Università degli studi di Milano–Bicocca
Scuola di Economia e Statistica

Corso di Laurea in
Scienze Statistiche ed Economiche

I salari dei calciatori europei: indagine statistica e specificazione di un modello di stima.

Relatore: Dott. Tommaso Rigon

Tesi di laurea di:
Lorenzo Rossi
Matricola N. 870612

Anno Accademico 2022/2023

Indice

INTRODUZIONE

Nell'ambito sportivo, il calcio rappresenta non solo una passione che coinvolge miliardi di tifosi in tutto il mondo, ma anche un settore economico di rilevanza globale. La centralità del tema degli stipendi dei calciatori, al centro di molteplici dibattiti, emerge come una delle questioni più pressanti nell'ambiente calcistico contemporaneo. Negli ultimi anni, l'abisso economico tra le squadre di vertice e quelle di dimensioni minori si è ampliato in modo significativo. Le grandi squadre, avendo maggiore potenza economica, hanno la capacità di offrire salari più elevati ai calciatori, rendendo il panorama calcistico meno equilibrato e sempre più elitario. Queste dinamiche, combinate con spese eccessive, hanno messo a rischio la stabilità finanziaria di molte società calcistiche. Nonostante l'introduzione di misure come il Fair Play Finanziario, la situazione non ha mostrato segni di miglioramento sostanziale.

La problematica cardine di questo studio risiede nella definizione e creazione di un modello in grado di stimare gli stipendi dei calciatori basandosi esclusivamente su variabili come le prestazioni in campo, l'età e la squadra di appartenenza, escludendo altri fattori che attualmente influenzano la determinazione salariale.

Il principale obiettivo di questa ricerca è fornire una stima, pur consapevoli delle sue limitazioni, degli stipendi dei calciatori. Tale stima, quantitativa e basata esclusivamente sulle prestazioni sportive, intende proporsi come strumento di riferimento nella contrattazione salariale. La sua adozione potrebbe promuovere una maggiore meritocrazia, sostenibilità finanziaria e competitività nel mondo del calcio. Inoltre, il modello potrebbe servire come base per stimare lo stipendio di quei giocatori per i quali non si dispone di informazioni salariali precise e potrebbe gettare le basi per la futura implementazione di un "salary cap".

Dal punto di vista metodologico, si è optato per un approccio quantitativo. Inizialmente, si è ipotizzata l'esistenza di una correlazione lineare tra salari e prestazioni calcistiche, motivo per cui si è scelto di utilizzare un modello di regressione lineare. Tuttavia, per affinare ulteriormente l'analisi e

ridurre la dimensionalità dei dati, si sono successivamente applicati metodi come lo stepwise e la PCA.

Concludendo, questa tesi si propone di affrontare una delle questioni più spinose e dibattute nel mondo del calcio, offrendo una nuova prospettiva e un modello che, pur nelle sue limitazioni, potrebbe rappresentare un punto di partenza per future riflessioni e implementazioni nel campo della determinazione degli stipendi dei calciatori.

Capitolo 1

STIPENDI NEL CALCIO, CONTESTO E ANALISI QUANTITATIVA

1.1 Il Calcio Europeo in Cifre: Un'Analisi dei Ricavi e degli Stipendi

Il calcio con circa 3,5 miliardi di appassionati e circa 250 milioni di giocatori in oltre 200 paesi (Migliani,2023), è lo sport più popolare al mondo.

A causa della sua popolarità, negli ultimi decenni la domanda di top players è aumentata notevolmente, con casi di giocatori scambiati anche per importi superiori ai 200 milioni di euro. Questi numeri sono nettamente superiori alle cifre storiche degli scambi rispetto al normale tasso di inflazione. Ad esempio, Diego Armando Maradona, è stato ceduto nel 1984 dal Barcellona al Napoli per una cifra record di 10,48 milioni di dollari (Frick, 2007). Tale cifra, attualizzata a 22,7 milioni di euro nel 2023, è considerata oggi un importo relativamente basso, a cui vengono scambiati molti giocatori di livello medio.

L'aumento delle cifre nei trasferimenti dei calciatori trova una giustificazione nel flusso finanziario sempre più consistente che circola nell'ambito calcistico, catalizzato anche dall'effetto della globalizzazione. Nel corso del 2022, il mercato calcistico europeo ha raggiunto il suo apice storico, con ricavi totali stimati a 29,5 miliardi di euro (Deloitte's Sports Business Group 2023). Di questi introiti, la quota predominante, pari al 58%, ossia circa 17,2 miliardi di euro, è attribuibile ai cinque principali campionati europei.

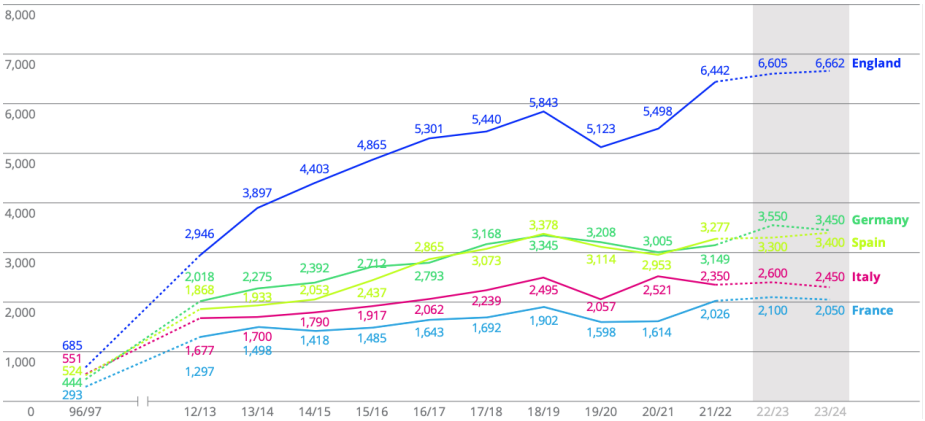

Figura 1: *Ricavi dei cinque maggiori campionati europei - 1996/97 e 2012/13 fino al 2023/24 (milioni di euro)(Fonte: Deloitte Annual review of football Finance, 2023).*

Nota: *I dati per il 2022/23 e il 2023/24 sono proiezioni. Le proiezioni per la Germania, la Spagna, l'Italia e la Francia sono arrotondate a all'approssimazione di 50 milioni di euro.*

L'analisi visiva dei dati rappresentati in questo grafico, dipinge una chiara immagine di un andamento crescente costante: i ricavi, con una regolarità quasi sorprendente, hanno manifestato una crescita incessante anno dopo anno. La forza motrice di questi ricavi è innanzitutto rappresentata dalla cessione dei diritti televisivi, che contribuiscono per approssimativamente il 50% del totale introiti di ogni campionato (Deloitte's Sports Business Group,2023). A seguire, troviamo i profitti derivanti dalle partite, gli sponsor e altre iniziative commerciali che integrano questo panorama finanziario.

In aggiunta, si evidenzia un trend interessante: nel corso degli anni, l'Inghilterra ha notevolmente consolidato la propria supremazia, quasi raddoppiando i ricavi della Germania, posizionata in seconda posizione, e triplicando quelli della Francia, che si piazza all'ultimo posto in questa graduatoria.

Naturalmente, l'aumento dei ricavi dei club e dei trasferimenti dei calciatori si ripercuote anche sui loro stipendi, che sono aumentati costantemente (Frick, 2007).

Analizzando i compensi delle squadre nei cinque principali campionati europei nell'arco dell'ultima decade, emergono chiare tendenze. Parallelamente all'incremento dei ricavi, si registra un costante aumento dei salari dei giocatori, con una notevole crescita del 56% dal periodo della stagione 2013/2014 a quello della stagione 2022/2023.

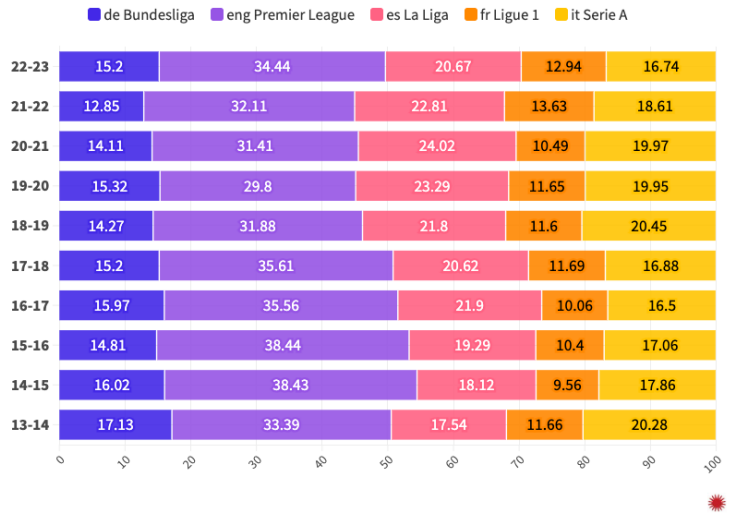

Figura 2: *Monte ingaggi dei cinque maggiori campionati europei - 2013/14 fino al 2022/23 (milioni di euro).*

Se, osservando il grafico dei ricavi, emerge il consolidamento della supremazia dell'Inghilterra rispetto agli altri campionati nel corso degli anni, dal grafico in esame si evince che, pur mantenendo un vantaggio netto in assoluto con retribuzioni considerevolmente più elevate rispetto agli altri campionati, la quota del monte ingaggi inglese è rimasta relativamente costante. Nel 2013, rappresentava il 33% del totale di tutti i campionati, e nel 2023, nonostante l'evidente incremento delle cifre, si attesta al 34% dell'ammontare complessivo. Questo trend è riscontrabile anche negli altri campionati, che hanno mantenuto sostanzialmente invariate le proporzioni, suggerendo che l'aumento delle retribuzioni sia avvenuto in maniera più uniforme, senza alterare significativamente le differenze tra i vari campionati.

Se è vero che le differenze tra i vari campionati si sono mantenute relativamente stabili, lo stesso non può essere detto per quanto riguarda le dinamiche interne a ciascun campionato. In particolare, nei campionati francese e tedesco, si è sviluppato un notevole divario tra le squadre di vertice, ovvero Paris Saint Germain e Bayern Monaco, e le restanti squadre del torneo nazionale. Questo divario è evidenziato dalla percentuale rappresentata da queste due squadre rispetto al totale dei salari nell'intero campionato. Nella stagione 2013/2014, Paris Saint Germain e Bayern Monaco costituivano rispettivamente il 21% e il 20% del monte ingaggi complessivo a livello nazionale. Tuttavia, nella stagione 2022/2023, queste percentuali sono balzate al 45% e al 29% del monte ingaggi complessivo del campionato.

Questa situazione può sollevare questioni riguardo alla competitività del campionato, in quanto tali squadre investono cifre notevolmente superiori rispetto alle altre per i salari dei giocatori, ottenendo i migliori talenti e inevitabilmente riducendo la parità in campo. Ciò trova conferma nei risultati delle partite, dove il Paris Saint Germain ha trionfato ben nove volte nelle ultime undici stagioni, mentre il Bayern Monaco ha vinto consecutivamente undici campionati.

Allargando l'analisi ai cinque maggiori campionati europei, emerge un altro dato significativo: sebbene la differenza tra i campionati sia rimasta pressoché costante, l'intervallo tra le squadre più ricche si è ampliato. Nel primo anno dell'indagine, la squadra con il monte ingaggi più elevato (il Barcellona) spendeva circa 25 volte in più rispetto alla squadra con il monte ingaggi più basso (il Guingamp). Tuttavia, nel 2023, il PSG investe circa 60 volte di più dell'Ajaccio. Questi dati mettono in luce un rafforzamento della posizione negoziale delle squadre di vertice, in cui le prime venti squadre di questa graduatoria contribuiscono al 60% del totale dei salari, mentre le ultime venti coprono appena il 3% dell'intero monte ingaggi.

1.2 Il Financial Fair Play: Tra Aspirazioni e Realizzazioni Pratiche

Nel contesto europeo del calcio professionistico, una delle maggiori sfide che ha afflitto l'industria negli anni precedenti l'introduzione del Fair Play Finanziario è stata la crescita esorbitante degli stipendi dei calciatori. Se da una parte, come analizzato precedentemente gli introiti, grazie a diritti televisivi, sponsorizzazioni e altre fonti di reddito, hanno conosciuto una crescita notevole, dall'altra i costi legati agli stipendi dei calciatori sono aumentati a un ritmo non proporzionale.

La spirale crescente degli stipendi può essere attribuita a vari fattori. In primo luogo, la crescente globalizzazione del gioco ha portato a una concorrenza sempre più intensa tra i club per assicurarsi i servizi dei migliori talenti. Questa "corsa agli armamenti" ha avuto come risultato un'escalation nei prezzi per l'acquisto di giocatori, ma soprattutto nel loro mantenimento attraverso contratti sempre più opulenti. Inoltre, l'intervento di potenti agenti calcistici e l'intensificarsi della rivalità tra club ha ulteriormente inflazionato gli stipendi.

L'analisi dei bilanci dei club ha mostrato che, nonostante la crescita dei ricavi, molti club spendevano una quota sproporzionata dei loro introiti in stipendi. Per mettere in prospettiva, come possiamo vedere dalla Figura 3

se nel 1996 il rapporto stipendi su ricavi oscillava intorno al 51%, alla vigilia dell'introduzione del FFP, nel 2008, tale percentuale era salita al 64%. Questo dato non era solo un campanello d'allarme sulle prospettive di sostenibilità finanziaria, ma anche un segno tangibile dell'urgente necessità di regolamentare e monitorare le spese dei club in maniera più rigorosa.

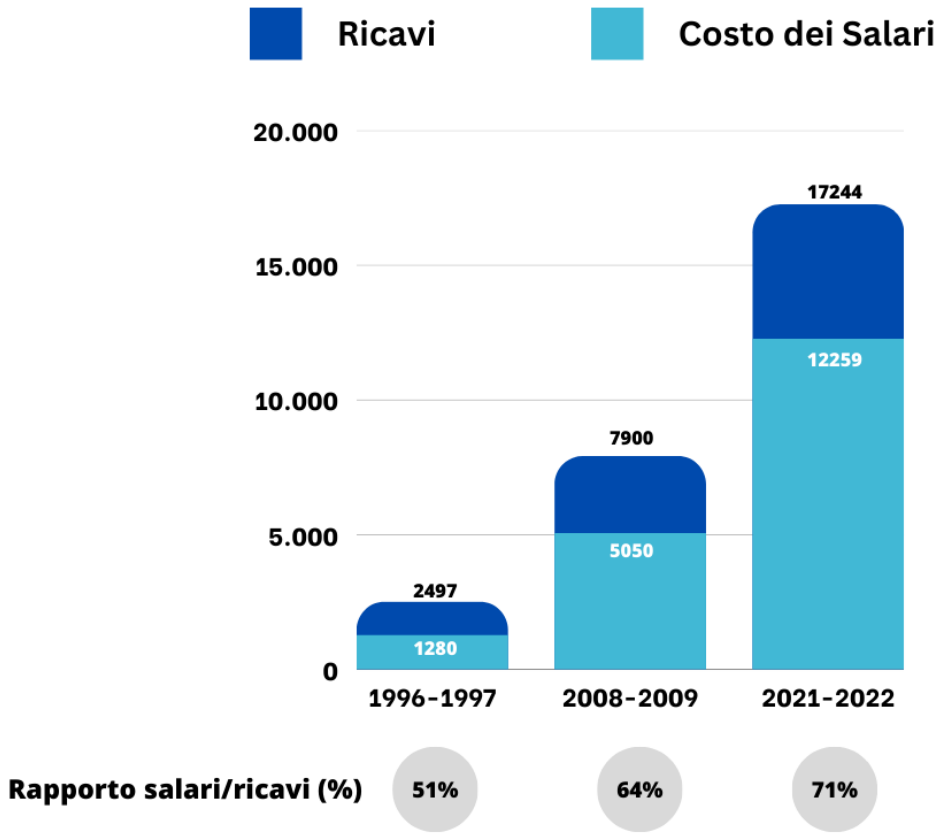

Figura 3: *Grafico comparativo dei salari e dei ricavi nelle stagioni calcistiche 1996-97, 2008-09 e 2021-22.*

Ulteriori dettagli sulla gestione finanziaria dei club emergono da uno studio condotto nel 2010 dalla società 'A.T. Kearney'. Questa analisi, focalizzata sulle operazioni di calciomercato dei cinque principali campionati europei, ha rivelato un bilancio negativo tra gli investimenti per l'acquisto di calciatori e gli incassi derivanti dalla loro cessione.

	Investimenti totali in calciatori (€ Mln)	**Incassi totali dalla cessione di calciatori** (€ Mln)	**Differenza tra acquisti e cessioni** (€ Mln)
Premier League	580	489	-91
Bundesliga	243	125	-118
Liga	502	245	-257
Serie A	498	460	-38
Ligue 1	270	208	-62

Tabella 1:*Riepilogo delle transazioni di calciomercato effettuate nei principali cinque campionati europei nel 2010.*

Il quadro finanziario generale dei club europei, però, diventa ancora più preoccupante quando si esaminano i dati pubblicati nel 2008 dall'UEFA (Union of European Football Associations), l'organo amministrativo, organizzativo e di controllo del calcio europeo. I dati sono i seguenti:

- l'insieme dei debiti dei club della Premier League ammontava a circa 4 miliardi di Euro (con particolare riferimento a Manchester United e Chelsea, finaliste di Champions League proprio nell'anno 2008 ed ai primissimi posti per indebitamento a livello mondiale);
- il 47% dei club europei aveva riportato perdite nell'anno di riferimento;
- il 22% di tali club aveva riportato perdite superiori al 20% del reddito, quindi definite 'rilevanti';
- il 35% dei club registrava un patrimonio netto negativo e per il 44% dei club, tale patrimonio netto negativo era in peggioramento rispetto agli anni precedenti;
- i costi erano aumentati del 9,3% con una notevole incidenza sulla redditività dei club;
- nel 2008, anno dello studio di riferimento, le squadre avevano registrato perdite di gestione per complessivi 578 milioni di Euro, nonostante un

aumento delle entrate del 10,6% rispetto all'anno precedente. Le perdite, nello stesso periodo, infatti, erano aumentate dell'11,6%, creando un importante scarto negativo di bilancio.

I numeri che emergono da questi studi evidenziano la piega di insostenibilità che aveva assunto la gestione dei club, governati con spirito mecenatesco da presidenti e magnati, che poco si curavano del lato economico, concentrandosi piuttosto sul ritorno d'immagine e popolarità che conseguiva dai risultati sportivi, inevitabilmente figli degli investimenti fatti sul mercato. Questa scarsa accortezza finanziaria ha contribuito in modo decisivo alla formazione di un crescente debito sempre più insostenibile, che ha posto le basi per la nascita del Fair Play Finanziario (FFP), introdotto dalla UEFA nel 2008, e adottato per la prima volta nel 2011.

Sostenendo l'idea del presidente Michel Platini di instaurare maggiore equilibrio finanziario e leale concorrenza tra i club. Il FFP, infatti, mira principalmente a garantire che i club non spendano più di quanto guadagnano, ponendo l'accento sul principio dell'autofinanziamento. In altre parole, i club sono incoraggiati, e in alcuni casi obbligati, a bilanciare le loro spese con le proprie entrate, evitando di affidarsi eccessivamente a finanziamenti esterni o a debiti insostenibili.

Nel cuore di questa regolamentazione c'è la regola del "break-even", che richiede ai club di bilanciare le loro entrate e uscite. Oltre a ciò, il FFP prevede una serie di altre misure, come la limitazione dell'indebitamento, l'obbligo di pagare i debiti nei tempi previsti e l'imposizione di sanzioni ai club che non rispettano tali regole.

Il Fair Play Finanziario rappresentava quindi un tentativo di riequilibrare il campo di gioco nel calcio europeo, assicurando che la ricchezza e il potere finanziario non compromettano la competitività e la sostenibilità a lungo termine del bellissimo gioco. Con l'introduzione del FFP, la UEFA ha mirato a creare un ambiente in cui la progettazione a lungo termine e la responsabilità finanziaria vengano premiate, piuttosto che una mentalità a breve termine che potrebbe mettere a rischio l'intera struttura del calcio europeo.

Il Financial Fair Play (FFP) ha portato significativi cambiamenti nel panorama calcistico europeo, manifestando tangibili benefici per i bilanci dei club. Tra il 2008 e il 2011, i club dei principali campionati europei avevano accumulato perdite complessive per 5,1 miliardi di euro, culminando con un deficit record di 1,7 miliardi nel 2011. Questo quadro ha sottolineato l'urgente necessità di riforme nel settore. Con l'introduzione del FFP, in dieci anni, l'approccio gestionale dei club ha subito una profonda metamorfosi. Questa trasformazione è evidente nel contrasto tra i soli 300 milioni di perdite subite

nel 2022 e le perdite di 1,7 miliardi registrate dieci anni prima. Queste perdite sono sicuramente dovute anche alla pandemia del Covid e il periodo molto difficile in cui ha versato l'Europa in questi ultimi due anni, anche perché se si va a guardare i bilanci nell'ultima stagione prima della pandemia, non si erano registrate perdite, ma anzi profitti per circa 1.3 miliardi di euro.

Parallelamente, il patrimonio complessivo delle società europee è cresciuto da 1,9 miliardi di euro nel 2011 a 7,7 miliardi alla fine del 2017. Oltre ai vantaggi diretti del FFP, è essenziale considerare l'importanza dei diritti televisivi e degli investitori non europei, che hanno avuto un ruolo cruciale nell'amplificare la crescita finanziaria. L'esplosione dei prezzi dei diritti televisivi ha fornito ai club risorse significative, mentre gli investitori esterni, in particolare dall'Asia, hanno infuso capitali essenziali, trasformando il calcio in un'industria prospera. Il Financial Fair Play (FFP) è stato concepito come uno strumento per garantire equilibrio e sostenibilità nel calcio europeo. Tuttavia, nonostante le nobili intenzioni, la sua implementazione ha manifestato diverse criticità. Una delle principali preoccupazioni riguarda il sistema delle "Plusvalenze", che si riferisce alle plusvalenze realizzate attraverso la vendita di giocatori. Questo sistema, sebbene legittimo, può essere manipolato, come evidenziato dal ritorno delle "Plusvalenze fittizie", dove le transazioni possono essere gonfiate per mostrare profitti artificiali.

Un aspetto particolarmente critico è l'"Ossificazione del Sistema". Questo termine suggerisce un congelamento dello status quo nel calcio, dove le gerarchie esistenti vengono mantenute e rafforzate. In pratica, invece di ridurre il divario tra i club ricchi e quelli meno abbienti, come originariamente previsto dal FFP, la situazione potrebbe essere peggiorata. L'ossificazione ha potenzialmente ampliato il divario economico e tecnico tra le squadre di punta e quelle in difficoltà, creando un ambiente in cui le squadre ricche diventano sempre più forti, mentre quelle con risorse limitate faticano a competere. Questo è esattamente l'opposto di ciò che il FFP intendeva realizzare.

Inoltre, ci sono stati dubbi applicativi, in particolare riguardo alle sponsorizzazioni. Il caso del Paris-Saint Germain serve come esempio emblematico, dove le sponsorizzazioni potrebbero essere utilizzate come mezzo per aggirare le regole del FFP.

Un dato ulteriormente preoccupante è l'aumento del rapporto stipendi/ricavi dal 2008, anno di introduzione del FFP, che è cresciuto dal 64% al 71%. Questa tendenza suggerisce che i club continuano a spendere in modo sproporzionato rispetto ai loro ricavi.

In sintesi, mentre l'idea di base del FFP era lodevole, la sua realizzazione ha lasciato spazio a diverse lacune e manipolazioni. Queste criticità sottolineano la necessità di rivedere, affiancare o sostituire l'attuale FFP con

un sistema che affronti in modo più efficace le sfide del settore calcistico, in particolare in termini di competitività e sostenibilità finanziaria.

1.3 Fattori che determinano il salario di un giocatore

Il salario di un calciatore è influenzato da vari fattori, tra cui le sue abilità, le prestazioni passate, l'età, il potenziale di miglioramento, la durata del contratto e persino la personalità. Questi elementi dovrebbero convergere nella definizione degli stipendi attraverso una valutazione puntuale ed oggettiva.

Infatti, determinare il compenso di un calciatore costituisce una sfida complessa nell'ambito sportivo, modellata da una combinazione di fattori interconnessi.

Nell'era pre-digitale, data la mancanza di raccolta sistematica di dati statistici sui giocatori, tale valutazione si basava principalmente su analisi qualitative rendendo difficile una valutazione quantitativa delle loro performance e abilità (Frick, 2006).

Ancora oggi la definizione dei salari dei calciatori dipende invece molto piu' dalle abilita' commerciali degli agenti, dalle relazioni tra i club ed i procuratori finanziariamente piu' potenti, che non da una valutazione oggettiva dell performance e potenzialita' dell'atleta calciatore. Tale approccio spesso provoca degli scompendi nei conti delle societa' poco giustificabili dalle prestazioni sul campo degli atleti, contribuendo ad aumentare la crisi finanziria in cui versa la mggior parte dei club in Europa.

Tuttavia, dalla fine degli anni '90, la raccolta e la completezza dei dati relativi al calcio sono aumentate costantemente e ci consentono innanzitutto di fare un'analisi piu' accurata per capire meglio l'interconnessione tra i compensi dei calciatori e le dinamiche di gestione sportiva degli atleti. Attualmente, dati pubblici includono una vasta gamma di informazioni sulle performance e gli stipendi degli atleti nei principali campionati (Frick, 2006). Spesso, le stelle del calcio guadagnano più degli altri giocatori, anche per il contributo finanziario che apportano attraverso biglietti, merchandising e accordi televisivi. Questo effetto è accentuato dalla scarsità di calciatori di élite, generando un aumento dei salari grazie a rendite di monopsonio (Garcia-del Barrio e Pujol, 2007). Questo concetto si verifica quando vari datori di lavoro competono per un gruppo limitato di dipendenti (Garcia-del Barrio e Pujol, 2007). Nell'ambito calcistico, i club competono per i servizi di un gruppo ristretto di atleti di spicco, spingendo al rialzo gli stipendi per attrarli.

Dal punto di vista del comportamento, le performance dei calciatori migliorano con l'aumento del reddito assoluto (Torgler e Schmidt, 2007). Al contrario, una marcata disparità salariale influisce negativamente sul rendimento della squadra, poiché aumenta la differenza di guadagno tra i giocatori e il resto del team (Torgler et al., 2006). Inoltre, lo stipendio del giocatore ha un evidente impatto sulle decisioni dell'allenatore, come dimostrato dal fatto che i giocatori con salari più alti vengono impiegati in modo sproporzionato rispetto alle loro performance in campo, a discapito di altri che ricevono compensi minori (Garcia-del Barrio e Pujol, 2009).

Mentre l'impegno profuso dal giocatore ha un effetto trascurabile sul salario (Wicker et al., 2013), lo studioso Frick ha proposto che la determinazione dello stipendio possa dipendere dalle performance del giocatore nella stagione precedente, dalle partite giocate a livello internazionale e dal numero di gol segnati (Frick, 2011). Questi fattori influenzano chiaramente la decisione sullo stipendio, ma è plausibile che altre variabili, come la precisione dei passaggi, dei calci di punizione, la velocità e l'abilità nei tackle, possano contribuire al calcolo. Queste diverse abilità lavorano sinergicamente per creare un quadro completo delle capacità del giocatore, e quindi il confronto di un insieme limitato di variabili individuali potrebbe non riflettere appieno il suo valore complessivo.

Come conseguenza della disponibilita' di tale mole di dati analitici, nell'era post-digitale, sono stati sviluppati sistemi e metodi per analizzare dati calcistici utilizzando statistiche computazionali e il riconoscimento di modelli (Lames et al., 2011). Questi approcci sono stati adottati per semplificare l'analisi dei dati dei giocatori e delle traiettorie del pallone (Feess et al., 2010), per rilevare automaticamente la posizione dei giocatori (Siegle et al., 2013), per creare ambienti interattivi di allenamento (Jensen et al., 2014) e per il monitoraggio delle partite (O'Donoghue e Robinson, 2009; Castellano et al., 2014).

1.4 Obiettivo e vantaggi dell'implementazione di un approccio quantitativo.

Come spiegato in precedenza, il salario di un giocatore di calcio è una combinazione di diverse variabili predittive e non, rendendo il tema estremamente complesso da affrontare e prevedere. Questo argomento coinvolge spiegazioni di natura economica, statistiche e comportamentale.

L'obiettivo dell'analisi è fornire un approccio quantitativo e oggettivo per stimare il salario dei giocatori di calcio basandosi sulle loro abilità e capacità

sul campo, tralasciando altri fattori che contribuiscono alla valutazione reale del salario. A tale scopo, utilizzeremo dati relativi alle abilità e alle performance dei giocatori durante le partite, per analizzare in maniera matematica il salario dei calciatori in rapporto alle loro abilità rispetto agli altri giocatori.

Questo metodo potrebbe avere diverse ricadute positive nell'economia del calcio, apportando un maggiore grado di trasparenza, obiettività e meritocrazia nella definizione dei contratti. Alcuni dei possibili vantaggi e contributi pratici di questo approccio metodologico potrebbero essere:

- **Assistenza nella negoziazione dei contratti:** Questo metodo potrebbe servire come supporto sia per i dirigenti dei club che per gli agenti dei giocatori durante le trattative contrattuali. In un ambiente speculativo, dove circolano informazioni spesso fuorvianti, avere una base oggettiva per iniziare le trattative è fondamentale. Inoltre, la previsione dei valori futuri potrebbe essere utile nella negoziazione di percentuali di vendita e commissioni aggiuntive.

- **Riduzione dell'influenza degli agenti:** Un'analisi oggettiva del salario potrebbe ridurre il potere eccessivo degli agenti, poiché si baserebbe su una valutazione oggettiva. Questo potrebbe limitare richieste eccessive da parte degli agenti per i loro assistiti. Questa pratica è spesso osservata, dove gli agenti richiedono somme esagerate per ottenere commissioni considerevoli. A conferma di quanto evidenziato si noti che, solo nel 2022 è stata spesa dai club l'incredibile cifra di 622 milioni di euro per le commissioni agli agenti(Ramazzotti, 2023).

- **Promozione della meritocrazia:** Con una valutazione oggettiva, i giocatori sarebbero incentivati a dimostrare sul campo il loro reale valore, poiché la valutazione sarebbe basata su abilità e prestazioni calcistiche, escludendo altri fattori estranei. Ciò stabilirebbe criteri oggettivi per la determinazione dei salari, riducendo il rischio di disparità salariali ingiustificate tra giocatori con abilità e ruoli simili.

- **Pianificazione finanziaria più precisa per i club:** Questo approccio aiuterebbe i club a pianificare con precisione il budget necessario per i salari dei giocatori, favorendo una gestione finanziaria più efficiente. Ciò contribuirebbe a evitare sopravvalutazioni o sottovalutazioni per i giocatori, promuovendo una distribuzione più equa delle risorse finanziarie.

- **Possibile introduzione di un salary cap:** Nel lungo termine, questo metodo potrebbe essere la base per l'implementazione di un tetto

salariale, definito e regolamentato dalla UEFA, limitando la spesa dei club e promuovendo una maggiore parità tra le squadre, una competizione più equa ed assicurando un maggiore spettacolo per il pubblico. Un esempio di successo si è visto negli Stati Uniti, dove nel 1994 la National Football League (NFL) ha introdotto un tetto salariale rigoroso. Da allora, oltre l'84% delle squadre NFL ha concluso almeno una stagione con uno dei sei migliori record nella lega. Nel medesimo periodo, dalla fondazione della English Premier League due anni prima, meno del 30% dei club che vi hanno partecipato ha concluso una stagione tra i primi quattro classificati. (Considerando che la NFL ha un numero maggiore di squadre, i primi sei classificati della NFL sono paragonabili ai primi quattro della EPL).

Capitolo 2

MATERIALI

2.1 Descrizione del dataset

Nel contesto di questa ricerca, sono stati adoperati due insiemi di dati a cui per semplicità daremo il nome `dataset_salari`, e `dataset_statistiche`. Questi sono stati successivamente integrati al fine di costituire un dataset consolidato, idoneo per l'analisi e l'implementazione del modello studiato.

Il primo insieme di dati comprende 2.891 osservazioni suddivise in 7 variabili. Questo incorpora informazioni demografiche relative ai calciatori e ai loro emolumenti lordi per la stagione sportiva 2022/2023, estratte dal database fornito da Capology (https://www.capology.com/). Quest'ultima, quale ente, sfrutta una vasta rete di professionisti direttamente coinvolti nelle negoziazioni contrattuali, nonché un ampio spettro di pubblicazioni a livello globale. La retribuzione di un calciatore è designata come "verificata" (contrassegnata con un distintivo verde) qualora sia stata direttamente rilasciata dal club o dall'agente, o confermata da almeno due fonti indipendenti. Tale distintivo verde da parte di Capology serve come indicatore di fiducia, attestando che sono state intraprese misure rigorose per garantire la massima precisione dei dati. Si stima che una quota significativa, superiore al 30%, dei calciatori attivi a livello internazionale rientri sotto questa categorizzazione verificata. Nei casi in cui le informazioni presenti non siano esaurienti, gli algoritmi sviluppati da Capology intervengono, offrendo una stima calibrata, basata su un insieme di oltre 20 variabili e caratteristiche.

Il secondo dataset, diversamente dal primo, consta di 2.891 osservazioni suddivise in 126 variabili. Sebbene includa informazioni demografiche e di associazione analoghe al primo set di dati, questo dataset si caratterizza maggiormente per l'ampia inclusione di variabili statistiche che illustrano le prestazioni sportive dei calciatori per la stagione sportiva 2022/2023. Queste in-

formazioni sono state acquisite dal database di Fbref (https://fbref.com/it/), un portale web specializzato nell'offrire statistiche dettagliate relative a squadre e calciatori di calcio a livello globale.

È importante sottolineare che FBref è riconosciuto come un leader nel campo dell'analisi calcistica, vantando una copertura estensiva grazie alla collaborazione con il partner, Opta, per più di 20 competizioni. Opta si dedica alla raccolta di dati in tempo reale, rendendoli poi disponibili ai propri clienti attraverso un'ampia varietà di feed. Questi feed sono strutturati in modo da offrire differenti gradi di dettaglio, permettendo di spaziare da semplici commenti testuali in tempo reale fino ad analisi storiche dettagliate e statistiche aggregate stagionali.

2.2 Descrizione delle variabili presenti nei dataset

Il dataset_salari, è composto da 9 variabili:

- **Rk:** Questa variabile serve come identificativo numerico per ogni osservazione, procedendo in un ordine sequenziale.

- **Player:** Questa variabile contiene il nome completo del calciatore, permettendo una chiara identificazione del soggetto in esame.

- **Nation:** Si riferisce alla nazionalità del giocatore, fornendo contesto sulla sua origine.

- **Position:** Indica la posizione principale in cui il calciatore viene solitamente schierato. Questa variabile può assumere quattro valori distinti: "GK" per i portieri, "DF" per i difensori, "MF" per i centrocampisti e "FW" per gli attaccanti.

- **Squad:** Segnala la squadra di calcio di cui il calciatore fa parte durante la stagione in esame.

- **Age:** Questa variabile ci dice quanti anni aveva il calciatore all'inizio della stagione, specificatamente il primo agosto.

- **Weekly wages:** Fornisce l'importo dello stipendio lordo che il calciatore riceve su base settimanale. É importante sottolineare che questa cifra non include eventuali bonus o incentivi.

- **Annual Wages:** Simile alla variabile precedente, ma riferito all'arco dell'intero anno. Anche in questo caso, gli stipendi base non sono comprensivi di bonus o incentivi.

- **Notes:** Questa variabile fornisce ulteriori dettagli, in particolare in merito alla verifica o stima dello stipendio riportato.

Il dataset denominato "dataset_statistiche" comprende 129 variabili. Di queste, molte coincidono con quelle presenti nel "dataset_salari", ad eccezione di quelle specificamente orientate alla remunerazione. Predominantemente, "dataset_statistiche" si focalizza su parametri relativi alle prestazioni calcistiche dei giocatori, inclusi attributi anagrafici quali età e anno di nascita. Al fine di garantire una categorizzazione sistematica e coerente, le variabili con caratteristiche affini saranno aggregate in sottogruppi omogenei come segue:

- **Tempo di Gioco (4 variabili):** Questa sezione offre informazioni relative ai minuti giocati, al numero totale di partite giocate e alle partite iniziate da ciascun giocatore.

- **Rendimento (8 variabili):** Questo segmento concentra le variabili correlate alle performance complessive di un calciatore, incluse le reti realizzate, gli assist forniti e le sanzioni disciplinari ricevute.

- **Prestazione Prevista (4 variabili):** Questo insieme di variabili, basate su dati analitici avanzati forniti da Opta, forniscono proiezioni probabilistiche, come gli "expected goals", derivanti dalla qualità di un tiro e dai contesti in cui è stato eseguito. Queste metriche permettono un confronto con dati effettivi per valutare l'overperformance o l'efficacia del giocatore.

- **Per 90 Minuti (10 variabili):** Tale sezione standardizza le variabili sopra menzionate su un arco temporale di 90 minuti. Si precisa che nella graduatoria sono considerati solo i giocatori con almeno 30 minuti di gioco.

- **Tiri (21 variabili):** Questo comparto aggrega le metriche relative ai tiri, incluse le statistiche sui tiri in porta, i goal per tiro in porta e le proiezioni predittive, come i gol previsti escludendo i rigori.

- **Passaggi (21 variabili):** All'interno di questo insieme sono racchiuse le statistiche sui passaggi, includendo dettagli sui passaggi effettuati, tentati e sulla distanza dei passaggi. Vi sono, inoltre, specifici indicatori relativi ai passaggi chiave e alle proiezioni.

- **Creazione di Gol e Tiri (15 variabili):** Questo gruppo offre insight sulla capacità di un giocatore di creare opportunità, mettendo in luce le azioni pericolose generate che portano a tiri o passaggi decisivi.

- **Possesso Palla (22 variabili):** Qui si analizzano le metriche legate al possesso palla del giocatore, comprendendo dettagli come i tocchi, i dribbling tentati e riusciti e le statistiche legate al movimento del giocatore con il pallone.

- **Azioni Difensive (16 variabili):** Questa raccolta si concentra sulle prestazioni difensive, offrendo dati sui tackle, i blocchi, le intercettazioni e altre metriche difensive chiave.

Capitolo 3

PRE PROCESSING

3.1 Pulizia dei dati

Nell'ambito della preparazione dei dati per l'implementazione del nostro modello statistico avanzato, è stata condotta una rigorosa fase di integrazione e pre-elaborazione dei dataset. Il primo passo ha riguardato l'incorporazione della variabile "stipendi annuali" dal ***dataset_salari*** al ***dataset_statistiche***. In questo contesto, è stato essenziale affrontare le questioni legate alla qualità dei dati. Le osservazioni classificate come *'Unverified estimation'* all'interno della variabile notes rappresentano stime generate da algoritmi propri di Capology, la cui attendibilità non è stata corroborata. Questi valori sono stati, pertanto, imputati come mancanti per assicurare un'analisi robusta e priva di possibili artefatti.

Una fase successiva ha riguardato l'analisi della variabile Annual.wages*text*, che, essendo classificata come character, conteneva stipendi rappresentati in diverse valute (euro, dollari, sterline). Al fine di assicurare omogeneità e accuratezza nelle successive analisi econometriche, si è adottato l'euro come valuta di riferimento, e mediante operazioni di parsing e conversione, abbiamo trasformato *'Annual.wages'* in un formato numerico.

L'operazione di integrazione tra i dataset ha seguito un criterio di join basato sulle variabili *'Player'* ed *'Age'*. Nel caso di corrispondenza tra le osservazioni, si è mantenuto il dato di *Annual.wages.* Tuttavia, laddove mancava una corrispondenza, si è optato per l'imputazione di un valore mancante (NA).

Dopo tale integrazione, una variazione nel conteggio delle osservazioni ha suggerito la presenza di osservazioni multiple in ***dataset_salari***. Una dettagliata analisi ha evidenziato la presenza di 122 duplicati. Una possibile ipotesi attribuisce questa presenza ai movimenti di mercato dei giocatori nella sta-

gione 2023/2024, il che ha comportato diversi stipendi all'interno del medesimo periodo. Al fine di affrontare tale sfida, si è optato per un'aggregazione basata sulla media ponderata degli stipendi.

Successivamente, si è eseguito un merge tra i dataset, rimuovendo le variabili ridondanti. La variabile posizione (pos), che indicava ruoli multipli dei giocatori, dato che alcuni giocatori hanno assunto ruoli multipli durante la stagione è stata binarizzata per assicurare un'interpretazione chiara e univoca delle posizioni in campo.

Concludendo, sono state identificate 322 osservazioni duplicate nel dataset aggregato. Queste rappresentavano una porzione significativa, circa il 10%, del nostro dataset finale. Data la sfida interpretativa e metodologica di queste osservazioni, specialmente in relazione all'aggregazione delle variabili, ho scelto di escluderle per garantire un'analisi finale ottimizzata e priva di ambiguità.

3.2 Valori Mancanti

Nel panorama dell'analisi dei dati, affrontare l'incidenza di dati incompleti, in particolare la presenza di valori mancanti, rappresenta una sfida metodologica di rilevante importanza. Tali mancanze possono insorgere da una molteplicità di fattori: possono derivare da errori nella fase di raccolta, dalla mancata risposta degli intervistati in sondaggi strutturati o da discrepanze nella fase di registrazione. E' di cruciale importanza identificare e comprendere il meccanismo che ha portato a tali omissioni per garantire l'accuratezza dell'analisi. Tradizionalmente, questi meccanismi di mancanza vengono categorizzati come MCAR (Missing Completely At Random), in cui la mancanza è puramente casuale; MAR (Missing At Random), dove l'omissione potrebbe essere legata ad altre variabili note; e NMAR (Not Missing At Random), dove l'omissione potrebbe dipendere dai valori mancanti stessi.

Nel dataset in esame, si è identificato un notevole volume di valori mancanti. A parte la variabile Annual.wages, che funge da target, è emerso che una proporzione significativa delle variabili presentava questa problematica. La natura dei modelli statistici, in particolare della regressione lineare multipla, richiede idealmente un dataset completo per produrre stime accurate e robuste. Di fronte a questa necessità, è stata intrapresa un'analisi esplorativa dei dati per valutare l'entità e la distribuzione dei valori mancanti. Da tale analisi, si è rivelato che il 74% delle variabili era affetto da almeno un valore mancante. Ulteriormente, esplorando la distribuzione di questi valori mancanti tra le osservazioni, si è scoperto che tre specifiche osservazioni erano particolarmente problematiche, presentando oltre il 60% di valori mancanti

ciascuna. Queste sono state rimosse dal dataset attraverso un approccio di "Threshold Deletion" per preservare l'integrità complessiva dell'analisi.

Una successiva analisi dettagliata ha evidenziato che ben 26 variabili presentavano un solo dato mancante. Questa singolarità ha portato a un'indagine più approfondita, identificando specifiche osservazioni, in particolare le righe 1670 e 2129, come responsabili. La decisione è stata quella di rimuovere tali osservazioni per preservare la completezza delle variabili affette.

Nonostante ciò, la questione non ha trovato completa risoluzione in questa fase. Data la vastità del dataset, la rimozione pura e semplice di tutte le osservazioni con valori mancanti avrebbe compromesso la sua ricchezza, dato che rappresentavano il 72% del totale. Ciò ha posto un dilemma metodologico: procedere con l'imputazione dei dati mancanti o adottare un approccio di "stepwise selection" variabile per variabile.

Sebbene l'imputazione possa offrire una soluzione rapida, porta con sé diversi rischi. Innanzitutto, l'incorporazione di valori imputati può introdurre un bias nell'analisi, in particolare se le assunzioni sottostanti alla tecnica di imputazione non sono soddisfatte. Inoltre, l'imputazione può portare a una riduzione artificiale della varianza dei dati. Contrariamente, il "stepwise selection" prevede una rimozione graduale delle variabili in base alla quantità e natura dei valori mancanti. Questo assicura che qualsiasi analisi successiva sia basata su dati autentici e non su valori stimati, evitando potenziali distorsioni.

Dopo aver ponderato i pro e i contro di ciascun approccio, si è scelto il "stepwise selection" come metodo più rigoroso e metodologicamente robusto per questo specifico contesto. Questa scelta è nata dalla volontà di mantenere l'autenticità dei dati e di evitare potenziali errori introdotti da stime imprecise.

Al termine di questo processo, il dataset, ora privo di valori mancanti, è composto da 2554 osservazioni e 118 variabili, su un totale di, rispettivamente, 2890 e 126, offrendo una base solida e integra per analisi successiva.

3.3 Divisione in training e test

La divisione del dataset in training e test è un passo fondamentale nell'ambito del machine learning. Serve innanzitutto a valutare le prestazioni del modello in modo obiettivo, poiché il set di test rappresenta dati sconosciuti su cui il modello non è stato addestrato o validato. Questa valutazione aiuta a determinare se il modello è in grado di generalizzare efficacemente su nuovi dati. Inoltre, la divisione dei dati è cruciale per prevenire *l'overfitting*, una

condizione in cui il modello si adatta troppo strettamente ai dati di addestramento, ma non riesce a generalizzare su dati inediti. Inoltre, la divisione dei dati fornisce una stima realistica dell'errore futuro del modello quando viene implementato in produzione, poiché il test set rappresenta dati sconosciuti.

Se ci troviamo in una situazione ricca di dati, l'approccio migliore quindi è quello di dividere il dataset in due parti in modo casuale:

- **Training set:** è il set di dati che utilizziamo per addestrare il nostro modello. È questo set di dati che il nostro modello utilizza per apprendere eventuali schemi o relazioni sottostanti che consentiranno di fare previsioni in seguito. Il set di addestramento deve essere il più rappresentativo possibile della popolazione che stiamo cercando di modellare. Inoltre, è necessario prestare attenzione e assicurarsi che sia il più imparziale possibile, poiché qualsiasi pregiudizio in questa fase può essere propagato a valle durante l'inferenza.

- **Test set:** viene utilizzato per approssimare le prestazioni reali del modello in natura e per valutare l'errore del modello finale scelto. Dovrebbe essere controllato solo come forma finale di valutazione, dopo che il set di validazione è stato usato per identificare il modello migliore.

Non esiste una regola precisa su come selezionare le dimensioni di questi set, si è deciso di dividerlo come mostrato nella figura, ovvero il 67 delle osservazioni appartiene al training set, e il restante 33% al test set.

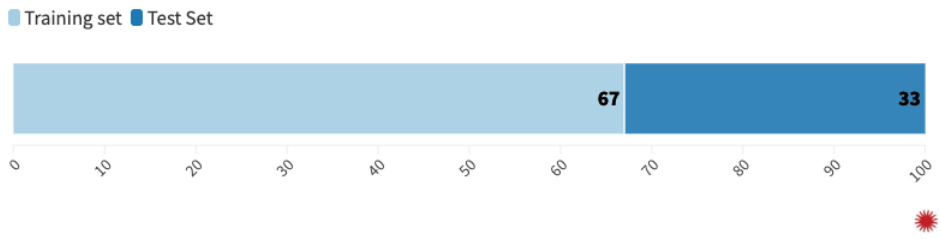

Figura 4: *Percentuali di divisione del dataset iniziale*

3.4 Analisi e Trattamento delle Distribuzioni delle Variabili

Come quarta fase si sono analizzati i boxplot e gli istrogrammi delle variabili. Si è notato dai boxplot di *PK* e *Def.1* e (Figura 5) la presenza di valori anomali, questo perché si hanno elevate percentuali di valori pari a 0 rispetto alle osservazioni del training piccolo come si può vedere nella Tabella 2.

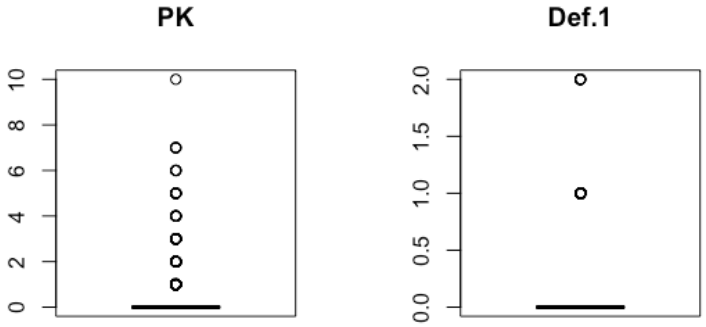

Figura 5: *Boxplot delle variabili "PK" e "Def.1".*

Variabile	**Percentuali di 0 presenti**
PK	92%
Def.1	96%

Tabella 2: *Percentuali di valori pari a 0 rispetto alle osservazioni del training piccolo.*

Data la situazione ho è scelto di escludere dall'analisi queste due variabiabili poiché con il valore 0 così predominante non vengono trasmesse delle informazioni significative.

Con il fine di studiare le distribuzioni delle variabili, notiamo come diverse covariate seguano una distribuzione molto simile tra di loro. Si può infatti notare dalla Figura 6 che le distribuzioni sono fortemente asimmetriche.

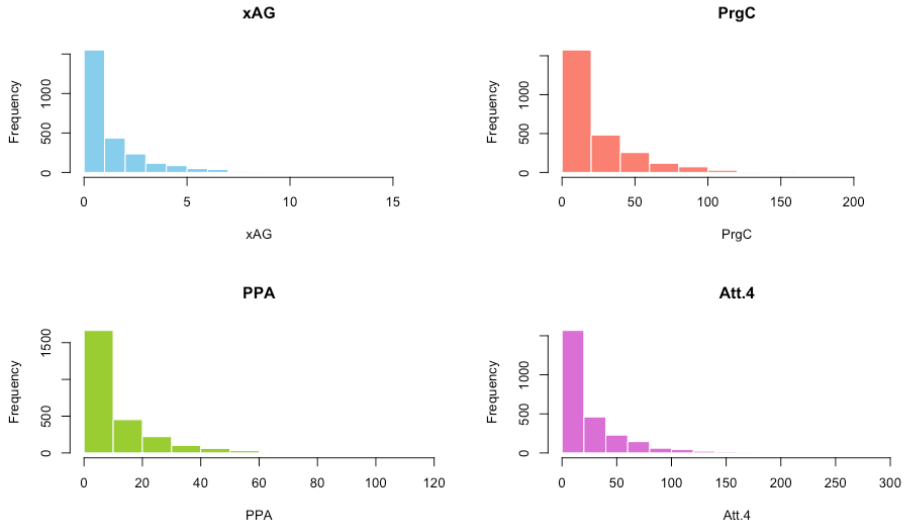

Figura 6: *Istogrammi delle frequenze delle variabili xAG, PrgC, PPA e Att.*

Una possibilità per risolvere questo problema è applicare una trasformazione logaritmica, ma anche in questo modo, le distribuzioni non risultano più simmetriche. (Figura 7).

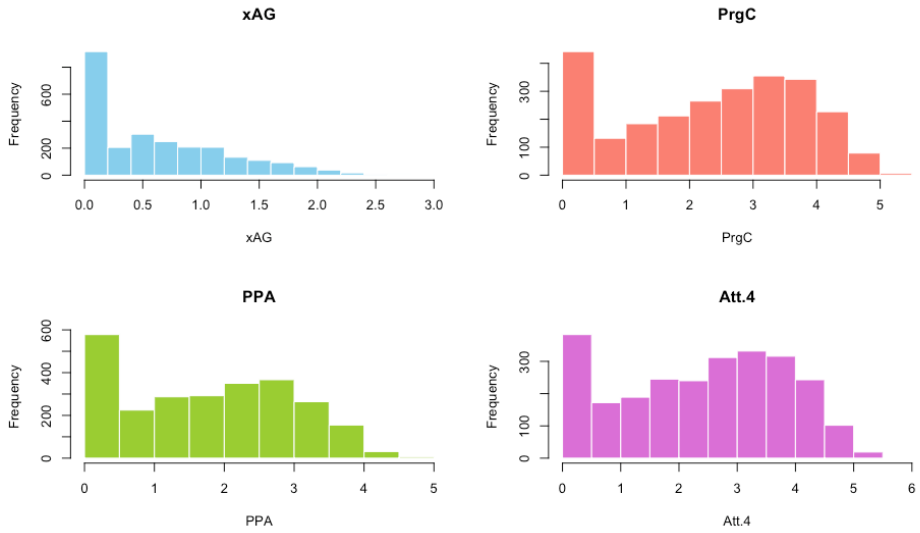

Figura 7: *Istogrammi delle frequenze delle variabili xAG, PrgC, PPA e Att. trasformate.*

Capitolo 4

ANALISI DEI SALARI TRAMITE METODI DI REGRESSIONE LINEARE MULTIPLA

4.1 Regressione Lineare Multipla

La regressione lineare rappresenta uno degli approcci più fondamentali nel campo della statistica e dell'apprendimento automatico per modellizzare e analizzare le relazioni lineari tra variabili. In un modello lineare la variabile di risposta Y_i, è correlato alle covariate attraverso la funzione

$$\mathbb{E}(Y_i) = f(x_i; \beta) = \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} = x_i^T \beta$$

dove $x_i = (x_{i,1}, ..., x_{i,p})^T$ è un vettore di covariate e $\beta = (\beta_1, ..., \beta_p)^T$ è il corrispondente vettore di coefficienti.

I coefficienti sono stimati attraverso il metodo dei minimi quadrati (Ordinary Least Squares, OLS), che mira a minimizzare la somma dei quadrati dei residui:

$$Q(\beta) = \sum_i (y_i - (\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p}))^2$$

Dove:

- n è il numero di osservazioni
- y_i rappresenta l'osservazione i-esima della variabile dipendente.

- x_{ij} rappresenta l'osservazione i-esima della j-esima variabile indipendente.

L'intuizione dietro l'OLS è semplice: vuole trovare la "migliore" retta (o piano, in casi multi-dimensionali) che si adatta ai dati. "Migliore", in questo contesto, significa che la somma delle distanze verticali quadrate tra i dati osservati e i dati stimati dal modello (i residui) è la più piccola possibile. Questo comporta che i coefficienti sono stimati in modo che minimizzino la somma dei quadrati dei residui, da cui il nome "minimi quadrati".

Quando parliamo delle assunzioni fondamentali della regressione lineare, ci riferiamo a una serie di presupposti che il modello fa sul comportamento dei dati e degli errori. Se queste assunzioni non sono soddisfatte, la validità delle stime e delle inferenze ottenute dal modello potrebbe essere messa in discussione. Le assunzioni più comuni ed importanti sono:

1. **Linearità:** Si assume che ci sia una relazione lineare tra le variabili indipendenti e la variabile dipendente. Questo significa che, ad esempio, raddoppiando una variabile indipendente (mantenendo costanti tutte le altre), l'effetto sulla variabile dipendente sarà un cambiamento di una quantità fissa.

2. **Indipendenza:** Si assume che gli errori (o residui) siano indipendenti tra loro. Questo significa che l'errore associato a una particolare osservazione non dovrebbe influenzare l'errore di un'altra osservazione.

3. **Omoschedasticità:** Gli errori hanno varianza costante, il che significa che la dispersione degli errori non varia al variare dei valori delle variabili indipendenti. Questo è in contrasto con l'eteroscedasticità, dove la varianza degli errori varia al variare dei valori delle variabili indipendenti.

4. **Errori normalmente distribuiti:** Si presume che gli errori siano distribuiti normalmente. Questa assunzione è particolarmente importante quando si fanno inferenze statistiche, come i test d'ipotesi.

Infine, per valutare la bontà dell'adattamento, possiamo calcolare il coefficiente di determinazione oppure il cosiddetto coefficiente di determinazione aggiustato: R^2 **(R squared)**
Il coefficiente di determinazione, R^2, è una misura statistica che rappresenta la proporzione di varianza della variabile dipendente che è spiegata dalle variabili indipendenti in un modello di regressione. È un indicatore della bontà

di adattamento del modello ai dati osservati.
La formula per calcolare R^2 è:

$$R^2 = 1 - \left(\frac{SSres}{SStot}\right)$$

Dove:

- $SSres$ è la somma dei quadrati dei residui, calcolata come:

$$SSres = \sum_i (y_i - \hat{y}_i)^2$$

- $SStot$ è la somma totale dei quadrati, calcolata come:

$$SStot = \sum_i (y_i - \bar{y}_i)^2$$

Qui, y_i rappresenta il valore osservato, $\hat{y}_i$ è il valore predetto dal modello, e $\bar{y}$ è la media dei valori osservati.

(R squared adjusted)
Mentre R^2 fornisce una misura di quanto bene le variabili indipendenti spiegano la variabilità della variabile dipendente, ha un problema: R^2può aumentare semplicemente aggiungendo più variabili al modello, indipendentemente dal fatto che queste variabili abbiano un effettivo significato predittivo.
Per ovviare a questo problema, si utilizza R^2_{adj}. L'adjusted R squared penalizza l'aggiunta di variabili non informative al modello.
La formula per calcolare R^2_{adj} è:

$$R^2_{adj} = 1 - (1 - R^2) \cdot \frac{n-1}{n-p-1}$$

Dove:

- n è il numero totale di osservazioni.

- p è il numero di variabili indipendenti nel modello.

Questi indicatori sono fondamentali quando si valuta la qualità di un modello di regressione.
Mentre un alto R^2 può suggerire che il modello si adatta bene ai dati, è sempre bene controllare anche R^2_{adj} per assicurarsi che non si stiano aggiungendo variabili inutili che potrebbero portare a un modello sovradimensionato.

4.2 Applicazione del modello di regressione lineare multipla

Per prima cosa si è deciso di andare ad analizzare la variabile 'Squad', l'unica covariata di tipo factor, che ha 97 livelli, ovvero tutte le squadre dei 5 campionati europei. Si è pensato che l'aggiunta di questa variabile dummy possa portare il modello all'*overfitting*, ovvero il caso in cui il modello si adatta troppo bene ai dati del training, dove lo addestriamo, e si adatta male nel test. Si è pensato dunque di andare a dividere le squadre in 10 gruppi, in base al loro budget per le spese salariali. In questo modo, si è ridotta la dimensionalità della variabile, passando da 97 a 10 livelli complessivi.

Un vantaggio significativo dei modelli lineari è che possono descrivere relazioni non lineari attraverso trasformazioni delle variabili come polinomi, logaritmi, ecc.

Si è quindi pensato di fare una trasformazione logaritmica della variabile risposta 'Annual.Wages' (log(Annual.Wages))

Questa specificazione è lineare nei parametri, risolve i problemi di dominio ed è anche più coerente con la natura dei dati: le previsioni non possono essere negative, trattandosi di salari.

Inizialmente si è quindi andati a stimare un modello lineare sul *training_set* composto da 1417 osservazioni e 86 variabili indipendenti più ovviamente la variabile risposta. Il modello trovato ha i seguenti risultati:

- Il modello lineare stimato ha un indice di determinazione lineare R^2 di 0.6515 e un valore di 0.6266 per l'adjusted R^2. Quindi potremmo dire che la capacità descrittiva del modello sembra relativamente buona.

- Il risultato della statistica F di Fischer è 26.21 con un p-value associato 2.2e-16, e pertanto viene rifiutata l'ipotesi nulla (R^2)=0 suggerendo che il modello di regressione fornisce un adattamento migliore dei dati rispetto a un modello nullo.

Questo risultato però non ci porta alla conclusione che il modello sia un buon modello di previsione perché l'indice di determinazione trovato potrebbe essere soggetto al problema dell'overfitting perché siamo in presenza di 93 esplicative.

Difatti l'aggiunta di variabili nel modello porta a valori di R^2 sempre più elevati e nel nostro caso sono presenti quaranta variabili esplicative.

Ciò è associato al problema della dimensionalità, cioè, che sopra una certa soglia, aggiungere esplicative non migliora il modello selezionato ma aumenta solo R^2. Per questo motivo sarà necessaria la selezione delle variabili, per rimuovere le variabili irrilevanti.

Inoltre, si verifica, attraverso dei grafici (Figura 8), le assunzioni del modello lineare appena stimato per decidere se accettare questo modello.

Il primo grafico in alto a sinistra confronta i valori previsti con i residui e serve a fare diagnostica rispetto all'ipotesi di linearità. La linea rossa ci aiuta ad interpretare il grafico e la vorremmo parallela all'asse delle ascisse. Nel caso in questione la linearità sembra essere abbastanza rispettata.

Il secondo grafico in alto a destra confronta i quantili teorici di una normale con i quantili osservati e va a verificare l'ipotesi di normalità. Nel nostro caso la normalità della variabile 'log(Annual.Wages)' non sembra essere rispettata.

Per conferma di ciò che è stato appena informato si effettua il test di Shapiro-Wilk che pone come ipotesi nulla la normalità dei residui del modello.

Il p-value associato al test è 1.793e-09 che porta a rifiutare l'ipotesi nulla e quindi si può affermare che l'ipotesi di Gaussianità non è rispettata.

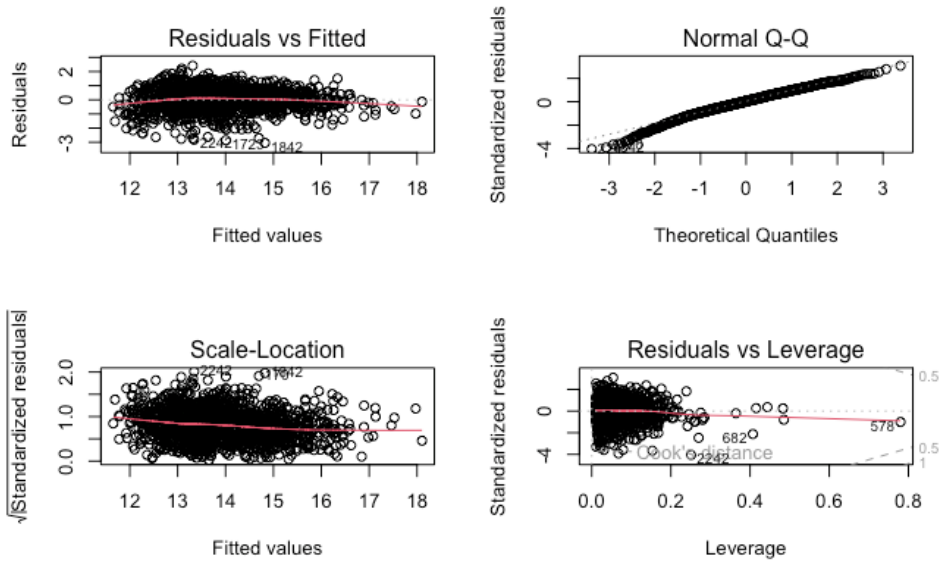

Figura 8: *Diagnostica sul modello lineare*

Il terzo grafico in basso a sinistra confronta i valori previsti con la radice dei residui standardizzati di Pearson e testa la funzione di varianza. Dato che vengono considerati i residui standardizzati ci aspettiamo una dispersione dei punti costante. In questo caso non siamo in presenza di omoschedasticità.

Il quarto grafico è il grafico che va valutare la presenza di punti influenti.

I punti estremi rispetto all'asse delle ascisse sono potenzialmente punti di leva, quelli estremi rispetto all'asse delle ordinate sono potenziali outliers e quelli esterni dalle bande tratteggiate hanno una distanza di Cook elevata.

Vengono evidenziate le osservazioni 578, 2242 e 682 che potrebbero essere outliers, si procede quindi con delle ulteriori verifiche.

Si esegue la diagnostica sulle osservazioni per averne conferma, il primo grafico rappresentato (Figura 9) valuta per tutte le osservazioni la distanza di Cook, residui studentizzati, i p-value con correzione di Bonferroni e gli hatvalues. Si nota che l'osservazione 2242 ha distanza di cook più elevata di tutti, residuo studentizzato elevato. Quindi si afferma che l'osservazione 2242 (Samuele Vignato) è un punto influente, un outlier per il nostro modello. Viene evidenziata anche l'osservazione 1842(Gio Reyna) come outlier, le osservazioni 578 e 1890 come punti di leva (Sacha Delaye e Connor Ronan) e l'osservazione 578 (Sacha Delaye) come punto influente.

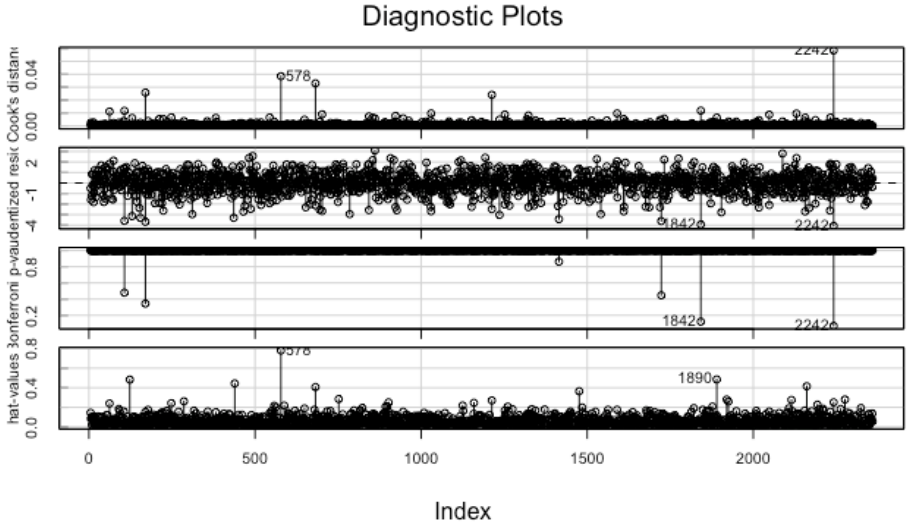

Figura 9: *Grafici sulla diagnostica delle osservazioni valuta per tutte le osservazioni la distanza di Cook, residui studentizzati, i p-value con correzione di Bonferroni e gli hatvalues.*

Il seguente grafico (Figura 10) conferma tutto ciò, in quanto osservazioni estreme rispetto all'asse delle ascisse sono potenziali punti di leva, osservazioni estreme rispetto all'asse delle ordinate sono potenziali outliers e ogni osservazione è racchiusa in un cerchio di area proporzionale alla distanza di Cook.

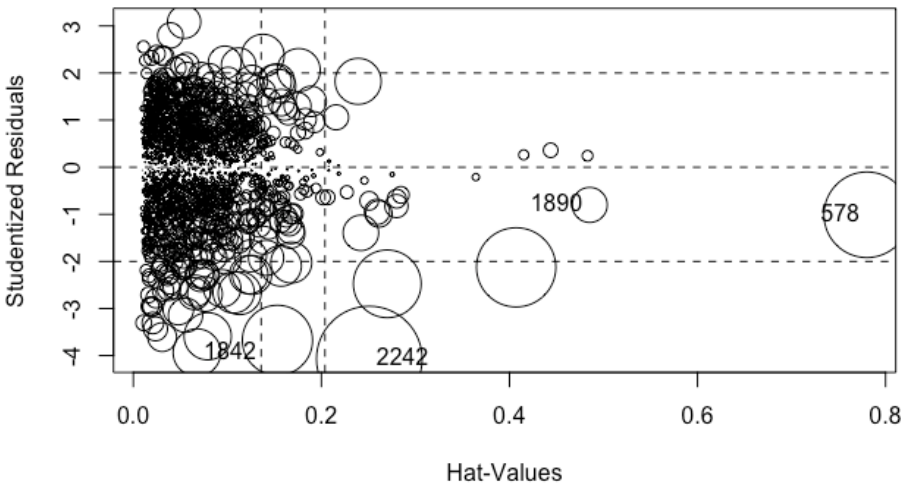

Figura 10: *Le osservazioni estreme rispetto all'asse delle ascisse sono potenziali punti di leva, osservazioni estreme rispetto all'asse delle ordinate sono potenziali outliers e ogni osservazione è racchiusa in un cerchio di area proporzionale alla distanza di Cook.*

4.3 Modello lineare tramite stepwise selection

Nel modello lineare iniziale proposto, si è osservato un soddisfacente coefficiente di determinazione R^2. Tuttavia, alcune assunzioni fondamentali del modello lineare sembravano essere violate. Specificamente, è stata rilevata la presenza di eteroschedasticità e una deviazione dall'assunzione di normalità degli errori. L'eventuale presenza di overfitting poteva inoltre compromettere la generalizzabilità delle stime. Notabilmente, alcune variabili, come il numero di tiri effettuati (Sh), presentavano un'elevata significatività statistica. Tuttavia, questa rilevanza poteva essere dovuta alla loro diversa scala di variazione rispetto ad altre variabili, influenzando così il coefficiente stimato e potenzialmente portando a interpretazioni fuorvianti.

Per ovviare a queste limitazioni, è stata proposta una revisione del modello. Come primo passo, si è proceduto con la standardizzazione delle variabili, trasformando ognuna di esse in modo da avere una media di zero e una deviazione standard di uno. Questa procedura ha lo scopo di armonizzare le scale delle variabili, minimizzando l'influenza di quelle con ampie variazioni. Sebbene l'analisi esplorativa abbia indicato che non tutte le variabili seguono una distribuzione normale, la standardizzazione è stata ritenuta appropriata date le assunzioni sottostanti alla regressione lineare.

Successivamente, per affrontare il rischio di includere predittori irrilevanti che potrebbero gonfiare inutilmente il modello, si è optato per una procedura di selezione variabile attraverso l'approccio stepwise backward. Questa tecnica iterativa inizia con un modello che include tutte le variabili candidate e, ad ogni iterazione, rimuove la variabile che offre il minor contributo informativo, basandosi su criteri predeterminati.

Il criterio adottato per questa selezione è stato l'Akaike Information Criterion (AIC), un indice che quantifica la bontà di adattamento di un modello tenendo conto della complessità del modello stesso. Matematicamente, l'AIC è definito come

$$\text{AIC} = 2k - 2ln(\hat{L}),$$

dove k rappresenta il numero di parametri nel modello e $\hat{L}$ è la massima verosimiglianza del modello. In pratica, modelli con valori di AIC più bassi sono considerati superiori.

Il modello derivante da questo processo rigido e sistematico è come segue:

	Estimate	Std. Error	t value	Pr(>|t|)
(intercept)	13,28633	0,03177	418,177	< 2e-16
rk	0,05643	0,02173	2,597	0,009503
Age	0,45241	0,02253	20,085	< 2e-16
Ast	0,09293	0,04927	1,886	0,059479
CrdY	0,06889	0,02996	2,300	0,021614
xG.1	0,14202	0,04048	3,508	0,000465
Sh.90	-0,09108	0,04018	-2,267	0,023575
FK	0,06670	0,03054	2,187	0,028946
PrgDist	-0,33584	0,14982	-2,242	0,025150
Att.1	-0,56174	0,16333	-3,439	0,000601
Att.3	-0,15445	0,09920	-1,557	0,119730
CrsPA	-0,11850	0,03686	-3,214	0,001338
SCA	-1,24192	0,39603	-3,136	0,001750
SCA90	-0,05062	0,02253	-2,247	0,024811
PassLive	1,03961	0,28574	3,638	0,000285
PassDead	0,27217	0,09697	2,807	0,005074
TO	0,14811	0,05655	2,619	0,008907
Sh.1	0,15273	0,04753	3,213	0,001344
GCA	-0,11378	0,07063	-1,611	0,107450
GCA90	0,07112	0,03191	2,229	0,025974
Def.3rd	0,80539	0,16192	4,974	7,40e-07
Mid.3rd	0,81743	0,18406	4,441	9,68e-06
Att.3rd	0,55336	0,14045	3,940	8,56e-05
Carries	-0,44083	0,11752	-3,751	0,000183
CPA	0,07882	0,04912	1,605	0,108792
Mis	-0,13059	0,05119	-2,551	0,010848
Lost	-0,13443	0,03970	-3,386	0,000728
Blocks	0,08887	0,05225	1,701	0,089200
Int	-0,11126	0,04736	-2,349	0,018963
Clr	-0,17503	0,07144	-2,450	0,014418
Squad_group2	0,90360	0,05517	16,378	< 2e-16
Squad_group3	1,52602	0,07628	20,005	< 2e-16
Squad_group4	1,77814	0,13192	13,479	< 2e-16
Squad_group5	1,62448	0,12919	12,574	< 2e-16
Squad_group6	2,02250	0,20343	9,942	< 2e-16
Squad_group7	2,35545	0,16395	14,367	< 2e-16
Squad_group8	2,09104	0,24081	8,684	< 2e-16
Squad_group9	2,11824	0,16599	12,762	< 2e-16
Squad_group10	2,19502	0,18498	11,866	< 2e-16

Per l'interpretazione dei coefficienti, bisogna fare attenzione, in quanto un modello log-lin è un tipo particolare di modello di regressione in cui la variabile dipendente (risposta) è logaritmizzata, mentre le variabili indipendenti (spiegative) rimangono nella loro forma originale. In questo caso dunque, l'interpretazione del coefficiente di una variabile indipendente è la variazione percentuale nella variabile dipendente per un aumento di una unità in quella variabile indipendente, tenendo tutte le altre variabili costanti. Più precisamente: Variazione percentuale di $y = (e\beta - 1) \times 100\%$.

Da questo modello si evince che le variabili più significative ai fini di percepire un salario maggiore sono l'età, difatti, per un aumento di un anno di età, il salario aumenta di circa il 57.2% a parità di tutto il resto. Invece, con l'aumento di un'unita nei passaggi che hanno portato ad un tiro, si ha addirittura un aumento del 180%.

Inoltre più gol previsti, più tiri effettuati, più tocchi palla in difesa, a centrocampo o in attacco portano a percepire un salario più elevato.

Come si potrebbe prevedere, un incremento di un'unità nei palloni persi comporta una riduzione percentuale del salario del 12%. Curiosamente, un aumento di un'unità nelle azioni che conducono a un tiro determina una decrescita salariale del 70%. Questo fenomeno potrebbe essere attribuito alle intercorrelazioni tra le variabili considerate.

Anche questo modello lineare stimato ha un indice di determinazione lineare R^2 buono simile all'altro di 0.6424 e quindi la capacità descrittiva del nuovo modello sembra buona ma uguale all'altra. Questo è dovuto alla differenza del numero di esplicative dei due modelli, come si è visto l'R^2 del primo modello è influenzato dall'elevato numero di esplicative rispetto al numero di osservazioni.

Si procede quindi, con un confronto dell'AIC tra i due modelli e dell'indice di determinazione corretto. L'AIC del nuovo modello è inferiore (3384.085) rispetto a quello del primo modello (3458.303) e inoltre l'R^2 corretto è superiore (0.6324) anche se di poco rispetto a quello del primo modello (0.6266).

Il nuovo modello individuato è migliore del primo in termini di AIC e di Adjusted R-squared, ed è un risultato aspettato considerando anche l'utilizzo del metodo stepwise con cui sono state rimosse cinquantacinque esplicative e la standardizzazione delle variabili osservate.

Eseguendo la diagnostica sulle osservazioni, (figura 12), si nota che 88 (Julian Kristoffersen) è un punto influente, e outlier del nostro modello, mentre 416 (Jean-Charles Castelletto) è solamente un outlier. Come punto di leva, ovvero che hanno un hat-value elevato abbiamo 673 e 1379 (Domingos Duarte, José Mari).

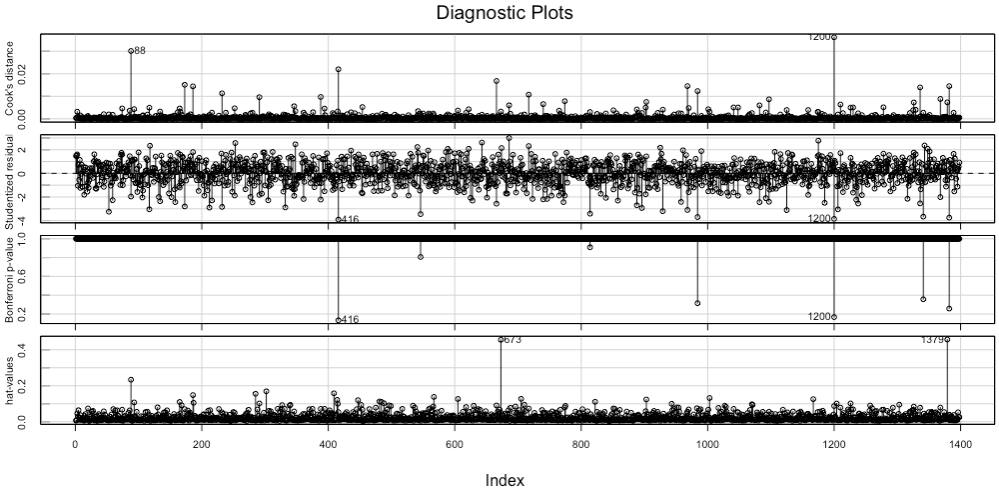

Figura 12: *Grafici sulla diagnostica delle osservazioni valuta per tutte le osservazioni la distanza di Cook, residui studentizzati, i p-value con correzione di Bonferroni e gli hatvalues.*

Si è verificato che anche questo nuovo modello lineare non rispetta completamente le assunzioni ma è presente un miglioramento in termini di indice di bontà corretto e nell'AIC. Si decide quindi di passare ad un modello più complesso, per ridurre la dimensionalità come l'analisi PCA.

Prima di eseguire tale modello si descrive in forma teorica che cos'è e come funziona l'analisi delle componenti principali.

4.4 Analisi delle Componenti Principali (PCA) e Regressione sulle Componenti Principali (PCR)

Nell'ambito della statistica, quando ci troviamo di fronte ad un'elevata dimensionalità delle covariate, si presentano numerosi problemi. Una delle soluzioni proposte per affrontare questa problematica è l'utilizzo della PCA, un metodo che cerca di comprimere le informazioni contenute in un insieme di covariate attraverso una rappresentazione di dimensione inferiore, conservando al contempo la maggior parte delle informazioni. La PCA opera attraverso la determinazione di un nuovo set di variabili, le componenti principali Z, che sono combinazioni lineari delle variabili originali X. La matrice delle covariate X viene sostituita dalla matrice Z di dimensione ridotta k, che mantiene la maggior parte della varianza dei dati originali.

Formalmente, il processo di centramento dei dati comporta la sottrazione della media, cioè $x_{i,j} - \bar{x}_j$ per le covariate e $y_i - \bar{y}$ per la variabile dipendente. Questa operazione di centramento è cruciale in quanto la PCA si basa sulla decomposizione dei valori singolari (SVD) di una matrice centrata.

Data una matrice X di dimensione $n \times p$, la sua SVD è espressa come

$$X = U \cdot D \cdot V^T,$$

dove D è una matrice diagonale che contiene i valori singolari, mentre U e V sono matrici ortogonali.

La matrice $X^T \cdot X$ può quindi essere espressa come

$$X^T \cdot X = V \Delta^2 V^T,$$

dove Δ^2 è la matrice diagonale dei quadrati dei valori singolari.

Le colonne di $Z = XV$ sono le componenti principali, che sono ortogonali tra loro. Queste componenti rappresentano combinazioni lineari delle variabili originali come $z_{i,j} = x_i^T \cdot v_j$.

Le componenti principali non sono scelte a caso: la prima componente principale è la combinazione lineare delle variabili originali con la massima varianza, ovvero

$$v_1 = \arg\max \frac{1}{n} v^T \cdot X^T \cdot X v,$$

soggetta alla condizione $v^T \cdot v = 1$.

Seguendo questa logica, le componenti successive sono scelte per massimizzare la varianza rimanente e sono ortogonali alle componenti già selezionate.

Una volta ottenute le componenti principali, è possibile utilizzare la PCR per la previsione. In PCR, le componenti principali sono utilizzate come nuove covariate in una regressione lineare. L'ortogonalità delle componenti principali rende semplice la stima dei coefficienti di regressione. Il parametro k, ovvero il numero di componenti principali da utilizzare, è fondamentale in PCR e può essere scelto attraverso criteri informativi o validazione incrociata.

In conclusione, la PCA e la PCR sono strumenti potenti che offrono soluzioni al problema della multicollinearità e dell'elevata dimensionalità in statistica, consentendo di rappresentare e prevedere i dati in uno spazio di dimensione ridotta, mantenendo la maggior parte delle informazioni originali.

4.5 Applicazione Analisi Componenti Principali (PCA) e Regressione sulle Componenti Principali (PCR)

Nel contesto dell'analisi multivariata, l'Analisi delle Componenti Principali (PCA) rappresenta una tecnica fondamentale per la riduzione dimensionale e la semplificazione della complessità intrinseca dei dati. Al fine di preparare il dataset per la PCA, abbiamo effettuato delle trasformazioni preliminari: la variabile categorica 'Squad_Group' è stata codificata come dieci variabili binarie distinte, rendendo così ogni livello un indicatore binario di appartenenza a un particolare gruppo.

Questa codifica è cruciale perché la PCA, per sua natura, richiede un input costituito esclusivamente da variabili numeriche. Tuttavia, un passo metodologico ulteriore riguarda la rimozione della variabile risposta. Questa decisione si basa su solide motivazioni tecniche: la PCA è una tecnica di analisi non supervisionata, designata per esplorare la struttura intrinseca delle sole variabili esplicative, indipendentemente dalla variabile target. Integrare la variabile risposta avrebbe potuto compromettere la qualità e l'interpretazione delle componenti estratte, facendo perdere l'essenza puramente esplorativa e non supervisionata della PCA.

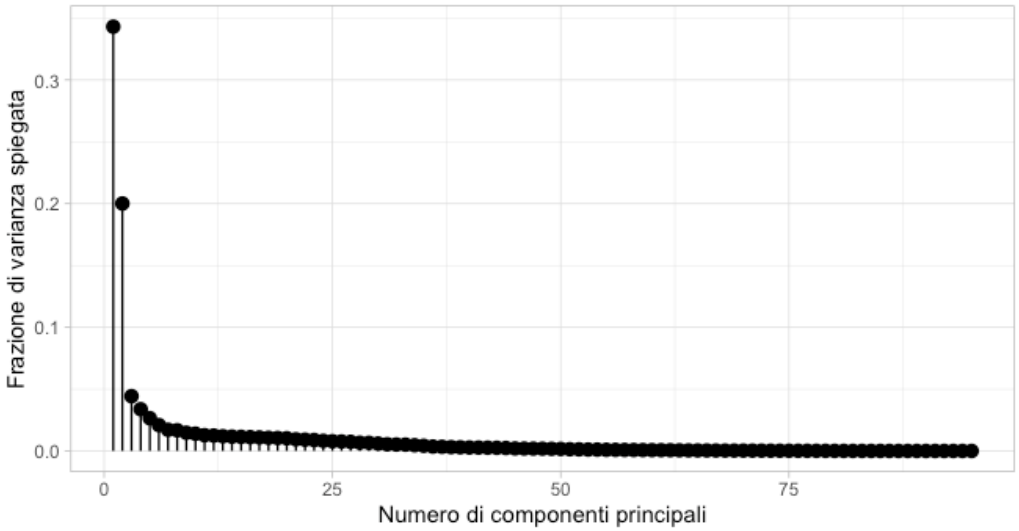

Figura 13: *Percentuale di varianza spiegata da ogni componente principale.*

È stato osservato che la prima componente spiega circa il 40% della varianza totale. Sebbene una prassi comune possa essere quella di selezionare componenti basandosi sulla copertura di una certa proporzione della varianza complessiva (ad esempio, il 90%), in questo studio abbiamo optato per un approccio diverso.

La strategia adottata consiste nell' applicare la regressione sulle componenti principali (PCR), come già spiegato precedentemente, e quindi stimare modelli sequenziali, iniziando con un modello che incorpora solo la prima componente principale e calcolando, per ciascuno, l'Errore Quadratico Medio (MSE) - una metrica che quantifica l'accuratezza di un modello. Questo processo iterativo prosegue incorporando una componente alla volta, re-estimando il modello e confrontando gli MSE risultanti.

Questa procedura iterativa prosegue fino a includere l'intero insieme delle componenti principali. Una volta raggiunto questo stadio, il modello essenzialmente equivale alla matrice dei dati originale, ma trasformata in un nuovo sistema di coordinate ortogonali. Tale modello, in assenza di qualsiasi riduzione dimensionale, produrrà risultati identici a quelli di una regressione lineare ordinaria (OLS) sul set di dati originale.

Da un'analisi approfondita emerge che il modello che incorpora le prime 31 componenti principali è il più performante, con un MSE di 6.36×10^{12}, indicando così il giusto equilibrio tra complessità del modello e capacità predittiva.

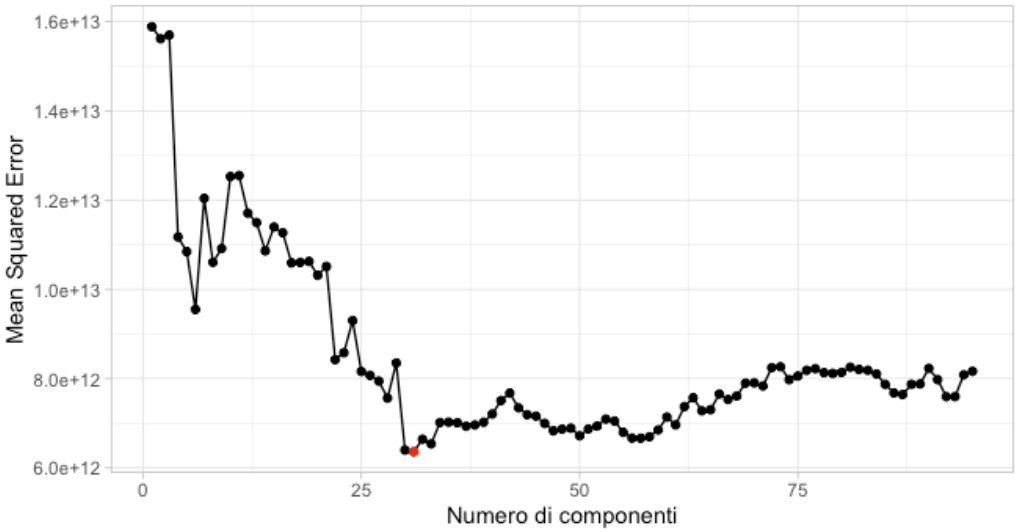

Figura 14: *Variazione dell' MSE, in scala originale, in relazione all'incremento delle componenti principali nella PCR.*

Capitolo 5

RISULTATI

5.1 Confronto tra modelli

Nel presente capitolo, viene condotta un'analisi comparativa dei modelli di regressione al fine di selezionare il modello ottimale, la cui accuratezza sarà ulteriormente indagata. Tra i due modelli di regressione lineare proposti, il secondo, ottenuto attraverso la procedura stepwise, si è distinto per la sua superiorità in termini di indice di determinazione corretto, valore dell'informazione di Akaike e modificazioni operate sui dati originali.

Si è poi passati all'analisi comparativa tra il modello di regressione lineare ottenuto mediante procedura stepwise e quello derivante dalla regressione sulle componenti principali (PCR) nell'ottica di identificare il modello più adatto alla previsione degli stipendi nel *test_set*.

Nella fase successiva dell'analisi, si è optato per l'impiego congiunto dei due modelli stimati al fine di predire gli stipendi del *test_set* e sottoporre tali stime a un confronto con i valori effettivamente osservati, avvalendosi dell'Errore Quadratico Medio (MSE) come principale metrica di accuratezza. Il MSE rappresenta, per sua natura, la media delle differenze al quadrato tra le osservazioni effettive e quelle previste dal modello. È d'importanza cruciale sottolineare come l'obiettivo primario di qualsiasi modello di regressione sia quello di fornire stime il più possibile aderenti ai dati effettivi, allo scopo di ottimizzare la precisione della previsione. Da un punto di vista pratico, un MSE inferiore è indice di un modello predittivo di maggior qualità.

Il campione preso in considerazione per questa fase dell'analisi comprende le unità che, in seguito alla divisione casuale del dataset originario in training e *test_set*, sono state assegnate a quest'ultimo, per un totale di 688 osservazioni. Le previsioni sugli stipendi, ottenute tramite i due modelli in questione, sono state confrontate con i dati effettivi mediante il calcolo dell'MSE. Que-

st'ultimo è stato determinato utilizzando la formula: $1/n \sum_i (Y_i - \hat{Y}_i)^2$, dove n denota la numerosità del campione, Y_i gli stipendi effettivamente percepiti dai 688 giocatori e $\hat{Y}_i$ gli stipendi previsti dai modelli.

Prima di procedere con la valutazione dei modelli su dati non visionati, è stata necessaria un'accurata fase di standardizzazione dei dati presenti nel test_set. È d'importanza cruciale menzionare che, avendo precedentemente proceduto alla standardizzazione dei dati nel set di addestramento, è imprescindibile applicare la stessa trasformazione anche al test_set per garantire coerenza e corretta interpretazione dei risultati.

Tuttavia, un aspetto fondamentale da tener presente in questa fase riguarda la metodologia di standardizzazione: è essenziale che le statistiche utilizzate per la standardizzazione (specificamente la media e la deviazione standard) siano quelle derivanti dal set di addestramento e non quelle calcolate direttamente sul test_set. Tale approccio è motivato dalla necessità di testare il modello in condizioni che riflettano fedelmente l'ambiente in cui è stato calibrato. Adottando le statistiche del set di addestramento, si garantisce una simulazione realistica dell'applicazione del modello a nuovi dati, replicando un contesto in cui le statistiche dei dati freschi non sono a priori disponibili.

Si è notato che entrambi i valori, calcolati su scala originale, sono molto elevati, il risultato non sorprende vista la grandezza dei valori della variabile stipendi. Inoltre, i due risultati sono molto simili, difatti, nel modello lineare il valore è leggermente maggiore ed è uguale a 1.25×10^{13} mentre il valore dell'MSE nel modello stimato con PCR corrisponde a 1.12×10^{13}.

Come ulteriore paragone, anche dal punto di vista grafico sono rappresentati i grafici di dispersione tra i valori predetti dai modelli e i valori osservati cioè gli stipendi reali della stagione 2022/2023. Più il modello è accurato più i punti del grafico dovrebbero distribuirsi lungo la bisettrice rappresentata in nero.

Sia i salari stimati con il secondo modello lineare rappresentati in blu nel grafico 16 che i salari stimati attraverso la PCR rappresentati in rosso nella figura 15 seguono una distribuzione simile. Infatti, i punti nei grafici non si discostano in modo significativo dalla bisettrice del quadrante. Si nota che in corrispondenza dei salari più bassi i modelli hanno una capacità previsiva minore.

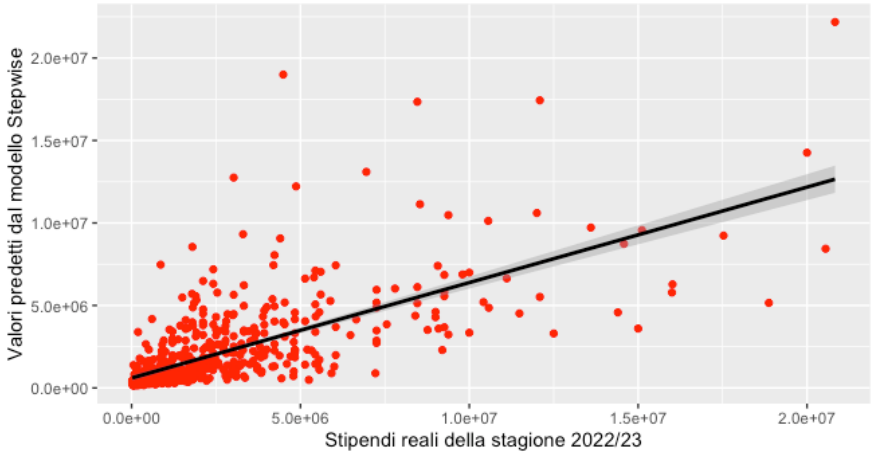

Figura 15: *Sull'asse delle ascisse i salari reali e sull'asse delle ordinate i salari previsti dal secondo modello lineare.*

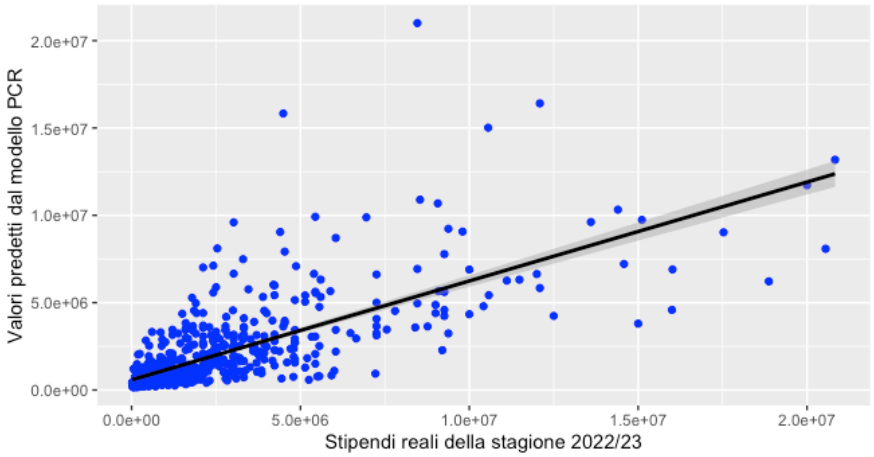

Figura 16: *Sull'asse delle ascisse i salari reali e sull'asse delle ordinate i salari previsti dal modello PCR.*

Come possiamo vedere anche graficamente non si nota tanta differenza tra i 2 modelli.

Dall'analisi di questi grafici, dal confronto dell'indice di determinazione, e dal paragone dei due errori quadratici medi si è osservato come non è possibile giungere alla conclusione di quale modello di previsione sia migliore per lo studio effettuato sugli stipendi.

Si decide di scegliere il modello PCR perché ritenuto un miglioramento rispetto al secondo modello lineare in quanto riduce la dimensionalità, e uti-

lizza 31 componenti principali, invece delle 38 utilizzate dal secondo modello lineare.

Infine, si descrivono le principali esplicative con relativi commenti ai coefficienti stimati del modello di regressione sulle componenti principali.

5.2 Variabili influenti

In questa sezione, si è proceduto all'analisi dell'incidenza di alcune variabili esplicative sulla variabile dipendente, quantificando la variazione percentuale della variabile di risposta associata a un incremento unitario di una specifica covariata, mantenendo costanti le altre condizioni e considerando la possibile interdipendenza con altre variabili, alcune delle quali potrebbero essere correlate. È interessante notare che alcuni coefficienti di regressione congiunta possono manifestare un segno opposto rispetto alla loro valenza marginale. Si è condotto uno studio approfondito sulle variabili che inducono una considerevole variazione percentuale nel modello e su altre variabili che, a priori dell'analisi, potrebbero sembrare significative, quali ad esempio il numero di tiri, gol realizzati o assist forniti.

Come già detto precedentemente dato che si è utilizzata la variabile risposta logaritmica, un coefficiente beta non rappresenta un cambiamento assoluto nella variabile dipendente per un cambiamento unitario nella variabile indipendente. Invece, rappresenta il cambiamento percentuale. Specificamente, un aumento di una unità in una variabile indipendente risulterà in un cambiamento del $100 \times \beta\%$ nella variabile dipendente.

Età-Age

L'età emerge come una variabile predictor di rilievo in tutti i modelli analizzati. Un incremento unitario dell'età, rispetto alla media, si traduce in un aumento salariale approssimativo del 53%. Questa associazione può essere attribuita a diversi fattori. L'esperienza accumulata con l'avanzare dell'età potrebbe migliorare le performance del giocatore sul campo, giustificando una remunerazione più elevata. Inoltre, i club tendono a offrire salari più generosi ai giocatori che hanno dimostrato nel tempo le loro capacità e che hanno guadagnato una reputazione consolidata nel mondo del calcio.

Gruppo di squadra-Squad_Group

La variabile 'Squad_Group', analogamente alla variabile 'Age', si manifesta come un fattore cruciale nella previsione salariale dei calciatori. Trattan-

dosi di una variabile categorica, presenta diverse sfumature nei coefficienti, oscillando tra una riduzione del 35% per i giocatori del gruppo 1 fino ad un incremento del 130% rispetto a quelli del gruppo 1, per quelli del gruppo 10, comprendendo club di prestigio come il Paris Saint Germain e il Real Madrid. Queste discrepanze sottolineano le profonde iniquità nel potere d'acquisto tra le squadre di élite e quelle di rango inferiore. La capacità delle squadre di punta di attrarre e remunerare generosamente i migliori talenti amplifica ulteriormente queste disuguaglianze. L'importanza di questa variabile nel modello sottolinea come la filiazione di un giocatore ad una specifica squadra sia un elemento chiave nella determinazione del suo salario: un calciatore d'élite in una squadra di rango inferiore, come nel caso ipotetico di Messi all'Ajaccio (gruppo 10), percepirebbe un salario notevolmente inferiore al suo potenziale di mercato.

Tempo di gioco-MP, Starts, Min

Le variabili 'MP', 'Starts' e 'Min' quantificano rispettivamente le partite giocate, le partite iniziate come titolari e i minuti totali trascorsi in campo. Sebbene 'Starts' non mostri una significatività nell'analisi, le variabili 'Min' e 'MP' presentano una correlazione parzialmente significativa con il salario. In particolare, un incremento unitario rispetto alla media dei minuti giocati conduce ad un aumento salariale del 1%. Allo stesso modo, disputare una partita in più rispetto alla media si traduce in un incremento salariale del 2%. Questa tendenza si giustifica considerando che i giocatori di maggiore talento o con un ruolo cruciale all'interno della formazione sono più propensi ad essere impiegati regolarmente e, di conseguenza, a ricevere una remunerazione maggiore.

Goals-Gls, Gls.1, xG, xG.1, npxG.xAG

Queste variabili sono indici dei goal realizzati o dei goal attesi da un calciatore. Emergono come variabili significative all'interno del modello. Concretamente, marcare un gol in più rispetto alla media porta ad un incremento salariale del 2,5%. Nonostante la rilevanza intrinseca della capacità di segnare nel calcio, è sorprendente notare che queste variabili siano meno incisive rispetto all'età o alla squadra di appartenenza nell'influenzare il salario. Questa peculiarità potrebbe derivare dalla diversità dei ruoli dei giocatori inclusi nel dataset. Se l'analisi fosse stata circoscritta ai soli attaccanti, è probabile che tali variabili avrebbero manifestato una maggiore significatività.

Assist-Ast, Ast.1, xAG, xAG.1, xG.xAG

Le variabili che riguardano i passaggi che culminano in un goal, o gli assist attesi che potrebbero condurre ad un goal, assumono un'importanza sorprendentemente elevata nel modello. Contrariamente alle aspettative iniziali, queste variabili mostrano un impatto sul salario comparabile a quello dei goal. Specificamente, fornire un assist in più rispetto alla media implica un incremento salariale analogo a quello generato da un goal, ovvero circa il 2,5%. Questo risultato suggerisce che nel contesto moderno del calcio, la capacità di creare opportunità di goal, quantificata attraverso gli assist, è valutata dalle società in maniera simile all'abilità di finalizzazione.

Ruoli-FW, MF, DF

Le variabili associate alla posizione del giocatore in campo rivestono un'importanza peculiare nel modello. Mentre la posizione di centrocampista non mostra un impatto significativo sulla remunerazione, risultati inattesi emergono per quanto riguarda difensori e attaccanti. Sorprendentemente, il modello indica che un difensore, rispetto alle altre posizioni, può aspettarsi un incremento salariale del 2,6%. Ancora più notevole è il dato relativo agli attaccanti, che evidenzia una diminuzione salariale del 3,5% rispetto agli altri ruoli.

Passaggi-Cmp.1, Cmp.2, Cmp.3 Att.1, Att.3

Le variabili in esame forniscono informazioni riguardo i passaggi tentati (indicati come "Att") e quelli completati (indicati come "Cmp"), nonché la loro lunghezza: con "var.1" si intendono passaggi corti, con "var.2" quelli di lunghezza media e con "var.3" i passaggi lunghi. La nostra analisi suggerisce che i passaggi corti tentati e riusciti rivestono una significatività notevole nel determinare il salario dei giocatori. Specificamente, per ogni passaggio corto completato in eccesso rispetto alla media, si osserva un incremento salariale del 7%. Al contrario, i passaggi di lunghezza media non mostrano un impatto distintivo sul salario. Sorprendentemente, i passaggi lunghi, sia tentati che completati, portano a una diminuzione salariale del 4% per ogni unità in eccesso rispetto alla media. Questa tendenza può essere interpretata in vari modi. Una possibile spiegazione è che il modello sembra favorire giocatori che adottano uno stile di gioco più cauto e prediligono passaggi corti, minimizzando il rischio di perdite di palla. Al contrario, chi si impegna in passaggi più lunghi e potenzialmente rischiosi potrebbe essere penalizzato in termini di remunerazione.

5.3 Giocatori sopravvalutati e sottovalutati

In questo sottocapitolo, analizzeremo l'intero dataset attraverso l'applicazione del modello PCR (Principal Component Regression), precedentemente identificato come il modello ottimale, per determinare quali giocatori vengono sovrastimati o sottostimati in termini salariali. L'obiettivo è identificare quei giocatori che, secondo l'analisi, ricevono un salario che non corrisponde alle loro prestazioni o caratteristiche misurate dalle variabili nel modello.

L'analisi si basa sulla differenza tra il salario previsto dal modello e il salario reale percepito dal giocatore. Un valore positivo di questa differenza indica che il giocatore potrebbe essere sottopagato rispetto a quanto previsto dal modello, mentre un valore negativo suggerisce che il giocatore potrebbe essere sovrastimato.

Giocatore	**Squad_Group**	**Differenza salariale**
Karim Benzema	10	29946777
Riyad Mahrez	7	22032239
Thiago Silva	7	20104842
Mohamed Salah	6	17292541
Rodri	7	17046925
Robert Lewandowski	9	15867147
Olivier Giroud	3	14259595
Eric Maxim Choupo-Moting	9	10859479
Ashley Young	3	10522527
Zlatan Ibrahimović	3	10279865
Francesco Acerbi	5	10087352
Bruno Fernandes	8	9670955
James Milner	6	9088277
Chris Smalling	3	8922972
İlkay Gündoğan	7	8431298
Martin Ødegaard	4	7142285

Edin Džeko	5	6853309
Andrew Robertson	6	6549924
Érik Lamela	3	6508547
Marcus Rashford	8	6283173
Raphinha	9	6214922
Granit Xhaka	4	6153155
Jesús Vallejo	10	6003732
Daniel Parejo	2	5591439
Bernardo Silva	7	5424228

Tabella 3: *Elenco dei 25 calciatori maggiormente sottovalutati in termini salariali secondo le stime del modello.*

Sorprendentemente, tra i giocatori ritenuti sottostimati dal modello, troviamo alcune figure di primo piano nel panorama calcistico. Per esempio, Karim Benzema, recente vincitore del Pallone d'Oro, risulta essere il giocatore più sottostimato nonostante un salario annuale di 24 milioni di euro; il modello suggerisce che dovrebbe percepire una cifra significativamente superiore, quasi il doppio. Altri esempi notevoli includono i campioni europei Bernardo Silva e Mahrez, i quali, con salari attuali di 9 e 9,5 milioni di euro, dovrebbero ricevere rispettivamente un incremento salariale di 5,5 milioni e quasi 22 milioni secondo le stime.

Analizzando la distribuzione per ruolo, tra i 25 giocatori più sottostimati, 11 sono attaccanti, 5 difensori e 9 centrocampisti. Questa distribuzione potrebbe riflettere l'importanza di certi ruoli all'interno del campo, ma è anche un segnale dell'eterogeneità dei salari nel mondo del calcio e di come vari fattori, oltre alle sole prestazioni in campo, possano influenzare la remunerazione di un giocatore.

Giocatore	Squad_Group	Differenza salariale
Kylian Mbappé	10	- 57.856.621
Neymar	10	- 43.009.874
Frenkie de Jong	9	-33.221.863
Eden Hazard	10	-22.713.764
Gerard Piqué	9	-18.050.725
Cristiano Ronaldo	8	-17.294.572
Toni Kroos	10	-17.287.161
Jadon Sancho	8	-13.631.322
Niklas Süle	5	-13.020.251
David Alaba	10	-12.741.392
Raphaël Varane	8	-10.893.616
Lucas Hernández	9	-10.862.362
Erling Haaland	7	-10.422.937
Achraf Hakimi	10	-10.384.183
Gabriel Jesus	4	-9.848.509
Kingsley Coman	9	-9.678.426
Marco Verratti	10	-9.604.189
Serge Gnabry	9	-9.561.530
Matthijs de Ligt	9	-9.092.701
Dušan Vlahović	5	-8.980.099
Aurélien Tchouaméni	10	-8.887.997
Reece James	7	-8.746.979
Lionel Messi	10	-8.737.826
Jules Koundé	9	-8.674.515
Koke	4	-8.669.925

Tabella 4: *Elenco dei 25 calciatori maggiormente sopravvalutati in termini salariali secondo le stime del modello.*

Nell'analisi dei giocatori sovrastimati, risaltano immediatamente due figure emblematiche del panorama calcistico mondiale: Kylian Mbappè e Neymar, stelle luminose del Paris Saint-Germain (PSG). Il modello suggerisce che, rispetto alle performance e variabili considerate, entrambi i giocatori dovrebbero percepire compensi notevolmente inferiori: 57 milioni in meno per Mbappè e 43 milioni in meno per Neymar. Questo dato, pur essendo notevole, non sorprende del tutto, considerando le consolidate tendenze del PSG a offrire retribuzioni generose, spesso al di sopra delle stime di mercato, ai suoi giocatori chiave.

Un'osservazione rilevante emerge analizzando la distribuzione delle squadre di appartenenza dei calciatori sovrastimati: solo 6 su 25 non fanno parte delle squadre classificate nei gruppi 8, 9 e 10, ovvero le categorie che raggruppano i club economicamente più potenti a livello mondiale. Questa constatazione mette in luce un evidente squilibrio tra il salario reale e quello previsto dal modello per i giocatori militanti in questi club d'élite. Pur riconoscendo che l'appartenenza a una squadra di alto profilo tende ad elevare la retribuzione, sembra che in realtà l'entità di questa differenziazione salariale sia superiore alle stime del modello. Ciò potrebbe suggerire che vi sono ulteriori variabili non considerate nell'analisi che influenzano significativamente la retribuzione dei giocatori di questi top club.

Infine, osservando la distribuzione per ruolo, tra i calciatori sovrastimati abbiamo 11 attaccanti, 9 difensori e 5 centrocampisti. Questo dato potrebbe offrire spunti di riflessione sull'effettiva remunerazione relativa ai ruoli in campo, tenendo conto delle variabili considerate nel nostro modello di regressione.

CONCLUSIONI

Nel corso di questo studio, si è investigato il panorama salariale dei calciatori che militano nei cinque principali campionati europei, con l'intento di elaborare una stima affidabile dei salari basata su variabili statistiche. L'analisi si è focalizzata sulle prestazioni in campo dei calciatori nella stagione 2022/2023, integrando altresì variabili anagrafiche come l'età, la squadra di appartenenza e il campionato di militanza. Questa indagine aveva lo scopo di fornire uno strumento quantitativo su cui giocatori, società e procuratori potessero fare riferimento per delineare le fasce salariali, al di là di fattori esterni.

Dopo aver delineato il quadro economico e finanziario del settore calcistico, ponendo particolare enfasi sulla questione salariale e sulle iniziative, come il Fair Play Finanziario, mirate a regolare l'escalation retributiva, si è proceduto con l'acquisizione e la preparazione dei dataset. Questi ultimi comprendevano sia le statistiche legate alle prestazioni dei giocatori che i dettagli retributivi.

L'approccio metodologico ha privilegiato inizialmente il modello di regressione lineare, partendo dall'ipotesi di una correlazione lineare marcata tra salari e prestazioni. Al fine di perfezionare l'analisi e mitigare il rischio di overfitting, sono stati integrati il metodo stepwise e la PCA. I risultati dei modelli hanno evidenziato valori di MSE molto simili (1.25×10^{13} e 1.12×10^{13}.). Sebbene non si sia giunti a una distinzione netta sulla superiorità di uno dei due modelli, si è optato per il modello PCA per le sue qualità sintetiche e un lieve vantaggio in termini di MSE.

È essenziale riconoscere che, nonostante l'accuratezza dei modelli elaborati, esistono delle limitazioni intrinseche. Al di là delle variabili quantitativamente misurabili, elementi come la reputazione, la visibilità mediatica e la fan base di un calciatore giocano un ruolo cruciale nella determinazione dei salari. La localizzazione geografica delle squadre, l'interesse degli sponsor e l'analisi degli esperti del settore sono ulteriori componenti che possono influenzare le decisioni salariali. La complessità del panorama salariale va oltre le sole statistiche.

Pertanto, per affinare ulteriormente le previsioni salariali, sarebbe opportuno incorporare variabili che tengano conto di questi fattori. Un ulteriore spunto di riflessione potrebbe riguardare l'introduzione di un "salary cap", ispirandosi ai modelli statunitensi, al fine di rendere il calcio più sostenibile, meritocratico e competitivo, garantendo nel lungo termine uno spettacolo di qualità per gli appassionati.

BIBLIOGRAFIA

- Ace Advisory Zrt. (2019). *Broadcasting revenue landscape – big money in the "big five" leagues.* Football Benchmark.
- Ace Advisory Zrt. (2022). *Is the english premier league the european super league already?.* Football Benchmark
- Chenyao, L. & Kampakis, S. & Treleaven, P. (2022). *Machine Learning Modeling to Evaluate the Value of Football Players*
- Deloitte's Sports Business Group. (2023). *Annual Review of Football Finance 2023*
- Feess, E., Gerfin, M., & Muehlheusser, G. (2010). *The incentive effects of long-term contracts on performance-evidence from a natural experiment in european soccer.* Technical Report, Mimeo: Berlin.
- Fifa. (2022). *Intermediaries in international transfers 2021.*
- Franceschi, M. & Brocard, J.-F. & Follert, F. & Gouguet, J.-J. (2023). *Determinants of football players' valuation: a systematic review.* Journal of Economic Surveys
- Frick, B. (2006). *Salary determination and the pay-performance relationship in professional soccer: Evidence from germany. Sports Economics After Fifty Years: Essays in Honour of Simon Rottenberg.* Oviedo: Ediciones de la Universidad de Oviedo, 125– 146.
- Frick, B. (2007). *The football players' labor market: empirical evidence from the major european leagues.* Scottish Journal of Political Economy
- Frick, B. (2011). *Performance, salaries, and contract length: empirical evidence from german soccer.* International Journal of Sport Finance, 6(2), 87.

- Garcia-del Barrio, P., & Pujol, F. (2007). *Pay and performance in the spanish soccer league: who gets the expected monopsony rents.* Technical report, University of Navarra, Spain
- Garcia-del Barrio, P., & Pujol, F. (2009). *The rationality of underemploying the bestperforming soccer players.* Labour, 23(3), 397–419.
- Houmes, J. (2008). *European Football... Its Time for a Salary Cap—Part One*
- Lames, M., McGarry, T., Nebel, B., & Roemer, K. (2011). *Computer science in sportspecial emphasis: Football* (dagstuhl seminar 11271). Dagstuhl Reports, 1(7).
- Mignani, L. (2023). *Quali sono gli sport più popolari e praticati al mondo?.* Men's Health
- Montaño Casillas, J. (2020). *Soccer Analytics: Prediction of salary and market value using machine learning* (1/3). Analytics Vidhya
- O'Donoghue, P. & Robinson, G. (2009). *Validity of the prozone3 r player tracking system: A preliminary report.* International Journal of Computer Science in Sport, 8(1), 37–53.
- Poli, R. & Ravenel, L. & Besson, B. (2020). *Scientific evaluation of the transfer value of football players.* CIES Football Observatory Monthly Report
- Ramazzotti, A. (2023). *Quanti affari: solamente nel 2022 pagati 622 milioni ai procuratori.* Gazzetta
- Siegle, M., Stevens, T., & Lames, M. (2013). *Design of an accuracy study for position detection in football.* Journal of Sports Sciences, 31(2), 166–172.
- Torgler, B. & Schmidt, S. L. (2007). *What shapes player performance in soccer? empirical findings from a panel analysis.* Applied Economics, 39(18), 2355–2369.
- Torgler, B., Schmidt, S. L., & Frey, B. S. (2006). *Relative income position and performance: an empirical panel analysis.*
- Wicker, P., Prinz, J., Weimar, D., Deutscher, C., & Upmann, T. (2013). *No pain, no gain? effort and productivity in professional soccer.* International Journal of Sport Finance, 8(2), 124.

• Yaldo, L. & Shamir, L. (2017). *Computational Estimation of Football Player Wages.* International Journal of Computer Science in Sport

SITOGRAFIA

- https://elearning.unimib.it/course/view.php?id=43429
- https://fbref.com/it/comp/Big5/stats/calciatori/Statistiche_di_I_5_campionati_europei_piu_importanti
- https://highpaycentre.org/wp-content/uploads/2020/09/hpc_06_07.pdf
- https://tommasorigon.github.io/datamining/slides/un_A.html
- https://tommasorigon.github.io/datamining/slides/un_B.html
- https://tommasorigon.github.io/datamining/slides/un_C.html
- https://www.footballbenchmark.com/library/broadcasting_revenue_landscape_big_money_in_the_big_five_leagues
- https://www.footballbenchmark.com/library/is_the_english_premier_league_the_european_super_league_already
- https://www.statlect.com/fundamentals-of-statistics/multicollinearity
- https://www.capology.com/