggio diventa più informativa.

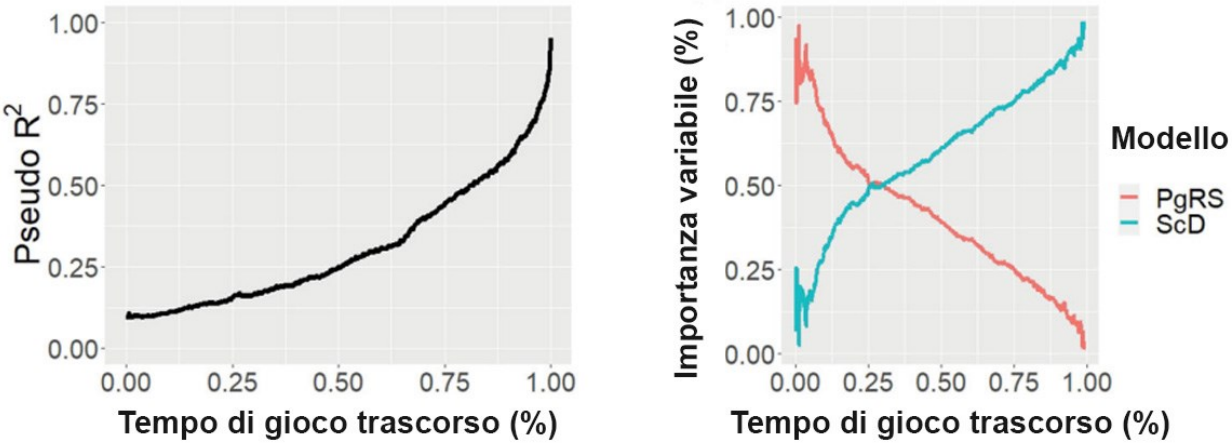

Figura 3.1 *A sinistra: lo pseudo R2 del modello di regressione logistica PgRSScD, che utilizza le covariate forza relativa e differenza di punteggio, in funzione del tempo di gioco. A destra: l'importanza di ciascuna covariata per i modelli PgRS e ScD, in base a quanto contribuisce allo pseudo R2.* (Yeh et al., 2021)

[4] Questo indice funziona adeguatamente per la regressione logistica in quanto la log-verosimiglianza è sempre negativa ed è pari a 0 in caso di adattamento perfetto.

Capitolo 4

Valutazione della qualità delle previsioni

4.1 Taratura nelle previsioni continuamente aggiornate

In questo paragrafo ci si focalizzerà sul compito di valutare la taratura per un dato insieme di previsioni aggiornate in modo continuo $\hat{p}_i(t)$ con risultati realizzati $Y(i), i = 1, \dots N$. Come accennato nell'introduzione, tradizionalmente, quando si valutano tali previsioni, spesso si considerano quelli che vengono chiamati diagrammi di calibrazione (*calibration plots*) o diagrammi di affidabilità (*reliability diagrams*). Un diagramma di calibrazione è un diagramma di previsioni probabilistiche raggruppate in base alla frequenza condizionata degli eventi associata alle rispettive previsioni in un determinato intervallo. Poiché per le previsioni ben calibrate la frequenza degli eventi corrisponde alla probabilità prevista, la taratura può essere misurata confrontando questi punti con una linea di riferimento diagonale di 45 gradi. Grandi scostamenti da questa linea indicano quindi una cattiva calibrazione (cfr. Dawid 1986; Murphy e Winkler 1992; Ranjan e Gneiting 2010).

Un metodo chiaro per verificare la calibrazione delle previsioni continuamente aggiornate è quello di produrre un grafico di calibrazione per ogni $t \in [0, 1]$ basato sulle coppie $(\hat{p}_i(t), Y_i)$. Sebbene questo sia in sostanza ciò che viene proposto, ci sono due sfide principali da affrontare.

(i) Tradizionalmente, quando si producono grafici di calibrazione, le probabilità previste vengono intervallate in intervalli fissi, di solito in decili. Ad esempio, spesso la frequenza degli eventi corrispondente a tutte le probabilità previste tra [0; 0,1] sono confrontate con 0,05, analogamente per [0.1, 0.2] a 0.15 e così via. Con le previsioni aggiornate continuamente, come per le previsioni ESPN, è tipico che le previsioni fluttuino, in modo che per certi punti temporali t le previsioni si raggruppino intorno ad alcuni valori fissi, e sono quindi ben lontani dall'essere uniformemente distribuite in tali intervalli fissi. Nel caso delle previsioni ESPN, verso la fine della partita la maggior parte delle previsioni sono raggruppate intorno a 0 e 1. Gli intervalli fissi hanno spesso il problema che le previsioni al loro interno non sono uniformemente distribuite.

(ii) Dopo aver costruito i grafici di calibrazione per ogni t, è necessario esaminare un gran numero di tali grafici per individuare se un metodo sembra essere ben calibrato, o per diagnosticare se c'è un sottoinsieme di tempi t in cui il metodo è più o meno tarato rispetto ad altri. Un semplice riepilogo dei molti grafici di calibrazione prodotti sarebbe utile.

Per risolvere il problema (i), si propone l'utilizzo di intervalli adattabili nella costruzione dei grafici di calibrazione. In particolare, per ogni t, si suppone di voler costruire M intervalli per le previsioni $\hat{p}_i(t)$. Calcolando le previsioni ordinate $\hat{p}_{(i)}(t)$, $i = 1,\dots,N$, si possono raggruppare in M intervalli, in modo che $\hat{p}_{(i)}(t)$ si trovi nell'intervallo j se $([N/M](j-1)+1) \leq i < [N/M]j$, dove per $[N/M]$ si intende la parte intera di N/M. L'insieme di $\hat{p}_i(t)$ nel j-esimo intervallo è indicato come Bin_j. In termini più semplici, le previsioni in un determinato momento t sono raggruppate in M intervalli in base al loro rango. Come punto di riferimento o riepilogo delle previsioni nel j-esimo intervallo, utilizziamo $\widetilde{p_j}(t)$ = Mediana $(\hat{p}_{(i)}(t)$, $([N/M](j-1)+1) \leq i < [N/M]j)$.

Si costruisce quindi un grafico di calibrazione al tempo t confrontando $\widetilde{p_j}(t)$ con $\overline{Y}_j(t)$ = Media(Y_i, tale che $\hat{p}_i(t), \in Bin_j$). Lasciando che j indichi il numero di previsioni in Bin_j, un intervallo di confidenza del 95% per la media degli eventi in Bin_j è costruito come:

$$\frac{n_j\bar{Y}_j(t)+k^2/2}{n_j+k^2} \pm \frac{k\, n_j^{\frac{1}{2}}}{n_j+k^2}\left(\bar{Y}_j(t)\left(1-\bar{Y}_j(t)\right)+k^2/(4n_j)\right)^{1/2} \qquad (7)$$

dove $\kappa = z_{\alpha/(2M)}$, e z_β indica il quantile β della distribuzione normale standard.

α è tipicamente considerato pari al 5%, in modo da poter calcolare intervalli di confidenza del 95%.

A questo punto sorge il seguente problema: quando vengono verificate più ipotesi, aumenta la possibilità di osservare un evento raro e, quindi, aumenta la probabilità di rifiutare erroneamente un'ipotesi nulla (tasso di errore familiare).

Perciò, nell'equazione (7) è stata applicata una correzione di Bonferroni al livello di significatività in base al numero di intervalli utilizzati M. La correzione compensa l'aumento di tale probabilità verificando ogni singola ipotesi a un livello di significatività di α/m, dove α è il livello di significatività statistica e m è il numero di ipotesi. Tuttavia, nel caso in cui siano presenti molti test, la correzione di Bonferroni

ha il costo di ridurre la potenza statistica. Quindi, prima di procedere, si nota che un'interessante alternativa sarebbe stata quella di applicare la correzione di Šidák (Šidák, 1967) che, in genere, è considerata essere più potente.

L' intervallo (7) è spesso indicato come l'intervallo di Wilson (Wilson 1927).

La scelta di usare questo intervallo, piuttosto che il semplice intervallo di Wald, è dovuta al fatto che, come è stato dimostrato da Brown, Cai e DasGupta (2001), la copertura di quest' ultimo, può essere significativamente minore anche quando si hanno dimensioni campionarie piuttosto elevate, e questo accade in modo imprevedibile e piuttosto casuale, quindi si necessitava di un intervallo di confidenza più preciso.

Inoltre, è ampiamente riconosciuto che l'effettiva probabilità di copertura dell'intervallo standard è scarsa per p, probabilità di successo, prossima a 0 o 1. In una serie di articoli interessanti, per quanto non recenti, è stato sottolineato che le proprietà di copertura dell'intervallo standard possono essere errate anche se p non è vicino ai confini; si vedano, ad esempio, Vollset (1993), Santner (1998).

Perciò, sono stati notati miglioramenti significativi utilizzando l'intervallo di Wilson in questa impostazione, a causa delle frequenze degli eventi che si avvicinano a 0 e a 1 verso la fine del gioco.

Inoltre, per evitare che gli intervalli contengano prevalentemente previsioni di probabilità 0 o 1, sono state scartate tutte le previsioni, per produrre un grafico di calibrazione in un determinato t, e gli eventi corrispondenti, se $\hat{p}_i(t) < 0{,}005$ o $\hat{p}_i(t) > 0{,}995$. La calibrazione per quelle previsioni verrà analizzata separatamente.

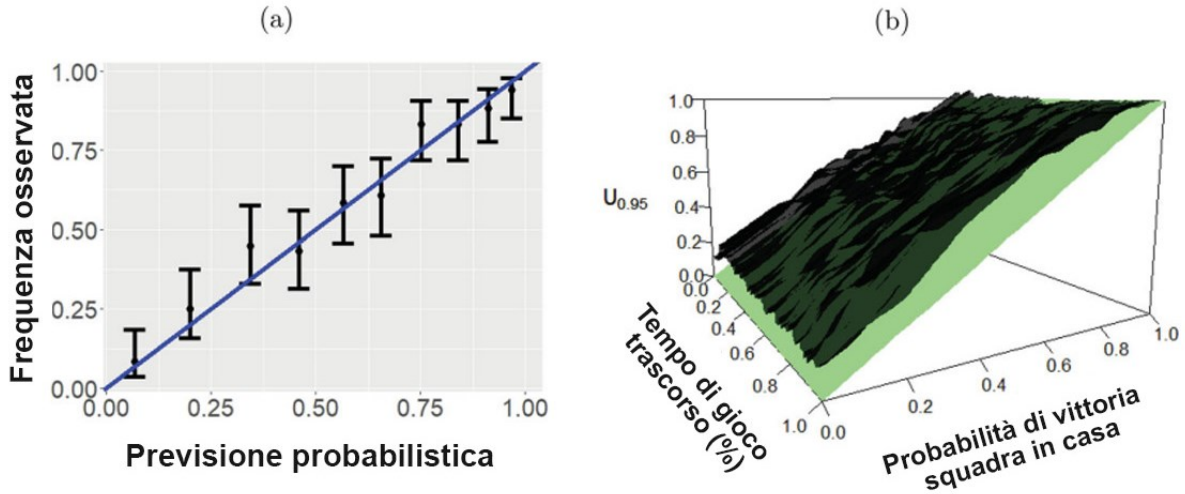

Figura 4.1 *A sinistra: un grafico di calibrazione delle previsioni ESPN al punto temporale* $t = 0{,}5$. *A destra: una superficie di calibrazione superiore con il piano di riferimento f(t, p) = p per le previsioni ESPN aggiornate in modo continuo, ottenute interpolando i limiti superiori di ciascun intervallo di confidenza dei grafici di calibrazione su* $t \in [0,1]$. (Yeh et al., 2021)

Si può quindi realizzare un grafico di calibrazione tracciando gli intervalli di riferimento $\widetilde{p_j}(t)$ contro gli intervalli di confidenza di cui sopra e confrontando questi intervalli con la linea di riferimento $y = x$. Il metodo è considerato ben calibrato al livello di significatività dato, se la linea di riferimento passa generalmente attraverso ogni intervallo. Un esempio di questo è mostrato sulla base delle previsioni ESPN per $t = 0{,}5$ utilizzando $M = 10$ intervalli nel pannello sinistro della *Figura 4.1(a)* al 95% del livello di confidenza. Interpolando linearmente i limiti superiori di questi intervalli, considerano sia diverse proporzioni di riferimento $\widetilde{p_j}(t)$ che diverse t, possiamo costruire una "superficie di calibrazione superiore", $U_{1-\alpha}(t,p)$.

La superficie di calibrazione superiore al 95% $U_{0.95}(t,p)$ è visualizzata nel pannello di destra della *Figura 4.1(b)* per le previsioni ESPN della stagione 2018-2019. Una superficie inferiore $L_{1-\alpha}(t,p)$ può essere costruita in modo simile interpolando linearmente i limiti inferiori. Un metodo di previsione continuamente aggiornato può quindi essere considerato ben calibrato a un determinato livello di significatività se il piano di riferimento $f(t,p) = p,\ per\ t,p \in [0,1]$, è contenuto tra entrambe le superfici.

Sebbene questi diagrammi di superficie siano informativi, può essere difficile dedurre rapidamente, sulla base di questi diagrammi, se un dato metodo sembri essere calibrato. Per produrre un riepilogo più facilmente interpretabile di tali diagrammi di superficie di calibrazione, si consideri invece un grafico per ogni t della distanza minima di p tra il piano di riferimento $f(t,p) = p$ e le superfici di calibrazione superiore e inferiore. In particolare, consideriamo i grafici delle funzioni $U_{1-\alpha}^{min}(t) = min_{1\le j\le M}\, U_{1-\alpha}(t,\widetilde{p_j}(t)) - \widetilde{p_j}(t)$ e $L_{1-\alpha}^{max}(t) = max_{1\le j\le M} L_{1-\alpha}(t,\widetilde{p_j}(t)) - \widetilde{p_j}(t)$ rispetto t.

Se le superfici di confidenza superiore e inferiore contengono il piano di riferimento $f(t,p) = p$, allora $U_{1-\alpha}^{min}(t)$ dovrebbe essere sempre superiore a zero e $L_{1-\alpha}^{max}(t)$ dovrebbe essere sempre inferiore a zero. I punti rispetto a t in cui questo non vale, possono essere utilizzati per identificare i tempi per i quali un determinato metodo non appare ben calibrato. Notiamo che in questo caso, a causa dell'alto grado di fluttuazione delle previsioni probabilistiche aggiornate in modo continuo per la previsione del basket, è stato ritenuto utile, per aiutare l'interpretazione di questi grafici, lisciarli rispetto a t utilizzando una semplice media mobile sul 5% dei tempi di gioco.

Questi grafici riassuntivi calcolati in base alle previsioni dell'ESPN nonché sui modelli di riferimento HomeWP, ScDnoInt e PgRSScD sono mostrati nella *Figura 4*.2. Da questi si nota che il modello semplice HomeWP, che predice che la squadra di casa vincerà ogni partita in tutti i momenti con il tasso di vittorie storico di 10 anni precedente delle squadre di casa nell'NBA, è ben calibrato, come previsto. Tracciati simili per il metodo CF[5], che prevede semplicemente che la squadra di casa vincerà con una probabilità del 50%, mostrano che questo metodo non è ben calibrato. Considerando questo grafico per il metodo ScDnoInt (Pannello (b) nella Figura 4.2), si vede che le previsioni sono scarsamente tarate all'inizio del gioco, ma la calibrazione migliora verso la fine della partita. Questo è atteso, poiché il modello logit corrispondente è considerato privo di un termine di intercetta e quindi il modello non è in grado di catturare il vantaggio della squadra di casa che dovrebbe costringere le probabilità di previsione a favorire la squadra di casa all'inizio della partita. Sia le previsioni di ESPN sia quelle del modello PgRSScD, che incorpora la forza della squadra, oltre alla differenza di punteggio, hanno dimostrato in generale una buona calibrazione per tutti i tempi di gioco.

Infine il tasso di vittoria empirico della squadra di casa, quando la previsione di probabilità in un momento della partita aveva superato il valore di 0,995 o era inferiore a 0,005 per le previsioni di ESPN e per le previsioni basate sui modelli PgRSScD e ScDnoInt, è simile alle probabilità previste, suggerendo che le previsioni per ciascuno di questi metodi sono ragionevolmente ben calibrate a questi livelli estremi di probabilità di previsione (*Tabella 4.1*).

	ESPN		PgRSScD		ScDnoInt	
$\hat{p}_i(t)$ for some t	> 0.995	< 0.005	> 0.995	< 0.005	> 0.995	< 0.005
Total Games	555	343	591	358	589	374
Home team wins	553	1	588	0	587	0
Proportion	0.9964	0.0029	0.9949	0.0000	0.9966	0.0000

Tabella 4.1 *Il numero di partite in cui una previsione di probabilità ha superato lo 0,995 o è stata inferiore allo 0,005 ad un certo punto per ciascuna delle previsioni ESPN, PgRSScD con legame logit e ScDnoInt con legame logit, nonché il numero e la proporzione tra queste partite in cui la squadra di casa ha vinto* (Yeh et al., 2021).

[5] Il modello, non riportato nel grafico, viene chiamato Coin Flip perché le probabilità di vittoria per una squadra di basket sono le stesse di ottenere testa o croce lanciando una moneta.

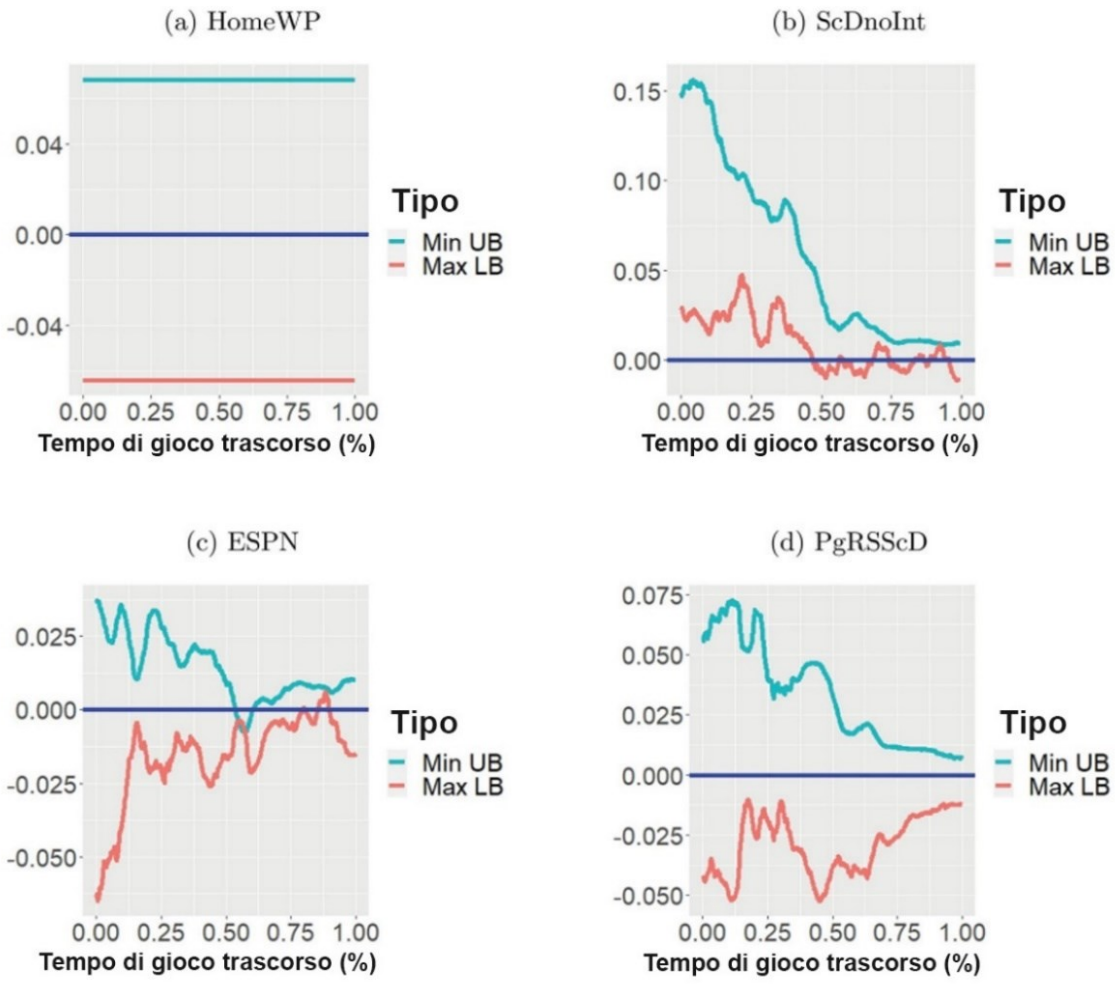

Figura 4.2 *Grafici di* $U_{1-\alpha}^{min}(t)$ *e* $L_{1-\alpha}^{max}(t)$ *rispetto a* t *sulla base di M = 10 intervalli: Metodi (a) probabilità storica di vittoria della squadra di casa (HomeWP); (b) GLM che usa la differenza di punteggio senza intercetta; (c) previsioni ESPN; (d) GLM che utilizza la forza relativa e la differenza di punteggio prima della partita (PgRSScD).* (Yeh et al., 2021)

4.2 Abilità nelle previsioni continuamente aggiornate

In questa sezione, si considerano i metodi per produrre intervalli di confidenza puntuale per la differenza di abilità, misurata dal punteggio di Brier, tra due metodi in competizione come funzione del parametro *intra-game* t, che può essere utilizzato per identificare i momenti t in cui un metodo sembra avere prestazioni significativamente migliori di un *benchmark*. Sarà introdotto anche un metodo per valutare la significatività statistica delle differenze di abilità aggregate in tutti i tempi *intra-game*. Come descritto nel paragrafo §1.2, l'abilità di un metodo di previsione probabilistico si riferisce in genere alla sua acutezza rispetto a un metodo concorrente o di riferimento. Formalmente questo può essere misurato definendo una funzione di perdita, o regola di punteggio, utilizzata per misurare l'accuratezza di una determinata previsione probabilistica basata sugli eventi realizzati. La funzione di perdita più frequentemente utilizzata è $L(a,b) = (a-b)^2$ e definisce il punteggio di Brier (Brier 1950). Di seguito verrà utilizzata questa funzione di perdita, ma i seguenti risultati si generalizzano a qualsiasi funzione di perdita che abbia un equivalente lineare, il che significa che

(1) $L'(x,b)$ è lineare in x;

(2) $L'(x,b) - L(x,a)$ non dipende da x.

Seguendo il lavoro di Lai, Gross e Shen (2011), idealmente, qualsiasi metodo di previsione di $p_i(t)$ dovrebbe essere tale che $L(p_i(t), \hat{p}_i(t))$ sia piccolo, e inoltre, quando si fa la media di tutte le previsioni, dovrebbe minimizzare:

$$L_N(t) = \frac{1}{N}\sum_{i=1}^{N} L(p_i(t), \hat{p}_i(t)). \quad (8)$$

Poiché le probabilità reali sottostanti $p_i(t)$ non sono osservabili, una stima sensata di $L_N(t)$ si ottiene sostituendo queste probabilità con le loro stime puntuali basate sulle realizzazioni Y_i per produrre

$$\hat{L}_N(t) = \frac{1}{N}\sum_{i=1}^{N} L\big(Y_i, \hat{p}_i(t)\big). \quad (9)$$

La quantità $\hat{L}_N(t)$ cattura l'abilità del metodo di previsione in un determinato punto temporale t, come descritto in precedenza, in quanto a un a un determinato metodo di

previsione vengono attribuite perdite generalmente più basse, o punteggi più alti, se $\hat{p}_i(t)$ è più vicino a Y_i, dove quest'ultimo assume i valori 0 e 1.

Si supponga di voler confrontare due metodi, denominati A e B, per produrre previsioni probabilistiche continuamente aggiornate. Denotando tali previsioni con $\hat{p}_i^A(t)$ e $\hat{p}_i^B(t)$, le confrontiamo in base ai corrispondenti eventi realizzati Y_i, $1 \leq i \leq N$. Questo può essere fatto confrontando le loro perdite medie definite nell'Equazione (9). In particolare, si consideri la funzione di t:

$$\hat{\Delta}_N(t) = \frac{1}{N}\sum_{i=1}^{N}\left[L\left(Y_i, \hat{p}_i^A(t)\right) - L\left(Y_i, \hat{p}_i^B(t)\right)\right], \tag{10}$$

che può essere visto come una stima della differenza di perdita reale

$$\Delta_N(t) = \frac{1}{N}\sum_{i=1}^{N}\left[L\left(p_i(t), \hat{p}_i^A(t)\right) - L\left(p_i(t), \hat{p}_i^B(t)\right)\right]. \tag{11}$$

I valori maggiori di zero di $\hat{\Delta}_N(t)$ favoriscono il metodo B al dato t, mentre i valori negativi mostrano il favore per il metodo A. Per misurare la significatività statistica di qualsiasi deviazione di $\hat{\Delta}_N(t)$ da zero si può considerare $\hat{\Delta}_N(t)$ come uno stimatore di $\Delta_N(t)$. Costruendo intervalli di confidenza adeguati per $\Delta_N(t)$ basati su $\hat{\Delta}_N(t)$, può essere valutato se le deviazioni osservate suggeriscono la superiorità di un modello rispetto a un altro e poi costruire semplici riepiloghi grafici che illustrano l'abilità di un modello rispetto a un altro in funzione di t. Per costruire tali intervalli di confidenza, si definisce innanzitutto la varianza di $\hat{\Delta}_N(t)$ come

$$s_N^2(t) = \frac{1}{N}\sum_{i=1}^{N}\delta_i^2(t)p_i(t)\big(1 - p_i(t)\big), \tag{12}$$

dove $\delta_i(t) = \left[L\left(1, \hat{p}_i^A(t)\right) - L\left(0, \hat{p}_i^A(t)\right)\right] - \left[L\left(1, \hat{p}_i^B(t)\right) - L\left(0, \hat{p}_i^B(t)\right)\right]$.

Il seguente risultato è dimostrato in Lai, Gross e Shen (2011) e viene affermato per ogni $t \in [0, 1]$.

Teorema 2, Lai, Gross e Shen (2011): Si supponga che per ogni $t \in [0, 1]$ che $s_N^2(t)$ converga in probabilità ad una costante positiva quando $N \to \infty$, e che le variabili $A_i(t) = L\left(Y_i, \hat{p}_i^A(t)\right) - L\left(p_i(t), \hat{p}_i^A(t)\right)$ e $B_i(t) = L\left(Y_i, \hat{p}_i^B(t)\right) - L\left(p_i(t), \hat{p}_i^B(t)\right)$, siano differenze di martingala. Allora, per ogni t,

$$\frac{\hat{\Delta}_N(t)-\Delta_N(t)}{s_N(t)} \xrightarrow{D} N(0,1), \quad (13)$$

dove $\xrightarrow{D}$ denota la convergenza in distribuzione e $N(0,1)$ indica la distribuzione normale standard.

Le due condizioni principali del teorema di cui sopra sono che (1) $s_N^2(t)$, la varianza di $\Delta_N(t)$, dovrebbe, per N grandi, comportarsi come una costante positiva, e (2) che le differenze di perdita di previsione si comportino come una differenza di martingala (*MDS*).

La prima condizione può essere considerata come una condizione di non degenerazione: questo risultato vale solo se le previsioni dei due metodi da confrontare non coincidono completamente.

Non è valido per due metodi che producono previsioni equivalenti o quasi equivalenti. Questo è, in generale, un presupposto ragionevole se le previsioni da confrontare provengono da modelli completamente diversi, o se una o entrambe le serie di previsioni da confrontare provengono da modelli sconosciuti, come nel caso dei dati di ESPN, poiché in questo caso è improbabile che producano previsioni che coincidono. Questo presupposto è in discussione quando si confrontano le previsioni di modelli annidati, per esempio, confrontando due modelli GLM la cui unica differenza è l'inclusione o l'esclusione di una covariata (Clark e McCracken, 2015).

Per quanto riguarda la seconda condizione, questa è quasi sempre soddisfatta quando si utilizza una funzione di perdita con un equivalente lineare e costruendo dei veri e propri metodi di previsione che devono basarsi sulle informazioni disponibili (passate), piuttosto che sulle informazioni del futuro sconosciuto, come discusso in Lai, Gross e Shen, pag. 2361 (2011).

I risultati di cui sopra suggeriscono di costruire un intervallo di confidenza del $100(1-\alpha)\%$ per $\Delta_N(t)$ come

$$\hat{\Delta}_N(t) \pm z_{1-\alpha/2}\frac{s_N(t)}{\sqrt{N}}. \quad (14)$$

Si noti che poiché $p_i(t)(1-p_i(t))$ nella definizione di $s_N^2(t)$ è non osservato, possiamo sostituirlo con il limite superiore 1/4 in entrambe le equazioni (13) e (14) per ottenere un intervallo di confidenza conservativo che può essere usato per valutare la relativa abilità di un modello comparato ad un altro. I punti t, associati a intervalli di

confidenza al livello $1 - \alpha$ per $\hat{\Delta}_N(t)$ che non contengono lo zero, indicando un miglioramento significativo al livello α della perdita media di un metodo rispetto ad un altro. Esempi di grafici si trovano nella Figura 5.2. Si noti, ancora una volta, che a causa dell'elevato grado di fluttuazione delle previsioni probabilistiche aggiornate continuamente per la previsione della pallacanestro, è stato utile, per facilitare l'interpretazione di questi grafici, smussarli rispetto a t utilizzando un semplice media mobile sul 5% dei tempi di gioco.

4.2.1 Test per la misurazione dell'abilità di aggregati rispetto a **t**

Sebbene gli intervalli di confidenza di cui sopra possano essere utilizzati per valutare se due metodi presentano un'abilità simile o significativamente diversa in un determinato momento t, spesso è anche interessante valutare se due modelli aggiornati continuamente hanno un potere predittivo approssimativamente uguale quando le discrepanze tra loro sono aggregate in $t \in [0, 1]$. Per esempio, potrebbe essere che un metodo mostri un'abilità simile ma leggermente migliore in ogni momento di gioco che, se considerato in modo aggregato, suggerisce la superiorità di un modello rispetto a un altro. Al contrario, è anche possibile che un metodo mostri prestazioni apparentemente migliori in un singolo momento di gioco t_0, anche se aggregato rispetto $t \in [0, 1]$ questo miglioramento può apparire piuttosto insignificante.
Per essere più precisi, si formula l'ipotesi nulla di uguale potere predittivo aggregato su $t \in [0,1]$ di due metodi come:

$$H_0: ||\Delta_N(t)||^2 = 0, \tag{15}$$

dove $||\cdot||^2$ è la norma quadratica standard L^2 di una funzione tale che $\left||f|\right|^2 = \int_0^1 f^2(t)dt$.

L'ipotesi H_0 prevede quindi che i due metodi da confrontare mostrino approssimativamente in media un'abilità uguale. Una misura della discrepanza globale nel tempo tra i due metodi di previsione può essere ottenuta considerando $||Z_N||^2$.

$$Z_N(t) = \sqrt{N}\hat{\Delta}_N(t). \tag{16}$$

Al fine di determinare le proprietà asintotiche di Z_N per poter così valutare i livelli di significatività per i test di H_0 basati su $||Z_N||^2$, si utilizza il seguente risultato che viene affermato in modo rigoroso e dimostrato nel materiale supplementare dell'articolo Yeh et al. (2021)

Teorema 1. Sotto H_0 e condizioni analoghe a quelle di Lai, Gross e Shen (2011, teorema 2), esiste una sequenza infinita di costanti $\{\lambda_i, i \geq 1\}$ che soddisfano $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$, e $\sum_{i=1}^{\infty} \lambda_i < \infty$
In modo tale che

$$||Z_N||^2 \xrightarrow{D} \sum_{i=1}^{\infty} \lambda_i \chi_i^2(1), \tag{17}$$

dove $\chi_i^2(1)$, $i = 1, 2, \ldots$ sono variabili casuali indipendenti e identicamente distribuite χ^2 con un grado di libertà. Inoltre, le costanti $\{\lambda_i, i \geq 1\}$ possono essere stimate in modo conservativo dagli autovalori della funzione

$$\hat{C}_{cons}(t,s) = \frac{1}{N}\sum_{i=1}^{N}[\hat{p}_i^A(t) - \hat{p}_i^B(t)][\hat{p}_i^A(s) - \hat{p}_i^B(s)]. \tag{18}$$

Vale a dire con $\{\hat{\lambda}_i$, i = 1, ... N$\}$ definiti in modo tale che $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_N \geq 0$ e che soddisfano la condizione che esistono funzioni $\widehat{\varphi}_i(t)$, $i = 1, \ldots N, t \in [0,1]$ con $||\widehat{\varphi}_i||^2 = 1$, tali che

$$\hat{\lambda}_i \widehat{\varphi}_i(t) = \int_0^1 \hat{C}_{cons}(t,s)\ \widehat{\varphi}_i(s)ds, \tag{19}$$

allora per ogni fissato $j \geq 1, P(\hat{\lambda}_j \geq \lambda_j) \to 1$ quando $N \to \infty$.
Questo risultato suggerisce un modo semplice per condurre un test approssimativo e conservativo dell'ipotesi H_0.

Fase 1: Valutare $||Z_N||^2$.

Fase 2: stimare $\hat{C}_{cons}$ e gli autovalori che soddisfano la (19).

Fase 3: stimare la distribuzione della variabile casuale $Q_D = \sum_{i=1}^{D} \hat{\lambda}_i \chi_i^2(1)$ dove D è un numero grande (di seguito prendiamo $D = 10$, scelta ritenuta generalmente adeguata). Questo può essere fatto facilmente utilizzando la simulazione Monte Carlo.

Fase 4: Calcolare un p-value approssimativo e conservativo del test di H_0 come $p = P(Q_D \geq ||Z_N||^2)$.

Questo p-value, combinato con gli intervalli di confidenza nell'Equazione (14), consente una valutazione dettagliata, sia in particolari momenti di gioco t che in tutti i $t \in [0,1]$, dell'abilità relativa dei metodi concorrenti continuamente aggiornati. E'stato studiato questo test e gli intervalli di confidenza puntuali introdotti nel paragrafo §4.2 in un ampio studio di simulazione di dati sintetici di partite di pallacanestro ed è stato scoperto che entrambi i metodi si sono comportati generalmente come previsto e hanno mostrato un forte potere di differenziare i modelli con diversi livelli di abilità. Una descrizione di queste simulazioni e i relativi risultati sono disponibili nel capitolo successivo.

Capitolo 5

Applicazioni alle partite della NBA

5.1 Simulazione di una partita di Basketball

In questo paragrafo si spiega come sono stati generati dati che assomigliano a quelli delle partite dell'NBA descritti nel secondo capitolo.

Ciò comporta la generazione di quantità casuali che fungono da differenza di punteggio e forza relativa iniziale delle squadre in campo. Siano $\{W_i(t), t \in [0,1]\}, i = 1, \ldots, N$ moti browniani[6] standard indipendenti, dove N è il numero totale di partite. Per rappresentare la forza relativa della squadra, consideriamo $RSi \sim a \times Unif(-1,1) + c$, dove $Unif(-1,1)$ denota una variabile casuale uniforme su $[-1,1]$ e $a, c \in R$ sono costanti che verranno utilizzati per ponderare i dati simulati. Con questa definizione, la differenza di punteggio è modellata come $ScD_i(t) = tRS_i + W_i(t)$, le variabili indicatrici (analogamente alla vittoria della squadra di casa) vengono definite come $Y_i = 1\ se\ ScD_i(1) > 0$, e $Y_i = 0$ altrimenti. In altre parole, la differenza di punteggio è modellata come un moto browniano con *drift* determinata da RS_i; una forza relativa positiva facilita la vittoria della squadra di casa, mentre una forza relativa negativa ha l'effetto opposto. Le costanti a e c che definiscono RS_i sono state scelte in modo che la probabilità di vittoria della squadra di casa sia approssimativamente del 59%, in modo da corrispondere alla percentuale di vittorie storiche della squadra di casa negli ultimi 10 anni nella NBA. Modelli simili e semplici di moti browniani per i punteggi NBA sono stati ampiamente studiati; si veda ad esempio Chen e Fan (2018).

Per ogni impostazione di $N = 100, 250\ e\ 500$ e per il confronto dei modelli considerati, sono stati simulati i dati in modo indipendente per 1000 volte ed è stato applicato il test di H_0 descritto per il confronto delle previsioni nel paragrafo §4.2.1. I risultati in termini di tassi empirici di rifiuto di H_0 sono riassunti per ogni confronto effettuato nella *Tabella 5.1*.

[6] Una sequenza di variabili casuali B(t) è un moto browniano se $B(0) = 0$, e per tutti i t, s tali che $s < t$, $B(t) - B(s)$ è normalmente distribuito con varianza ts e la distribuzione di $B(t) - B(s)$ è indipendente da B(r) per $r \leq s$.

Competing Models	N=100			N=250			N=500		
	10%	5%	1%	10%	5%	1%	10%	5%	1%
PgRSScD v.s. PgRS	1.000	0.998	0.972	1.000	1.000	1.000	1.000	1.000	1.000
PgRSScD v.s. ScD	0.510	0.377	0.176	0.831	0.745	0.509	0.995	0.982	0.907
PgRSScD v.s. LS	0.795	0.704	0.415	0.990	0.967	0.898	1.000	1.000	1.000
PgRSScD v.s. PgRSLS	0.820	0.648	0.253	1.000	0.999	0.958	1.000	1.000	1.000

Tabella 5.1 *Tassi di rifiuto empirici con livelli nominali del 10%, 5% e 1% per il test* $H_0 : ||\Delta_N||^2 = 0$ *in 1000 simulazioni indipendenti.* (Yeh et al., 2021)

Per quanto riguarda il confronto tra i modelli GLM proposti, nella *Figura 5.1* sono riportati grafici rappresentativi che confrontano il modello PgRSScD correttamente specificato con diversi concorrenti più semplici. Confrontando PgRSScD con modelli che non aggiustano le loro previsioni con il progredire della partita, è evidente che il test sviluppato è sempre stato in grado di distinguere PgRSScD come più abile. L'abilità relativa di PgRSScD migliora nel corso della partita rispetto ai modelli che non incorporano informazioni sul punteggio, mentre per i modelli che incorporano quest'ultima informazione il miglioramento relativo di PgRSScD diminuisce con il progredire della partita. Nella *Tabella 5.1* vediamo che una volta che la dimensione del campione raggiunge i 500 il test proposto per H_0 è generalmente in grado di distinguere tra i modelli correttamente specificati e quelli semplici con una potenza empirica prossima a uno.

Nel complesso, il test proposto sembra funzionare bene e come previsto in molti esempi controllati, tende ad essere conservativo. Sebbene il test sia certamente potente per differenziare i modelli con scarse prestazioni in quanto dotati di bassa abilità rispetto ai concorrenti, potrebbe faticare a differenziare i modelli competitivi senza un campione di grandi dimensioni ($N \geq 500$ negli esempi considerati). Questo, insieme agli strumenti grafici proposti, permette di identificare facilmente l'abilità relativa di due modelli di previsione probabilistici in competizione e continuamente aggiornati.

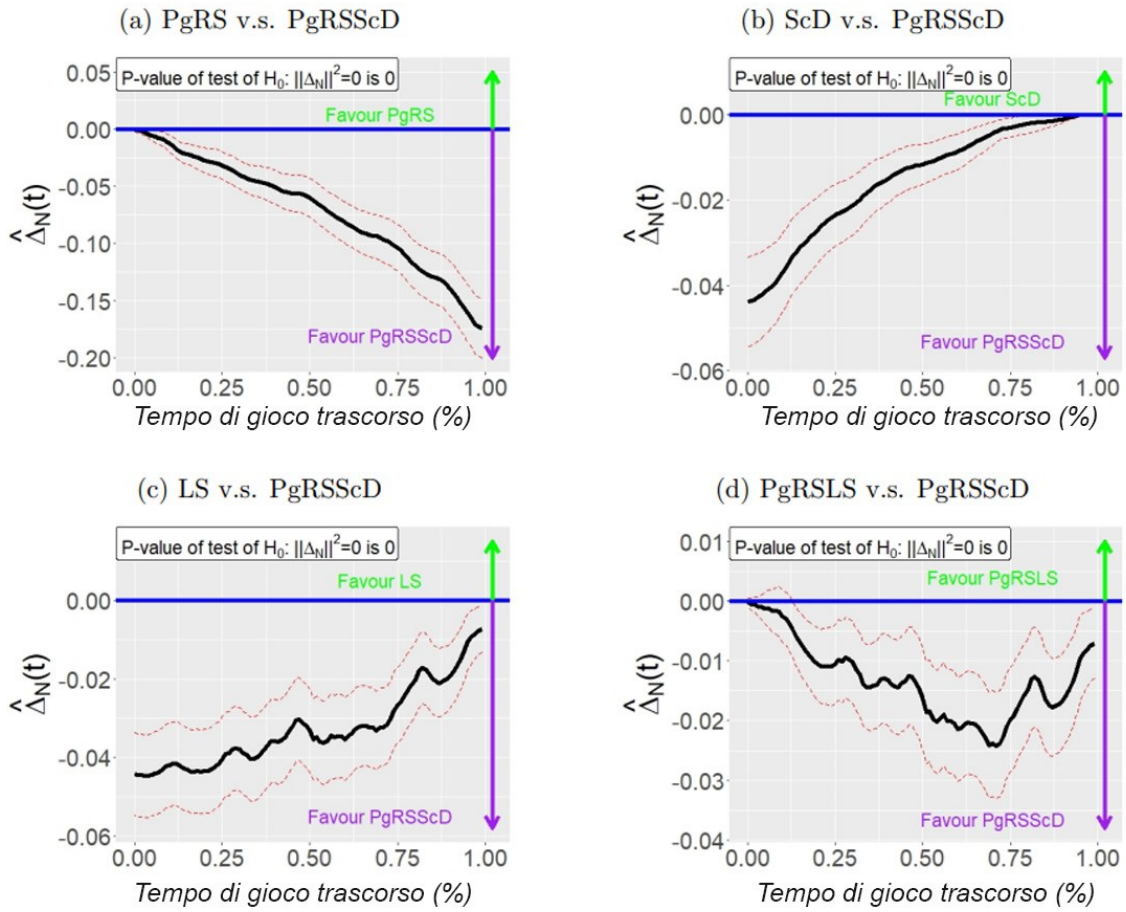

Figura 5.1 *Grafici rappresentativi di* $\hat{\Delta}_{1000}(t)$ *derivati da dati simulati con intervalli di confidenza al 95% e valori di p approssimati per i test di* H_0 *per (a) PgRSScD rispetto a PgRS (b) PgRSScD rispetto a ScD (c) PgRSScD rispetto a LS (d) PgRSScD rispetto a PgRSLS* (Yeh et al., 2021)

5.2 Valutazione dell'abilità delle previsioni di ESPN

In questo paragrafo, vengono applicati i metodi descritti nel paragrafo §4.1 per valutare l'abilità delle previsioni probabilistiche continuamente aggiornate di ESPN. La *Figura 5.2* mostra i grafici di $\hat{\Delta}_{1213}(t)$ con intervalli di confidenza conservativi al 95%, nonché i p-value approssimativi del test di H_0 per i confronti delle previsioni di ESPN con i modelli semplici PgRS, ScD, LS e PgRSLS. I grafici indicano anche i punti specifici della partita in cui le previsioni di ESPN presentano un'abilità superiore rispetto a modelli *benchmark*. Per i modelli che utilizzano la forza relativa della squadra, codificata dalla probabilità di vittoria iniziale in casa di ESPN come covariata, l'abilità relativa è simile alle previsioni di ESPN all'inizio della partita, e allo stesso

modo quelli che fanno uso della differenza di punteggio migliorano rispetto alle previsioni di ESPN verso la fine della partita. Ad esempio ScD, il modello basato solo sulla differenza di punteggio, è fortemente superato dalle previsioni di ESPN nelle prime fasi della partita, ma le loro previsioni hanno un'abilità indistinguibile verso la fine della partita.

Infine, sono confrontate le previsioni di ESPN con quelle del modello PgRSScD, più esperto, che utilizza un legame logit. In termini assoluti, l'abilità stimata misurata dal punteggio Brier ha generalmente favorito quest'ultimo modello, che utilizza come variabile esplicativa la differenza di punteggio e la forza relativa, ad eccezione degli ultimi momenti della partita. Tuttavia, da questa analisi si vede che la differenza non appare statisticamente significativa al livello del 5% in qualsiasi momento della partita, in base alle stime conservative degli intervalli di confidenza, né è significativa quando la differenza viene aggregata tra i punti temporali.

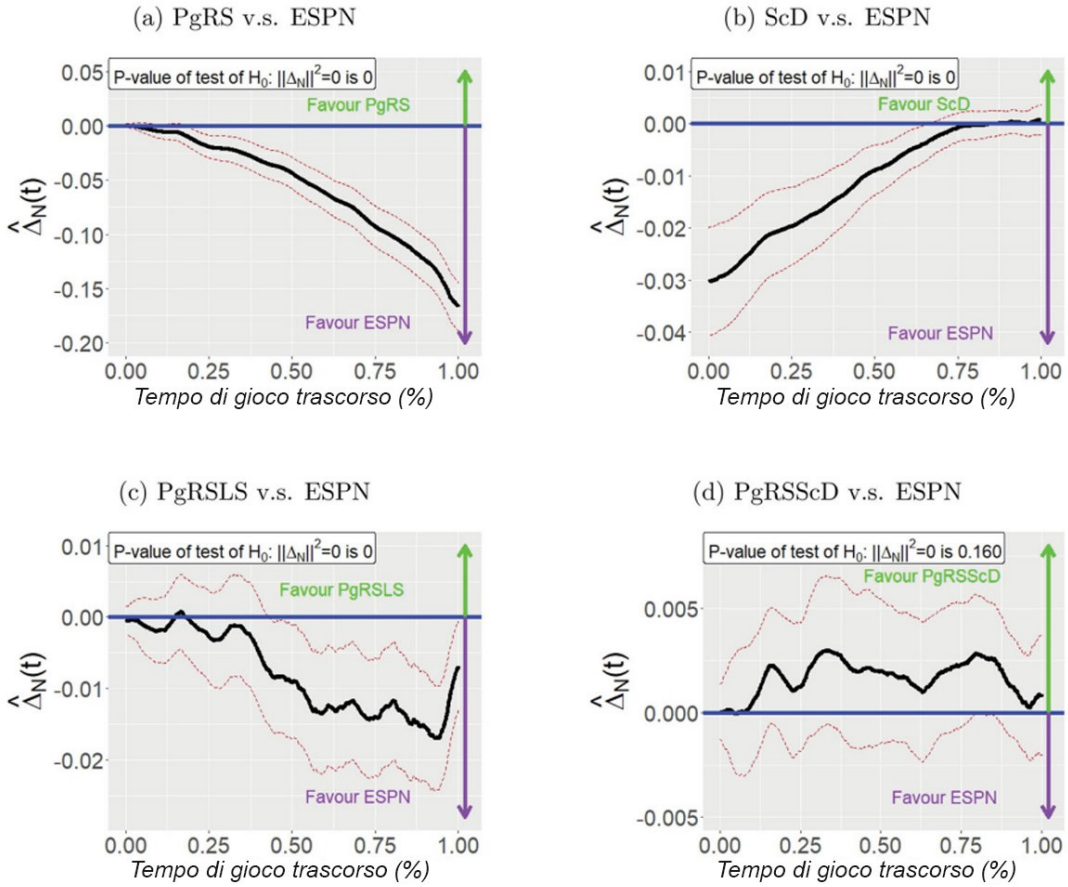

Figura 5.2 *Grafici di* $\hat{\Delta}_{1213}(t)$ *basati sulle previsioni della stagione 2018-2019 con intervalli di confidenza al 95% e p valori approssimativi per i test di* H_0 *per (a) ESPN contro PgRS; (b) ESPN contro ScD; (c) ESPN contro PgRSLS; (d) ESPN contro PgRSScD.* (Yeh et al., 2021)

Conclusioni

L'obiettivo di questo lavoro è stato la valutazione delle previsioni delle partite di basket dell'NBA; sono stati sviluppati strumenti grafici e test statistici per valutare la calibrazione e l'abilità relativa delle previsioni probabilistiche continuamente aggiornate. Questi sono stati elaborati attraverso uno studio di simulazione di partite di basketball sintetiche, e sono stati applicati per valutare e confrontare le previsioni pubblicate su ESPN e una serie di modelli concorrenti. In termini di taratura, le previsioni di ESPN, così come le previsioni prodotte da semplici modelli di regressione logistica che utilizzano la differenza di punteggio durante la partita e/o la forza relativa delle squadre prima della partita come variabili ausiliari, appaiono ragionevolmente ben tarati. In termini di abilità, le previsioni di ESPN hanno mostrato un'abilità significativamente più elevata rispetto ai modelli più semplici ma non hanno dimostrato una superiorità rispetto ai semplici modelli di regressione logistica basati sulla differenza di punteggio e sulla forza relativa delle squadre. E' interessante il fatto che il sofisticato modello proprietario di ESPN, che apparentemente fa uso di informazioni più sfumate sullo stato della partita e di modelli più sofisticati, non ha superato in modo significativo un semplice modello di regressione logistica. Si potrebbe trarre la conclusione che qualsiasi informazione aggiuntiva utilizzata dal modello di ESPN per produrre queste previsioni non è chiaramente vantaggiosa ai fini della previsione, tranne forse nei momenti finali della partita.

In conclusione, si osserva che lo studio dell' abilità dei vari modelli, riportato in questa tesi di laurea, poteva essere condotto in maniera diversa, rendendo gli intervalli di confidenza definiti nell' equazione (14) più stretti e meno conservativi, utilizzando informazioni ausiliarie per migliorare l'approssimazione della sostituzione di $p_i(t)(1 - p_i(t))$ con il limite superiore 1/4.

Potrebbe essere interessante adattare gli intervalli e i test corrispondenti, a questo caso, per esempio, utilizzando metodi simili a quelli descritti in Clark e McCracken (2015).

Bibliografia

Brier, G. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78.

Brown, L. C. (2001). Interval Estimation for a Binomial Proportion. *Statistical Science*, 16, 101-117.

Chen, T. a. (2018). A functional data approach to model score difference process in professional basketball games. . *Journal of Applied Statistics*, 45, 112-127.

Clark, T. E. (2015). Nasted Forecast Model Comparison: A New Approach to Testing Equal Accurancy. *Journal of Econometrics*.

Dawid, A. P. (1986). *Probability Forecasting* (Vol.7). In *Encyclopedia of Statistical Sciences* (pp. 210-218). New York: S.Kotz, N.L. Johnson and C.B. Read.

Elo, A. (1978). *The Rating of Chessplayers, Past and Present.* New York: Arco Pub.

ESPN. (2020). *ESPN Internet Ventures, National Basketball Association Teams, Scores, Stats, News, Standings, Rumors*. Retrieved from www.espn.com/nba/.

Fox, C. a. (2002). Forecasting trial outcomes: Lawyers assign higher probability to possibilities that are described in greater detail. *Law Human Behavior*, 26, 159-173.

Gneiting, T. a. (2014). Probabilistic Forecasting. *The Annual Review of Statistics and its Applications*, 1, 125-151.

Gneiting, T. B. (2007). Journal of the Royal Statistical Society, Series B. *Probabilistic Forecasts, Calibration and Sharpness*, 69, 24-268.

Hari, P. N.-J.-O. (2009). is the international staging system superiror to the Durie-Salmon stating system? A comparison in multiple myeloma patients undergoing autologous trasplant. *leukemia*, 23, 1528-1534.

Lai, T. G. (2011). Evaluating Probability Forecasts. *The Annals of Statistics*, 39, 2356-2382.

McCullagh, P. a. (1989). *Generalized Linear Models*, 2nd ed. . New York: Chapman and Hall/CRC.

McFadden, D. (1973). *Conditional Logit Analysis of Quantitative Choice Behaviour. In Frontiers in Econometrics* (pp. 105-142). New York: Academic Press.

Murphy, & A.H. and Winkler, R. (1992). Diagnostic Verification of Probability Forecasts. *International Journal of Forecasting*, 7, 435-455.

Murphy, A. (1998). The Early History of Probability Forecasts: Some Extensions and Clarifications. *Weather and Forecasting*, 13, 5-15.

Murphy, A. a. (1987). A General Framework for Forecast Verification. *Monthly Weather Review*, 115, 1330-1338.

Ranjan, R. a. (2010). Combinig Probability Forecast. *Journal of the Royal Statistical Society, Series B*, 72, 71-91.

Šidàk, Z. (1967). Rectangular confidance region for the means of multivariate normal distribution. *Journal of the American Statistical Association*, 62, 626-633.

Silver, N. a.-B. (2015). *How we calculate NBA ELO Ratings*. Retrieved from https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/

Silver, N. B. (2019). *How our NBA Predictions Work*. Retrieved from https://fivethirtyeight.com/methodology/how-our-nba-predictions-work/

T.J., S. (1998). A note on teaching binomial confidence intervals. *Teaching Statistics*, 20-23.

Vollset, S. (1993). Confidance intervals for a binomial pro-portion. *Statistics in Medicine*, 809-824.

Wilson, E. B. (1927). Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Associations*, 22, 209-212.

Yeh, C. R. (2021). Evaluating Real-Time Probabilistic Forecasts With Application to National Basketball Association Outcome Prediction. *The American Statistician*.

Yeh, C. R. (2021). Supplementary Material for Evaluating real-time probabilistic forecasts with application to National Basketball Assiociation outcome prediction. *The American Statistician*.