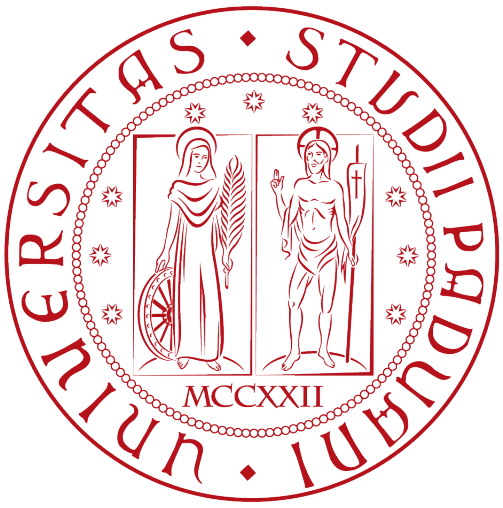
Università degli Studi di Padova

Dipartimento di Matematica "Tullio Levi-Civita"

Corso di Laurea Magistrale in Informatica



Modellazione di un modello Bradley-Terry per l'individuazione delle variabili significative per l'esito di una partita di calcio nella Seria A italiana

Tesi di laurea magistrale

Relatore

Prof. Annamaria Guolo

Laureando

Federico Perin

Anno Accademico 2022-2023

Abstract

Utilizzo del modello Bradley-Terry per individuare quali variabili possono influenzare l'esito di una partita di calcio della Serie A italiana TO DO.

"If something's important enough, you should try. Even if the probable outcome is failure."

— Elon Musk

Ringraziamenti

Innanzitutto, vorrei esprimere la mia gratitudine al Prof.Annamaria Guolo, relatrice della mia tesi, per l'aiuto ed il sostegno fornitomi durante tutto il lavoro.

Desidero ringraziare con affetto i miei genitori per il sostegno, per il grande aiuto che mi hanno dato e per essermi stati vicini in ogni momento durante gli anni di studio.

Voglio inoltre ringraziare i miei amici per questi tre bellissimi anni trascorsi assieme e per avermi sempre sostenuto anche nei momenti più difficili.

Padova, Febbraio 2023 Federico Perin

Indice

Elenco delle figure

Elenco delle tabelle

1 | Introduzione

MEMO: Spiegazione teorica/matematica del modello per la comparazione a coppie, cosa andrò a fare, esposizione struttura della tesi(capitoli) TO DO

1.1

1.2

1.2.1

2 | Serie A 2021/2022 dataset

Nel seguente capitolo verrà descritto in dettaglio la raccolta dati effettuata per costruire il dataset riguardante le partite di calcio della Serie A italiana della stagione 2021/2022 e di come tale dataset è strutturato descrivendone le variabili e i dati al suo interno, utilizzati per l'analisi descritta precedentemente.

2.1 Serie A 2021/2022

L'analisi che è stata effettuata ha preso in considerazione le partite della Serie A italiana della stagione 2021/2022. La Serie A è un torneo che comprende 20 squadre sparse per tutta l'Italia, alcune anche della stessa città ad esempio, Milan e Inter sono due squadre di Milano. Tale torneo è organizzato con una struttura Double-Round-Robin, dove ogni squadra affronta due volte le altre 19 avversarie del torneo. Vi è quindi una partita di andata e una di ritorno che in base al sorteggio della creazione del calendario delle partite decide quale delle due partite sarà giocata in casa oppure fuori casa (in casa dell'avversario). Tale torneo nella stagione 2021/2022 è iniziato il 22 Agosto con Inter - Genoa e si è concluso il 22 Maggio con le partite Salernitana - Udinese e Venezia - Cagliari, per un totale 380 partite giocate suddivise in 38 turni dove ogni turno è composto da 10 partite.

2.1.1 Ranking

Le squadre di calcio sono classificate in base all'ordine dei punti che hanno totalizzato al termine della stagione. In un torneo calcistico, per ogni partita vinta la squadra vincente guadagna 3 punti, per ogni pareggio le due squadre avversarie guadagnano entrambe un punto, mentre per ogni sconfitta la squadra perdente non guadagna punti. Nel torneo della Serie A chi guadagna più punti vince il campionato, mentre chi si classifica tra le ultime tre retrocede alla lega inferiore, la Serie B, dove il posto delle tre squadre retrocesse verrà presso da tre squadre della Serie B che hanno guadagnato la promozione alla Serie A.
La classifica della stagione 2021/2022 è mostrata nella Tabella 2.1.

2.2 Costruzione del dataset

Al giorno d'oggi, nelle partite di calcio professionistico viene raccolta un'enorme quantità di variabili. Ad esempio, per ogni squadra è noto il tempo in percentuale del possesso della palla o il numero di tiri in porta prodotto dalla squadra in una determinata partita. L'obiettivo principale di questo lavoro è determinare l'influenza di queste variabili specifiche della partita. Per creare il dataset per tale scopo, sono state raccolte un gran numero di variabili che a primo avviso possono essere significative, tali dati sono stati offerti dal sito web FBref (https://fbref.com/).

FBref è un sito web dedicato al tracciamento delle statistiche relative ai calciatori e alle squadre di calcio di tutto il mondo.

FBref mette a disposizione i dati sotto forma di tabelle che possono essere modificate per mantenere solo i dati di nostro interesse, in più per rendere più facile l'esportazione, tali tabelle possono essere convertite in formato di CSV per poter essere poi trasportate in un file Excel.
Quindi per ogni squadra che ha partecipato alla stagione 2021/2022 di Serie A si è esportato per ogni partita giocata alcune variabili che ci interessavano, selezionando per prima cosa la macro aree dove si trovavano le variabili d'interesse e poi, modificando le tabelle per ottenere solo i dati di tali variabili. Ogni tabella generata veniva poi riconvertita in CSV per essere poi unita con tutte le altre in un file Excel che una volta completato, divenne il dataset per le nostre analisi. Per rendere più leggibile il file Excel, dato che le stringhe in CSV separavano i dati con il carattere separatore virgola, si è utilizzata la funzione di Excel "trasforma testo in colonne" per inserire tutti i dati in modo ordinato nelle celle del foglio Excel.

2.2.1 Struttura dataset

Il dataset risultate dalla raccolta dati è composto da 760 righe e 35 colonne. Ogni riga riguarda una specifica partita di calcio giocata dalla squadra indicata nella colonna "Team" contro la squadra indicata nella colonna "Vs". Ogni riga perciò contiene informazioni riguardati solo la squadra indicata in "Team" fatta eccezioni per la data della partita ("Date"), il turno ("Round"), e gli spettatori ("Spec"). Quindi per ogni partita esistono due righe, una per ognuna delle due squadre coinvolte. Perciò ogni squadra appare nella colonna "Team" 38 e dato che si hanno 20 squadre si hanno perciò 760 righe totali. Per quanto riguarda le colonne se ne discuterà nella prossima sotto sezione. (TO DO metti foto esempio dataset)

2.2.2 Covariate

2.3 Adattamento dataset al modello Bradley-Terry

Posizione	Squadra	Punti	% casa
1	Milan	86	0.47
2	Inter	84	0.54
3	Napoli	79	0.46
4	Juventus	70	0.50
5	Lazio	64	0.56
6	Roma	63	0.57
7	Fiorentina	62	0.66
8	Atalanta	59	0.33
9	Hellas Verona	53	0.57
10	Torino	50	0.58
11	Sassuolo	50	0.48
12	Udinese	47	0.53
13	Bologna	46	0.61
14	Empoli	41	0.42
15	Sampdoria	36	0.58
16	Spezia	36	0.50
17	Salernitana	31	0.48
18	Genoa	30	0.50
19	Cagliari	28	0.61
20	Venezia	27	0.52

Tabella 2.1: La tabella mostra i punti guadagnati da ogni squadra con il loro piazzamento. Viene mostrata la percentuale di punti guadagnati in casa

3 | Modeling Paired Comparisons

TO DO

3.1

4 | Conclusioni

MEMO Riassunto del lavoro/risultati ottenuti, possibili estensione e migliorie che possono essere apportate. Sottolineare che alcune variabili possono avere un peso differente a seconda della lega in cui si svolge la partita, (ad esempio Premier league è un campionato più fisico con alti ritmi rispetto alla Serie A che è più "tattica") TO DO

Bibliografia