Università degli Studi di Padova
Dipartimento di Scienze Statistiche

Corso di Laurea Triennale in
Statistica per le Tecnologie e le Scienze

Relazione Finale

Analisi end-to-end dei dati delle leghe calcistiche virtuali: modelli e applicazioni

Relatore Prof. Adriano Paggiaro
Dipartimento di Scienze Statistiche

Laureando Filippo Bedin
Matricola 2086328

Anno Accademico 2025/2026

Indice

1 Introduzione

L'avvento della *sports analytics* ha profondamente trasformato i paradigmi di valutazione della performance calcistica, favorendo il passaggio da un'analisi puramente osservazionale a un approccio probabilistico orientato alla misurazione dei processi di gioco. In questo contesto, metriche avanzate come gli *expected goals* (xG) [1, 2] consentono di distinguere tra il risultato osservato e la qualità delle occasioni prodotte, offrendo una rappresentazione più stabile e informativa della prestazione.
Il presente lavoro propone un'analisi delle ultime stagioni di Serie A e della lega calcistica virtuale ad essa associata, con l'obiettivo di confrontare il potere esplicativo e predittivo delle metriche avanzate rispetto alle statistiche tradizionali (quali gol e assist). In particolare, si intende valutare se l'inclusione di indicatori attesi e di variabili di contesto legate alla squadra migliori la capacità di spiegare e prevedere la variazione della quotazione assegnata ai giocatori dall'inizio alla fine della stagione nel contesto fantacalcistico, provando a isolare parte del rumore presente nella valutazione.
L'analisi segue un'impostazione metodologica di tipo end to end che parte dall'integrazione e dalla pulizia dei dataset, prosegue con la costruzione di un archivio coerente e comparabile, sviluppa un'esplorazione delle principali relazioni empiriche e si conclude con la modellazione statistica. In tutte le fasi è stata mantenuta un'attenzione specifica alla validazione dei risultati e alla loro interpretazione in un'ottica applicativa coerente con il contesto del progetto.
Prima di entrare nel dettaglio dell'analisi, è necessario chiarire i riferimenti teorici e operativi su cui si basa il lavoro.

1.1 Expected goals

Il concetto di gol attesi compare per la prima volta in letteratura nel 1993 [3], in uno studio sull'effetto del manto erboso sintetico sulla performance della squadra casalinga. Nel 2004, Jake Ensum, Richard Pollard e Samuel Taylor [4], attraverso un modello di regressione logistica, individuarono cinque fattori significativi nella valutazione della pericolosità di un tiro: distanza dalla porta, angolazione, presenza di un difensore entro un metro, tiro immediatamente successivo a un cross e numero di giocatori tra il tiratore e la porta. Gli expected goals (gol attesi) sono una me-

trica che assegna a ogni tiro una probabilità di trasformarsi in rete sulla base delle sue caratteristiche osservabili. La loro somma, per un giocatore o per una squadra, su partita o stagione, consente di misurare la qualità delle occasioni create e concesse, attenuando la componente aleatoria del risultato. È importante tenere in considerazione che gli xG rappresentano una misura *model-based*, prodotta tramite statistical o machine learning, e che quindi non esiste una probabilità "vera" assegnata a un tiro, ma stime che possono variare in funzione delle variabili incluse e della metodologia adottata. A partire dagli xG si è sviluppata un'intera famiglia di indicatori (che verranno definiti nel capitolo successivo), tra cui gli expected assists, i non-penalty expected goals, gli expected goals chain e gli expected points. Tali metriche ampliano l'analisi oltre i soli eventi osservati, permettendo di descrivere in modo più completo la fase di costruzione e concessione delle occasioni e, più in generale, lo stile di gioco.

1.2 Dinamiche del fantacalcio

Il Fantacalcio [5] è una competizione virtuale in cui più partecipanti, organizzati in leghe generalmente composte da otto o dieci squadre, simulano un campionato parallelo a quello reale di Serie A per l'intera durata della stagione. Ogni partecipante assume il ruolo di fantallenatore e costruisce una propria rosa selezionando calciatori effettivamente impegnati nel campionato.
All'inizio della stagione si svolge un'asta iniziale in cui tutti i fantallenatori dispongono della medesima dotazione di crediti; con tali risorse devono completare la rosa, solitamente composta da venticinque giocatori suddivisi in tre portieri, otto difensori, otto centrocampisti e sei attaccanti. Il sistema presenta vincoli stringenti; per esempio, non è possibile esaurire i crediti prima di aver completato la rosa, e ogni calciatore può appartenere a una sola squadra all'interno della lega. Si configura quindi un mercato chiuso, con risorse limitate e forte competizione tra i partecipanti. La fase d'asta rappresenta un momento decisivo, poiché condiziona l'intera stagione. Le valutazioni iniziali sono spesso influenzate dalla squadra di appartenenza del giocatore o dalle performance dell'annata precedente, che possono aver incluso rendimenti superiori alle attese. Tali dinamiche generano potenziali distorsioni nella formazione dei prezzi, con il rischio di sovrastimare alcuni profili e sottovalutarne altri.
Durante la stagione, in corrispondenza delle giornate di Serie A, le squadre si affrontano in scontri diretti. Il punteggio di ciascun team è determinato dalla somma dei voti assegnati ai giocatori schierati, corretti da bonus e malus legati a eventi di gioco come gol, assist, ammonizioni o espulsioni. Al superamento di determinate soglie di punteggio complessivo viene assegnato un gol virtuale e l'esito della partita è stabilito dal numero di reti ottenute, secondo una logica analoga a quella del calcio

reale.
Dopo la sessione invernale di calciomercato si svolge generalmente un'asta di riparazione, che consente di modificare parzialmente la rosa sulla base delle informazioni accumulate nella prima parte di stagione. In questa fase, la disponibilità di dati aggiornati rende più strutturata la valutazione dei calciatori e apre lo spazio a decisioni maggiormente fondate su indicatori oggettivi di rendimento.

1.3 Metodologia di ricerca

Con analisi full stack si intende un percorso completo che prende avvio dalla raccolta di dati grezzi e imperfetti e conduce alla produzione di un risultato interpretabile, verificato e comunicabile. Non si tratta di un singolo momento analitico, ma di una sequenza coerente di passaggi interdipendenti in cui ogni fase condiziona la solidità di quella successiva. L'intero processo inizia con l'individuazione delle fonti informative, che devono essere affidabili, coerenti con l'obiettivo della ricerca e adeguate in termini di copertura e granularità. Segue la fase di pulizia e strutturazione del dato, in cui vengono definiti encoding, tipi di variabili e gestione dei valori mancanti, così da ottenere un archivio consistente e utilizzabile per le analisi successive.
Una volta costruito il dataset, l'analisi esplorativa consente di comprendere la forma empirica del problema attraverso l'osservazione di distribuzioni, range, eventuali outlier e relazioni preliminari tra variabili, oltre a suggerire trasformazioni utili per la modellazione. La fase successiva consiste nella definizione di un quadro formale che colleghi le variabili osservate alle quantità di interesse, rendendo esplicite le ipotesi sottostanti e la struttura delle relazioni stimate. La validazione rappresenta un momento cruciale, poiché permette di verificare se i risultati ottenuti si mantengono anche al di fuori del campione utilizzato per la stima e di valutare la robustezza delle conclusioni rispetto a differenti specificazioni. Il percorso si conclude con l'interpretazione e la comunicazione dei risultati, che richiede la traduzione delle evidenze statistiche in un linguaggio coerente con il contesto applicativo, e con l'attenzione alla riproducibilità dell'analisi, garantendo che le procedure adottate possano essere replicate in momenti successivi, su dati differenti e da parte di altri utenti.

2 Gestione dei dati

Volendo spiegare le fantaquotazioni, il fantavalore e il fantavoto dei giocatori attraverso statistiche al di fuori di quelle messe a disposizione dalla redazione di fantagazzetta; la mia idea è stata quella di unire ai dati estraibili dal sito ufficiale di Fantaleghe altri dataset che contenessero quante più informazioni sofisticate possibili sulle performance individuali e di squadra. Con l'obiettivo di costruire un unico dataset coerente su cui sviluppare le analisi successive, ho utilizzato Talend for Data Integration per gestire l'unione delle diverse fonti informative. In una prima fase, sono state integrate le tabelle relative alle quotazioni e ai fantavoti; successivamente, è stato completato il quadro unendo i dati fantacalcistici con le altre metriche attese individuali e collettive.

Per connettere i diversi dataset sono state usate, a seconda del contesto, sia delle *Left Outer Join*, sia delle *Inner Join*. La funzione di Join [6] in SQL consente di combinare i dati tra due tabelle di un database relazionale tramite una colonna comune denominata chiave di join. Questo meccanismo permette di raccordare i record presenti in differenti set di dati sfruttando le relazioni esistenti tra Primary Key e Foreign Key, garantendo così la coerenza informativa tra le diverse fonti. La differenza del tipo di Join più adeguato è data da quali record delle due fonti di input si vogliono far rientrare nel dataset in uscita. L'Inner Join è un'operazione che restituisce esclusivamente le righe presenti in entrambe le tabelle, risultando la scelta ottimale quando si desidera isolare solo i casi che presentano una corrispondenza biunivoca, escludendo i record privi di match. La Left Outer Join, al contrario, preserva l'integrità della tabella di sinistra (main table) restituendo tutti i suoi record indipendentemente dalla presenza di un riscontro; a questi vengono associati i dati della tabella di destra (look-up table) dove disponibili, mentre per i casi mancanti vengono generati dei valori nulli.

L'intero processo, ad esclusione del download preliminare dei dati e della predisposizione della struttura ad albero nel file system, è stato progettato in modo automatizzato. Attraverso un'unica esecuzione del main job, il workflow itera sui file di input relativi alle diverse stagioni scaricate e genera i dataset finali, salvandoli nei percorsi corretti. In questo modo si garantisce coerenza tra le stagioni analizzate e si riduce il rischio di errori manuali nelle fasi di integrazione; predisponendo inoltre il flusso per lavori futuri.

Nelle sezioni che seguono illustrerò nel dettaglio le procedure operative adottate, indicando gli strumenti impiegati e la logica che ha guidato ogni passaggio metodologico. Saranno inoltre approfondite le modalità con cui sono state affrontate le criticità e i limiti dei dati per garantire la massima trasparenza e solidità scientifica ai risultati ottenuti.

2.1 Analisi delle sorgenti dati

Il primo passo è stato quello di acquisire i dati ufficiali forniti dalla redazione di Fantacalcio attraverso la sezione risorse del portale. Tale operazione ha permesso di estrarre i dataset relativi alle quotazioni e alle statistiche di gioco, integrando indicatori economici quali la quotazione finale, la variazione di prezzo (ΔQ) e il FantaValore di Mercato (FVM) con fattori di blocco determinanti come il ruolo e il club di appartenenza. Per arricchire l'analisi con metriche di football analytics avanzate, la scelta è ricaduta sul provider Understat [7], individuato come la fonte più idonea per la granularità delle informazioni sia a livello individuale che collettivo.
L'interfaccia, inoltre, ha facilitato l'estrazione di dataset su finestre temporali mirate, permettendo di isolare le prime 27 giornate fino al 4 marzo, momento in cui il numero di partite disputate risultava allineato per tutti i club; e le restanti 11 partite della stagione. L'acquisizione di questi blocchi temporali separati è fondamentale per testare l'efficacia dei modelli predittivi in un contesto che rispecchia fedelmente lo stato del campionato al momento della pubblicazione di questo studio. Questa metodologia consente di validare la precisione delle proiezioni statistiche finali utilizzando una base di conoscenza reale per stimare l'evoluzione delle performance nelle restanti partite.
L'ampia disponibilità di variabili permette di quantificare quanto la varianza delle quotazioni dipenda dall'assetto tattico della squadra e dalle metriche di rendimento atteso. Il dataset offre gli strumenti per verificare se il valore di mercato stia riflettendo una reale efficienza realizzativa o se sia influenzato da distorsioni contestuali legate al sistema collettivo. Attraverso l'incrocio tra statistiche avanzate e dinamiche economiche, è possibile isolare il contributo tecnico individuale rispetto al valore aggiunto derivante dai volumi di gioco prodotti dalla squadra.
L'architettura del dataset presenta tuttavia vincoli strutturali legati alla natura aggregata delle informazioni, che impediscono lo sviluppo di analisi basate su serie storiche. L'assenza di dati match-by-match non permette di mappare i picchi di forma o l'effettiva distribuzione del minutaggio nel corso dei mesi. Allo stesso modo, rimangono ignote variabili quali l'esatto momento delle cessioni, l'incidenza degli infortuni e le conseguenze di eventuali cambi tecnici o societari sul rendimento dei singoli. Queste lacune circoscrivono la ricerca alla dimensione dei volumi medi sta-

gionali e accettano la perdita di dettaglio sulla volatilità di breve periodo a favore di una maggiore stabilità statistica complessiva.

2.1.1 Dataset statistiche

Dal database di Fantacalcio vengono estratte le statistiche fondamentali della piattaforma, quali il voto e il fantavoto, unitamente ai parametri che determinano bonus e malus. Il dataset include inoltre informazioni essenziali come il ruolo e le partite a voto necessarie per inquadrare il rendimento e la partecipazione di ogni calciatore.
I campi selezionati per la costruzione del dataset finale sono riepilogati nella Tabella 2.1.

Variabile	**Descrizione**	**Valore Bonus/Malus**
Id	Chiave univoca del calciatore (Fantacalcio)	-
R	Ruolo del calciatore (P, D, C, A)	-
Pv	Partite valide ai fini del voto	-
Mv	Media dei voti puri assegnati	-
Fm	Media comprensiva di bonus e malus	-
Rc	Rigori totali calciati	-
Gf	Gol e rigori segnati	+3.0
Rp	Rigori parati	+3.0
Ass	Assist forniti	+1.0
Amm	Sanzione disciplinare (ammonizione)	-0.5
Esp	Espulsioni	-1.0
Gs	Gol subiti	-1.0
Au	Autorete realizzata	-2.0
Rf	Rigore calciato e fallito	-3.0

Tabella 2.1: Descrizione delle variabili del dataset statistiche Fantacalcio

Per la redazione, un giocatore riceve un voto se gioca almeno 15 minuti in una partita e viene sufficientemente coinvolto, a meno di eventi notevoli come gol, assist, autogol o espulsioni.

2.1.2 Dataset quotazioni

Il dataset delle quotazioni fornisce le variabili fondamentali per definire il valore economico e il FantaValore di Mercato (FVM) dei calciatori.
Le informazioni estratte per l'analisi sono riepilogate nella Tabella 2.2.

Variabile	Descrizione
Id	Chiave univoca del calciatore (Fantacalcio)
R	Ruolo assegnato nel sistema di gioco classico
Qt.I	Quotazione iniziale stabilita all'apertura delle liste
Qt.A	Quotazione attuale o valore finale raggiunto a termine stagione
Diff	Variazione assoluta della quotazione (ΔQ) nel periodo considerato
FVM	FantaValore di Mercato espresso su un budget di 1000 crediti

Tabella 2.2: Descrizione delle variabili economiche del dataset quotazioni Fantacalcio

Scaricando i dati dal pulsante di download proposto dal sito, i dati vengono salvati in un file *.xlsx* diviso in più fogli: Tutti, Portieri, Difensori, Centrocampisti, Attaccanti, Ceduti.
L'isolamento dei calciatori ceduti permette di monitorare la stabilità delle quotazioni dopo l'abbandono del campionato. Tale verifica risulta determinante per stabilire se includere questi record nel training del modello o escluderli per evitare possibili distorsioni nei dati.

2.1.3 Dataset teams

Le metriche collettive di Understat forniscono il contesto tattico necessario per valutare come l'assetto di squadra influenzi i volumi di rendimento individuali. Oltre ai dati tradizionali, la base dati integra indicatori di pressione e pericolosità offensiva utili a misurare la qualità reale del gioco prodotto rispetto ai risultati effettivamente maturati sul campo.
Le variabili selezionate per descrivere il rendimento dei club sono riepilogate nella Tabella 2.3.

Variabile	Descrizione
number / team	Identificativo numerico e nome della squadra
wins / draws / loses	Numero di vittorie, pareggi e sconfitte
goals / ga	Gol segnati e gol subiti (*goal against*)
points / xPTS	Punti reali ottenuti e punti attesi
xG / NPxG	*Expected Goals* totali e al netto dei calci di rigore
xGA / NPxGA	*Expected Goals Against* totali e al netto dei rigori subiti
NPxGD	Differenza tra *Non-Penalty xG* fatti e subiti
ppda	Intensità del pressing (*passes per defensive action*)
ppda_allowed	Pressing subito dalla squadra avversaria [8]
deep / deep_allowed	Passaggi completati o concessi entro i 18 metri dalla porta

Tabella 2.3: Descrizione delle variabili collettive del dataset teams Understat

2.1.4 Dataset players

Le statistiche individuali di Understat permettono di analizzare la performance tecnica del singolo attraverso i volumi di gioco e le metriche attese. Questi dati inte-

grano le informazioni di base con indicatori avanzati sulla qualità delle conclusioni e sul coinvolgimento nella manovra offensiva.
Le variabili selezionate per il dataset individuale sono riepilogate nella Tabella 2.4.

Variabile	**Descrizione**
number / player	Identificativo univoco (Understat) e nome del calciatore
team / apps / min	Squadra, presenze effettive e minuti totali giocati
goals / a	Gol e assist realizzati
NPG	Gol realizzati al netto dei calci di rigore (*non-penalty goals*)
xG / NPxG / xA	*Expected Goals* (totali e *non-penalty*) ed *Expected Assists* [9]
xGChain	Somma del valore xG di ogni azione in cui il calciatore partecipa alla manovra, inclusi il tiratore e l'assistente [10]
xGBuildup	Somma del valore xG di ogni azione in cui il calciatore partecipa alla manovra, esclusi il tiratore e l'assistente
...90	Versioni normalizzate delle metriche precedenti sui 90 minuti
xG90xA90	Somma dei contributi attesi (xG + xA) già normalizzati

Tabella 2.4: Descrizione delle variabili individuali del dataset players Understat

2.2 Data processing

Una volta scaricati, i dati sono stati organizzati all'interno di una struttura di directory coerente e gerarchica. La cartella principale *data* costituisce la radice dell'archivio e si articola in due sottocartelle, *input* e *output*: nella prima sono collocati i dataset grezzi, mentre nella seconda vengono salvati i file prodotti a valle delle trasformazioni. All'interno di entrambe le directory è stata creata una sottocartella per ciascuna stagione, così da mantenere separazione temporale, tracciabilità delle fonti e maggiore controllo sulle versioni dei dati. Questa struttura ad albero facilita l'integrazione progressiva di nuove stagioni e rende il flusso di lavoro più ordinato e replicabile. L'integrazione e la trasformazione dei dati sono state realizzate tramite Talend, uno strumento di data integration basato su un'architettura a job grafici, in cui i flussi informativi vengono costruiti attraverso componenti collegati tra loro (input, trasformazioni, join, output). Nel caso specifico, per avere una migliore tracciabilità delle trasformazioni, le operazioni di lettura, pulizia, unione e scrittura dei dataset sono state realizzate seguendo un'impostazione modulare articolata in tre job distinti. Un job principale (Figura 2.1) svolge una funzione di orchestrazione e richiama in sequenza due job operativi dedicati rispettivamente alla fusione in un unico dataset fantacalcistico e alla generazione del dataset finale integrato. Il job principale gestisce il flusso complessivo del processo, iterando sulle stagioni disponibili mediante un componente tLoop e, per ogni stagione al di fuori di quella attuale, ripete i passaggi per i dati completi e per la parte iniziale e finale di stagione.

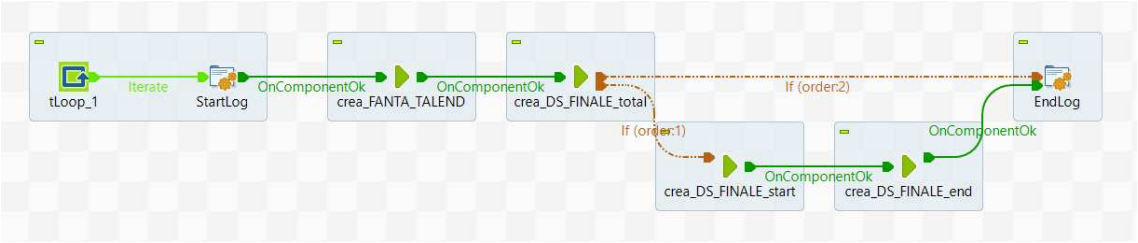

Figura 2.1: Job main 0.1

La parametrizzazione dei percorsi e delle variabili è stata gestita attraverso variabili di contesto e variabili globali, in modo da rendere l'esecuzione indipendente dalla stagione considerata e facilmente estendibile negli aggiornamenti successivi, giovando così alla riproducibilità del progetto. Le variabili di contesto sono il tipo di run, che indica se stiamo trattando una stagione intera o parziale; e l'anno, ovvero la stagione a cui i dati fanno riferimento. È stato necessario creare con il componente tSetGlobalVar due blocchi di variabili globali, in quanto nel primo sono stati inseriti i path utilizzati per i dati completi e per quelli troncati, dato che si differenziano per un pezzo di path ("_START" o "_END", Figura 2.2); nel secondo, le variabili globali comprendono in che contesto ci troviamo e selezionano il percorso corretto.

```
\\ players
context.tipo_run.equals("total") ? (((String)globalMap.get("filepath_players")).replaceAll("<anno>",context.anno)) :
 (context.tipo_run.equals("start") ? (((String)globalMap.get("filepath_players_START")).replaceAll("<anno>",context.anno)) :
  (((String)globalMap.get("filepath_players_END")).replaceAll("<anno>",context.anno)))


\\ teams
context.tipo_run.equals("total") ? (((String)globalMap.get("filepath_teams")).replaceAll("<anno>",context.anno)) :
 (context.tipo_run.equals("start") ? (((String)globalMap.get("filepath_teams_START")).replaceAll("<anno>",context.anno)) :
  (((String)globalMap.get("filepath_teams_END")).replaceAll("<anno>",context.anno)))


\\ scarti
context.tipo_run.equals("total") ? (((String)globalMap.get("filepath_scarti")).replaceAll("<anno>",context.anno)) :
 (context.tipo_run.equals("start") ? (((String)globalMap.get("filepath_scarti_START")).replaceAll("<anno>",context.anno)) :
  (((String)globalMap.get("filepath_scarti_END")).replaceAll("<anno>",context.anno)))


\\ finale
context.tipo_run.equals("total") ? (((String)globalMap.get("filepath_finale")).replaceAll("<anno>",context.anno)) :
 (context.tipo_run.equals("start") ? (((String)globalMap.get("filepath_finale_START")).replaceAll("<anno>",context.anno)) :
  (((String)globalMap.get("filepath_finale_END")).replaceAll("<anno>",context.anno)))


\\ fanta
((String)globalMap.get("filepath_fanta")).replaceAll("<anno>",context.anno)
```

Figura 2.2: Pipeline per selezionare il path corrispettivo

Nel primo dei due job operativi, crea_FANTA_talend 0.1 (Figura 2.3), sono stati caricati i due file *.xlsx* con il componente tFileInputExcel e, usando la colonna Id come chiave di join, sono stati uniti in un unico file, stampato in output e caricato nel seguente job. Prima di fondere i due dataset, tramite un tJavaRow, è stata aggiunta una colonna tra le variabili di quotazione per indicare se il giocatore è stato venduto o meno, in base all'origine del dato (foglio "tutti" o foglio "ceduti").

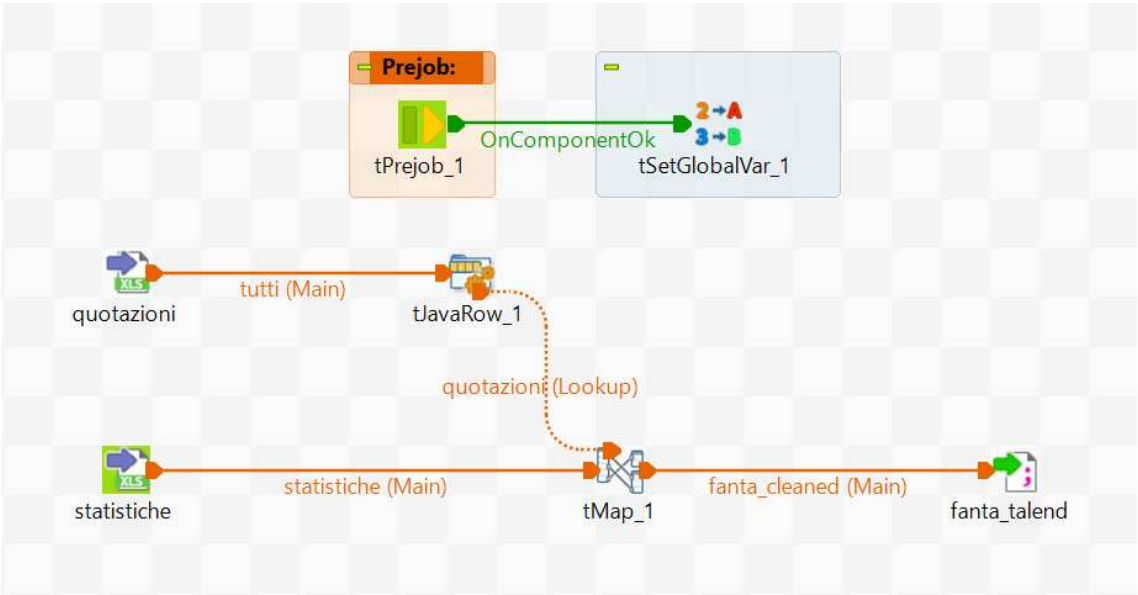

Figura 2.3: Job crea_FANTA_TALEND 0.1

Il job successivo (Figura 2.4) è incaricato di integrare le variabili fantacalcistiche precedentemente apparecchiate, il dataset player e quello team, per poi stampare tutto nel file che verrà caricato poi per le analisi esplorative e la modellazione.

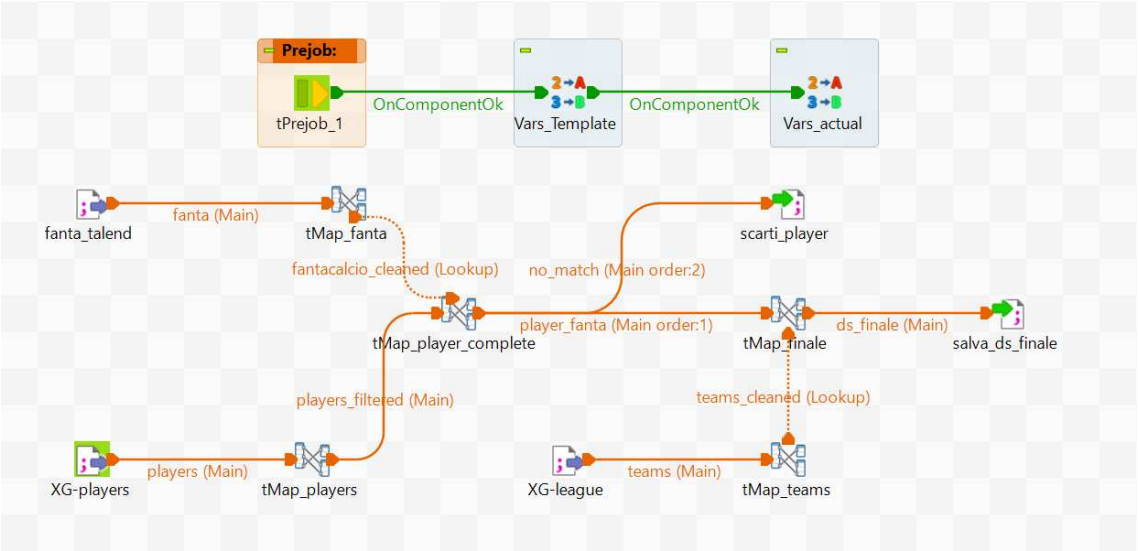

Figura 2.4: Job crea_DS_FINALE 0.1

L'ultimo passaggio di far combaciare i dataset tramite la variabile team ha richiesto solo un minimo lavoro di ridenominazione delle variabili (es. da xG a team_xG); diverso è il discorso per quanto riguarda la comunicazione tra i record dei giocatori nei diversi dataset. Difatti, è servito un complesso lavoro di encoding e di uniformazione dei tipi poichè non esisteva un Id univoco e quindi si è scelto di usare il nome come lookup, che però, se in un dataset era composto da nome e cognome del giocatore, nell'altro il formato era del tipo "cognome" + "iniziale del nome" + ".", oltre a svariate casistiche particolari (Figura 2.5). La gestione dei casi di mancata

corrispondenza (no_match) è stata separata in un flusso dedicato, consentendo di isolare osservazioni problematiche, ad esempio giocatori non presenti in una delle fonti, e verificarne l'impatto sul dataset finale.

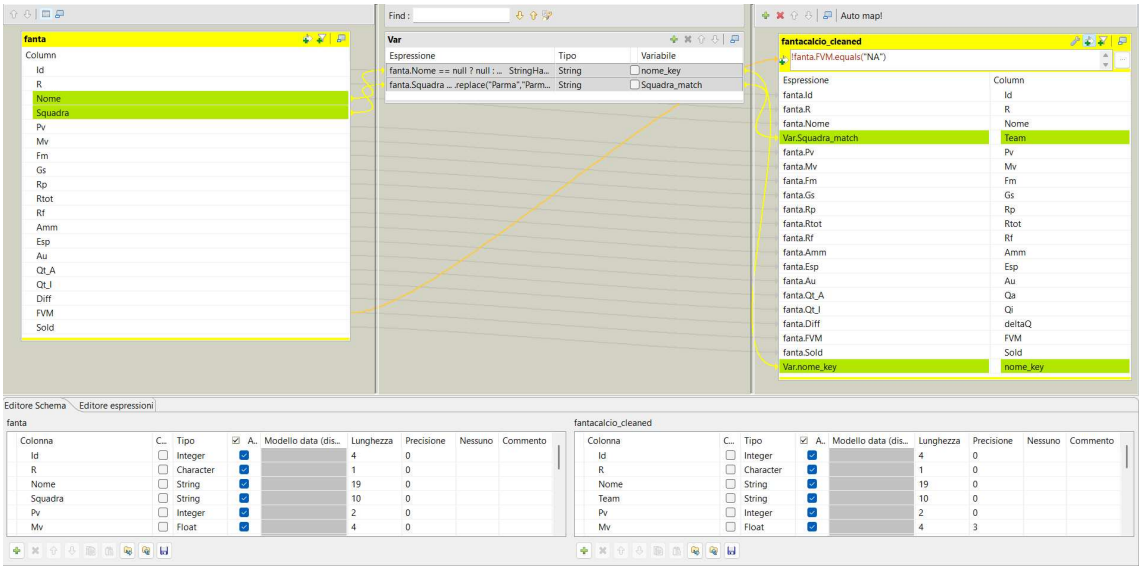

Figura 2.5: tMap usato per creare la variabile nome_key

3 Analisi esplorativa

L'analisi esplorativa dei dati rappresenta una fase diagnostica imprescindibile per validare la qualità della base informativa e orientare le scelte di modellazione statistica. In questa sezione, i dati vengono indagati nella loro struttura distributiva e nelle relazioni empiriche che intercorrono tra le metriche di rendimento e le valutazioni economiche. Il dataset in esame è quello della stagione 2024/2025, lo stesso su cui verranno addestrati successivamente i modelli.

3.1 Distribuzioni e trasformazioni delle variabili

Lo studio della morfologia del dataset ha richiesto inizialmente un intervento di filtraggio mirato a stabilizzare le analisi prestazionali, escludendo i calciatori caratterizzati da un impiego marginale. Attraverso l'osservazione del minutaggio stagionale e delle presenze a voto (Figura 3.1), si è scelto di mantenere nel database solo i profili con almeno 300 minuti disputati e 3 presenze, garantendo così che le medie prestazionali non fossero alterate da campioni di gioco troppo ridotti; passando così da un dataset di 547 record a uno di 423.

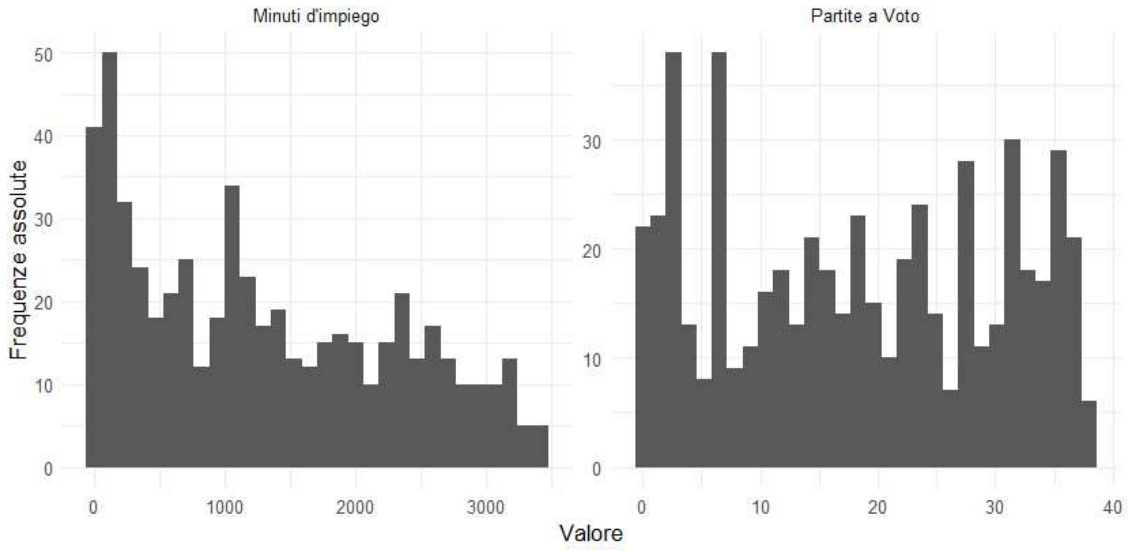

Figura 3.1: Distribuzioni dei minuti giocati e delle partite a voto

Una volta stabilizzato il campione tramite questo filtraggio, è stata verificata la consistenza delle metriche di rendimento (Figura 3.2). Tali variabili seguono distribu-

zioni tipiche degli eventi rari, con una forte asimmetria positiva e una concentrazione di frequenze in prossimità dello zero. Non essendo emersi outlier patologici o errori di rilevazione, si è scelto di mantenere tali variabili nella loro forma originaria per le analisi di efficienza.

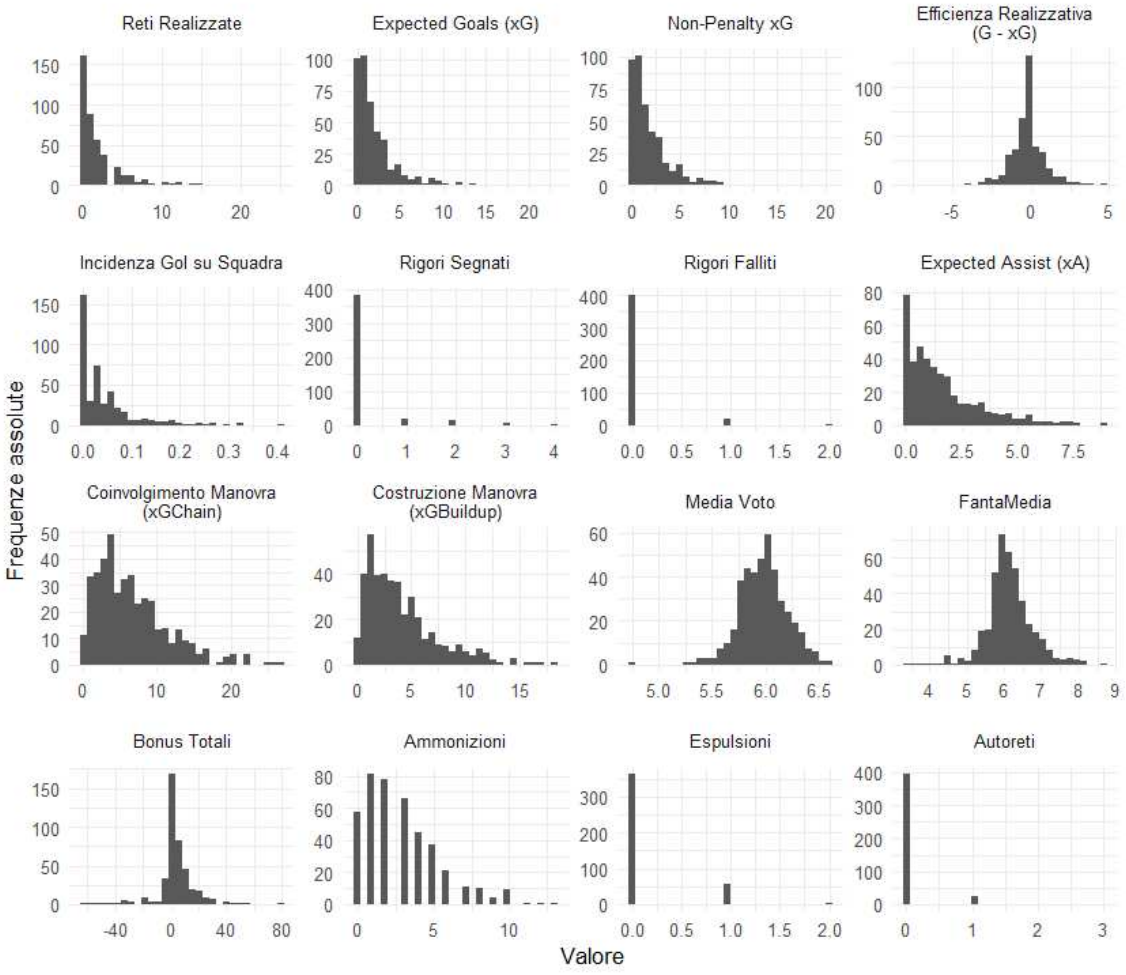

Figura 3.2: Distribuzioni delle metriche di rendimento

L'osservazione delle variabili economiche (Figura 3.3), quali il FantaValore di Mercato e la quotazione finale, rivela distribuzioni marcatamente asimmetriche dovute alla presenza di pochi profili d'élite con valutazioni sensibilmente distanti dalla massa del campione. L'applicazione della trasformazione logaritmica permette di normalizzare queste grandezze, rendendo variabili come $logFVM$ e $lograteoQ$ idonee a una modellazione lineare e riducendo l'impatto distorsivo dei valori estremi sulle stime complessive.

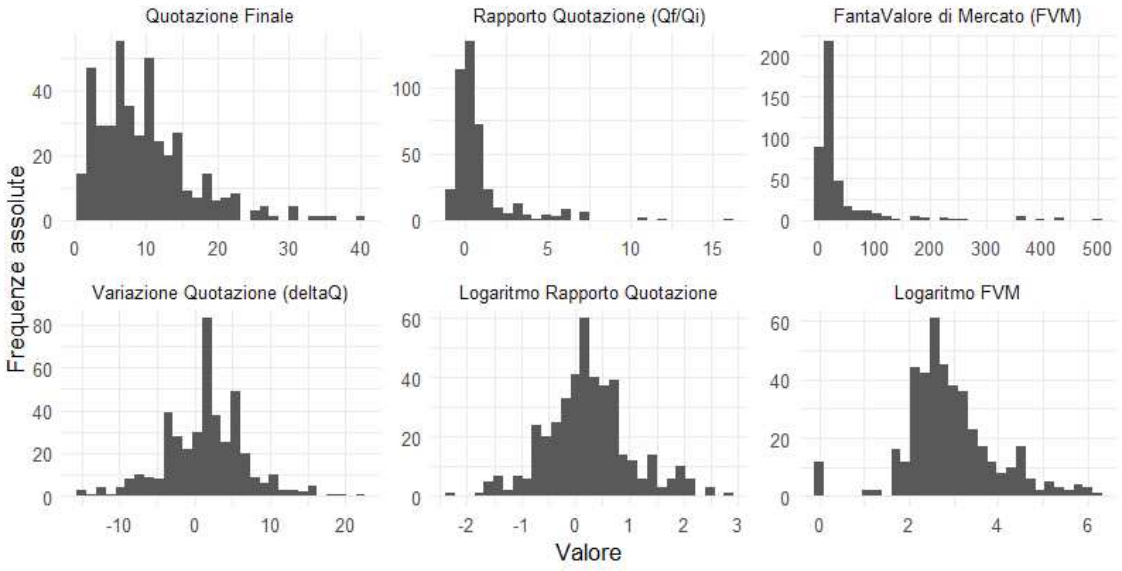

Figura 3.3: Distribuzioni delle variabili di interesse

Un ulteriore dubbio metodologico riguarda i calciatori ceduti durante la stagione, la cui permanenza nel dataset richiede una verifica sulla stabilità della loro valutazione economica post-trasferimento. È stato anzitutto verificato che le quotazioni non subissero azzeramenti o deprezzamenti tecnici post-vendita (Figura 3.4).

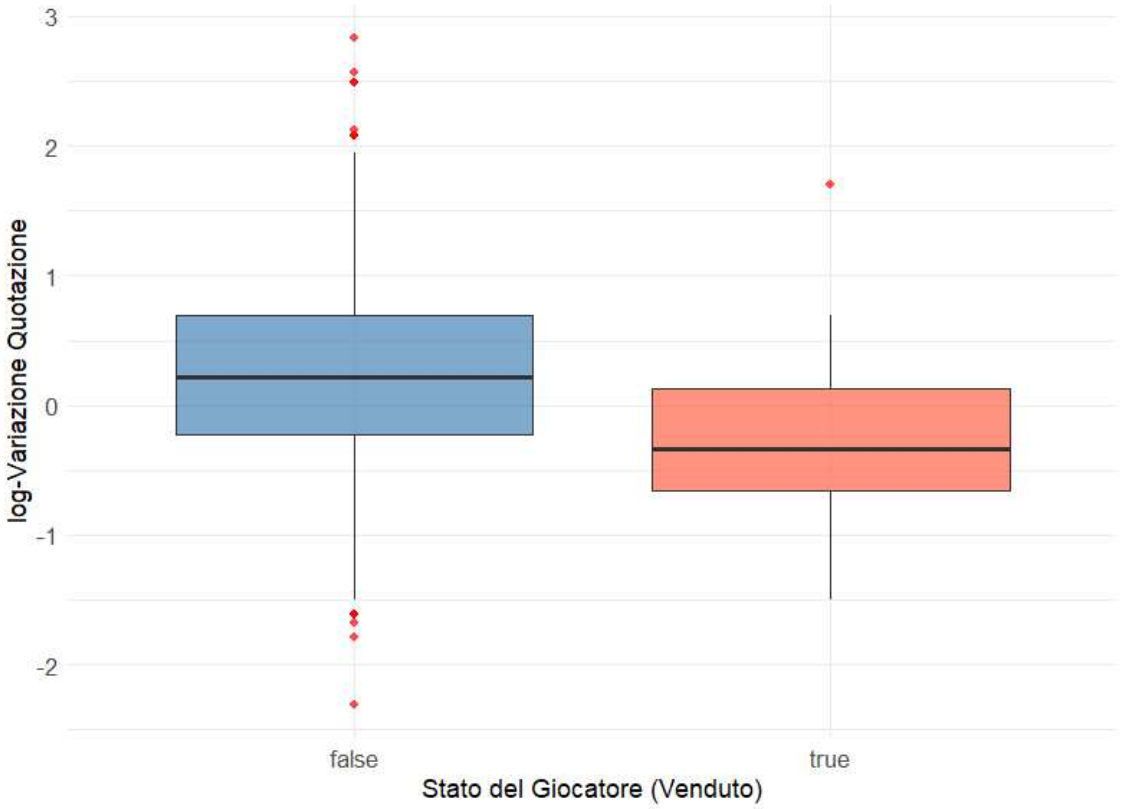

Figura 3.4: Impatto della cessione sul log-rapporto di quotazione (lograteoQ)

Nonostante il gruppo Sold mostri un log-rapporto medio inferiore, si è ipotizzata una causalità inversa: il trasferimento è spesso conseguenza del calo di rendimento

e non la causa del deprezzamento. Al fine di separare l'impatto della cessione dalle prestazioni effettive, si sceglie di modellare la valutazione del giocatore attraverso una regressione lineare che includa i principali indicatori di rendimento, spostando l'analisi sui residui della regressione (Tabella 3.1).

Variabile	**Coeff.**	**Err. Std**	**t-stat**	**p-value**
(Intercept)	-0.2559	0.1265	-2.022	0.0438*
goals	0.0218	0.0163	1.337	0.1818
assist	0.0074	0.0254	0.291	0.7715
RC	0.1841	0.1138	1.618	0.1065
RD	0.0566	0.1196	0.473	0.6365
RP	0.2890	0.1819	1.589	0.1129
Pv	0.0146	0.0046	3.158	0.0017**
R-squared: 0.0612				Adj. R-squared: 0.0477

Tabella 3.1: Risultati del modello di regressione lineare per la variabile *lograteoQ*.

Listing 3.1: Analisi dei residui in base allo stato del giocatore (Sold).

```
> t.test(residui ~ Sold, data = data)
        Welch Two Sample t-test

data:  residui by Sold
t = 1.5765, df = 18.755, p-value = 0.1316
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval: [-0.0901, 0.6381]
sample estimates:
mean in group false:  0.01166
mean in group true : -0.26234
```

Il t-test per il confronto tra medie con correzione della varianza di Welch (Listing 3.1) non ha rilevato differenze statisticamente significative tra i gruppi (p-value $= 0,1316$). Si è quindi scelto di mantenere i record nel dataset e di non escludere la variabile in fase di modellazione, poiché la quotazione riflette fedelmente il valore del giocatore al momento della cessione.

3.2 Produzione offensiva ed efficienza realizzativa

Passando all'analisi tecnica del campo, il primo passo è distinguere i gol segnati su azione dai tiri dal dischetto per valutare meglio il rendimento dei singoli (Figura 3.5). I dati mostrano che molti dei principali marcatori devono una fetta importante del loro punteggio proprio ai rigori. Se Mateo Retegui guida la classifica grazie a una produzione dominante su azione, profili come Riccardo Orsolini o Romelu Lukaku dipendono molto di più dai rigori per restare ai vertici. Isolare i gol su

azione permette di far emergere la reale pericolosità di giocatori come Moise Kean o Marcus Thuram, fornendo la base per confrontare i gol fatti con quelli attesi senza le distorsioni tipiche dei calci piazzati. Questa scelta metodologica porta a usare i non-penalty Goals come misura principale per studiare l'efficienza realizzativa nelle analisi successive.

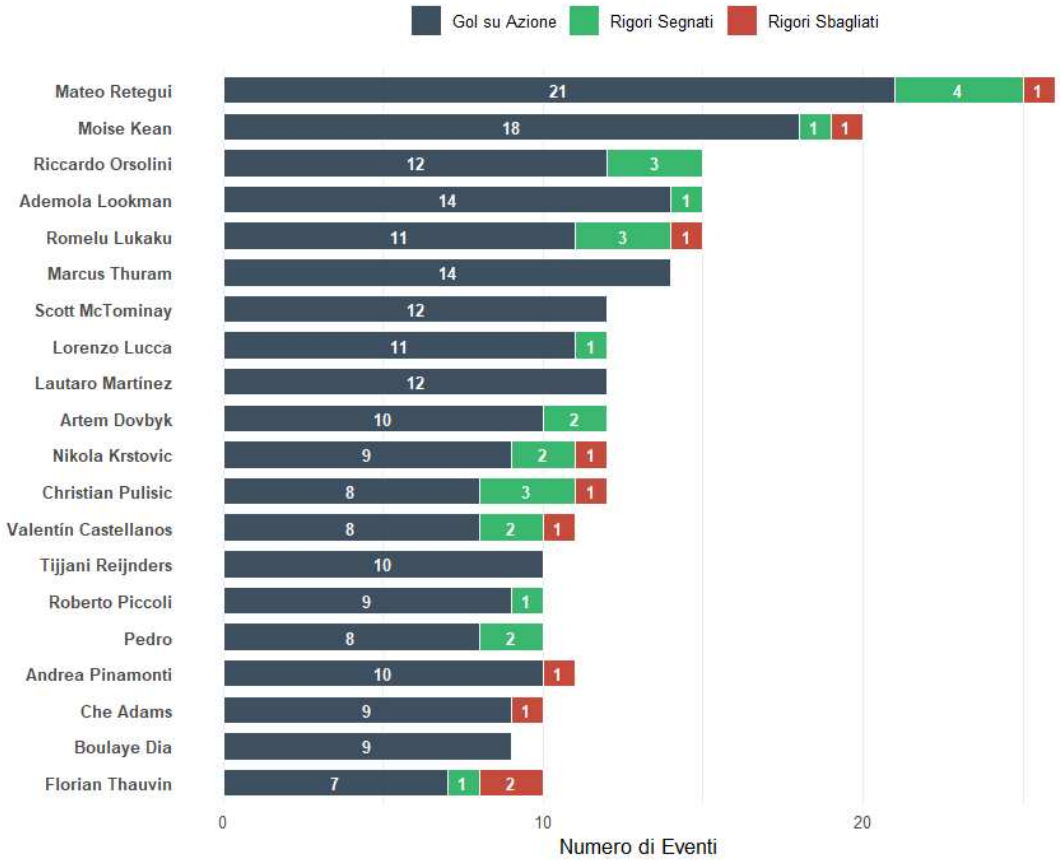

Figura 3.5: Analisi realizzativa: azioni, rigori ed errori

Il rapporto tra i gol segnati su azione e i relativi gol attesi permette di quantificare con precisione la qualità della conversione balistica attraverso i diversi reparti tattici (Figura 3.6). Osservando i singoli profili, si nota come i pochi difensori capaci di un buon bottino di reti, tra cui Nadir Zortea, Robin Gosens e Denzel Dumfries, presentino una discrepanza molto favorevole rispetto agli eventi attesi, seguiti a centrocampo dalla spiccata efficacia di Riccardo Orsolini e Tijjani Reijnders. Un'evidenza sistematica del periodo riguarda il posizionamento delle rette di regressione costantemente al di sotto della bisettrice tratteggiata, segnale di un deficit realizzativo generale o di una possibile sovrastima della metrica xG nel contesto della Serie A. Tale divario si riflette infine nei coefficienti di efficienza per ruolo (Tabella 3.2), che vedono gli attaccanti al vertice con un valore di 0,859, seguiti con distacchi regolari dai centrocampisti, attestati a 0,826, e dai difensori, che chiudono con 0,758.

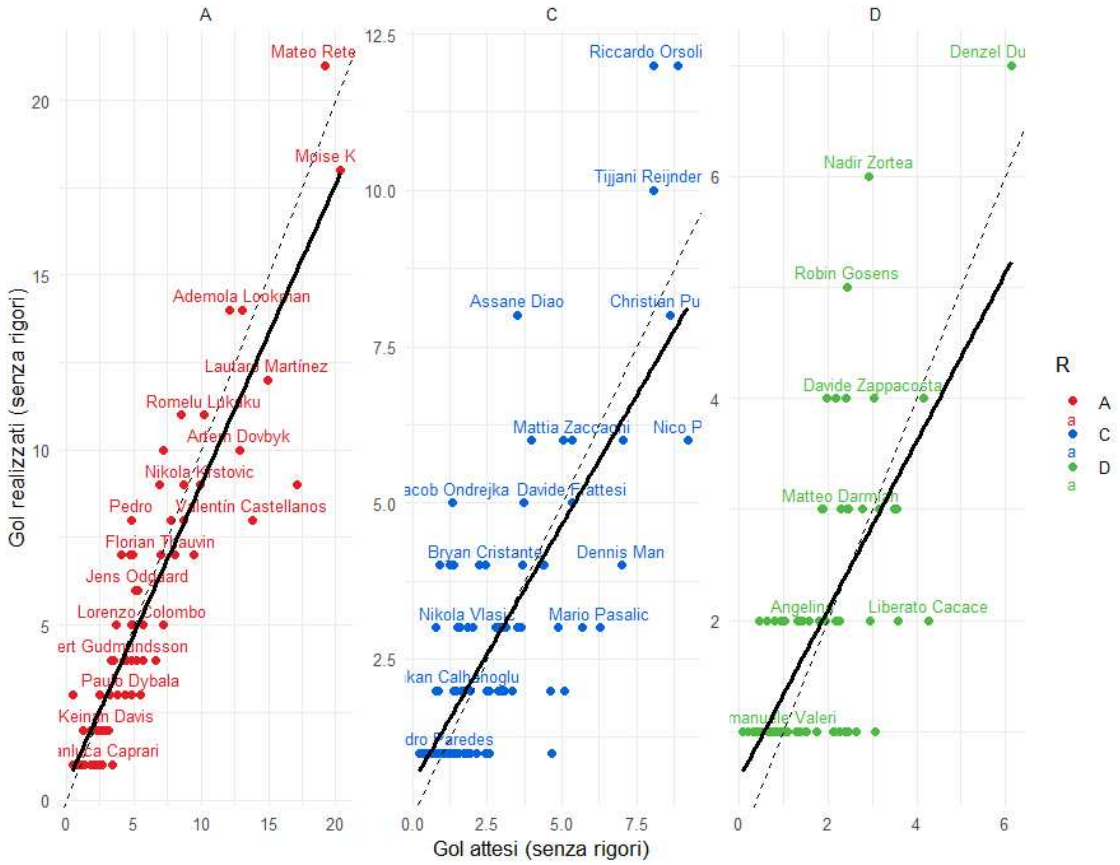

Figura 3.6: Confronto tra gol segnati e attesi per ruolo

Ruolo	**Coefficiente Angolare**
A (Attaccanti)	0.859
C (Centrocampisti)	0.826
D (Difensori)	0.758

Tabella 3.2: Efficienza realizzativa suddivisa per ruolo

Lo studio dei residui con una scala uniformata (Figura 3.7), permette di isolare i calciatori che più si allontanano dalla media, identificando i migliori e i peggiori nel concretizzare le chance avute. Oltre al già citato Riccardo Orsolini, Pedro e Scott McTominay spiccano come i trascinatori più cinici, capaci di segnare molto più di quanto i numeri lascerebbero presagire. Sul fronte opposto, il caso di Roberto Piccoli è il più evidente: l'attaccante si trova in fondo a questa classifica con un deficit di circa 8 gol rispetto al volume di gioco prodotto, segno di una difficoltà persistente nel capitalizzare i palloni ricevuti.

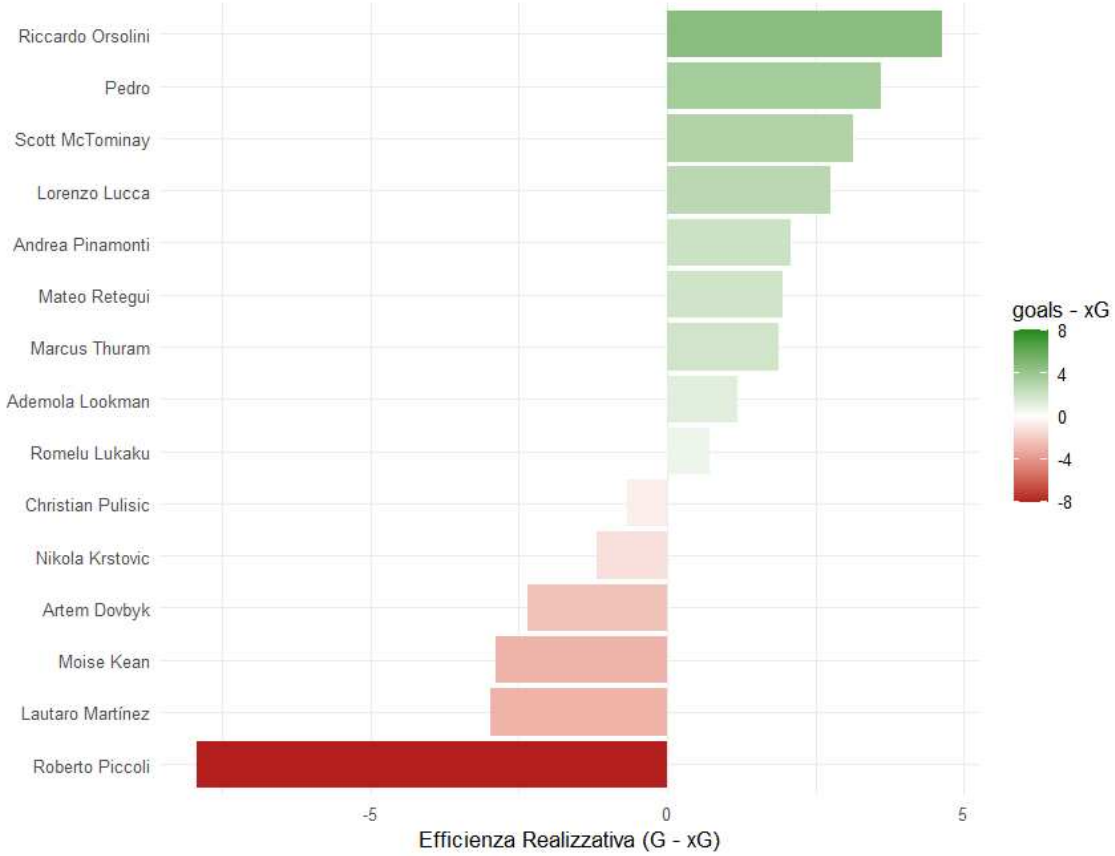

Figura 3.7: Gol su azione e i relativi gol attesi

L'analisi della produzione offensiva trova la sua sintesi finale nell'incrocio tra la pericolosità individuale e la capacità di generare occasioni per i compagni (Figura 3.8), una mappatura a due dimensioni che permette di definire con precisione il profilo tattico di ciascun giocatore. Attraverso questa visualizzazione si distinguono chiaramente i finalizzatori puri, come Moise Kean e Roberto Piccoli, che presentano volumi di tiro molto elevati a fronte di un contributo marginale alla manovra, e i creatori di gioco come Hakan Calhanoglu o Federico Dimarco, posizionati nel quadrante opposto per la loro spiccata attitudine all'assist. I profili di maggior valore strategico per il mercato virtuale sono tuttavia quelli che occupano il quadrante in alto a destra, capaci di combinare un'alta probabilità realizzativa con una costante rifinitura; in questa élite spiccano nomi come Mateo Retegui, Ademola Lookman e Christian Pulisic, i quali rappresentano i motori offensivi più completi e determinanti del dataset analizzato. Questa classificazione non solo aiuta a comprendere lo stile di gioco dei singoli, ma offre una guida oggettiva per pesare il potenziale di bonus complessivo di un calciatore prima di un investimento.

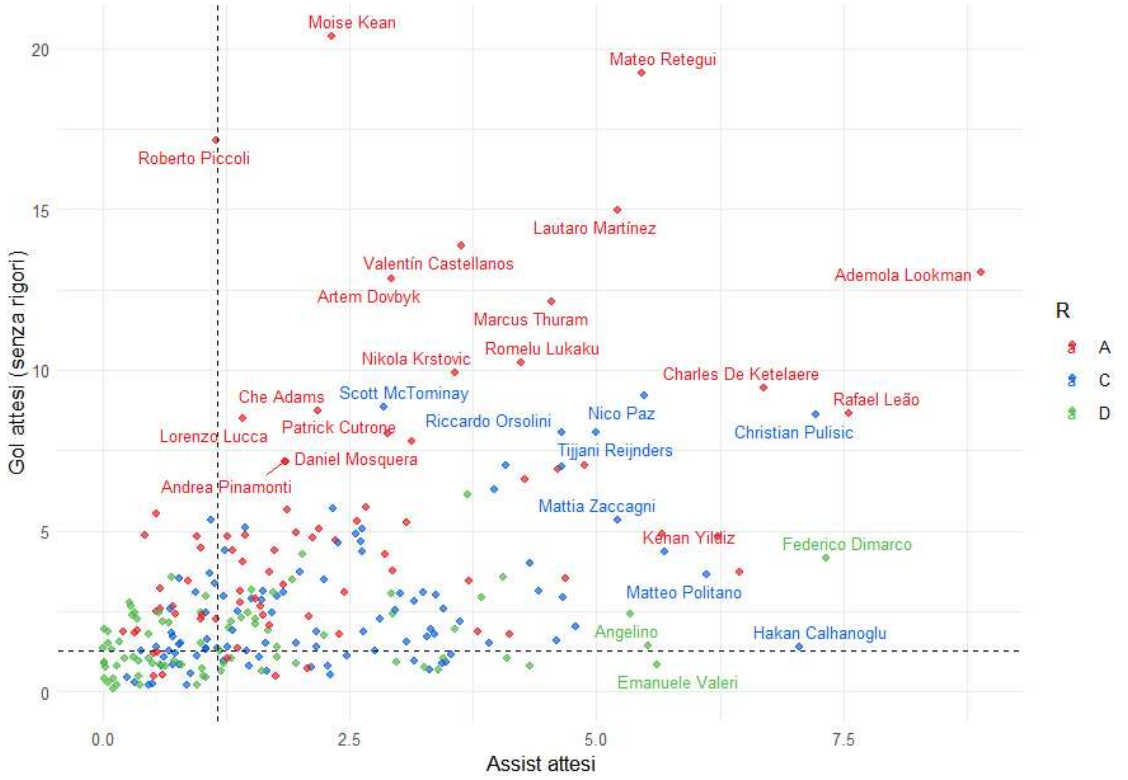

Figura 3.8: Produzione offensiva: non-penalty goals e assist attesi

3.3 Gestione delle rose e rotazioni della titolarità

Lo studio della continuità d'impiego è utile per valutare la stabilità dei profili nel dataset, dato che la regolarità del voto dipende molto dalle gerarchie interne di ogni squadra. Analizzando il numero di calciatori utilizzati (Figura 3.9), emergono approcci differenti tra i vari club: formazioni come Como e Udinese hanno impiegato un numero elevato di elementi (rispettivamente 25 e 24 con almeno 300 minuti in stagione), mentre Atalanta e Napoli si sono affidate a un gruppo più ristretto e definito. Oltre alle preferenze degli allenatori, questa minore rotazione può derivare da una rosa già equilibrata che non ha richiesto nuovi innesti a gennaio o da una stagione con pochi infortuni, fattori che facilitano il mantenimento di un assetto costante.

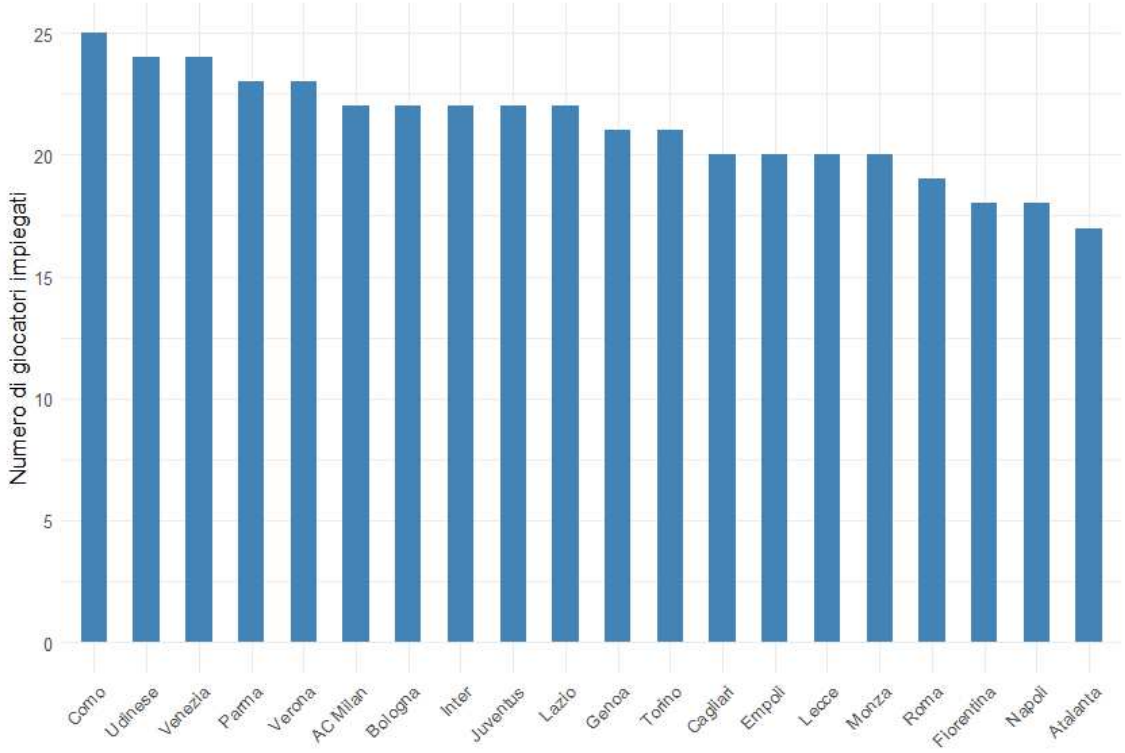

Figura 3.9: Numero di giocatori impiegati per squadra

Il minutaggio medio per partita permette di identificare meglio i titolari fissi all'interno di queste rotazioni. Come illustrato nella Tabella 3.3, si nota una riduzione dei giocatori disponibili all'aumentare della soglia di impiego: se quasi tutti i componenti di una rosa superano i 20 minuti medi, non tutte le rose, vedesi Bologna o Roma, riescono ad arrivare a una formazione di 11 giocatori con 70 minuti a partita, valore considerato limite per una titolarità certa. Le squadre di alta classifica, come Fiorentina e Atalanta, tendono a polarizzare l'impiego su questo nucleo di "titolarissimi", a differenza di formazioni con rotazioni più frammentate che rendono più complicata la stima del contributo stagionale dei singoli.

Squadra	**Impiegati**	**≥ 20'**	**≥ 45'**	**≥ 70'**	**≥ 85'**	**≥ 90'**
AC Milan	27	25	21	10	1	0
Atalanta	28	24	16	10	4	3
Bologna	25	22	18	7	4	0
Cagliari	25	22	17	8	3	1
Como	35	27	21	12	3	1
Empoli	28	22	18	11	6	3
Fiorentina	23	23	17	9	6	3
Genoa	34	31	21	9	5	2
Inter	23	22	16	8	2	2
Juventus	25	24	21	11	3	3
Lazio	24	22	17	9	4	2
Lecce	29	25	17	10	5	3
Monza	26	22	18	8	1	0
Napoli	22	18	14	12	5	1
Parma	30	27	22	10	4	2
Roma	25	24	16	7	3	2
Torino	25	23	18	10	4	2
Udinese	30	27	21	12	7	4
Venezia	33	27	22	13	5	2
Verona	30	26	19	9	5	2

Tabella 3.3: Numero di giocatori per squadra suddivisi per soglie di minutaggio medio a partita.

3.4 Analisi delle variabili di mercato per ruolo

L'indagine si sposta ora sulla distribuzione dei calciatori per ruolo, fattore che influenza direttamente la disponibilità numerica e le dinamiche di prezzo nel mercato virtuale. Osservando la ripartizione dei record (Figura 3.10), si nota una proporzionalità coerente con le necessità di composizione di una rosa standard nelle leghe virtuali (3 portieri, 8 difensori, 8 centrocampisti, 6 attaccanti), con una netta prevalenza di difensori e centrocampisti rispetto ad attaccanti e portieri. Questo equilibrio numerico fa sì che non ci sia un effetto moltiplicativo sul prezzo dovuto alla scarsità di giocatori in una categoria.

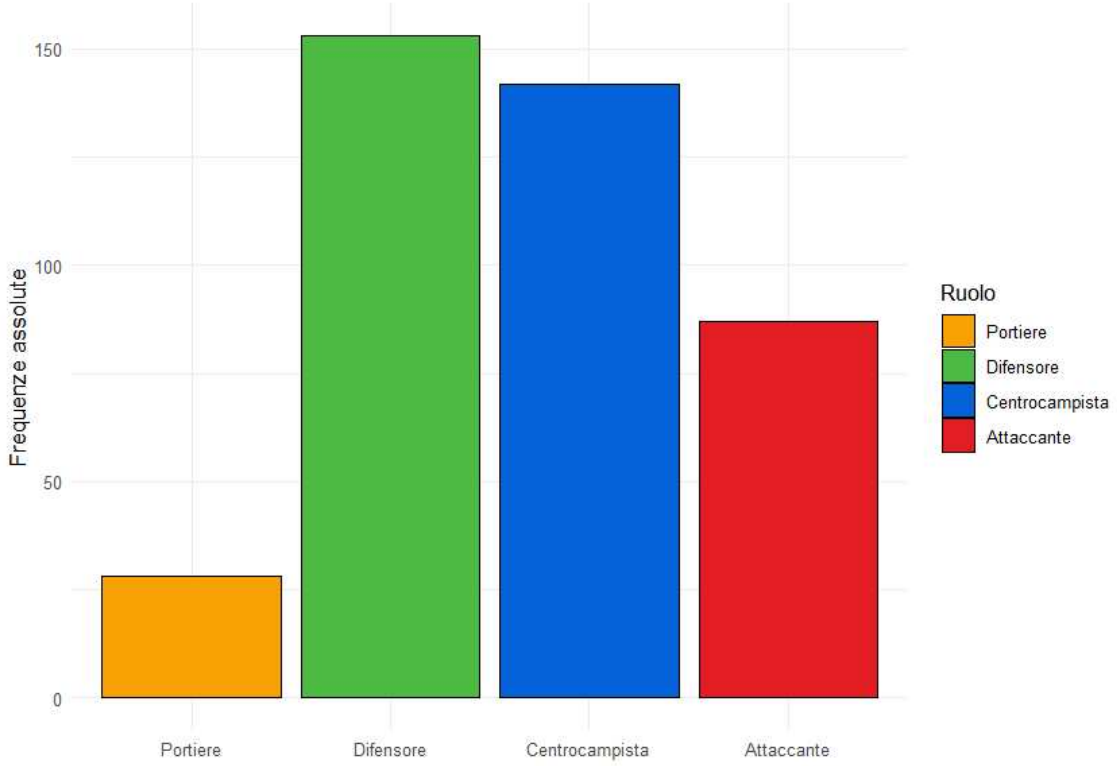

Figura 3.10: Frequenze osservate per ruolo

L'analisi della Media Voto (Figura 3.11) mette in luce quanto sia difficile ottenere la sufficienza costante in alcuni comparti tattici. Se i portieri mantengono una mediana rassicurante, i difensori rappresentano il gruppo più penalizzato, con una massa critica di voti che resta stabilmente sotto il 6.0. Questa evidenza aumenta il valore di un difensore capace di registrare prestazioni da voti alti con costanza. Lo stesso discorso potrebbe valere per i centrocampisti, mascherati da numerosi outlier presenti nella parte alta del grafico

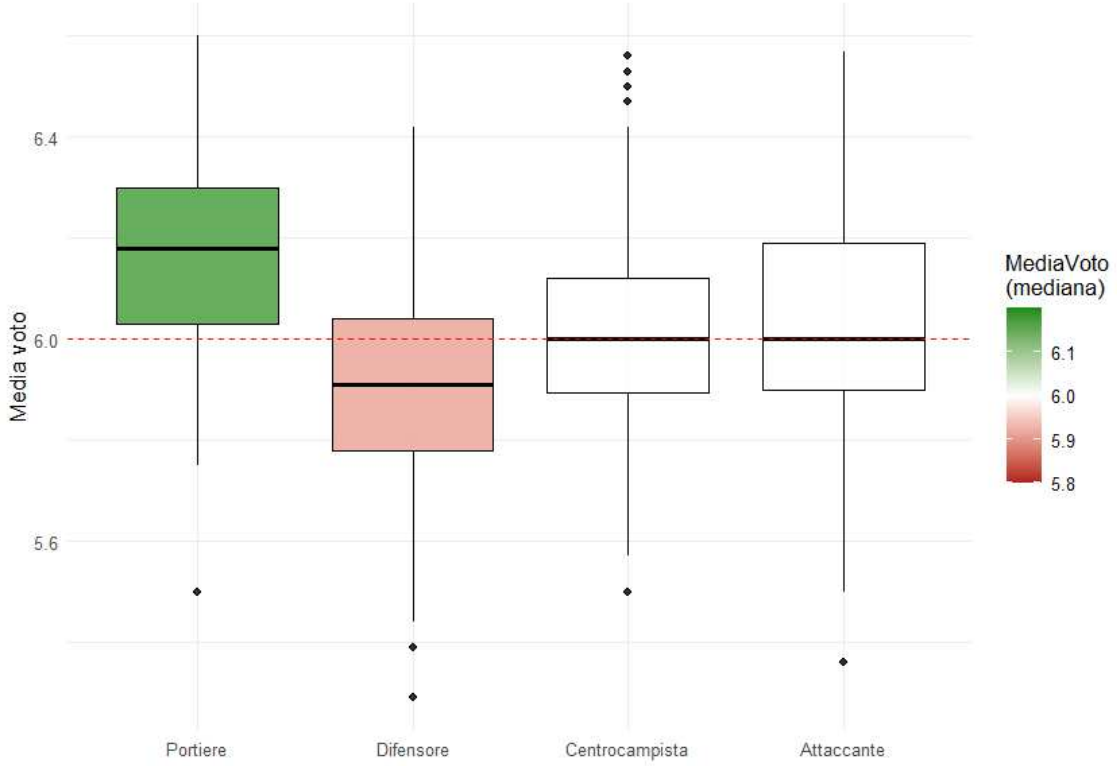

Figura 3.11: MediaVoto per ruolo

Spostando l'attenzione sul mercato, la crescita delle quotazioni (Figura 3.12) e il FantaValore di Mercato (Figura 3.13) non rispondono a un generico rendimento, ma sono guidati principalmente dalla caccia ai bonus offensivi. Attaccanti e centrocampisti mostrano una volatilità estrema e reagiscono con forti scatti ai gol segnati, portando a valutazioni molto più instabili rispetto a quelle dei difensori, che tendono a rimanere ancorati alle prestazioni difensive pure. La distribuzione del FVM per le punte evidenzia infine una asimmetria totale dovuta a pochi nomi che drenano la maggior parte dei crediti. La presenza di numerosi outlier con prezzi oltre le 400 unità conferma che il budget dei fantallenatori converge quasi interamente sui pochi finalizzatori capaci di garantire i bonus pesanti e trasforma l'attacco nel vero motore finanziario del gioco.

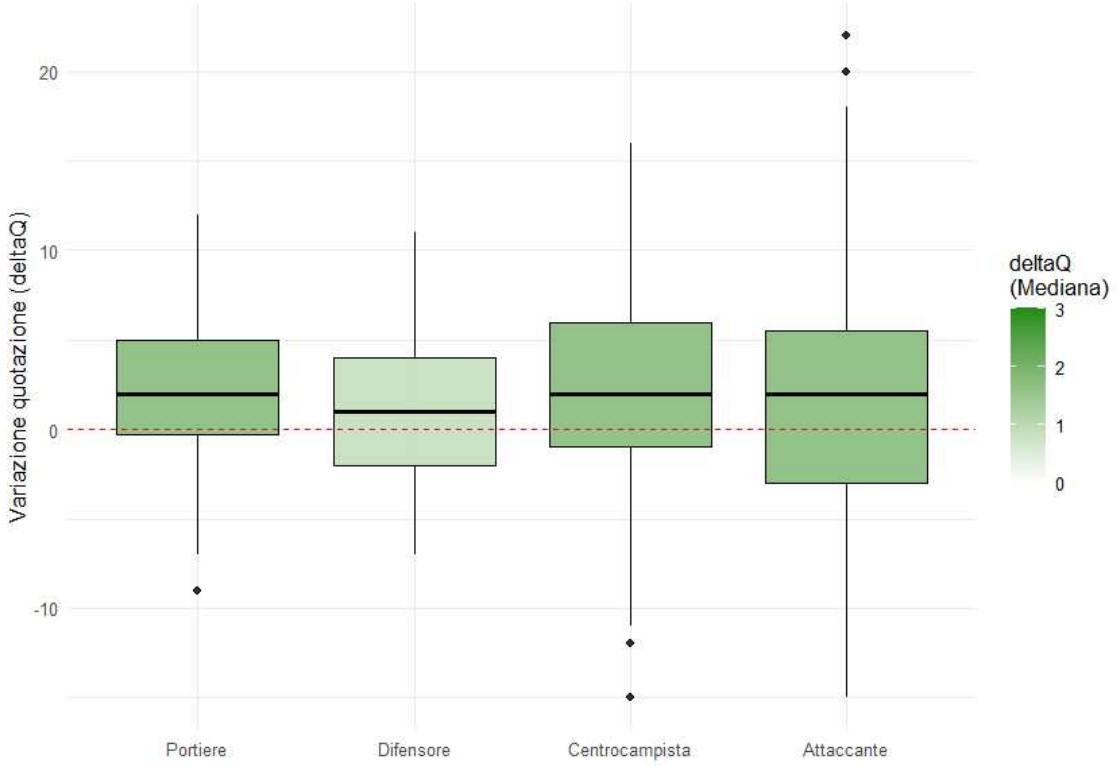

Figura 3.12: Variazione quotazione per ruolo

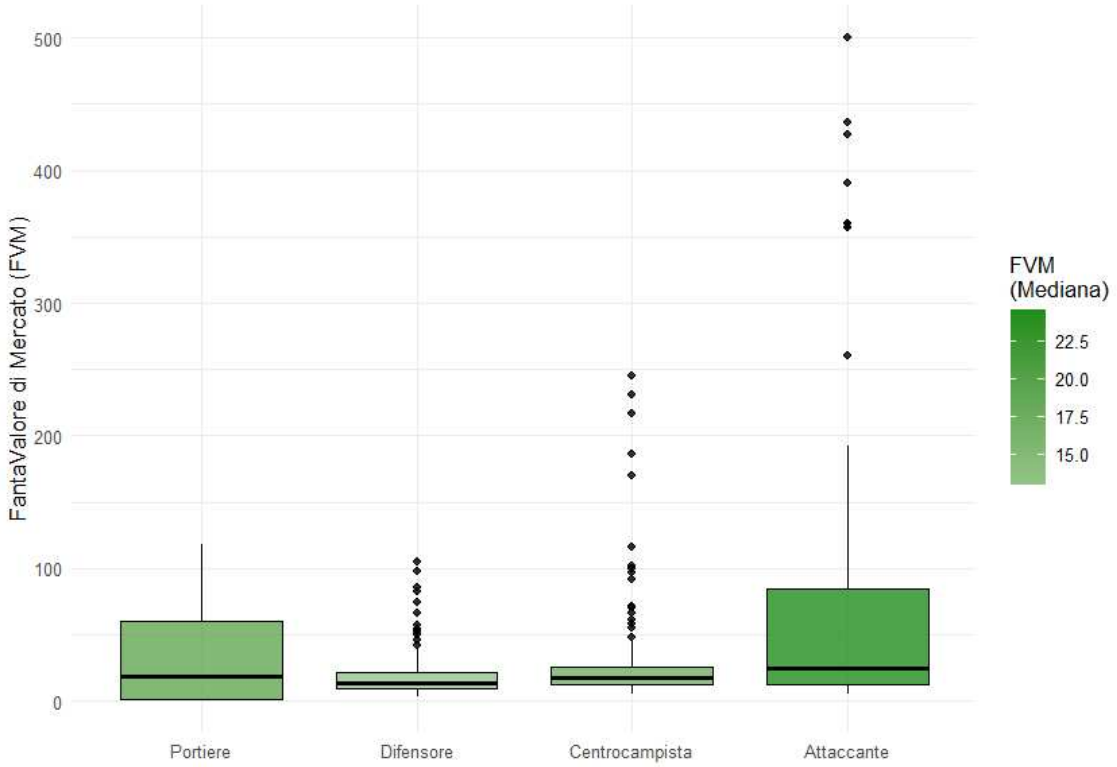

Figura 3.13: FantaValore di Mercato per ruolo

3.5 Analisi del contesto collettivo e impatto del sistema squadra

L'ultima fase dell'analisi prestazionale si sposta dalla dimensione individuale a quella collettiva, quantificando come l'ambiente tattico e i risultati del club agiscano da acceleratori o freni per le statistiche dei singoli calciatori.
Il rapporto tra la produzione offensiva totale di un team e la quota spettante al suo miglior realizzatore (Figura 3.14) permette di mappare il grado di dipendenza tattica di ogni formazione. Squadre che occupano il quadrante in alto a sinistra del grafico, come il Lecce con Nikola Krstovic o il Cagliari con Roberto Piccoli, mostrano una centralizzazione estrema della pericolosità, dove il leader offensivo si assume oltre il 30% degli Expected Goals dell'intera squadra. Questa condizione evidenzia contesti che faticano a generare soluzioni alternative, rendendo il rendimento del singolo estremamente vulnerabile allo stato di forma del club.
Al contrario, i top club come Inter e AC Milan si posizionano nel quadrante in basso a destra, caratterizzato da volumi di gioco elevati ma distribuiti in modo corale. Nonostante Lautaro Martínez e Christian Pulisic rimangono i riferimenti principali, la loro incidenza percentuale è più contenuta rispetto ai colleghi delle squadre minori perché inserita in un sistema capace di produrre minacce da più fronti, garantendo una maggiore stabilità ai bonus stagionali.
Da qui potremmo dedurre che i giocatori di squadre centralizzate, al di fuori del proprio capocannoniere, avranno un prezzo relativamente più basso, in quanto meno propensi a bonus; ragionamento inverso per le squadre più corali.

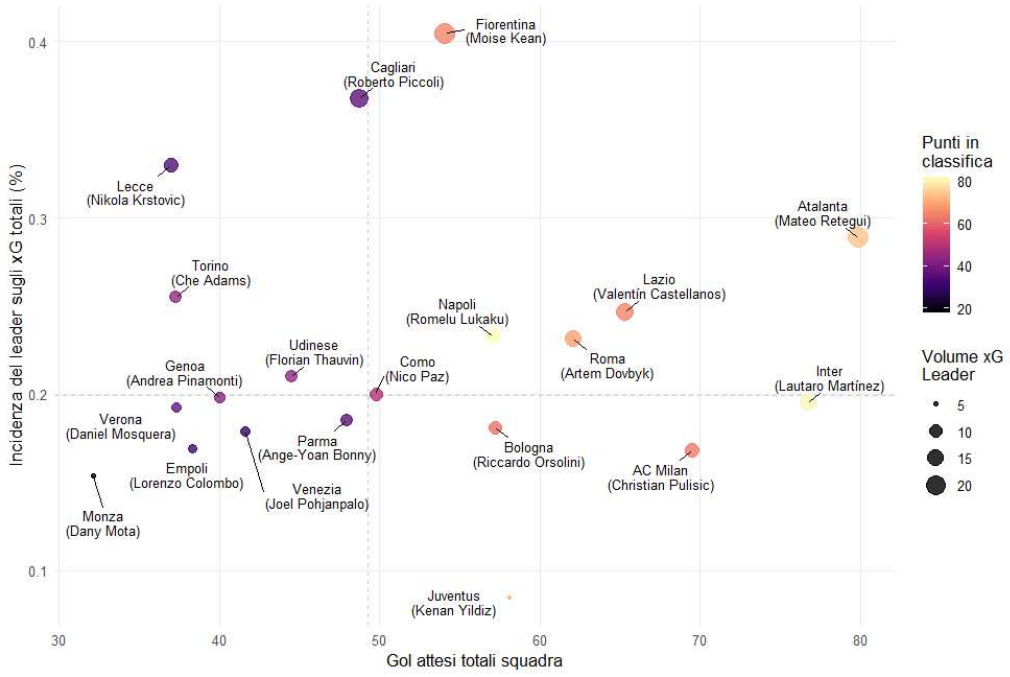

Figura 3.14: Incidenza degli xG del leader sulla squadra

La capacità di un club di raccogliere punti rispetto al volume di gioco prodotto, espressa dalla metrica del *deltaPTS* (Figura 3.15), definisce il perimetro entro cui si muovono le prestazioni dei singoli interpreti. Osservando la distribuzione dei club rispetto alla retta di regressione, squadre come Napoli e Fiorentina emergono per la spiccata abilità nel capitalizzare le occasioni create, posizionandosi stabilmente nell'area di sovraperformance del grafico. All'estremo opposto, Monza, Venezia ed Empoli evidenziano una mancata capacità di portare dalla loro parte i risultati attesi; squadre che di fatto sono retrocesse al termine del campionato.

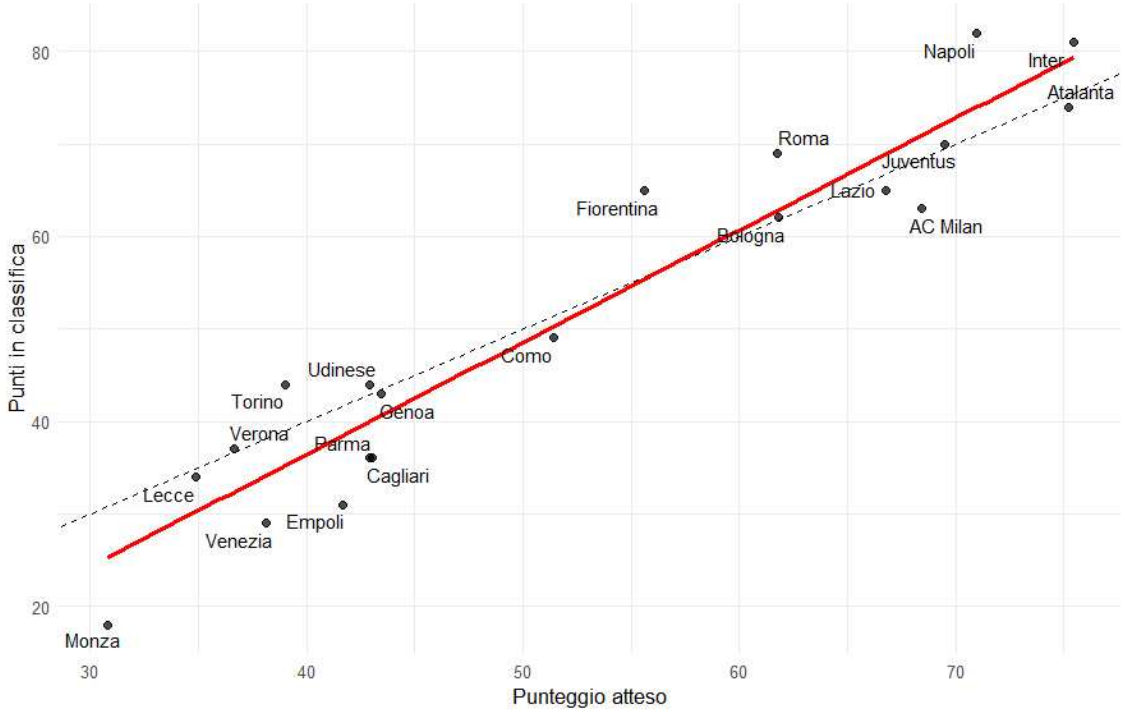

Figura 3.15: Punti guadagnati vs punti attesi

Questo posizionamento collettivo sembra generare un effetto trascinamento sulla MediaVoto, dove il giudizio dei pagellisti risulta pesantemente condizionato dai risultati della squadra (Figura 3.16). I calciatori appartenenti ai club caratterizzati da un alto *deltaPTS* godono infatti di distribuzioni di voto sensibilmente spostate verso l'alto; in particolare, si può notare come Fiorentina e Ac Milan, rispettivamente con importanti rendimenti superiori e inferiori alle attese, ma con un punteggio a fine stagione molto simile, differiscano notevolmente nella loro MediaVoto mediana, trovandosi addirittura in parti opposte rispetto alla soglia critica.

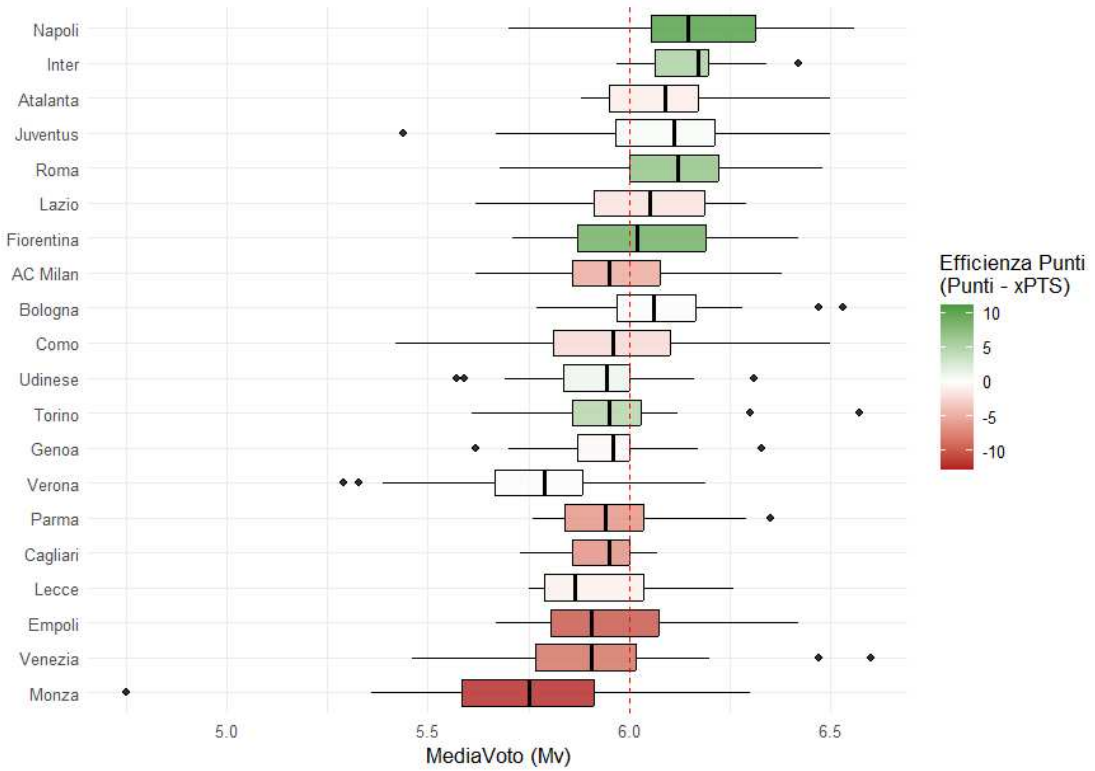

Figura 3.16: MediaVoto per squadra

Un aspetto controintuitivo dell'analisi economica riguarda il fatto che le squadre capaci di una stagione d'élite in termini di punti non sempre garantiscono una rivalutazione fanta-monetaria dei propri interpreti (Figura 3.17). Al contrario, si osserva spesso un calo generale nelle quotazioni dei calciatori appartenenti ai top club, mentre le rose delle squadre di bassa classifica o retrocesse mostrano una variazione mediana positiva. Questo paradosso deriva dal fatto che i prezzi di partenza per i giocatori delle "grandi" sono già estremamente elevati e scontano in anticipo le massime aspettative di rendimento, lasciando poco margine per ulteriori incrementi. Al contrario, chi gioca in contesti meno nobili parte da basi d'asta minime, rendendo molto più semplice generare una plusvalenza anche a fronte di prestazioni collettive deficitarie.

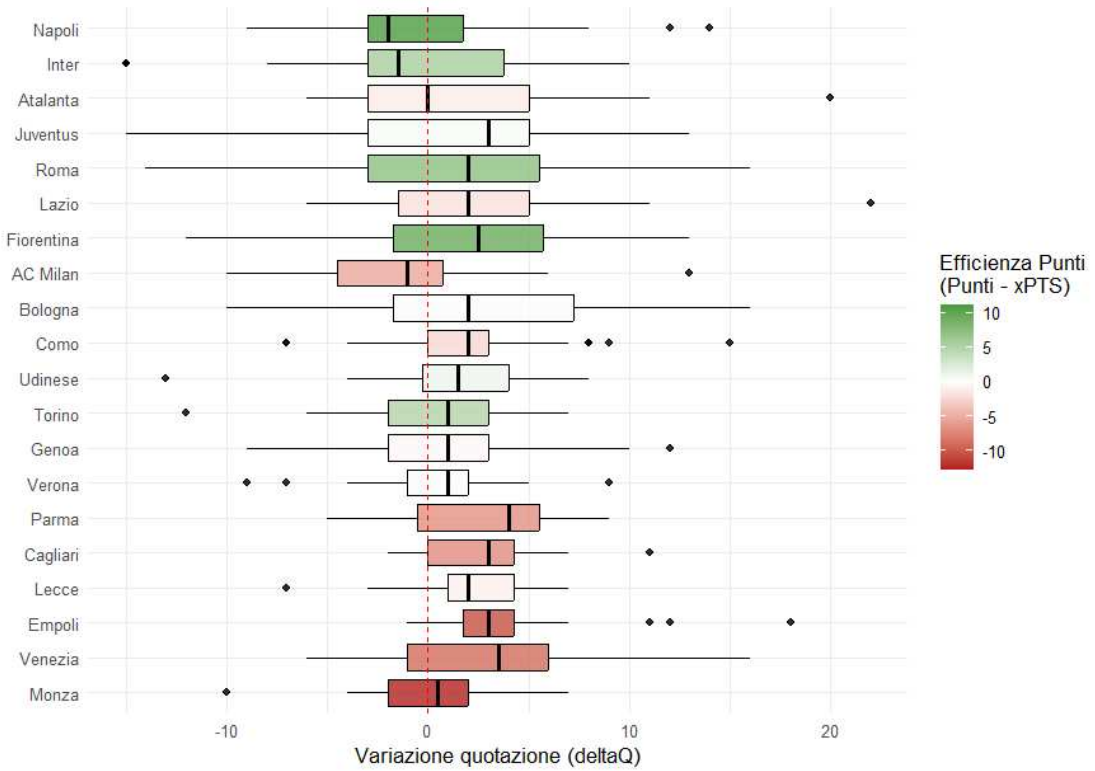

Figura 3.17: Delta quotazione per squadra

Approfondendo la distribuzione del log-FantaValore di Mercato (Figura 3.18) per ogni team, emerge una distorsione riconducibile al fattore "grande città" e al peso delle rispettive tifoserie. I calciatori di Inter e Ac Milan risultano sensibilmente più costosi rispetto alle prestazioni effettive se confrontati con i colleghi di Atalanta o Fiorentina, evidenziando una sorta di premio sul prezzo dettato dal blasone e dalla domanda mediatica. È interessante notare come le distribuzioni di Atalanta e Fiorentina siano fortemente influenzate verso l'alto dai propri leader offensivi, i quali agiscono come outlier trascinando le medie di reparto, mentre il resto della rosa mantiene quotazioni più razionali e accessibili rispetto ai giganti milanesi. Tale evidenza suggerisce che il mercato virtuale tenda a sovrastimare sistematicamente i giocatori delle squadre con più tifosi, creando inefficienze che possono essere sfruttate puntando su rose altrettanto competitive ma meno esposte alla pressione del marchio.

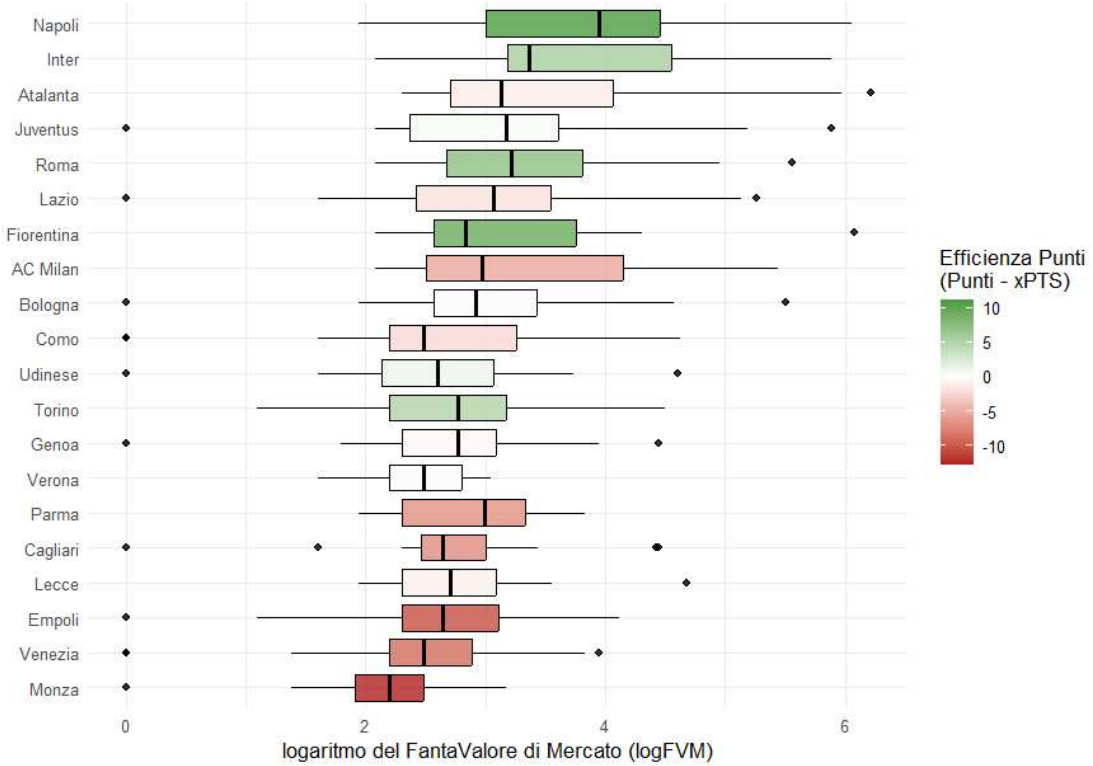

Figura 3.18: log-FantaValore di Mercato per squadra

3.6 Sintesi delle Interdipendenze

L'analisi esplorativa si chiude con la matrice di correlazione clusterizzata (Figura 3.19) che sintetizza visivamente l'intera rete di relazioni tra performance atletica, contesto tattico e valutazioni di mercato. Il raggruppamento gerarchico delle variabili evidenzia tre macro-blocchi fondamentali che guideranno la fase di modellazione. Il primo blocco conferma la coerenza tra produzione attesa e realizzata mentre il secondo mostra il legame indissolubile tra il successo del club e il valore economico dei singoli calciatori. Infine le aree di correlazione negativa associate alle prestazioni positive di squadra e ai gol subiti chiudono il quadro diagnostico.

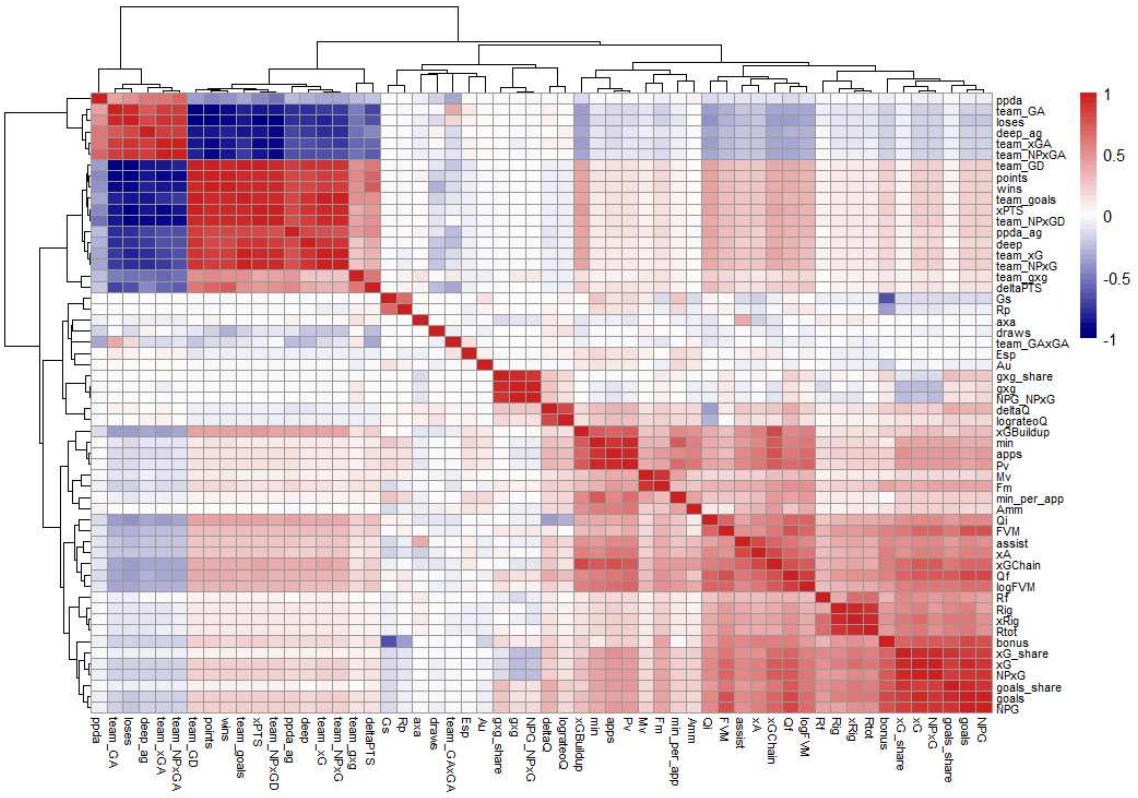

Figura 3.19: Matrice di correlazione (clustered)

Sulla base della matrice di correlazione ho definito un nuovo set di variabili per mitigare la collinearità e isolare i segnali più robusti. Ho scelto di escludere le macro-variabili di squadra ridondanti come i punti o i gol subiti privilegiando metriche avanzate più stabili tra cui gli Expected Goals collettivi, il deltaPTS e l'indice di pressing PPDA. Per quanto riguarda la performance individuale ho sostituito il dato grezzo dei gol totali con una scomposizione più informativa che separa il volume di occasioni create (NPxG) dall'efficacia realizzativa (gxg) e dalla gestione dei rigori. Nonostante questa selezione, permangono alcune sovrapposizioni tra variabili che verranno comunque approfondite durante la fase di modellazione nel prossimo capitolo. Infine, ho mantenuto temporaneamente alcuni parametri soggetti a data leakage al solo scopo di istruire un modello baseline di riferimento e allo studio di diverse variabili risposta.

4 Modellazione

Il presente capitolo segna il passaggio dall'analisi descrittiva alla costruzione dei modelli, con l'obiettivo di formalizzare in termini quantitativi le prestazioni in chiave di voto medio e le dinamiche che regolano la variazione di valore dei calciatori nel mercato simulato del Fantacalcio. Attraverso l'applicazione della regressione lineare multipla (OLS), si intende isolare l'impatto marginale delle singole metriche di performance, sia tradizionali che avanzate, sulla fluttuazione delle quotazioni e sulla media voto dei giocatori.
La modellazione è stata sviluppata a partire dal dataset della stagione 2024/25, precedentemente filtrato escludendo i portieri e i giocatori con minutaggio limitato, così da ridurre la variabilità non informativa. Sono state considerate quattro variabili dipendenti, corrispondenti alla variazione assoluta di quotazione, alla crescita percentuale in forma logaritmica, al logaritmo del FantaValore di Mercato e alla Media Voto. Le specificazioni includono indicatori di produzione offensiva, scarti tra rendimento reale e atteso, quota di partecipazione alla manovra della squadra, minutaggio, ruolo e quotazione iniziale, oltre ad alcune interazioni tra centralità offensiva e ruolo.

4.1 Analisi della variazione di quotazione assoluta e percentuale

Partendo dalla quotazione, l'obiettivo è spiegare come e perché il prezzo di un giocatore cambi durante la stagione. Inizialmente, l'approccio metodologico prevedeva lo sviluppo di due modelli indipendenti: uno dedicato alla variazione assoluta in crediti ($deltaQ$) e uno rivolto alla redditività percentuale ($lograteoQ$). Tuttavia, il confronto tra le performance statistiche ha evidenziato una netta superiorità del modello assoluto, capace di spiegare l'85,4% della varianza totale rispetto al più modesto 49,1% del modello log-percentuale (Tabella **??**).
Questa marcata differenza nella capacità esplicativa è riconducibile alla natura intrinsecamente più rumorosa delle variazioni relative. Le percentuali di crescita sono infatti estremamente sensibili ai valori estremi e alle speculazioni su profili a basso costo, elementi che introducono una volatilità difficilmente catturabile da un modello lineare tradizionale. Per garantire la massima solidità scientifica e una coerenza

matematica interna all'elaborato, si è scelto di eleggere il modello *deltaQ* come unico pilastro predittivo. La variazione percentuale (*RateoQ*) viene pertanto derivata in seconda istanza tramite una trasformazione deterministica basata sul rapporto tra il delta stimato e la quotazione iniziale (Q_i).

Variabile	delta Q.	log-rapporto di Q.
(Intercept)	-44.036*** (5.684)	-5.893*** (1.266)
axa	0.452*** (0.096)	0.039. (0.020)
xA	0.523*** (0.093)	0.056** (0.020)
min	0.001*** (0.000)	0.000*** (0.000)
Mv	6.712*** (0.998)	1.061*** (0.223)
goals_share	38.376*** (6.027)	3.158*** (0.749)
team_xG	0.175*** (0.032)	0.005* (0.002)
Qi	-0.791*** (0.057)	-0.091*** (0.008)
RC	4.590** (1.444)	-0.113 (0.083)
RD	5.807*** (1.447)	-0.258** (0.099)
R-quadrato	0.854	0.491

Tabella 4.1: Confronto dei coefficienti: variazione assoluta vs percentuale

L'analisi dei coefficienti del modello sulla variazione assoluta (Tabella **??**) mette in luce dinamiche di mercato estremamente solide. Il minutaggio totale (0.001) e il peso del giocatore sui gol della propria squadra ($goals\_share = 38.376$) si confermano i fattori principali che spingono al rialzo il valore nominale. In sostanza, chi gioca con continuità e partecipa attivamente alla produzione offensiva vede la propria quotazione salire in modo costante. Un ruolo centrale è rivestito dalla Media Voto (6.712), la quale agisce come catalizzatore del valore economico e valida l'ipotesi che il mercato del fantacalcio tenda a premiare la qualità della prestazione pura. Specularmente, il coefficiente relativo alla quotazione iniziale ($Q_i = -0.791$) risulta fortemente negativo e significativo. Questo dato conferma una regola logica fondamentale del mercato, per cui i calciatori che partono da valutazioni d'élite hanno margini di crescita assoluta molto più ridotti rispetto alle scommesse a basso costo. Si osserva quindi un fenomeno di saturazione della quotazione per i top player, le quali tendono a stabilizzarsi o a subire lievi correzioni verso la media quando non sono supportate da bonus eccezionali.

Per garantire la validità delle stime prodotte ho sottoposto il modello ai test diagnostici classici. Sebbene i residui (Figura 4.1) non presentino una distribuzione perfettamente normale si tratta di un esito ampiamente giustificabile all'interno di un dataset calcistico. La natura stessa della disciplina caratterizzata da variabili imprevedibili quali exploit improvvisi di giovani talenti o crisi di rendimento inaspettate genera inevitabilmente delle risposte anomale nei dati. Il problema statistico principale è tuttavia rappresentato dall'eteroschedasticità. Al fine di evitare che la varianza non costante degli errori potesse rendere inaffidabili i test di significatività ho ricalcolato gli errori standard attraverso lo stimatore robusto $HC1$. I risulta-

ti derivanti da questa correzione confermano la solidità dei regressori selezionati e consentono di procedere con sicurezza verso la fase di validazione operativa.

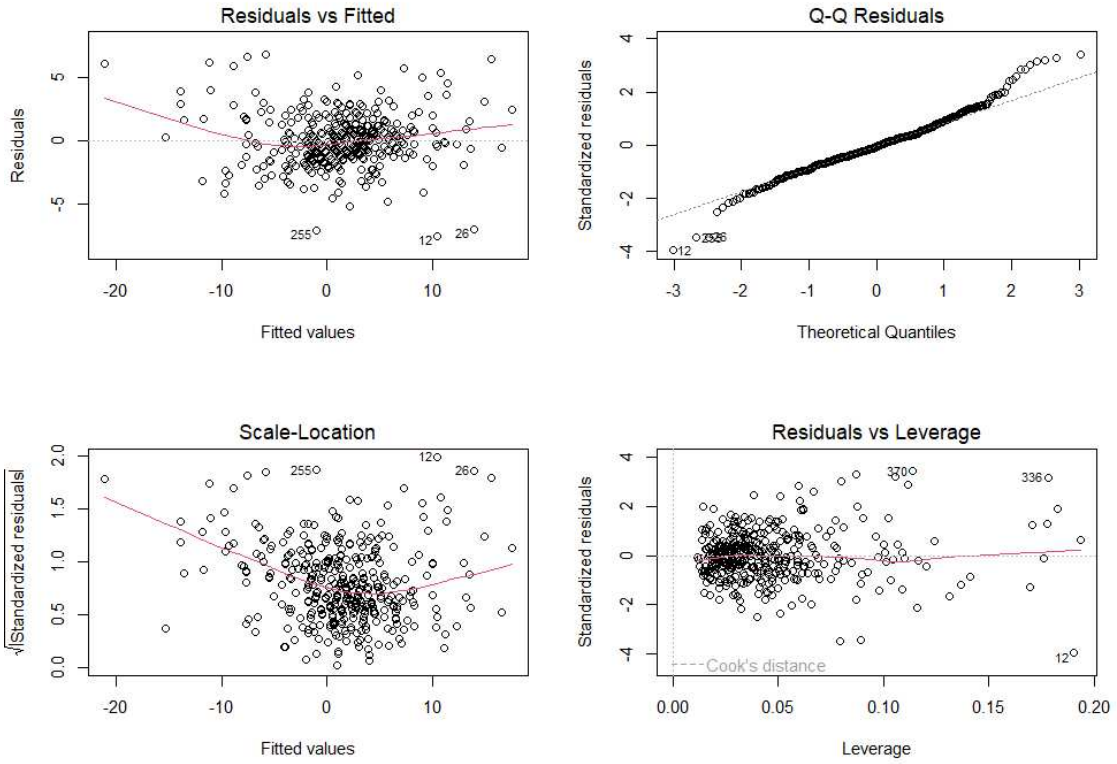

Figura 4.1: Diagnostica dei residui per il modello delta Q.

Dall'analisi grafica del modello *deltaQ* si nota la presenza di alcuni outlier sulle code del Q-Q Plot insieme a una leggera dispersione a imbuto nel grafico dei residui. Tali evidenze suggeriscono che mentre la parte centrale della distribuzione segue correttamente le assunzioni del modello lineare le fluttuazioni più estreme del mercato richiedano una cautela interpretativa maggiore.

4.2 Analisi del FantaValore di Mercato

Dopo aver analizzato le fluttuazioni delle quotazioni ufficiali, l'indagine si è spostata sul FantaValore di Mercato, un altro dato di natura editoriale, legato al prezzo d'asta di un calciatore su un budget standard di 1000 crediti.
La necessità di studiare l'FVM come variabile dipendente separata emerge chiaramente dall'analisi della matrice delle correlazioni (Tabella 4.2), che rivela dinamiche quasi antitetiche. Mentre la quotazione iniziale (Q_i) ha una correlazione fortemente negativa con la variazione di valore durante l'anno ($deltaQ = -0,459$), la stessa Q_i è legata all'FVM da un coefficiente positivo e molto solido ($0,684$).
Ci troviamo di fronte a un paradosso molto interessante, dato che i giocatori più costosi sono statisticamente quelli più soggetti a perdere valore nel tempo (confer-

mando la regressione verso la media discussa nel paragrafo precedente); eppure, il mercato editoriale continua a percepirli come asset di alto valore, come dimostrato dalla correlazione di 0,778 tra FantaValore di Mercato e quotazione finale (Q_f). Al contrario, il legame tra FVM e plusvalenza netta ($deltaQ$) è pressoché nullo (0,068). Questo suggerisce che l'FVM non sia minimamente influenzato dal potenziale di crescita della quotazione, ma sia quasi interamente guidato dal blasone del giocatore e dal suo rendimento atteso.

	Qf	Qi	deltaQ	FVM	Mv
Qf	1.000	0.669	0.353	0.778	0.708
Qi	0.669	1.000	-0.459	0.684	0.439
deltaQ	0.353	-0.459	1.000	0.068	0.294
FVM	0.778	0.684	0.068	1.000	0.488
Mv	0.708	0.439	0.294	0.488	1.000

Tabella 4.2: Matrice di correlazione tra le variabili di valore e rendimento

Proprio a causa di queste divergenze, i risultati del modello logaritmico (Tabella 4.3) evidenziano driver strutturalmente differenti rispetto a quelli che governano il $deltaQ$. Per questa analisi, ho scelto deliberatamente di escludere le quotazioni (Qi, Qf) dal set di variabili indipendenti. Inserirle avrebbe garantito un potere predittivo artificialmente elevato (data l'alta correlazione), ma avrebbe impedito di isolare il reale apporto delle metriche di rendimento. L'obiettivo è infatti far emergere il peso specifico delle variabili tecniche che verrebbero altrimenti oscurate dalla dominanza del prezzo storico.
Nonostante l'esclusione delle quotazioni, il modello mantiene un ottimo adattamento ai dati ($R^2 = 0,767$). L'impatto più massiccio è esercitato dalla Media Voto: il coefficiente di $1,373$ indica che, applicando la trasformazione ($e^{1,373} - 1$), un incremento unitario del voto medio è associato a un aumento del valore di mercato stimato del $294,7\%$. Questo dato conferma che l'FVM è, prima di tutto, un indicatore di qualità prestazionale costante.
Anche la centralità offensiva gioca un ruolo cruciale; difatti, la variabile goals_share ($5,553$) evidenzia come il peso del giocatore nella produzione di reti della squadra sia il fattore che più influisce sull'investimento percepito. Parallelamente, la metrica degli Expected Assists (xA: $0,072$) risulta statisticamente molto significativa, segno che il mercato tende a incorporare e prezzare la pericolosità potenziale e la capacità latente di creare occasioni da gol. Infine, i coefficienti negativi per centrocampisti ($RC : -0,184$) e difensori ($RD : -0,207$) confermano l'esistenza di un "premio di status" per gli attaccanti. A parità di rendimento, un attaccante manterrà sempre un valore superiore poiché l'FVM prezza la rarità della risorsa "bonus".

Variabile	Stima (Coeff.)	Std. Error ($HC1$)
(Intercept)	-6.450***	(1.148)
Mv (Media Voto)	1.373***	(0.204)
goals_share	5.553***	(0.689)
xA (Expected Assist)	0.072***	(0.019)
min (Minuti giocati)	0.0002***	(0.000)
team_xG	0.012***	(0.002)
RC (Centrocampista)	-0.184*	(0.092)
RD (Difensore)	-0.207*	(0.101)
R-quadrato	0.767	

Significatività: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; . $p < 0.1$

Tabella 4.3: Risultati del modello sul logaritmo del FantaValore di Mercato

Analogamente al modello precedente, la diagnostica dei residui (Figura 4.2) mostra una distribuzione tendenzialmente normale, con leggere deviazioni sulle code del Q-Q Plot. Essendo confermata una chiara eteroschedasticità dal grafico Scale-Location, anche per questo modello le stime riportate in tabella sono state calcolate ricorrendo agli errori standard robusti (correzione $HC1$), assicurando così la piena validità statistica dell'inferenza.

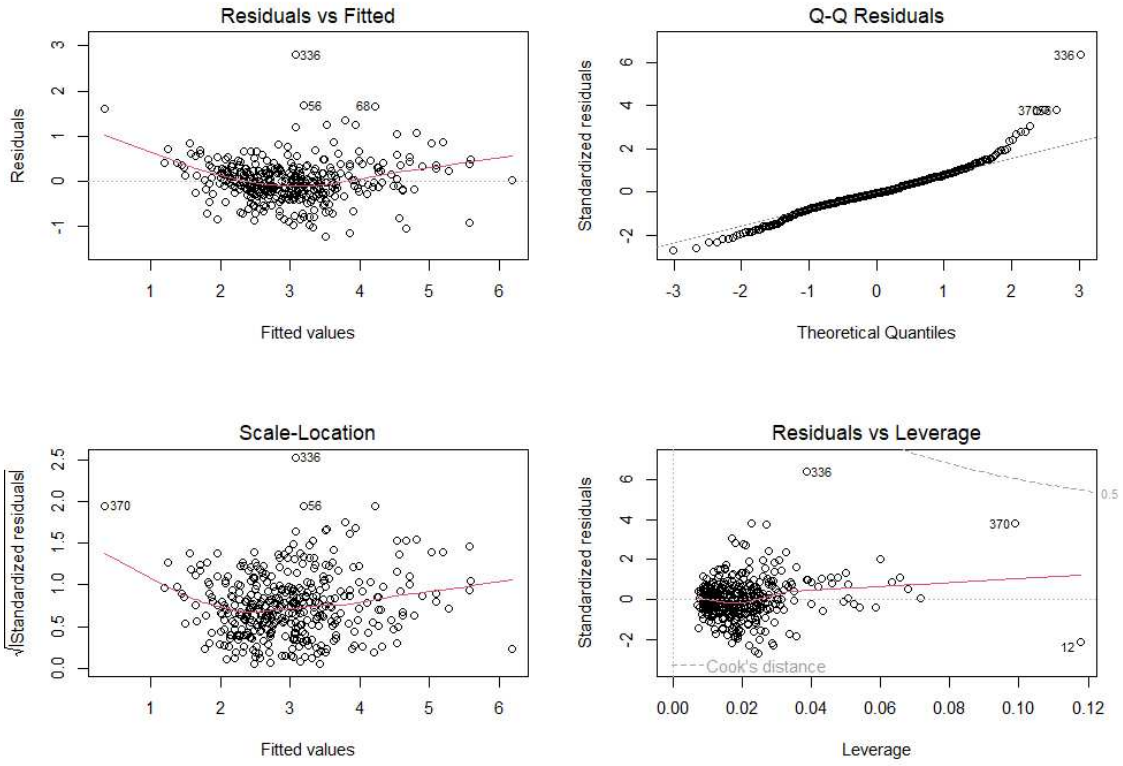

Figura 4.2: Diagnostica dei residui per il modello log-FantaValore di Mercato

4.3 Analisi della Media Voto

L'ultimo modello analizza le determinanti della Media Voto. Questa analisi parte con pretese a priori strutturalmente diverse rispetto a quelle sulle quotazioni. Il voto in pagella, infatti, dipende in larga misura dal giudizio soggettivo dei giornalisti e da innumerevoli micro-eventi in campo (dribbling, contrasti, posizionamento) che sfuggono alle metriche puramente offensive utilizzate in questo studio. Di conseguenza, era lecito attendersi un potere esplicativo inferiore. L'obiettivo non è prevedere il voto esatto, bensì quantificare quanto l'efficienza realizzativa e il contesto di squadra riescano a influenzare la valutazione soggettiva e la costanza di rendimento.
Come previsto, il modello spiega (Tabella 4.4) circa il 47% della varianza ($R^2 = 0,469$). L'intercetta, pari a 6,11, fissa una "base" di partenza perfettamente coerente con la classica sufficienza fantacalcistica. Analizzando i coefficienti, emerge chiaramente come i pagellisti premino la partecipazione attiva alla manovra offensiva ($xGChain$: $0,026$) e, soprattutto, l'efficienza rispetto alle aspettative: le overperformance sui gol (gxg: $0,041$) e sugli assist (axa: $0,018$) garantiscono incrementi di voto statisticamente molto significativi.Un aspetto particolarmente interessante riguarda l'impatto del contesto tattico. Il coefficiente negativo dell'intensità difensiva ($ppda$: $-0,027$) indica che un valore basso di questa metrica, sintomo di un pressing di squadra molto aggressivo, è associato a voti più alti, suggerendo che i giocatori militanti in squadre con un approccio proattivo ricevano valutazioni mediamente migliori. Tra i fattori penalizzanti, spicca il forte impatto negativo degli autogol (Au: $-0,074$), mentre il coefficiente dei difensori (RD: $-0,057$) conferma una lieve, ma strutturale, maggiore severità di giudizio verso il reparto arretrato rispetto agli attaccanti.

Variabile	Stima (Coeff.)	Std. Error (HC1)
(Intercept)	6.111***	(0.078)
xGChain	0.026***	(0.003)
xGBuildup	-0.011*	(0.005)
ppda (Intensità Pressing)	-0.027***	(0.005)
deep (Passaggi chiave)	0.0003*	(0.000)
Au (Autogol)	-0.074*	(0.031)
gxg (Overperformance Gol)	0.041***	(0.006)
axa (Overperformance Assist)	0.018**	(0.006)
team_GAxGA	-0.006**	(0.002)
RC (Centrocampista)	0.016	(0.023)
RD (Difensore)	-0.057*	(0.027)
R-quadrato	0.469	

Significatività: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; . $p < 0.1$

Tabella 4.4: Risultati del modello sulla Media Voto

Analogamente alle analisi precedenti, l'ispezione grafica dei residui ha confermato la presenza di eteroschedasticità. Per tale motivo, al fine di garantire l'affidabilità

dell'inferenza statistica, le stime e i livelli di significatività riportati in (Tabella 4.4) sono stati calcolati applicando gli errori standard robusti.

4.4 Validazione dei Modelli

Per valutare le capacità predittive dei modelli e la loro robustezza, devo accertarmi che l'alto potere esplicativo sul set di addestramento non derivi da *overfitting*, ovvero che non abbia inglobato anche il rumore del campione, imparando a memoria i dati. Per escludere questa eventualità, la fase di validazione è stata strutturata in due step.
In primis, è stata applicata una *Cross-Validation* a k *fold* (con $k = 10$) all'interno dello stesso set di dati. Matematicamente, questa tecnica suddivide casualmente il dataset in 10 sottoinsiemi (fold) di eguale dimensione. Il modello viene addestrato iterativamente su 9 fold e testato sul decimo fold escluso, calcolando l'errore di previsione. Il processo viene ripetuto k volte, permettendo a ogni fold di fungere da test set esattamente una volta. La media degli errori calcolati fornisce una stima non distorta delle performance del modello su dati non visti.
Tuttavia, il contesto calcistico è per sua natura un ambiente caratterizzato da un elevatissimo grado di rumore intrinseco. Solo testando l'algoritmo su un contesto temporale estraneo alla fase di stima è possibile accertare la reale generalizzabilità delle logiche individuate; per questo motivo, ho preferito effettuare un secondo controllo con la validazione out-of-sample su una stagione calcistica diversa (Serie A 2023/24).

4.4.1 Validazione interna: k-fold Cross-Validation

Per la validazione interna, tutti e quattro i modelli di base sono stati ri-addestrati utilizzando la 10-fold Cross-Validation. Per valutare la bontà degli adattamenti, sono state estratte tre metriche standard: la Radice dell'Errore Quadratico Medio ($RMSE$), l'Errore Assoluto Medio (MAE) e il coefficiente di determinazione ricalcolato (R^2).

Modello	RMSE	R^2 (CV)	MAE
deltaQ (Variazione assoluta)	2.208	0.838	1.663
logFVM (FantaValore Mercato)	0.369	0.851	0.283
Mv (Media Voto)	0.164	0.456	0.122

Tabella 4.5: Risultati della 10-fold Cross-Validation sui quattro modelli in esame

I risultati ottenuti (Tabella 4.5) offrono riscontri molto positivi e confermano l'assenza di overfitting. I valori di R^2 generati dalla Cross-Validation sono infatti pressoché identici a quelli emersi dalla regressione OLS originaria. Il modello sulla variazione

assoluta si conferma estremamente solido, spiegando quasi l'84% della varianza su dati incrociati, con un errore medio assoluto (MAE) di circa 1,66 unità che, considerando l'alta volatilità delle quotazioni, è un margine di errore decisamente contenuto. Un'attenzione particolare merita l'interpretazione del MAE per il modello con variabile dipendente trasformata ($logFVM$). In questi casi, il MAE non indica un errore assoluto, ma riflette uno scarto logaritmico. Applicando la trasformazione esponenziale ($e^{MAE} - 1$), è possibile interpretare questo valore come un errore percentuale medio. Ad esempio, il MAE di 0,283 per il modello sul FantaValore di Mercato (che vanta un ottimo R^2 di 0,851) si traduce in un errore medio di previsione di circa il $32,7\%$: un margine del tutto fisiologico, considerando l'estrema componente soggettiva ed emotiva che regola le dinamiche delle aste.
Infine, per quanto riguarda la Media Voto, pur mantenendo un R^2 atteso più basso (0,456), il MAE si attesta a soli 0,122 punti. Questo significa che, in media, il modello è in grado di stimare il voto di un calciatore sbagliando di poco più di un decimo di punto rispetto alla pagella reale, confermando come le metriche di efficienza offensiva e il contesto di squadra catturino una fetta sostanziale del giudizio soggettivo dei giornalisti.

4.4.2 Validazione Out-of-Sample: test sulla stagione precedente

I modelli, regrediti e calibrati sui dati della stagione 2024/25, sono stati utilizzati per prevedere i valori di un dataset completamente estraneo alla fase di addestramento, costituito dai dati storici della stagione 2023/24. Il dataset storico è stato sottoposto alle medesime procedure di pulizia e filtraggio (esclusione dei portieri, minutaggio superiore a 300 minuti e almeno 3 presenze a voto). Per valutare le performance predittive su questo nuovo ambiente, sono stati calcolati l'RMSE (Radice dell'Errore Quadratico Medio) e l'indice di correlazione lineare di Pearson tra i valori reali osservati nella passata stagione e quelli previsti dall'algoritmo.
I risultati ottenuti (Tabella 4.6) certificano in modo inequivocabile la robustezza dell'impianto teorico.

Modello	RMSE (Out-of-Sample)	Correlazione
deltaQ (Variazione assoluta)	2.422	0.898
logFVM (log-FantaValore Mercato)	0.503	0.859
Mv (Media Voto)	0.161	0.670

Tabella 4.6: Risultati della validazione Out-of-Sample sulla stagione 2023/24

Il risultato più rilevante riguarda la variazione assoluta: la straordinaria correlazione di 0,898 e un errore medio di soli 2,4 crediti dimostrano che i driver individuati costituiscono regole di valutazione strutturali e non anomalie di una singola stagione. Il

modello sul log-FantaValore di Mercato mostra eccellenti capacità di generalizzazione, inquadrando fedelmente le gerarchie d'asta con una correlazione che sale a 0,859. Tuttavia, l'RMSE di 0,503 su scala logaritmica si traduce in un errore percentuale medio di circa il 65% ($e^{0,503} - 1$). Si tratta di un margine di errore considerevole, che evidenzia i limiti del modello nel catturare fattori esogeni per prevedere i valori assoluti.
Anche il modello sulla Media Voto conferma la propria stabilità. L'RMSE out-of-sample (0,161) ricalca fedelmente l'errore della validazione interna. Pur con una correlazione di 0,670, fisiologicamente limitata dalla soggettività delle pagelle, l'algoritmo stima il voto reale con uno scarto di soli 0,16 punti. Le metriche oggettive tracciano quindi una solida baseline valutativa. In conclusione, il test out-of-sample certifica che le dinamiche individuate non derivano da overfitting, ma rappresentano regole statistiche in grado di generalizzare su stagioni diverse.

4.4.3 Previsione intra-stagionale

L'impostazione di un modello predittivo intra-stagionale richiede una rigorosa attenzione alla sequenzialità temporale per scongiurare il fenomeno del *Data Leakage* (contaminazione dei dati), noto in letteratura anche come *look-ahead bias* (distorsione anticipatrice). Questo grave errore metodologico si verifica qualora si inseriscano nel set di informazioni a disposizione del modello delle variabili che, nel mondo reale, non sarebbero ancora note al momento della previsione. Nel contesto simulato dell'asta di riparazione invernale, l'utilizzo della Media Voto reale di fine campionato come regressore per stimare le quotazioni future comporterebbe un inquinamento dei dati, poiché si sfrutterebbe un'informazione non ancora osservata per prevedere il futuro stesso, invalidando di fatto l'intera architettura predittiva.
Per ovviare a questo limite strutturale e mantenere la solidità dell'analisi, si è adottato un approccio a cascata basato sull'utilizzo di regressori generati (*Generated Regressors* [11]). La procedura si articola in due stadi sequenziali. In un primo momento, si elabora una stima della Media Voto attesa ($\hat{Mv}$) impiegando esclusivamente le metriche tattiche e di efficienza offensiva rilevate alla data di chiusura del campione. Successivamente, questa variabile generata viene sostituita alla variabile reale e impiegata come regressore nei due modelli economici principali, delegati a prevedere la variazione assoluta di quotazione e il FantaValore di Mercato.
Dal punto di vista statistico, l'impiego di una stima empirica al posto di un dato osservato introduce inevitabilmente una componente di rumore aggiuntivo, poiché l'incertezza fisiologica del primo modello si propaga a cascata sui successivi. In un contesto di pura inferenza econometrica, tale procedura richiederebbe complesse correzioni degli errori standard, come l'applicazione dello stimatore di *Murphy-Topel*, per non sovrastimare la significatività dei p-value.
Questo approccio replica fedelmente le condizioni operative di un fantallenatore du-

rante il campionato. Poiché il rendimento finale di un calciatore non è noto in anticipo, il modello deve basarsi su una stima prudenziale fondata sulle prestazioni rilevate fino a quel momento. L'incertezza derivante da questa previsione parziale viene inclusa correttamente nei risultati finali, garantendo che le valutazioni sull'errore medio e sulla capacità predittiva siano realistiche e prive di distorsioni temporali.

Modello	RMSE (Previsione Finale)	Correlazione
deltaQ (Variazione assoluta)	7.000	0.566
logFVM (log-FantaValore Mercato)	1.086	0.662
Mv (Media Voto)	0.212	0.622

Tabella 4.7: Risultati della previsione precoce a due stadi

Come ampiamente preventivabile, l'adozione di un regressore generato ($\hat{Mv}$) comporta una flessione delle performance complessive (Tabella 4.7). Scontare l'incertezza fisiologica di una stima all'interno di un'altra stima restituisce un quadro predittivo più conservativo, ma metodologicamente inattaccabile. Il modello sulla variazione assoluta registra un errore medio di 7 crediti e una correlazione di 0,566. Pur perdendo precisione puntuale, l'algoritmo mantiene la capacità di intercettare i macro-trend di mercato utili per orientare le scelte strategiche all'asta di riparazione.
L'impatto maggiore della correzione si riversa strutturalmente sul log-FantaValore di Mercato. La correlazione di 0,662 certifica la tenuta delle gerarchie complessive, ma l'RMSE di 1,086 si traduce in un errore percentuale ($e^{1.086} - 1$) vicino al 196%. Questo dato sancisce definitivamente come il valore editoriale assoluto sia una metrica fortemente esogena, plasmata da logiche inflazionistiche e mediatiche impossibili da prevedere con esattezza a mesi di distanza tramite le sole statistiche parziali.
Infine, il modello sulla Media Voto, fungendo da primo stadio della cascata, mantiene intatte le proprie proprietà predittive con un errore di soli 0,212 punti. Validato l'impianto teorico e accertata l'assenza di distorsioni temporali, i modelli sono ora pronti per la loro reale applicazione operativa.

4.5 Applicazione operativa: previsioni per la stagione 2025/26

L'atto conclusivo della ricerca consiste nell'applicazione dell'architettura a due stadi sui dati della stagione in corso, aggiornati alla 26ª giornata. Utilizzando i coefficienti validati nella sezione precedente, l'obiettivo è generare una proiezione della variazione assoluta e relativa finale attesa per maggio 2026. Tali proiezioni assumono un valore strategico per la gestione della rosa nella fase finale del campionato, poiché permettono di individuare potenziali "plusvalenze latenti" (giocatori il cui valore è

destinato a salire drasticamente) o, al contrario, di identificare atleti in una fase di sopravvalutazione statistica. In primo luogo, mostro quali sono i giocatori con la maggiore differenza assoluta, positiva e negativa, della quotazione.
L'analisi della variazione assoluta evidenzia una netta polarizzazione geografica e tecnica. Tra i profili con i maggiori incrementi previsti (Tabella 4.8) spiccano i blocchi di Inter (Bonny, Zielinski, Esposito) e Como (Douvikas, Nico Paz), segno di una produzione offensiva corale che il mercato non ha ancora pienamente prezzato. Si possono riconoscere però dei limiti nel modello; per esempio, nei casi più estremi sottostima la crescita, tranne nel caso di Nico Paz, forse dando un eccessivo peso alla media voto attesa o non tenendo in corretta considerazione la quotazione di partenza.
Anche nella lista delle maggiori svalutazioni attese (Tabella 4.9), il modello è più parsimonioso rispetto alla realtà, tranne per Gimenez, giocatore di un top club e con una media voto attesa bassa, a confermare le ipotesi precedenti. In generale, le maggiori svalutazioni sono dominate da attaccanti di prima fascia militanti in squadre come Milan, Fiorentina e Lazio (Giménez, Kean, Dia). Questa dinamica suggerisce che, partendo da quotazioni iniziali molto elevate, tali profili subiscono una violenta regressione verso la media qualora il rendimento realizzativo non sostenga i volumi di spesa iniziali, portando a una rapida erosione della quotazione.
L'unica vera divergenza si riscontra in **Kephren Thuram** , il quale ha mantenuto la propria quotazione stabile nonostante una proiezione fortemente negativa. Tale discrepanza suggerisce che per alcuni profili di squadre "top" intervengano fattori di resistenza alla svalutazione non puramente statistici, come il prestigio del club o la garanzia della titolarità, elementi che l'attuale configurazione del modello tende a sottopesare a favore dei dati di rendimento puro.

Giocatore	team	R	ΔQ Attesa	ΔQ Oss.	Mv Att.
Piotr Zielinski	Inter	C	+13.70	+21	6.31
Tasos Douvikas	Como	A	+12.20	+16	6.34
Federico Bonazzoli	Cremonese	A	+11.60	+18	6.02
Nico Paz	Como	C	+10.90	+7	6.46
Ange-Yoan Bonny	Inter	A	+10.60	+11	6.28
Ismaël Koné	Sassuolo	C	+9.58	+10	6.06
Ruslan Malinovskiy	Genoa	C	+9.53	+9	6.08
Stefano Moreo	Pisa	A	+8.90	+12	6.05
Caleb Ekuban	Genoa	A	+8.18	+8	6.00
Federico Dimarco	Inter	D	+7.80	+14	6.59

Tabella 4.8: Top 10 incrementi di quotazione previsti (Stima al 27 Febbraio 2026)

Giocatore	team	R	ΔQ Att.	ΔQ Oss.	Mv Att.
Santiago Giménez	AC Milan	A	-17.00	-13	5.86
Ademola Lookman	Atalanta	A	-15.70	-15	6.09
Kephren Thuram	Juventus	A	**-14.40**	**+1**	6.08
Moise Kean	Fiorentina	A	-13.40	-14	5.99
Mattia Zaccagni	Lazio	C	-11.60	-13	5.97
Loïs Openda	Juventus	A	-11.50	-15	6.02
Paulo Dybala	Roma	A	-11.20	-13	6.18
Boulaye Dia	Lazio	A	-11.00	-16	5.85
Dusan Vlahovic	Juventus	A	-10.30	-11	6.04
Denzel Dumfries	Inter	D	-10.00	-10	6.14

Tabella 4.9: Top 10 decrementi di quotazione previsti (Stima al 27 Febbraio 2026)

Passando alle variazioni relative (Tabella 4.10), il focus si sposta sull'individuazione di giocatori non tipicamente di principale interesse per il pubblico dominio, ma che nel corso della stagione sono diventati nomi da tenere in considerazione. Per massimizzare l'utilità di questa indagine, la valutazione della variazione relativa è stata segmentata in due set basati sul valore iniziale. Nel primo troviamo i giocatori con quotazione iniziale inferiore o uguale a 5, quelle che potrebbero essere definite "Scommesse" vinte; nel secondo, invece, i profili che comunemente vengono già presi in considerazione dai fantallenatori e che hanno registrato un grande incremento di valore percentuale. Questa distinzione permette di separare le anomalie statistiche estreme dai reali consolidamenti di mercato.
Nel gruppo delle consolidazioni spicca l'attacco del Cagliari grazie alle performance di Semih Kiliçsoy e Sebastiano Esposito i quali hanno generato incrementi percentuali superiori al cento per cento. Risulta tuttavia evidente un pattern di conservatorismo statistico poiché per quasi tutti i profili analizzati in questa fascia il rateo atteso a fine stagione si posiziona su livelli sensibilmente più bassi rispetto a quello già osservato alla ventisettesima giornata. Questa discrepanza suggerisce che l'algoritmo stia operando una sottostima strutturale o che preveda un fisiologico rallentamento delle performance nella fase finale del campionato. La fluttuazione della quotazione tende spesso a sovrastimare il periodo positivo di forma di un giocatore, mentre il modello fondandosi su basi regressive sembra anticipare una stabilizzazione dei prezzi verso valori più cauti una volta riassorbito il rumore dell'exploit recente. Spostando lo sguardo sul cluster delle scommesse il fenomeno della sottostima appare ancora più marcato. Numerosi calciatori con una quotazione iniziale minima hanno registrato crescite esplosive nonostante una media voto attesa che si attesta su valori prossimi alla sufficienza o poco al di sotto. Profili come Tiago Gabriel o Toma Basic mostrano incrementi reali che superano abbondantemente le proiezioni dell'algoritmo indicando che la variabile della continuità di impiego stia spingendo il valore di mercato molto più in alto di quanto le sole metriche di rendimento tecnico lascerebbero presagire. La capacità di garantire una presenza

costante in campo sembra quindi agire come un moltiplicatore di valore che, pur riconoscendone l'importanza, il modello tende a valutare con eccessiva prudenza. Questa sottovalutazione sistematica nelle previsioni finali evidenzia come la certezza della titolarità possa generare plusvalenze che vanno oltre la pura qualità tecnica espressa nei pagellini giornalistici.

Giocatore	team	R	Rateo Oss.	Rateo Att.	Mv Att.	Qi
Cluster Scommesse ($Q_i \leq 5$)						
Federico Bonazzoli	Cremonese	A	+1800%	+1160.0%	6.02	1
Toma Basic	Lazio	C	+1100%	+545.0%	5.91	1
Tiago Gabriel	Lecce	D	+900%	+340.0%	5.78	1
Lorenzo Bernasconi	Atalanta	D	+900%	+486.0%	6.06	1
Jeff Ekhator	Genoa	A	+800%	+657.0%	5.95	1
Antonio Vergara	Napoli	C	+700%	+449.0%	5.96	1
Adrian Benedyczak	Parma	A	+700%	+444.0%	5.89	1
Piotr Zielinski	Inter	C	+525%	+344.0%	6.31	4
Danilo Veiga	Lecce	D	+500%	+411.0%	5.92	1
Sascha Britschgi	Parma	D	+500%	+429.0%	5.95	1
Cluster Consolidazioni ($Q_i \geq 6$)						
Tasos Douvikas	Como	A	+200%	+152.0%	6.34	8
Francesco Pio Esposito	Inter	A	+200%	+95.2%	6.21	7
Arthur Atta	Udinese	C	+200%	+90.2%	6.08	6
Leonardo Spinazzola	Napoli	D	+186%	+74.5%	6.11	7
Ismaël Koné	Sassuolo	C	+167%	+160.0%	6.06	6
Semih Kiliçsoy	Cagliari	A	+167%	+80.1%	5.96	6
Nicola Zalewski	Atalanta	C	+167%	+98.1%	6.15	6
Sebastiano Esposito	Cagliari	A	+129%	+92.9%	6.00	7
Strahinja Pavlovic	AC Milan	D	+114%	+54.6%	6.05	7
Giovanni Simeone	Torino	A	+112%	+82.8%	6.09	8

Tabella 4.10: Analisi comparativa dei cluster Scommesse e Consolidazioni tra performance attuale e proiezione a fine stagione

A conclusione dell'analisi operativa, è possibile sintetizzare le dinamiche di mercato emerse attraverso quattro profili strategici che evidenziano la capacità del modello di adattarsi a diverse fasce di prezzo (Tabella 4.11). Il sistema rileva correttamente la saturazione dei calciatori di alto livello, il cui potenziale di crescita percentuale è limitato da una quotazione già elevata, e lo distingue dalle opportunità offerte dai profili a basso costo che garantiscono invece margini di crescita più ampi. L'integrazione delle metriche tattiche permette inoltre di validare le reali rivelazioni tecniche e di anticipare la flessione dei giocatori attualmente in una fase di sovrapprezzo. Questa mappatura complessiva conferma che l'analisi dei dati consente di superare le inefficienze del mercato e di prevedere con razionalità i valori che si stabilizzeranno entro il termine della stagione calcistica, una volta riassorbite le oscillazioni temporanee della quotazione attuale.

Giocatore	Tipologia	Mv Att.	deltaQ att.	RateoQ att.
Lautaro Martínez	Top di Gamma	6.59	+7.4	+1.5%
Nico Paz	Rivelazione Tecnica	6.46	+10.9	+54.5%
Federico Bonazzoli	Efficienza Low-Cost	6.02	+11.6	+1160%
Santiago Giménez	Sopravvalutato	5.86	-17.0	-68.0%

Tabella 4.11: Sintesi delle diverse dinamiche predittive identificate dal modello

5 Conclusioni

Il percorso metodologico sviluppato in questo studio ha preso le mosse dall'integrazione di dataset eterogenei provenienti da Fantacalcio.it e Understat, armonizzati attraverso pipeline ETL in Talend per garantire la coerenza tra variabili economiche e metriche avanzate quali xG, xA e $ppda$. La fase di modellazione iniziale ha previsto il confronto tra quattro diversi modelli di regressione OLS, permettendo di identificare nella variazione assoluta di quotazione (ΔQ) il target statistico più solido rispetto alla crescita percentuale, quest'ultima risultata eccessivamente volatile sui profili a basso costo. Parallelamente, lo studio del FantaValore di Mercato e della Media Voto ha permesso di quantificare il peso della centralità tecnica e dell'efficienza realizzativa rispetto al giudizio soggettivo dei pagellisti, evidenziando come i volumi di gioco siano i reali driver del valore nel lungo periodo.
Per rendere il modello operativo e superare i limiti legati alla disponibilità dei dati in corso di campionato, è stata implementata un'architettura a due stadi basata sulla teoria dei generated regressors: un primo stadio deputato alla generazione di una Media Voto attesa depurata da bias temporanei e un secondo stadio finalizzato alla previsione della variazione di valore finale. Tale struttura è stata sottoposta a rigorosi test di validazione, tra cui la 10-fold cross-validation e una verifica out-of-sample sulla stagione 2023/24, confermando l'assenza di overfitting e un'alta capacità di generalizzazione. Infine, l'applicazione dei modelli alla stagione 2024/25, con lo split temporale fissato alla 27ª giornata, ha permesso di mappare il mercato calcistico attraverso quattro profili strategici, isolando con precisione statistica le reali opportunità di investimento dalle bolle speculative dettate dall'emotività dei partecipanti.
Il lavoro svolto ha dimostrato che l'integrazione delle metriche avanzate nelle dinamiche del fantacalcio permette di superare i limiti delle statistiche tradizionali, offrendo una visione probabilistica più solida e meno soggetta all'emotività dei partecipanti. Attraverso l'applicazione del modello a due stadi, è stato possibile distinguere tra le reali crescite di valore supportate dai volumi di gioco e le oscillazioni temporanee dettate dalla percezione soggettiva del mercato. Nonostante la natura aggregata dei dati rappresenti un vincolo strutturale, poiché non consente di mappare i picchi di forma nel breve periodo, i risultati confermano che variabili come la partecipazione alle azioni offensive e l'efficienza realizzativa costituiscono driver affidabili per prevedere la stabilità del valore nel lungo periodo.

In merito agli sviluppi futuri, la ricerca potrebbe evolversi verso l'acquisizione di dati granulari per singola partita, permettendo così di studiare i momenti di forma atletica e la reattività delle quotazioni alle prestazioni settimanali. Un'altra strada promettente riguarda il passaggio da modelli di regressione lineare ad algoritmi di apprendimento automatico più complessi, capaci di intercettare relazioni non lineari tra le metriche fisiche e quelle tecniche. Inoltre, l'inclusione di variabili esterne legate al sentiment del mercato o alla popolarità mediatica dei calciatori consentirebbe di quantificare l'impatto dei pregiudizi cognitivi sulle quotazioni, rendendo lo strumento ancora più preciso nel segnalare le inefficienze speculative.

5.0.1 Gestione della riproducibilità

La riproducibilità dell'intero percorso analitico costituisce un elemento centrale della tesi, in quanto garantisce la validità e la trasparenza dei risultati ottenuti. Il flusso di gestione dei dati è stato automatizzato tramite pipeline sviluppate nel software Talend, che permettono di eseguire le operazioni di integrazione e pulizia in modo sistematico, eliminando la possibilità di errori derivanti da processi manuali. L'uso di variabili di contesto e la parametrizzazione dei processi rendono l'architettura flessibile e pronta per essere applicata a nuove stagioni senza dover modificare la logica di calcolo originaria. Questo metodo assicura la coerenza delle informazioni provenienti da diverse fonti e permette di rigenerare l'intera reportistica in modo rapido e verificabile, mantenendo un legame diretto e costante tra il dato grezzo e il risultato finale.

Bibliografia

[1] Wikipedia. *Expected Goals*. Disponibile online: `https://en.wikipedia.org/wiki/Expected_goals`.

[2] Understat. *Expected Goals (xG) Model Description*. Disponibile online: `https://understat.com/`.

[3] Barnett, V., e Hilditch, S. (1993). "The Effect of an Artificial Pitch Surface on Home Team Performance in Football (Soccer)". *Journal of the Royal Statistical Society. Series A (Statistics in Society)*

[4] Ensum, J., Pollard, R., e Taylor, S. (2004). "Applications of logistic regression to shots at goal in association football: calculation of shot probabilities, quantification of factors and player/team". *Journal of Sports Sciences*.

[5] Quadronica S.r.l. (2024). *Quotazioni e Statistiche Ufficiali Fantacalcio.it*. Disponibile online: `https://www.fantacalcio.it`, sezione 'Risorse'.

[6] GeeksforGeeks. *DBMS: Inner Join vs Outer Join*. Disponibile online: `https://www.geeksforgeeks.org/dbms/inner-join-vs-outer-join/`.

[7] Understat (2024). *Serie A Expected Goals and Advanced Metrics Database*. Disponibile online: `https://understat.com/league/Serie_A`.

[8] Footballizer. *OPPDA (Opponent Passes Allowed Per Defensive Action) - Academy*. Disponibile online: `https://www.footballizer.com/academy/oppda/`.

[9] Star Consulting. *Gli indici statistici nello sport: Expected Assist (xA)*. Disponibile online: `https://www.star-consulting.it/post/gli-indici-statistici-nello-sport-expected-assist-xa`.

[10] MartinOnData. *xGChain & xGBuildup 101*. Disponibile online: `https://www.pythonfootball.com/p/xgchain-and-xgbuildup-101`.

[11] Wikipedia. *Generated Regressors*. Disponibile online: `https://en.wikipedia.org/wiki/Generated_regressor`.