UNIVERSITÀ
DEGLI STUDI
DI BRESCIA

DIPARTIMENTO DI ECONOMIA E MANAGEMENT

Corso di laurea
in Economia e Gestione delle Imprese

RELAZIONE FINALE

Sport Analytics: Modelli di Regressione Lineare Applicati ai Principali Indicatori di Performance nella Pallacanestro

Relatore: Chiar.ma Prof. Zuccolotto Paola

Laureando: Ortensi Stefano

Matricola n. 719233

Indice

Introduzione

Capitolo 1

Big Data e Data Science: lo stato dell'arte

1.1. Megadati: Uno sguardo all'attualità……...7
1.2. Definizione di Big Data…….9
1.3. Caratteristiche: le "5V" ……...…10
1.4. Campi di utilizzo……..12
1.5. Data Science: Definizione e Processo …….16
1.6. Principali tecniche di modellazione ……18
1.7. Presentazione teorica del Morello Regressivo…….20

Capitolo 2

Sport Analytics e Pallacanestro

2.1. Introduzione alla statica nell'attività sportiva……..25
2.2. Gli albori della statistica applicata allo sport: il caso Oakland Athletics…….26
2.3. Il Binomio Statistica e Pallacanestro……27
2.4. Identificazioni delle Fonti e dei Box-Score…….29
2.5. Applicazioni dell'Analisi Statistica alla pallacanestro…….31
2.6. I Pionieri del Basketball Analytics……...33

Capitolo 3

Analisi Regular Season NBA 2021/2022

3.1 Introduzione alla Regular Season NBA 2021/2022…….44
3.2 Preparazione del modello…….46
3.3 Analisi dei Four Factors……...48
3.4 Modello di regressione lineare inerente all'Effective Field Goal Percentage…… 51

3.5 Modello di regressione lineare inerente al Free Throw Rate……...52
3.6 Modello di regressione lineare inerente ai Turnovers per Possession…….....54
3.7 Modello di regressione lineare inerente all'Offensive Rebounding Percentage……..55

Conclusione

Bibliografia

Sitografia

Introduzione

Il seguente elaborato mira ad introdurre, illustrare ed esemplificare le nozioni chiave alla base dell'analisi statistica nel mondo della pallacanestro, con particolare attenzione ai fenomeni legati alla National Basketball Association (NBA).
Il basket rappresenta una delle attività più popolari e praticate all'interno del panorama sportivo globale, ma è stato solamente negli ultimi due decenni che questo sport ha vissuto una vera e propria rivoluzione: introducendo l'uso di metodi matematici e statistici, combinati con l'enorme quantità di dati rilevati per partita, è stato finalmente possibile analizzare le performance di giocatori e franchigie da un punto di vista più oggettivo e razionale, individuare i modelli di successo da perseguire e predire risultati di match e competizioni

Ma com'è possibile misurare l'efficienza o l'efficacia di un giocatore? Perché una squadra è più competitiva rispetto ad un altra? Ed ancora, è meglio rischiare un tiro da tre punti oppure è preferibile penetrare in lunetta?

Queste domande fino a qualche decennio fa difficilmente avrebbero trovato una risposta concreta ed argomentata, ma oggigiorno, grazie allo studio di attenti ricercatori, allo sviluppo di internet ed alla tecnologia, possiamo tentare di fornire spiegazioni a ciò che prima sembrava frutto del caso.
Di seguito, verrà presentata una panoramica dei più comuni metodi di reperimento ed elaborazione dei dati cestistici, al fine di fornire un supporto decisionale razionale ad allenatori, staff-tecnici e sfere manageriali, attraverso l'utilizzo di indicatori di performance e modelli, ripercorrendo anche alcuni argomenti trattati durante il corso di Statistica.

Il primo capitolo introduttivo mira a definire e circoscrivere il fenomeno dei Big Data, descrivendo quali siano le caratteristiche principali ed i possibili campi di utilizzo, per poi approfondire il concetto di "scienza dei dati" e le tecniche più utilizzate per l'estrazione di conoscenza dai dati, concentrandosi in particolare sui modelli regressivi, trattati durante questo percorso di studi.

Il secondo capitolo descriverà come l'impiego dei dati abbia rivoluzionato di fatto la disciplina sportiva: dai primi impieghi nel baseball professionistico fino ad arrivare all'odierna pallacanestro, dove verranno presentati in maniera precisa e sistematica i

possibili utilizzi e scopi dell'analisi statistico-sportiva applicata al basket, gli indicatori di performance più significativi ed i protagonisti che hanno avuto un ruolo cruciale nella rivoluzione di questo sport.

Infine, il terzo capitolo ed ultimo verterà sul calcolo degli indicatori di performance descritti nel capitolo precedente riguardanti la Regular Season 2021/2022 e sull'implementazione di quattro modelli regressivi basati sui Four Factors of Basketball Success teorizzati dal pioniere del Basketball Analytics Dean Oliver.

Capitolo 1
Big Data e Data Science: lo stato dell'arte

1.1 Megadati: Uno sguardo all'attualità

La Harvard Business Review aveva definito già nel 2012 il *data scientist* (o scienziato dei dati) come*, "*il lavoro piu sexy del XXI secolo" [1], ma queste figure professionali chi sono e di cosa si occupano? In che settore operano? E perché sono tra le figure più richieste sul mercato?

Per rispondere a questa domanda è necessario fare un passo indietro e contestualizzare il periodo storico in cui viviamo.
Sempre più smart e sempre più connessi ad internet, quotidianamente utilizziamo dispositivi in grado di produrre ed archiviare dati: dai tradizionali computer, smartphone e tablet, che hanno ormai invaso le nostre abitazioni; ai dispositivi indossabili, come smartwatch o smart-ring, in grado di monitorare l'attività sportiva o il benessere del sonno; sino ad arrivare ai prodotti di domotica o alle automobili con diversi livelli di guida autonoma.
In qualsiasi momento della giornata stiamo continuamente fabbricando dati.
In virtù di ciò, l'industria tecnologica ha sviluppato memorie di massa più capienti e veloci, al fine di ospitare sempre più contenuti.
A conferma di tutto ciò, già nel settembre 2015, l'autore Bernard Marr pubblicò un interessante articolo sulla famosa rivista *Forbes*, sottolineando come, all'interno del biennio 2013/2015, l'umanità avesse prodotto più dati che in tutta la sua storia.[2]
Questo primato fu prevedibilmente sgretolato già negli anni successivi, come riportato dal data journalist, Martin Armstrong, in un suo articolo pubblicato sulla pagina *Statista.com*[3].
Attraverso la figura (1), si può notare immediatamente come la quantità di dati presenti nel 2015 (12 Zettabytes) non sia nemmeno la metà di quelli fatti registrare nel 2018 (33

[1] https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century
[2] Big Data: 20 Mind-Boggling Facts Everyone Must Read- Bernard Marr:"*The data volumes are exploding "more data" has been created in the past two years than in the entire previous history of the human race."*
[3] https://www.statista.com/chart/17727/global-data-creation-forecasts/

Zettabytes); inoltre secondo le previsioni, questo trend è destinato a crescere ad un ritmo esponenziale fino al 2035, con un'ipotetica quantità di dati circolanti di 2142 Zettabytes.[4] Ai meno avvezzi dell'ambiente tecnologico, queste cifre rappresentano solamente congetture numeriche, che non approfondiscono la portata del fenomeno e rischiano di perdere valore se decontestualizzate. Per rendere l'idea degli ordini di grandezza e della mole di dati prodotti, Armstrong fornisce un esempio molto utile per sintetizzare il fenomeno, ovvero per immagazzinare i 33 Zettabytes di dati circolanti nel 2018 sarebbero necessari la bellezza di 660 miliardi di dischetti Blu-ray[5] (con una capacità di archiviazione standard di 50 Gigabytes).

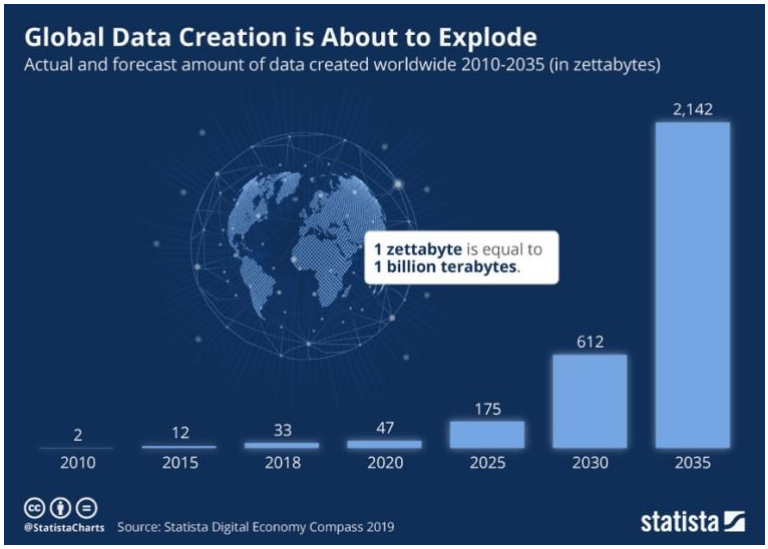


Figura 1: Martin Armstrong, Global Data Creation is About to Explode (fonte:https://www.statista.com/chart/17727/global-data-creation-forecasts/)

Solo nel 2018 Marr tornò sull'argomento pubblicando un nuovo articolo sempre su *Forbes*[6], correggendo il tiro, aggiornando i suoi studi ed analizzando statisticamente in dettaglio un fenomeno senza precedenti.

[4] Con il termine Zettabyte, si indichi uno spazio di archiviazione di un miliardo di Terabytes (un Terabyte indica mille Gigabytes),

[5] Global Data Creation is About to Explode-Martin Armstrong: "*if you were to burn all of the data created last year onto Blu-ray discs, you would need to invest in an astounding 660 billion - each with a standard capacity of 50 gigabytes."*

[6] How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read- Bernard Marr

Alcuni dei suoi rilievi sono particolarmente degni di nota: secondo Bernard infatti più di 3,7 miliardi di persone hanno accesso ad internet e circa metà delle nostre ricerche sono svolte da cellulare: ogni secondo vengono elaborate più di 40.000 ricerche solamente su *Google* (3,5 miliardi al giorno) escludendo dal calcolo gli altri motori di ricerca (poiché se dovessimo sommare le ricerche svolte con tutti i motori di ricerca disponibili sul mercato si supererebbero i 5 miliardi al giorno); ogni minuto inviamo 16 milioni di SMS e 156 milioni di e-mail (escludendo i quasi 104 milioni di spam).

Per giunta, la ricerca sottolinea come l'amore per i social network sia tra le cause scatenanti di questa corsa ai dati. Analizzando solamente il colosso *Facebook* (ad oggi ancora la più grande piattaforma social con 2 miliardi di utenti attivi) e trascurando gli altri competitors come *Instagram*, *Snapchat*, *Twitter* e simili, si è scoperto che ogni minuto vengono pubblicati 510.000 commenti e 293.000 stati, in aggiunta a più di 300 milioni di foto caricate ogni giorno.

1.2 Definizione di Big Data

Una frase molto inflazionata, che ognuno di noi ha sentito almeno una volta nella vita è "*viviamo nell'era dell'informazione*", al fine di sottolineare la facilità di accesso a qualsiasi nozione di carattere generale, tramite l'accesso al web. A portata di un semplice click possiamo ottenere un ventaglio di informazioni quasi sconfinate. È fondamentale però effettuare un chiarimento terminologico, infatti, molto spesso nel linguaggio di tutti i giorni i termini "dato" ed "informazione" assumono la stessa connotazione, quando in realtà rappresentano due concetti completamente diversi: un dato è un elemento grezzo, conosciuto e non interpretato, mentre un'informazione rappresenta l'elaborazione di uno o più dati al fine di ottenere conoscenza.

Alla luce delle considerazioni sopra indicate, possiamo inconfutabilmente affermare di non essere propriamente in un periodo in cui godiamo di informazioni, bensì di dati; quindi, è più corretto ribadire che da alcuni anni a questa parte, siamo entrati nell'epoca dei Big Data, ovvero un'era di vera e propria sovra-esposizione ai dati: sia utili e corretti, ma anche alle volte inutili, ridondanti o addirittura falsi.

Con la locuzione inglese Big Data (in italiano Megadati), *si fa riferimento ad un insieme di dati informatici, così esteso che non può essere archiviato, manipolato e/o analizzato attraverso gli strumenti più tradizionali, ma richiede necessariamente particolari conoscenze specifiche e dispositivi tecnologici avanzati per estrarne informazioni e conoscenze utili.*

Di per sé, questa definizione di Big Data può essere considerata forviante e creare ambiguità, perché può portare a pensare che intenda solamente grandi quantità dati salvati su piattaforme informatiche; in più, tenendo in considerazione il fatto che stiamo assistendo alla progressiva esplosione di questo fenomeno, non vi è una soglia quantitativa, oltre la quale un ammontare di dati prende il nome di Big Data, per cui la discriminante per circoscrivere il fenomeno è che tutti questi dati non rappresentano un cimitero composto da miliardi di Byte, dove, una volta memorizzati, non verranno mai più utilizzati o controllati; al contrario, rappresentano il punto di partenza per infinite possibili analisi, tramite algoritmi specifici, allo scopo di individuare pattern o informazioni, che prima erano celate o trascuravamo a causa della quantità e del rumore statistico dei dati.

1.3 Caratteristiche, le "5V" dei Big Data

Secondo un'analisi svolta nel 2001 dall'esperto Doug Laney, queste enormi masse di dati pur avendo origini diverse tra loro, presentano le medesime caratteristiche e seguono un modello di crescita detto "tridimensionale" (modello delle "3V"): cioè con l'inesorabile passare del tempo assistiamo ad una crescita in termine di volume, velocità e varietà dei dati:

1) *Il Volume:* come ricordato in precedenza, la quantità di dati prodotta da ognuno di noi è in continua crescita, per questo che non è possibile stimare una soglia dimensionale oltre la quale si può iniziare a parlare di megadati, senza poi cadere in errore negli anni a venire. Convenzionalmente si intendono nell'ordine dei Petabyte, ovvero migliaia di Terabyte.
2) *La Velocità:* attualmente i dati vengono generati e memorizzati sempre più rapidamente, la sfida tecnologica che siamo chiamati ad affrontare non è solamente l'acquisizione dei dati in tempo reale, ma anche la possibilità di analizzarli e di

fornire informazioni utili per poter prendere decisioni nel più breve tempo possibile; infatti, tanto più i dati sono "freschi" ed attuali, tanto più rappresentano un valore e sono utili.

3) *La Varietà:* con questa caratteristica si fa riferimento alla presenza di diverse tipologie di dati, provenienti dalle fonti più eterogenee. Ricordando che i dati possono essere classificati rispettivamente in: *strutturati*, quindi organizzabili ed ordinabili; *non strutturati*, ovvero conservati in maniera non schematizzabile; e *semi-strutturati*, cioè che presentano caratteristiche di entrambi i tipi sopracitati

Con il passare del tempo il paradigma è stato ampliato con altre due caratteristiche, così da completare il quadro delle caratteristiche che descrivono i Big Data ("5V"), sono rispettivamente:

4) *La Veridicità*: in riferimento alla qualità, all'affidabilità e alla completezza che i dati raccolti devono avere per rappresentare un buon punto di partenza per l'analisi. Costituisce un pilastro fondamentale e gli esperti de settore sono soliti ricordare che *"Bad data is worse than no data*", sottolineando come dati errati o imprecisi possano portare a conclusioni ambigue o errate.

5) *Il Valore:* il fine ultimo dello studio dei dati è la possibilità di trasformare gli stessi in informazioni, così da poter prendere delle decisioni guidate dalla conoscenza estrapolata e non frutto di un'interpretazione soggettiva di un fenomeno.

1.4 Campi di utilizzo

Già nel 2006, in completa controtendenza con l'epoca, il data scientist, Clive Humbly, coniò lo slogan "Data is the new oil"[7], questo parallelismo che inizialmente può sembrare fuori luogo e provocatorio, secondo il mio modesto parare, calza a pennello.
Infatti, così come il petrolio ha una scarsa utilità nel suo stato grezzo, anche i dati, per essere sfruttati al meglio, necessitano di una prima fase di "raffinazione"; in più, così come il petrolio ha assunto un ruolo determinante nello sviluppo economico e sociale nel XX secolo, anche i dati oramai hanno un impatto quasi determinante in qualsiasi ambito della nostra vita del XXI secolo: dall'attività lavorativa, al tempo libero, dal semplice intrattenimento, alla sanità.
Di seguito verranno elencati alcuni tra i settori maggiormente investiti dall'impiego di dati, così da fornire un quadro generale dei possibili utilizzi dei Big Data.

- ***Big Data e Produzione Industriale***

L'utilizzo di impianti di produzione connessi a sensori ed alla rete, raggiungibili quindi da remoto, ha dato il via alla quarta rivoluzione industriale, sancendo la nascita della cosiddetta industria 4.0. I megadati assumono quindi un ruolo fondamentale, essendo generati quotidianamente dai sistemi produttivi, possono essere sfruttati per anticipare possibili anomalie ed andare a minimizzare i costi riconducibili al processo produttivo, pur garantendo un prodotto dagli standard qualitativi elevati. Ad esempio, all'interno di un'impresa industriale digitalizzata è possibile monitorare lo stato di salute di uno specifico macchinario, attraverso l'utilizzo di sensori, sviluppando poi una strategia di manutenzione predittiva, in grado di anticipare le possibili rotture e risparmiando sensibilmente in termini di costi di sostruzione del macchinario, in costi di fermo della produzione ed in costi di manutenzione periodica, andando ad effettua interventi mirati solamente quando sono realmente necessari.

Inoltre, è possibile ottimizzare la catena dell'approvvigionamento e la gestione del magazzino, analizzando le variabili che influenzano il costo delle materie prime necessarie a svolgere l'attività ed anticipando le variazioni del prezzo.

[7] https://en.wikibooks.org/wiki/Lentis/%22Data_is_the_new_oil%22

- ***Big Data ed Attività di Marketing***

Uno degli aspetti da non trascurare nella vendita di un prodotto o di un servizio è sicuramente la pubblicità. Informando i potenziali clienti delle caratteristiche, delle funzionalità e dei benefici offerti, è possibile invogliarli all'acquisto anche se questo non strettamente necessario.

Il legame presente tra marketing e Big Data è evidente e rappresenta uno dei pochi campi di applicazione dei megadati di cui abbiamo maggior consapevolezza; infatti, questi, sono funzionali all'acquisizione di informazioni utili sulle nostre abitudini di consumo e capacità di spesa

Conoscendo le caratteristiche, i sentimenti e le esigenze degli acquirenti è possibile effettuare una comunicazione più mirata ed efficacie, facendo leva sui tratti psicologici dei compratori.

Il marketing e l'advertising on-line sono tra i settori che hanno tratto più giovamento nell'impiego dei Big Data, consentendo alle aziende di abbandonare le tradizionali tecniche di segmentazione del mercato (ad esempio demografiche) per adottare strategie "su misura" grazie ad una micro-segmentazione del mercato più efficiente.

Tra gli strumenti più efficaci in questo campo vi è il *Retargeting*, cioè il processo mediante il quale, ad ogni ricerca sul web, si è chiamati ad accettare o personalizzare dei cookie, i quali memorizzano i siti visitati e permettono una profilazione di massa, così da sviluppare campagne di marketing mirate nei confronti di clienti potenzialmente interessati.

- ***Big Data e Grande Distribuzione***

L'utilizzo dei big data nella grande distribuzione è dovuto alla volontà di ottimizzare l'efficienza della catena di approvvigionamento di supermercati e negozi, minimizzando le scorte e riducendo gli sprechi. Sfruttando i modelli previsionali è poi possibile stimare i consumi, basandosi sui dati storici e su altre variabili che prima non venivano prese in considerazione

In più, la disponibilità di un'enorme massa di dati sensibili (età, sesso, altezza, peso, interessi, abitudini) riguardanti i consumatori, permette di sviluppare dei modelli specifici sui loro comportamenti di acquisto. Un esempio molto famoso è lo studio sulla correlazione delle vendite tra birra e pannolini[8]: si è scoperto analizzando i dati demografici dei clienti, che gli uomini che acquistano pannolini nei supermercati, spesso comprano anche della birra; una volta scoperto ciò moltissimi supermercati hanno modificando la disposizione sugli scaffali dei due prodotti, molto spesso posizionandoli addirittura nella stessa corsia, al fine di incentivare ed aumentare gli acquisti.
Un supermercato che adotta queste pratiche è Walmart, la catena americana infatti, utilizza un sistema di riordino degli scaffali e di inventario dei prodotti basato sui dati, ad esempio analizzando le vendite medie dei rimedi per il raffreddore e dei prodotti antinfluenzali, è in grado, studiando i picchi di vendita, di capire quando è in corso un'epidemia e vederne la diffusione, così da rifornire i suoi negozi in modo appropriato.

- ***Big Data e Settore Automotive***

Tra le novità introdotte all'interno del settore automotive, oltre all'abbandono dei sistemi di propulsione endotermici e l'adozione di motori elettrici, vi sono tutta una serie di chicche tecnologiche, implementate in brevissimo tempo, derivanti dalla capacità di elaborazioni di enormi masse di dati da parte delle automobili.

Con l'equipaggiamento di internet, le vetture di fascia premium si sono trasformate in una sorta di "computer su ruote", in grado di fornirci informazioni in tempo reale sul traffico, sul percorso migliore da seguire, tracciando ogni tragitto intrapreso, ma non solo, hanno un'intelligenza per regolare autonomamente la climatizzazione in base alle condizioni rilevate dai sensori in dotazione ed ancora possono fornire quotidianamente un report sulle condizioni di usura delle parti meccaniche o dell'efficienza del motore.
In più, vi è l'introduzione di auto con diversi livelli di guida autonoma, la cui sicurezza dipende dalla capacità del computer a bordo di raccogliere ed elaborare in tempo reale tutti i dati acquisiti dai rilevatori interni ed esterni.

[8] Sports Data Mining -Robert P.Schumaker et al.

- ***Big Data e Settore Bancario***

Recentemente stiamo assistendo al progressivo disincentivo all'utilizzo di denaro contante, così facendo è possibile tracciare praticamente qualsiasi transazione di moneta, portando il numero delle operazioni bancarie a numeri senza precedenti. Il fatto di avere sempre più transazioni ha reso necessario lo sviluppo, da parte degli istituti bancari, di sistemi di gestione dei pagamenti in grado di rilevare ed identificare transazioni fraudolente da parte di malintenzionati e quindi garantire una maggiore sicurezza dei clienti.
Un secondo impiego nel settore bancario è l'utilizzo di megadati per sviluppare modelli in grado di prevedere l'andamento del valore di un prodotto finanziario, così da ottimizzare un portafoglio di investimenti.

Infine, le società che emettono mutui o carte di credito sono in grado di definire la capacità di spesa in relazione al reddito di ognuno di noi, stabilendo se un soggetto è finanziariamente affidabile o meno

- ***Big Data e Polizze Assicurative***

Storicamente le assicurazioni prima di fornire una qualsiasi polizza, effettuano una valutazione del profilo di rischio dei possibili nuovi clienti, commisurando il premio da pagare, così da non incappare in pratiche controproducenti.
La disponibilità di grandi quantità di dati permette una migliore la profilazione dei contraenti e quindi di sviluppare modelli previsionali molto più efficaci, da applicare indistintamente a qualsiasi polizza: dalle RCA fino alle polizze sulla vita.

- ***Big Data e Settore Sanitario***

L'utilizzo di dispositivi indossabili o wearable, come orologi od anelli, permettono di mantenere costantemente monitorati i parametri fisiologici di una persona. Banalmente questo continuo controllo biologico può essere utile nei processi di diagnosi delle malattie, ma anche dal punto di vista della prevenzione di alcune di esse; infatti, attraverso l'utilizzo di modelli predittivi, potremmo individuare dei campanelli d'allarme prim'ancora che una malattia si presenti, scongiurando conseguenze peggiori.

- ***Big Data e Settore Energetico***

Un settore poco menzionato in questo frangente, ma pur sempre fondamentale, è quello energetico, in cui l'acquisizione e l'elaborazione di dati sull'andamento della domanda e sulla disponibilità di energia, permette una programmazione più efficiente e mirata, che va a contrastare i costi economici ed ambientali di estrazione di petrolio e gas naturale.

- ***Big Data e Attività Sportiva***

Questo particolare settore sarà argomento di una trattazione più approfondita nei seguenti capitoli, per ora è sufficiente sottolineare come la pratica sportiva rappresenta una vera e propria miniera d'oro in termini di dati e come attualmente qualsiasi atleta, o team sia alla continua ricerca di esperti da ingaggiare in grado di estrarre informazioni utili da sfruttare durante le performance sportive, così da poter godere di anche solo un minimo vantaggio competitivo sui rivali.

1.5 Data Science: Definizione e Processo

Coniato per la prima volta dall'informatico danese Peter Naur (Concise Survey of Computer Methods, 1974), il termine inglese *Data Science* (in italiano Scienza dei Dati) indica l'insieme dei metodi e delle tecniche multidisciplinari tese ad estrarre ed interpretare informazioni utili dai dati.
I principi fondamentali su cui di basa tale scienza derivano da un ampio spettro di discipline: matematica, statistica, informatica, machine learning ed intelligenza artificiale; per questo il Data Scientist, ovvero colui che si occupa di Big Data e Data Science, è una figura altamente specializzata e poliedrica, che, oltre a godere di competenze statistico-matematiche molto ampie, dispone anche di un bagaglio culturale informatico non indifferente, padroneggiando linguaggi di programmazione e software dedicati come SQL,Python, R o Java.

Come riportato dal data scientist Cher Han Lau sulla pagina Medium/Towards Data Science[9], una volta definito il problema da affrontare, per risolverlo qualsiasi esperto di dati segue un processo lineare detto OSEMN, composto dalle seguenti fasi:

1 ***Obtain (ottenimento)***

Nonostante possa sembrare banale l'ottenimento e la raccolta dei dati, è uno degli step più delicati. Infatti, è fondamentale che i dati siano completi e affidabili per non vanificare sul nascere il lavoro, in più, è un passaggio dispendioso sia in termini di tempo che di danaro;

2 ***Scrub (pulizia)***

La pulizia dei dati è il processo di rimozione dei dati errati, duplicati, incompleti o formattati in modo scorretto;

3 ***Explore (esplorazione)***

L'esplorazione o l'analisi dei dati è una fase preliminare dove vi è una comprensione iniziale del contenuto del database, così da poter programmare in modo efficiente le strategie e i modelli;

4 **Model (modellazione)**

Rappresenta la fase cruciale del processo, dove vengono utilizzati algoritmi e tecniche di machine learning al fine di implementare un modello in grado di estrapolare concretamente le informazioni di dati.

5 ***Interpret (interpretazione)***

Una volta estrapolate le informazioni dal modello, è necessaria la presentazione della soluzione del problema iniziale ad un pubblico profano, attraverso schemi grafici o digrammi che siano comprensibili anche ad un non adetto del mestiere.

[9] https://towardsdatascience.com/5-steps-of-a-data-science-project-lifecycle-26c50372b492

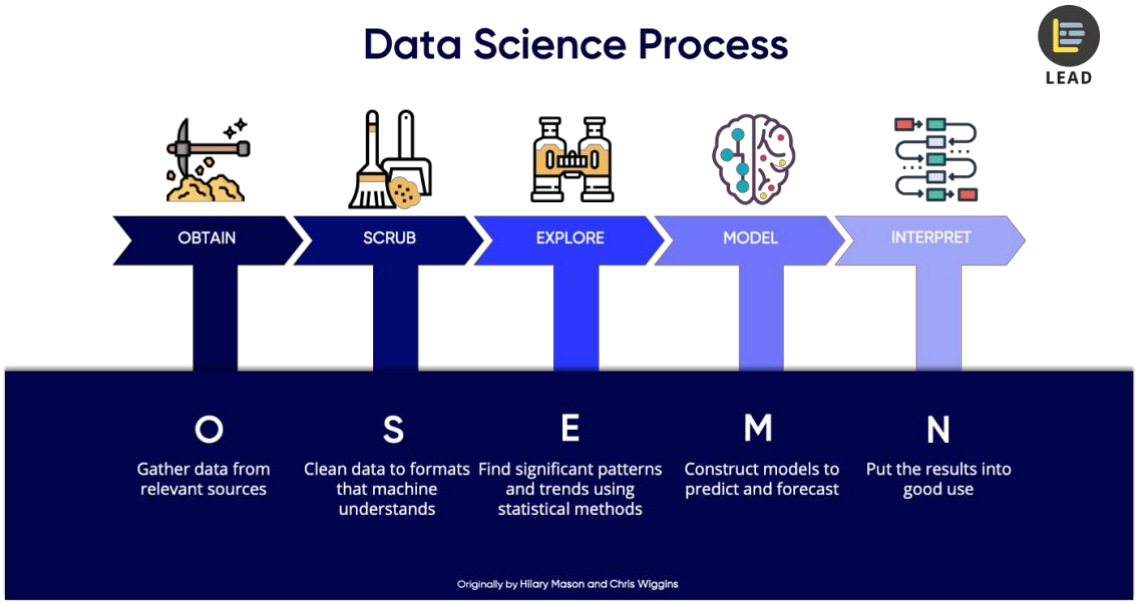


Figura 2: Il processo di Data Science, Cher Han Lau, 5 Steps of a Data Science Project Lifecycle

https://towardsdatascience.com/5-steps-of-a-data-science-project-lifecycle-26c50372b492

1.6 Principali tecniche di modellazione

Come riportato anche nel paragrafo precedente, la fase di modellazione è una delle più importanti nel processo di estrazione di informazioni.

Le scelte riguardanti tecniche di machine learning da applicare e quindi i modelli da implementare, dipendono sostanzialmente dal problema che si vuole risolvere e dalla tipologia di dati di cui è composto il database da interrogare.

⇒ Nel caso in cui si disponga di dati "etichettati", ovvero per i quali si conoscono già le variabili di input(X) e soprattutto di risposta target o output(Y), il ***modello di apprendimento automatico*** preferibile sarà quello ***supervisionato***, dove l'algoritmo del modello sarà addestrato dal supervisore, a sviluppare una mappatura dall'input all'output. Lo scopo è quello di sviluppare una funzione Y=f(X) che spieghi la relazione strutturale esistente tra le variabili di input e il target,, cosicché, nel caso in cui si disponga di nuovi dati di input(X), sia possibile generare delle previsioni sule variabili di output(Y) per quei dati.

A loro volta i modelli ad apprendimento supervisionato possono essere suddivisi, come illustrati nella figura (3), in problemi di Classificazione o Regressione.

- ***Classificazione***: l'algoritmo è istruito allo scopo di organizzare i dati e prevedere, per ogni elemento o individuo di una popolazione, a quale classe appartiene, tra un numero di classi prestabilite.
- ***Regressione***: utilizzata quando sia la variabile di input che di output esprimono valori quantitativi reali e si vuole ricercare una relazione tra essi.

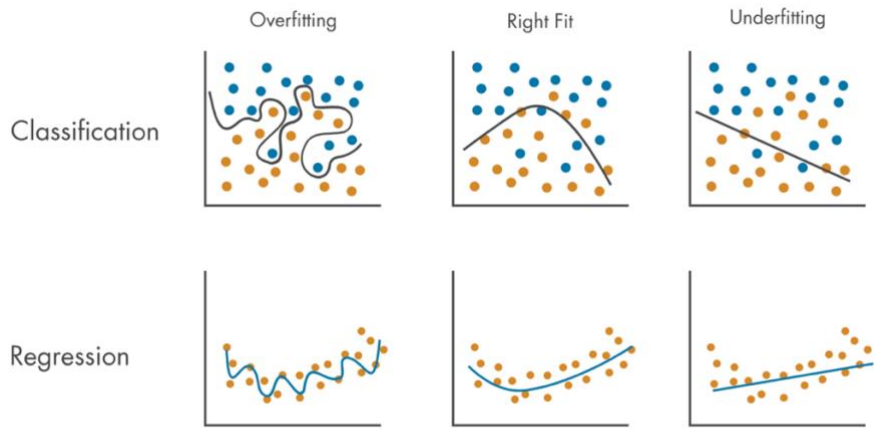


Figura 3 Esempi di Classificazione e Regressione
https://it.mathworks.com/discovery/overfitting.html

⇒ In caso contrario, se non si dispone di output etichettati, il ***modello di apprendimento automatico*** sarà ***non supervisionato***, per cui l'algoritmo lavorerà autonomamente al fine individuare eventuali schemi, tendenze o classificazioni tra i dati.
L'applicazione principale all'interno di questa tipologia di tecniche è il ***Clustering***, dove l'algoritmo viene istruito al fine di organizzare i dati in gruppi "probabili" (detti cluster) come riportato nella figura (4), dei quali non si aveva conoscenza prima dell'implementazione del modello, in quanto i dati in analisi - non essendo etichettati - non potevano essere classificati con precisione in categorie fisse.

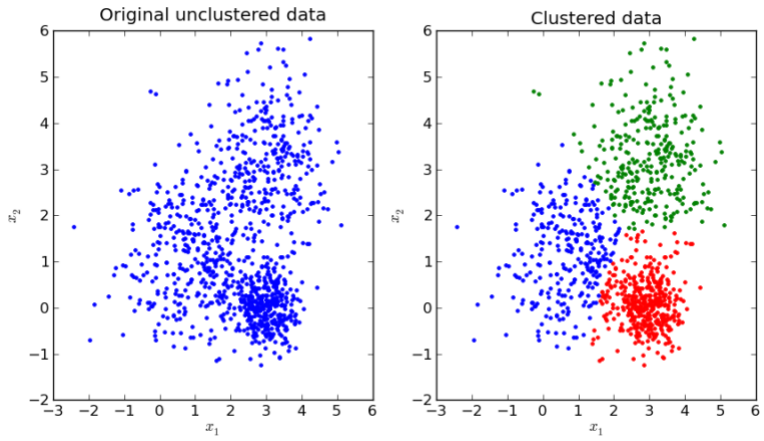


Figura 4: Esempio di Clustering https://i.stack.imgur.com/cIDB3.png

1.7 Presentazione teorica del modello di regressione

Nel seguente paragrafo verrà approfondito a livello teorico il concetto di regressione e lo sviluppo di modelli regressivi, trovando spunto dal materiale fornito dalla Prof. Zuccolotto Paola durante il corso di Statistica, svolto durante questo percorso di studi.

Come specificato in precedenza parlando delle tecniche si apprendimento supervisionato, l'andamento di un fenomeno e la relazione tra le variabili può essere espressa attraverso funzioni matematiche più o meno complesse.

Quando le variabili in esame sono di carattere quantitativo (nel caso in cui i caratteri sono espressi attraverso valori numerici) è possibile adottare un modello di regressione, teso a descrivere la relazione presente tra due variabili. Vi sono modelli di regressione multipla, quando la variabile target è spiegata utilizzando un set di variabili $X_1, X_2, \cdots, \mathrm{X_p}$. Invece, se si usa una sola variabile *X*, si parla di regressione semplice. In questo lavoro ci collochiamo nel cotesto della regressione semplice.

È possibile quindi studiare i valori di una variabile dipendente (o risposta), partendo da una variabile indipendente (o esplicativa); la scelta dell'una o dell'altra variabile come indipendente non è casuale: seguendo un criterio temporale la variabile indipendente deve

essere necessariamente antecedente rispetto a quella dipendente. Se non è possibile individuare una variabile che sia nota prima dell'altra, allora il modello non ha fini predittivi ma solo descrittivi, e viene studiata la correlazione tra le due variabili. In questo caso, la scelta tra variabile dipendente e indipendente è lasciata al ricercatore e dipende dal contesto e dagli obbiettivi dell'analisi.

Prima di effettuare la scelta del modello teorico da utilizzare per studiare la regressione è necessario costruire uno scatter plot (detto anche diagramma a dispersione o nuvola di punti), ovvero un piano cartesiano XY, in cui vengono riportati i valori della variabile indipendente sull'asse delle ascisse ed i valori della variabile dipendente sull'asse delle ordinate, mostrando un punto per ogni coppia di valori.

Una volta rappresentati i dati graficamente, è possibile sintetizzare la relazione tra le variabili attraverso una funzione matematica. Se questa funzione matematica è lineare, si parla di regressione lineare e si interpolano quindi i dati con una retta.

Posta in forma esplicita, la funzione sarà così articolata:

$$Y_i = \alpha + \beta\, X_i + \varepsilon_i$$

dove Y_i rappresenta la variabile dipendente, α l'intercetta, X_i la variabile indipendente, β il coefficiente angolare e ε_i l'errore casuale.

Si può notare come all'interno del modello di regressione ciascun valore assunto dalla variabile di risposta, sia esprimibile come funzione lineare del valore della variabile indipendente, più il termine di errore; la presenza di un termine che prenda in considerazione l'errore, è riconducibile all'incapacità del modello di sintetizzare perfettamente il fenomeno e di quanto il modello teorico si allontana dalla realtà.

La retta più adatta a rappresentare il fenomeno è quella che meglio sintetizza l'insieme dei punti della nuvola.

In merito al procedimento operativo per il calcolo dei parametri α e β, il metodo dei minimi quadrati è solitamente il più utilizzato: viene quindi individuata la retta che rende minima la somma dei quadrati dei residui ε_i; l'idea è di trovare la retta che minimizzi la distanza totale tra i valori osservati dai valori stimati, al quadrato, per eliminare il condizionamento del segno.

L'obbiettivo è calcolare i parametri dell'intercetta e del coefficiente angolare della retta per cui l'errore è minimo, per questo motivo all'interno dello scatter plot, la retta di regressione stimata passerà sempre per il punto avente come coordinate la media di X e la media di Y $(\bar{X}; \bar{Y})$.

Per chiarezza espositiva non verrà riportata l'intera dimostrazione del metodo dei minimi quadrati, svolta durante il corso di Statistica, ma dopo le considerazioni sopra indicate, è possibile affermare che i valori del coefficiente angolare e dell'intercetta della retta sono calcolati come segue:

$$\beta = \frac{\sum_{i=1}^{n}(X_i - \bar{X})\,(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sigma_{xy}}{\sigma_x^2}$$

$$\alpha = \bar{Y} - (\beta\,\bar{X})$$

In aggiunta, va ricordato, come sia sempre possibile individuare una retta con il metodo dei minimi quadrati, anche quando i punti all'interno del diagramma a dispersione non seguono una relazione lineare; in questi casi la retta di regressione non è in grado di sintetizzare l'andamento di un fenomeno e riassumere la relazione tra le variabili, rendendo inutili tutti gli sforzi.

Per verificare la validità del modello di regressione è possibile calcolare l'indice di bontà di adattamento R^2 (o indice di determinazione lineare), il quale è ottenuto rapportando la devianza spiegata alla devianza totale. Più precisamente:

$$R^2 = \frac{\sigma_s^2}{\sigma_t^2} = \frac{VAR(\bar{Y})}{VAR(Y)}$$

Ricordando che, la varianza totale può essere scomposta nella sommatoria tra la varianza spiegata e la varianza dei residui ($\sigma_T^2 = \sigma_s^2 + \sigma_R^2$), la formula può essere manipolata come segue:

$$R^2 = \frac{\sigma_s^2}{\sigma_t^2} = 1 - \frac{\sigma_R^2}{\sigma_T^2}$$

Si noti come l'indice possa assumere valori compresi tra 0 ed 1:

- 1 solamente quando tutti i punti dello scatter plot $(X_i; Y_i)$ giacciono sulla retta interpolante e la retta di regressione sintetizza perfettamente la variabilità di Y (σ_R^2=0 miglior adattamento possibile)
- 0 quando la retta di regressione non spiega variabilità di Y, per cui non vi è un legame lineare tra le variabili (σ_s^2=0 peggior adattamento possibile)

Riassumendo, tanto quanto più l'indice di determinazione si avvicinerà al suo valore massimo, tanto più il modello sarà attendibile nel riassumere la relazione tra le variabili, al contrario tanto quanto più si allontanerà dal suo massimo e si avvicinerà allo zero, tanto più il modello sarà inattendibile.

Inoltre si ricordi come, nei modelli di regressione lineare semplice, l'indice di determinazione R^2 sia uguale al quadrato del coefficiente di correlazione lineare di Pearson

r(X,Y), il quale misura la “direzione” e la “forza” di una relazione lineare tra due variabili quantitative.

Il coefficiente di correlazione lineare r può assumere valori compresi tra 1 e -1: nel caso in cui assuma un valore prossimo allo zero significa che non vi è alcuna relazione tra le due variabili, valori positivi vicini a 1 rappresentano una correlazione lineare positiva (se è esattamente uguale a 1 i punti giacciono su una retta ascendente), mentre valori negativi immediatamente prossimi a -1 indicano una correlazione lineare negativa (se è esattamente uguale a -1 i punti giacciono su una retta discendente).

Capitolo 2
Sport Analytics e Pallacanestro

2.1 Introduzione alla statica applicata all'attività sportiva

In un arco temporale che si snoda dalle prime Olimpiadi nell'antica Grecia sino ad arrivare ai giorni nostri, l'attività sportiva ha sempre avuto un ruolo fondamentale nella vita dell'uomo.

Il Consiglio d'Europa all'interno della Carta Europea dello Sport [10]ha speso una definizione alquanto accurata e "S*'intende per sport qualsiasi forma di attività fisica che, attraverso una partecipazione organizzata o non, abbia per obiettivo l'espressione o il miglioramento della condizione fisica e psichica, lo sviluppo delle relazioni sociali o l'ottenimento di risultati in competizioni di tutti i livelli."*

Nel corso del tempo però, soprattutto in epoca moderna, una pratica nata con finalità ludiche e salutari, si è trasformata in un macro-business, in grado di generare solamente in Italia, un volume d'affari di circa 78.8 miliardi di euro secondo un articolo pubblicato dal IlSole 24 Ore, con addirittura un impatto del 3% sul PIL.[11]

Per sintetizzare come la pratica sportiva si sia evoluta sino ad arrivare ad un vero e proprio settore trainante, basti pensare agli attori che popolano questo mondo: federazioni e squadre con indotti milionari grazie ai diritti televisivi e alla vendita di biglietti ai fan per assistere agli eventi, colossi economici che producono attrezzature sportive (vendesi Nike, Adidas, Puma, Under Armor e via discorrendo), aziende produttrici di integratori e bevande energetiche tese al miglioramento della prestazione, sino ad arrivare a cascata ai vari corsi offerti da professionisti ed istruttori all'interno di palestre e strutture dedicate.

In completa controtendenza con altre forme di intrattenimento, sempre più desuete ed in crisi di utenza, come la lettura, il cinema, il teatro; lo sport sembra non conoscere battute d'arresto, nemmeno durante la recente pandemia di COVID-19, dove una gran moltitudine di persone vedevano nella pratica sportiva una forma di distrazione.

[10] https://www.coni.it/images/documenti/Carta_europea_dello_Sport.pdf
[11] https://www.ilsole24ore.com/art/lo-sport-italia-vale-788-miliardi-pari-3percento-pil-AEG3V5NB

In questo contesto frizzante ed in continua espansione, diventa interessante analizzare come i principali attori di questo circus, ovvero gli atleti e le squadre, si stiano attrezzando al fine di rimanere competitivi e non a margine in un settore tra i più redditizi ed in crescita al momento.
In tempi recenti, tra le novità che stanno rivoluzionando questo settore, possiamo osservare l'introduzione su larga scala di metodi analitici e statistici, atti a valutare qualsiasi aspetto dell'universo sportivo: dalle performance fatte registrare in campo, alla tendenza agli infortuni di un atleta, fino allo studio economico del business e gli indotti derivati.
utilizzando l'enorme mole di dati questo sicuramente grazie anche all'avvento di internet ed alla disponibilità e facilità di reperimento di una smodata quantità di dati relativi ad ogni aspetto di quasi tutti gli sport.

2.2 Gli albori della statistica applicata allo sport: il caso Oakland Athletics

"Il primo che attraversa il muro è sempre insanguinato."

Disse John William Henry, proprietario dei Boston Red Sox (squadra di baseball professionistica americana) e del Liverpool Football Club (formazione di calcio militante in Premier League, massima competizione inglese), per sottolineare le pesanti critiche subite dai soggetti che per primi hanno introdotto l'analisi quantitativa nel mondo dello sport.
In merito alla possibilità d'impiego di analisi statistiche nella pratica sportiva, uno dei primi esempi fu senza dubbio proprio il baseball e tra i pionieri in questo campo è necessario citare lo scrittore e statista americano Bill James, il quale già a partire dal 1977, pubblicò le sue tesi inerenti al baseball, sugli annuali *Bill James Baseball Abstract*.
Queste teorie fondate sull'analisi dei dati ed applicate al baseball, passeranno poi alla storia come il termine *Sabermetrica*.
Completamente in controtendenza con l'epoca, le idee di James non furono immediatamente accettate dall'ambiente e dalla comunità sportiva in generale; dovettero attendere i primi anni del duemila e l'insediamento di Billy Beane come General Menager degli Oakland Athletics, per vedersi avvalorate.

Beane, applicando alla lettera i concetti di sabermetrica, riuscì a ri-organizzare una squadra mediocre con un budget ridotto, in una corazzata in grado di aggiudicarsi ben 20 partite di fila nell'anno 2002, fino ad allora un record per la Major League Baseball.
Il principio cardine su cui si fondava l'intero progetto era molto semplice: riuscire ad arrivare in base più volte degli avversari (vincendo la partita), spendendo però il meno possibile sull'acquisto del cartellino dei giocatori e sulla loro retribuzione.
Secondo il sistema messo in atto dal GM, la scelta di quali giocatori acquistare e quali mandare in campo fu guidata esclusivamente da indicatori statistici, come la *On-Base Percentage o la Slugging Percentage (Handbooks of Modern Statistical Methods and Analyses in Sports, Jim Albert et al. 2017),* per cui anche un giocatore considerato "scarso", "anziano", o semplicemente "acciaccato", ma dotato di alcune specifiche caratteristiche sottolineate dagli indicatori, poteva risultare nettamente più utile; di un giocatore più "bravo" (quindi più costoso e maggiormente retribuito), ma che non apportava il contributo richiesto alla squadra.
Nonostante la strabiliante stagione, gli Oakland Athletics non riuscirono ad aggiudicarsi il titolo, ma il loro esempio rappresentò un punto di svolta per questo sport, tant'è che negli anni successivi tutte le principali squadre adottarono metodi e modelli simili a quelli degli Oakland Athletics.
La vicenda divenne così popolare oltreoceano da essere narrata prima in un libro, "Moneyball: The Art of Winning an Unfair Game" di Michael Lewis (2003) e successivamente trasposta in un film con l'omonimo titolo, in cui la figura di Billy beane era interpretata di Brad Pitt (2011).

2.3 Il Binomio Statistica e Pallacanestro

La pallacanestro nasce nel dicembre 1891, grazie all'immaginazione e all'ingegno del canadese James Naismith.[12]
Docente presso la YMCA International Training School a Springfield, nel Massachusetts, James fu incaricato dal sovrintendente per l'educazione fisica della scuola, Dr.Luther

[12] https://it.wikipedia.org/wiki/Storia_della_pallacanestro

H.Gulick, di sviluppare un nuovo gioco indoor allo scopo di mantenere attivi i giovani studenti nei lunghi e freddi mesi invernali.
In particolare, Gulick lo delegò di sviluppare un gioco dal rapido apprendimento, da svolgere in un ambiente chiuso e che prediligesse l'abilità alla forza, ovvero con poche occasioni di contatto.

A questo punto sorge spontaneo chiedersi in che modo uno gioco per bambini si è trasformato in uno degli sport più praticati e studiati nel mondo? Cosa è cambiato? Perché ora è argomento di approfondite analisi statistiche?

Nel corso di un secolo le regole fondamentali di Naismith vennero rielaborate ed ampliate, andando a trasformare radicalmente il gioco da come era stato pensato in origine, soprattutto tenendo in considerazione l'evoluzione del ritmo di gioco, basti pensare che la prima partita ufficiale di pallacanestro della storia si concluse con un misero risultato di 1 a 0, ben lontano dai punteggi fatti registrare in epoca moderna.
In particolare, in origine non fu regolamentato un tempo limite per concludere un'azione, questo portò a delle dinamiche di gioco rallentate e viziate, in cui la squadra che era in vantaggio custodiva il pallone più a lungo possibile, senza concludere l'azione, così che la squadra avversaria non potesse entrare in possesso del pallone e quindi recuperare lo svantaggio.
Queste pratiche di gioco sfociarono nel paradosso ai campionati europei di Mosca del 1953, quando, durante la partita Jugoslavia-Israele, la nazionale slava fece "melina" per tutti i tempi supplementari rifiutandosi di attaccare o tirare, attendendo gli ultimi secondi disponibili per affondare a canestro. Durante la competizione andò in scena il medesimo copione nella partita URSS-Ungheria, offrendo uno spettacolo non proprio entusiasmante ai tifosi, che dovettero assistere ad uno sterile giropalla per tutta la partita, poiché la nazionale sovietica godeva di un piccolo margine di punteggio.
Questo portò la FIBA (Fédération Internationale de Basketball) nel 1956 ad imporre un tempo limite di 30 secondi per ogni possesso, poi ridotto a 24 secondi, così da rendere il basket uno sport più rapido, incalzante e frenetico, ma soprattutto aumentando lo spettacolo offerto.[13]

Circoscrivendo la pallacanestro ad uno sport di efficienza, dove le squadre tentano di concludere a canestro nel più breve tempo possibile, così da non dover cedere il possesso

[13] https://www.museodelbasket-milano.it/leggi.php?s=&idcontenuti=266

della palla prima ancora di aver tentato di tirare, è logico pensare che ogni possesso diventi rilevante ed ogni match oggetto di studio, in quanto il numero di eventi misurabili durante una partita (tiri, passaggi, dribbling, rimbalzi) sarà molto alto. In più, andando a considerare non solo la singola partita, ma più in generale l'intero campionato, l'insieme dei dati rilevabili sarà elevatissimo.

Alla luce delle considerazioni sopra indicate, appare scontato il binomio tra pallacanestro e statistica, in quanto quest'ultima rappresenta la disciplina che si occupa di predisporre metodi quantitativi utili ad esaminare fenomeni reali riproducibili, purché questi producano dati.

2.4 Applicazioni dell'Analisi Statistica alla pallacanestro

L'adozione di modelli e tecniche statiche all'interno dell'universo della pallacanestro è ormai diventata una pratica comune, tant'è che la quantità di libri, articoli, e siti web che trattano l'argomento, in maniera più o meno scientifica, è pressoché incalcolabile.

Durante la preparazione del seguente elaborato, sottoponendo la letteratura disponibile ad un'attenta valutazione e grazie al manuale Basketball Data Science (Zuccolotto e Manisera,2020) è stato possibile individuare una serie macro-obbiettivi dell'analisi statistica applicata alla pallacanestro:

- *La previsione dei risultati di un match o di una competizione*.[14]
- *L'individuazione di indicatori in grado di discriminare quali sono le franchigie di successo.*[15]
- *L'analisi di schemi ricorrenti o patterns nei punteggi dei match.*[16]

[14] Zuccolotto et al.,2020: *Predicting the outcomes of a game or a tournament (West, 2008; Loeffelholz et al., 2009; Brown and Sokol, 2010; Gupta, 2015; Lopez and Matthews, 2015; Ruiz and Perez-Cruz, 2015; Yuan et al., 2015; Manner, 2016; Vracar et al., 2016).*

[15] Zuccolotto et al.,2020: *Determining discriminating factors between successful and unsuccessful teams (Trninié et al., 2002; Sampaio and Janeira, 2003; Ibafiez et al., 2003; De Rose, 2004; Csataljay et al., 2009; Ibafiez et al., 2009; Koh et al., 2011, 2012; Garcia et al., 2013).*

[16] Zuccolotto et al.,2020: *Examining the statistical properties and patterns of scoring during the games (Gabel and Redner, 2012; Schwarz, 2012; Cervone et al., 2016).*

- *L'analisi delle performance di un giocatore e l'impatto di esse sulle possibilità di vittoria della squadra*.[17]
- *L'analisi dei patterns di gioco legata ai ruoli e alle posizioni in campo degli atleti*.[18]
- *L'analisi dei movimenti del corpo dei giocatori in relazione alle prestazioni offerte*.[19]
- *L'analisi dei movimenti dei singoli giocatori e delle linee di passaggio durante le azioni.*[20]
- *L'analisi delle tattiche adottate dalle franchigie e l'individuazione delle strategie di gioco ottimali.*[21]
- *L'analisi sulle prestazioni arbitrali alla ricerca di possibili errori sistematici o pregiudizi.*[22]
- *L'analisi di variabili psicologiche e l'impatto di esse sulle performance in gioco.*[23]

---

[17] Zuccolotto et al.,2020: *Analysing a player's performance and the impact on his team's chances of winning (Page et al., 2007; Cooper et al., 2009; Sampaio et al., 2010; Piette et al., 2010; Fearnhead and Taylor, 2011; Òzmen, 2012; Page et al., 2013; Erculj and Strumbelj, 2015; Deshpande and Jensen, 2016; Passos et al., 2016; Franks et al., 2016; Engelmann, 2017), also with reference to the so-called "hot hand" effect (Gilovich et al., 1985; Vergin, 2000; Koehler and Conley, 2003; Tversky and Gilovich, 2005; Arkes, 2010; Avugos et al., 2013; Bar-Eli et al., 2006) and with special focuses on the impact of high- pressure game situations (Madden et al., 1990, 1995; Goldman and Rao, 2012; Zuccolotto et al., 2018).*

[18] Zuccolotto et al.,2020: *Monitoring playing patterns with reference to roles (Sampaio et al., 2006), also with the aim of defining new playing positions (Alagappan, 2012; Bianchi et al., 2017).*

[19] Zuccolotto et al.,2020: *Designing the kinetics of players' body movements with respect to shooting efficiency, timing and visual control on the field (Miller and Bartlett, 1996; Okubo and Hubbard, 2006; de Oliveira et al., 2006; Aglioti et al., 2008).*

[20] Zuccolotto et al.,2020: *Depicting the players' movements, pathways, trajectories from the in-bounds pass to the basket and the network of passing actions (Fujimura and Sugihara, 2005; Perse et al., 2009; Skinner, 2010; Ther6n and Casares, 2010; Bourbousson et al., 2010b,c; Passos et al., 2011; Lamas et al., 2011; Piette et al., 2011; Fewell et al., 2012; Travassos et al., 2012; Shortridge et al., 2014; Ante et al., 2014; Clemente et al., 2015; Gudmundsson and Horton, 2016; Metulini et al., 2017a,b; Wu and Bornn, 2018; Bornn et al., 2017; Miller and Bornn, 2017; Metulini et al., 2018), the flow of events and the connected functional decisions which have to be taken by players during the game (Araujo et al., 2006, 2009; Araujo and Esteves, 2010; Perica et al., 2011; Araujo and Davids, 2016).*

[21] Zuccolotto et al.,2020: *Studying teams' tactics and identifying optimal game strategies (Annis, 2006; Zhang et al., 2013; Skinner and Goldman, 2017).*

[22] Zuccolotto et al.,2020: *Investigating the existence of possible referee biases (Noecker and Roback, 2012).*

[23] Zuccolotto et al.,2020: *Measuring psychological latent variables and their association to performance (Meyers and Schleser, 1980; Weiss and Friedrichs, 1986; Seifriz et al., 1992; Taylor, 1987; Maddi and Hess, 1992; Dirks, 2000; Heuzé et al., 2006; Bourbousson et al., 2010a).*

2.4 Identificazioni delle Fonti e dei Box-Score

Oggigiorno navigando in rete è possibile imbattersi in migliaia di pagine o forum che trattano esclusivamente argomenti sportivi dal punto di vista statistico: dai siti web generalisti come ESPN, SBNATION o Yahoo!Sports, a quelli più verticali su una singola disciplina (in particolare nel caso della pallacanestro sono degni di nota 82Games, Basketball Reference o DataBasketball).

Tutti i dati registrati per ogni partita di basket vengono poi organizzati in databases chiamati box scores, sia per quanto concerne le leghe professionistiche (come nel caso dell'NBA), sia per quelle universitarie o dilettantistiche.

Sono molto utili in quanto permettono di limitare il dispendio di tempo del reperimento dei dati grezzi inerenti al gioco, evitando di dover visionare sistematicamente tutte le partite. Sicuramente per la comprensione ottimale del fenomeno non si sostituiscono alla visione vera e propria dei match, ma permettono con un primo colpo d'occhio di farsi un'idea di ciò che è successo durante una partita o in un intero campionato.

All'interno di questi database è possibile trovare informazioni inerenti a qualsiasi aspetto del gioco, sia per quanto riguarda le performance di un'intera franchigia o per un singolo giocatore.

Nei box scores dei siti più sviluppati, come nel caso di quello offerto da NBA.com, è possibile interrogare direttamente il database attraverso delle maschere dedicate, in maniera tale da filtrare i risultati ed ottenere solamente dei dati utili all'analisi, senza perdere del tempo a manipolare e ripulire i risultati tramite programmi di terze parti.

Di seguito è fornito un glossario relativo alle variabili più utili e popolari, così da permettere la comprensione dell'elaborato anche ad un pubblico di neofiti:

- ***Minutes Played (MIN)***
  Sommatoria dei minuti giocati;
- ***Points (PTS)***
  Sommatoria di tutti i punti realizzati, indipendentemente dal fatto che il canestro sia stato prodotto da un tiro libero, un tiro da due punti oppure da tre;
- ***Field Goals Attempted (FGA)***

Sommatoria dei tiri dal campo tentati, indipendentemente dal fatto che questi siano tiri da due o da tre punti, non vengono considerati però i tiri liberi;

- ***Field Goals Made (FGM)***
  Sommatoria dei canestri realizzati dal campo;
- ***3 Point Field Goals Attempted (3PA)***
  Sommatoria dei tiri da tre punti tentati;
- ***3 Point Field Goals Made (3PM)***
  Sommatoria dei tiri da tre punti realizzati;
- ***Free Throws Attempted (FTA)***
  Sommatoria dei tiri liberi tentati;
- ***Free Throws Made (FTM)***
  Sommatoria dei tiri liberi realizzati;
- ***Offensive Rebounds (OREB)***
  Sommatoria dei rimbalzi offensivi, cioè i rimbalzi che sono stati recuperati nella lunetta avversaria in una fase offensiva;
- ***Defensive Rebounds (DREB)***
  Sommatoria dei rimbalzi difensivi, cioè i rimbalzi che sono stati recuperati nella propria lunetta durante un'azione offensiva avversaria;
- ***Total Rebounds (REB)***
  Sommatoria di tutti i rimbalzi conquistati, offensivi e difensivi;
- ***Assists (AST)***
  Sommatoria dei passaggi che hanno portato alla realizzazione di un canestro come conseguenza diretta di quel passaggio;
- ***Turnovers (TOV)***
  Sommatoria di tutte le palle perse concesse poi agli avversari;
- ***Steals (STL)***
  Sommatoria di tutte le palle rubate agli avversari senza che questi abbiano tentato un tiro;
- ***Blocks (BLK)***
  Sommatoria dei tiri avversari stoppati o bloccati;
- ***Blocks Against (BLKA)***
  Sommatoria di tutte le stoppate subite;
- ***Personal Fouls (PF)***
  Numero di falli fatti;

- ***Personal Fouls Drawn (PFD)***
  Numero di falli subiti;

Spesso, il semplice valore numerico riportato all'interno dei box scores può non fornire sufficienti informazioni per valutare le prestazioni di una franchigia o i un singolo giocatore, anzi, alle volte può risultare forviante ed allontanarsi dalla realtà se decontestualizzato.
Basti pensare ai box scores inerenti alle performance individuali di Lebron James, stella dei Los Angeles Lakers, sopra un parziale di 30 punti a partita durante la Regular Season 2021-2022. [24]
Nel caso di James, andando ad analizzare solamente le sue performance individuali si potrebbe essere portati a pensare che sia l'ennesima stagione da dominatore della lega, questo però - secondo alcuni - non corrisponderebbe esattamente alla realtà. Infatti, James è stato pesantemente criticato per il fatto di non giocare per e con la squadra, ma solamente allo scopo di migliorare le sue statistiche personali nel tentativo di superare più record possibili, e, secondo i sostenitori di tale tesi, la dimostrazione di questo sarebbe il fatto che la sua squadra è risultata persino esclusa dai playoff.
Per questo sono stati introdotti nel tempo degli indicatori più sofisticati e complessi per analizzare le prestazioni di atleti e franchigie, così da poter limitare il più possibile errori di valutazioni dovuti all'analisi di dati non completamente veritieri.

2.5 I Pionieri del Basketball Analytics

Tra i soggetti che hanno dato slancio all'applicazione dell'analisi statistica alla pallacanestro, dobbiamo sicuramente annoverare *Daryl Morey* e *Dean Oliver*, poiché grazie alle loro idee ed al loro lavoro pioneristico hanno di fatto rivoluzionato in meno di due decenni un intero sport.

Daryl Morey, dopo aver terminato gli studi in informatica alla Northwestern University ed aver conseguito un Master in Business Administration al MIT, approdò in NBA ai Boston

[24] https://www.nba.com/stats/players/traditional?PerMode=PerGame&sort=PTS&dir=-1&Season=2021-22&SeasonType=Regular+Season&TeamID=0

Celtics e successivamente diventò il General Manager degli Houston Rockets dal 2007 al 2020.

Fu proprio al timone della squadra texana che decise di applicare la sua filosofia di basket per la gestione di una franchigia NBA, dopo essere stato folgorato dal libro precedentemente citato "Moneyball: The Art of Winning an Unfair Game".

In un orizzonte temporale di tredici anni, Morey, plasmò gli Houston Rockets, ad immagine e somiglianza del proprio credo cestistico, che può essere riassunto in quattro concetti principali:[25]

- ***Alzare il ritmo del gioco***. *Alzando il numero di possessi per partita è alquanto probabile un incremento dei punti realizzati.*

- ***Abolizione del tiro dal midrange***. *Il tiro da media distanza è il tiro più contestato nella NBA, vale "solamente" due punti, ed ha la più bassa percentuale di realizzazione della lega. Molto meglio fare 1 o 2 passi indietro e prendersi un tiro da tre smarcato.*

- ***Il miglior tiro da tre è quello dagli angoli***. *Il tiro da tre meno contestato nella NBA è la cosiddetta "Corner Three" ed è il tiro da tre punti più ravvicinato dato che viene scagliato da 6.72 metri dal canestro anziché i canonici 7.25 metri.*

- ***Il tiro migliore è il tiro libero***. *Attaccare il canestro è cosa buona e giusta. Nella restricted area la percentuale di realizzazione è la più alta della lega e con maggiore frequenza si può guadagnare un fallo che porta a guadagnare un tiro libero. Ogni viaggio in lunetta permette di mettere punti a referto a gioco fermo.*

Queste teorie non portarono i Rockets alla vittoria dell'anello (riconoscimento per il titolo NBA), ma trasformarono la realtà di Houston, da città perdente per definizione ad una franchigia considerata tosta ed ostica da affrontare.

[25] https://www.dailybasket.it/nba/nba-stats-welcome-to-moreyball/

In più, le idee di Morey, furono riconosciute corrette da tutto l'ambiente NBA ed a prova di ciò, si può menzionare l'exploit dei Golden State Warriors: vincenti negli ultimi 8 anni di ben 4 titoli, anche attraverso la visione di gioco rinominata "Moreyball" per cui "Three is better than two" e che vede Stephen Curry il miglior interprete di questo credo, con all'attivo più di 3300 triple realizzate.

Non solo le squadre di vertice hanno modificato il loro modo di condurre e concludere il gioco, ma in generale tutta la lega. A dimostrazione di quanto riportato è interessante notare come sia cambiata la media dei tiri da 3 punti tentati per partita in NBA in poco più di 30 anni passando dai 5 tentativi per partita del 1980, ai 55 del 2017[26].

In più è impossibile non menzionare le tesi di Dean Oliver, statistico americano e pioniere dell'analisi sportiva, attuale assistente allenatore dei Washington Wizards.
Nel 2004, Oliver, scrisse "Basketball on Paper: Rules and Tools for Performance Analysis", ancora oggi ritenuto una vera e propria bibbia negli ambienti USA dell'analisi statistica legata alla pallacanestro.
Le idee e le analisi di Oliver si fondano sul concetto di "possesso", per poi concentrarsi a cascata su valutazioni offensive e difensive, plays, statistiche al minuto, percentuali di tiro effettive, tassi di rimbalzo, statistiche di plus/minus e percentuali di vittoria pitagorica.
In più, è l'ideatore dei cosiddetti *four factors*, i quali verranno utilizzati nel prossimo capitolo per implementare dei modelli regressivi atti a valutare la stagione regolare 2020-2021.

Il concetto di Possesso

La definizione di possesso fornita da Oliver apparentemente può sembrare banale, infatti, *inizia quando una squadra ottiene il controllo della palla e termina quando la stessa ne cede il controllo*, per sintetizzare un possesso termina quando:

- Una squadra realizza un canestro (attraverso un tiro libero o un tiro dal campo)
- Il team in attacco perde il pallone (turnover)
- La squadra in attacco fallisce un canestro e la squadra in difesa ne raccoglie il rimbalzo (difensivo)

[26] https://masterx.iulm.it/uncategorized/three-is-the-magic-number-levoluzione-recente-del-tiro-da-tre/

In più si noti come secondo la definizione di possesso fornita da Oliver, un rimbalzo offensivo non costituisce un nuovo possesso, in quanto il pallone è ancora nelle mani della squadra che ha raccolto il rimbalzo e che quindi può costruire une "giocate" (*Plays*).
I possessi rappresentano una buona base di partenza per valutare l'efficienza di una squadra o di un giocatore, essendo più o meno uguali per ogni squadra in un match. È superfluo sottolineare il fatto che la squadra vincente sarà quella che ha più punti rispetto alla squadra rivale, ma non è così immediato il fatto che ogni squadra tenterà di ottimizzare i propri possessi, cercando di segnare più punti per possesso rispetto agli avversari.

I possessi non vengono conteggiati all'interno dei box score comunemente disponibili, ma possono essere calcolati in modo approssimato tramite la formula che segue:

$$\boldsymbol{POSS_t} = (\boldsymbol{FGM_t} + \boldsymbol{\lambda FTM_t}) + \boldsymbol{\alpha}[(\boldsymbol{FGA_t} - \boldsymbol{FGM_t}) + \boldsymbol{\lambda}(\boldsymbol{FTA_t} - \boldsymbol{FTM_t}) - \boldsymbol{OREB_t}] + (\mathbf{1} - \boldsymbol{\alpha})\,\boldsymbol{DREB_o} + \boldsymbol{TO_t}$$

Dove :

- *FGA(t) sono i canestri dal campo tentati dal team(t)*
- *FTM(t) sono i tiri liberi realizzati dal team(t)*
- *FGM(t) sono i canestri dal campo realizzati dal team(t)*
- *FTA(t) sono i tiri liberi tentati dal team(t)*
- *OREB (t) sono i rimbalzi offensivi del team(t)*
- *DREB (t) sono i rimbalzi difensivi del team rivale(o)*
- *TO (t) sono le palle perse del team(t)*
- *λ è la frazione di tiri liberi che concludono il possesso*
- *α è un parametro tra 0 e 1..*

Si noti come i canestri dal campo realizzati (FGM), i tiri liberi segnati (FTM) e le palle perse (TO) costituiscono un possesso e quindi hanno un valore di possesso pari a 1. I tiri dal campo sbagliati e i tiri liberi sbagliati forniscono una quota di possesso pari ad *α,* mentre i rimbalzi difensivi (DREB) una quota (1 - *α*). I rimbalzi offensivi (OREB), come ricordato precedentemente, non costituiscono un nuovo possesso e quindi annullano i tiri dal campo e tiri liberi sbagliati, rappresentano un valore – *α.*
Attraverso studi empirici su una moltitudine di dati è stato possibile individuare dei valori plausibili per parametri α e λ: si è infatto riscontrato che se essi assumono i valori α =1 e

λ=0.44, il numero di possessi approssimato tramite la formula è molto vicino a quello effettivo.

$$\boldsymbol{POSS_t = FGA_t + 0.44\,x\,FTA_t - OREB_t + TO_t}$$

L'Offensive Rating ed il Difensive Rating

Oltre ad aver definito il concetto di possesso, Oliver introduce una serie di indicatori utili per sviscerare le performance in termini di efficienza.
I punti fatti per 100 possessi prendono il nome di Offensive Rating (ORtg), mentre con il termine Difensive Rating (DRtg) vengono indicati i punti concessi per 100 possessi.

$$\boldsymbol{Offensive\ Rating\ (ORtgt) = \frac{PTS_t}{POSS_t} x100}$$

$$\boldsymbol{Difensive\ Rating\ (DRtgt) = \frac{PTS_o}{POSS_o} x100}$$

Queste misure descrivono al meglio le qualità di attacco e difesa di una franchigia, infatti, si ricordi come vi siano squadre che impongono agli avversari ritmi elevati e squadre che, se non trovano un avversario "veloce" tengono di norma ritmi più lenti: alla fine del campionato il numero di possessi totale di due squadre può essere differente. Ecco che allora ha senso depurare i punti fatti e concessi per questa grandezza.
Si noti come la Figura (5) sintetizzi i rating offensivi e difensivi delle franchigie NBA nell'anno 2021-2022 Pre-All Star Game. Nel quadrante in alto a destra vi sono le squadre con un buon attacco, ma con una difesa da rivedere; mentre nel quadrante in basso a sinistra si trovano le squadre con una buona difesa ma con un attacco piuttosto innocuo. È possibile notare con una rapida occhiata, come le squadre che hanno raggiunto un buon piazzamento durante la Regular Season, come ad esempio i Boston Celtics o i Golden State Warriors, siano quelle che hanno saputo combinare un buon attacco, senza però concedere troppi punti alle rivali.

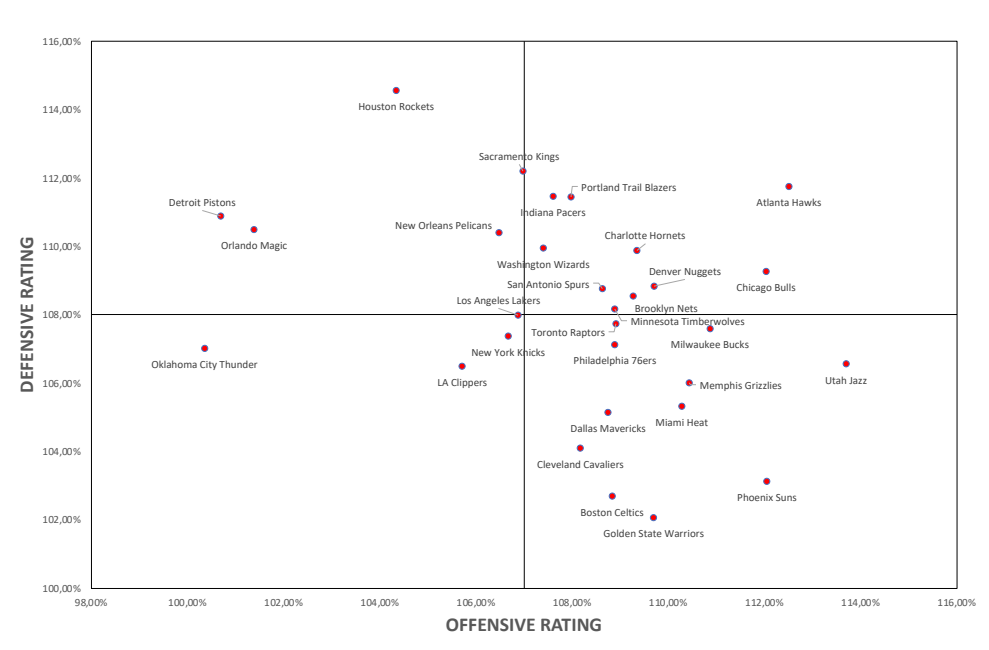


Figura 5: Offensive/Defensive Rating (Fonte: nostre elaborazioni)

Si può notare dalla figura (6) la costante crescita di efficienza nel campionato NBA; infatti, analizzando i rating offensivi e difensivi medi del campionato si può notare come negli anni ’70 erano tutto sommato bassi, per poi esplodere tra gli anni ’80 e ’90, raggiungendo fino ad allora i punti più alti mai toccati.

Tra il ’94 ed il ‘98 abbiamo assistito ad un tracollo delle valutazioni, queste, dopo aver perso circa 6 punti percentuali, hanno ricominciato a crescere in modo quasi continuativo fino alla stagione 2005-2006.

Negli ultimi anni il trend non si è mai fermato, facendo registrare valori che per le epoche passate sarebbero perfino fuori scala, soprattutto a livello di singola franchigia, come ad esempio i 112.7 punti per 100 possessi registrati dai Milwaukee Bucks nella stagione 2021/2022 o i 115.2 dei Brooklyn Nets nel 2020/2021.

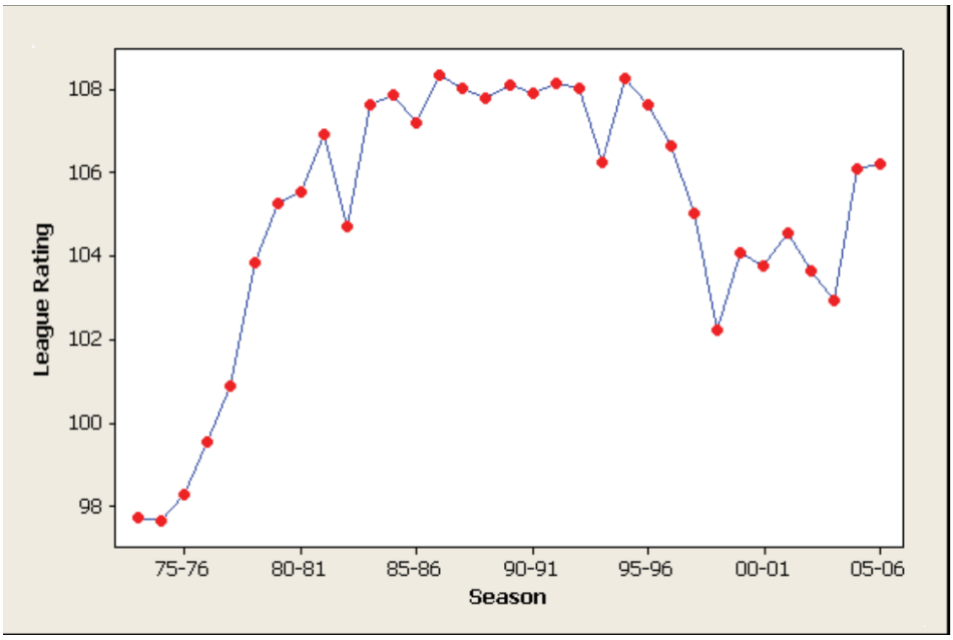


Figura 6: Legue Rating dalle stagioni 1973-74 a 2005-06, Journal of Quantitative Analysis in Sports

Plays

All'interno del suo elaborato, Oliver, chiarì anche il concetto di azione (play), spesso confusa con possesso. Le "azioni" o "giocate", infatti, sono molto simili ai possessi, con la differenza che i rimbalzi offensivi costituiscono una nuova azione, per cui una squadra può tirare, sbagliare, raccogliere il rimbalzo offensivo e dar vita a più azioni senza che la squadra rivale abbia un'azione.

Come si può ben capire, non essendo numericamente simili per ogni squadra durante una partita, quindi, non sono utili come i possessi per valutare l'efficienza di una squadra.

Nonostante ciò, Dean Oliver le stima come segue:

$$PLAYS_t = FGA_t + 0.44\ x\ FTA_t + TO_t$$

Statistiche per minuto

Un aspetto della pallacanestro che Oliver capì prima di altri fu l'importanza dell'introduzione di indicatori statistici basati sui minuti effettivamente giocati, piuttosto che su intere stagioni o serie di match, permettendo di paragonare le performance di giocatori che rappresentavo il quintetto iniziale di una squadra (i giocatori titolari che entravano in campo) a quelle delle riserve in panchina, che godevano di un minor minutaggio.

Tra gli indicatori più utilizzati troviamo:

$$\boldsymbol{PTS40_t = \frac{PTS_p}{MIN_p}\ x\ 48}$$

La Percentuale di Tiro Effettiva

Il semplice rapporto tra i tiri realizzati dal campo e quelli tentati (Field Goal Percentage FG%) non rappresenta una statistica molto precisa sull'efficienza di tiro, infatti non discrimina se questi siano tentativi da due o tre punti, andando a sfalsare un'ipotetica valutazione sulle performance.
Per ovviare a questo problema è possibile calcolare la *Effective Field Goal Percentage* (eFG%), un indicatore in grado di soppesare maggiormente il valore dei canestri da tre punti rispetto a quelli da due, così da garantire una valutazione corretta dell'efficienza di tiro.
Viene calcolata come segue:

$$\boldsymbol{eFG\% = \frac{FGM\ +\ 0,5\ x\ 3PM}{FGA}}$$

I Rimbalzi

I rimbalzi rappresentano uno degli aspetti cruciali nel gioco della pallacanestro, lo sa bene la leggenda Dennis Rodman, che dal 1991 al 1997 ha vinto la classifica come miglior rimbalzista NBA, rappresentando un punto di riferimento in quintetto per le franchigie in

cui ha militato, consentendogli di strappare contratti milionari nonostante il suo comportamento non proprio ortodosso. I rimbalzi possono essere difensivi (recuperando un tiro avversario sbagliato), dando l'opportunità di ripartire in contropiede e potenzialmente segnare un canestro facile a difesa avversaria sguarnita, oppure offensivo (recuperando la palla dopo un tiro sbagliato) concedendo alla squadra una seconda occasione di tiro ed escludendo la possibilità per gli avversari di ripartire a loro volta in contropiede.

Le statistiche legate ai rimbalzi sono pesantemente influenzate dal ritmo di gioco, dall'abilità di una squadra nel tirare a canestro e dal forzare gli avversari ad effettuare tiri difficili, così da aumentare la loro probabilità di sbagliare.
Raramente però una franchigia è ugualmente abile nei rimbalzi in entrambe le porzioni di campo e può rivelarsi utile effettuare una discriminazione tra rimbalzi offensivi e difensivi, per questo le rispettive percentuali di rimbalzo verranno calcolate come segue:

$$\boldsymbol{OREB\%_t = \frac{OREB_t}{OREB_t + DREB_o} x100}$$

rapportando i rimbalzi offensivi di squadra e la somma tra gli stessi ed i rimbalzi difensivi avversari (intesi come possibilità di rimbalzo offensivo non sfruttate dalla squadra e quindi conquistate dagli avversari);

$$\boldsymbol{DREB\%_t = \frac{DREB_t}{OREB_o + DREB_t} x100}$$

rapportando i rimbalzi difensivi di squadra e la somma tra gli stessi ed i rimbalzi offensivi avversari;
mentre la percentuale di rimbalzi totali viene così calcolata:

$$\boldsymbol{REB\%_t = \frac{TREB_t}{TREB_t + TREB_o} x100}$$

rapportando il totale dei rimbalzi della squadra con la somma tra gli stessi ed il totale dei rimbalzi avversari.

Plus/minus

La statistica di Plus/Minus venne introdotta per la prima volta nell'hockey sul ghiaccio, ma ben presto fu adotta anche nella pallacanestro.
Valuta il rendimento di un giocatore attraverso la differenza tra reti fatte e reti subite dalla squadra, quando il giocatore è sul parquet.
È in grado di tenere in considerazione delle sfumature di gioco che altrimenti non verrebbero prese in considerazione da nessun altro parametro, ad esempio l'abilità di un giocatore durante la partita di fare scelte efficaci in termini di squadra, alzando la qualità del gioco, oppure l'abilità di giocare una buona fase difensiva d'aiuto per la squadra.
È vero però, che le semplici statistiche plus/minus di un singolo giocatore possono essere influenzate, sia dalle performance dei suoi quattro compagni di squadra, sia dalle diverse qualità dei giocatori avversari che un quintetto è chiamato ad affrontare.

La Percentuale di Vittoria Pitagorica

Questo indicatore fu teorizzato per la prima volta da Bill James per essere utilizzato nel baseball, ripreso poi da Dean Oliver nei suoi studi nel basket, si fonda sul concetto abbastanza scontato per cui le percentuali di vittoria di una squadra si basano essenzialmente sui punti segnati e quelli concessi, per questo la percentuale di vittoria pitagorica viene calcolata nel seguente modo:

$$\boldsymbol{PYTH_t = \frac{PTS_t^x}{(PTS_t^x + PTS_o^x)}}$$

Dove il pedice "t" indica il team, il pedice "o" indica la squadra avversaria e l'apice "x" rappresenta un esponente determinato empiricamente.

Il termine pitagorico deriva da Bill James, il quale stabilì empiricamente che il valore corretto per l'esponente fosse x=2.

Oliver, per il campionato NBA, teorizzò un valore compreso tra 13 e 17, e solo dopo una collaborazione con ESPN propose il valore 16.5

Four Factors

Tra i contenuti più rivoluzionari proposti da Dean Oliver nei suoi studi, possiamo trovare un pacchetto di indicatori che vanno a riassumere il comportamento e le performance offensive-difensive di una franchigia, i rinomati Four Factors of Basketball Success, ovvero:

1) ***Effective field goal percentage*** (*eFG%* o percentuale effettiva di tiro)
2) ***Turnovers per possession*** (*TOVt/POSSt* o percentuale di palle perse per possesso)
3) ***Offensive rebounding percentage*** (OREBt% o percentuale di rimbalzi offensivi)
4) ***Free throw rate*** (FTMt/FGAt o percentuale di tiri liberi realizzati sui tiri totali tentati)

Questi fattori, durante la valutazione di una performance, non godono della medesima importanza secondo Oliver, che infatti li pesa come segue[27]:

1) *eFG%* (40%)
2) *TOVt/POSSt* (25%)
3) OREB%t (20%)
4) FTMt/FGAt (15%)

[27] https://www.basketball-reference.com/about/factors.html

Capitolo 3

Analisi Regular Season NBA 2021/2022

3.1 Introduzione alla Regular Season NBA

Prima di inoltrarci nella valutazione delle performance delle singole franchigie e nell'implementazione dei modelli regressivi, è bene dedicare il seguente paragrafo ad una breve descrizione sull'organizzazione della National Basketball Association, più comunemente nota come NBA, così da capire al meglio le dinamiche della principale competizione professionistica di basket a livello mondiale.
Nasce a New York il 6 giugno del 1946, con il nome di Basketball Association of America (BAA), ma è solamente nel 1949, a seguito del matrimonio con la lega rivale, la National Basketball League, che prenderà il nome con cui la conosciamo oggi.
Come la maggior parte delle leghe professionistiche americane non risponde ad alcuna federazione sportiva nazionale e non presenta meccanismi di retrocessione in serie minori o di promozione in serie maggiori.
Le squadre che compongono il campionato prendono il nome di franchigie e godono di una completa libertà per quanto riguarda simboli, colori o identità territoriale: ogni franchigia ha la possibilità, a discrezione della lega e degli investitori, di essere perfino trasferita in un'altra città, (il team conserva il nome, i colori, i simboli e persino i titoli vinti) per fini meramente economici e commerciali. Il trasferimento si concretizza in seguito ad una stima sul valore di mercato della nuova città ed i giovamenti che il ricollocamento potrebbe portare alla lega, in linea di massima vengono tenuti in considerazione fattori come: la posizione della città di destinazione, il numero degli abitanti, il nuovo impianto sportivo utilizzato per le gare casalinghe.

Le Franchigie che compongono attualmente l'NBA sono 30, rispettivamente suddivise in due Conference: Western Conference, composta dalle squadre che occupano la zona ovest del paese, ed Eastern Conference, formata dai team nella zona est degli Stati Uniti.
Inoltre, ogni Conference è composta da tre Division, ognuna delle quali conta cinque team, suddivisi come segue:

WESTERN CONFERENCE		
NORTHWEST DIVISION	**SOUTHWEST DIVISION**	**PACIFIC DIVISION**
Denver Nuggets	Dallas Mavericks	Golden State Warriors
Minnesota Timberwoles	Houston Rockets	LA Clippers
Oklahoma City Thunder	Memphis Grizzlies	Los Angeles Lakers
Portland Trail Blazers	New Orleans Pelicans	Phoenix Suns
Utah Jazz	San Antonio Spurs	Sacramento Kings

Tabella 1: Western Conference, fonte https://www.nba.com/

EASTERN CONFERENCE		
ATLANTIC DIVISION	**CENTRAL DIVISION**	**SOUTHEAST DIVISION**
Boston Celtics	Chicago Bulls	Atlanta Hawks
Brooklyn Nets	Cleveland Cavaliers	Charlotte Hornets
New York Knicks	Detroit Pistons	Miami Heat
Philadelphia 76ers	Indiana Pacers	Orlando Magic
Toronto Raptors	Milwaukee Bucks	Washington Wizards

Tabella 2: Eastern Conference, fonte https://www.nba.com/

Una normale stagione NBA è composta da quattro diverse fasi: Pre-Season, Regular Season, Play-off e Finals.

La Pre-Season consiste in un insieme di amichevoli, generalmente a settembre, che anticipano l'inizio del campionato effettivo, allo scopo di "concedere minutaggio" ai giocatori che giocano meno. Da molto tempo ci si interroga sull'utilità di queste partite, anche in virtù del fatto che inevitabilmente intaccano la preparazione dell'intera stagione, come ricordato nel 2016 dall'allenatore Michael Malone.

Una volta conclusa la Pre-Season, ha inizio il campionato vero e proprio, chiamato Regular Season, che si snoda da ottobre fino ad aprile, con una pausa nel mese di febbraio per gli NBA All-Star Game.

Durante la Regular Season ogni franchigia è chiamata a disputare 82 partite in totale:

- 4 sfide con ognuna delle 4 squadre della propria Division (16 partite)
- 4 sfide con le 6 franchigie delle altre Division della stessa Conference (24 partite)

- 3 sfide con le 4 franchigie delle altre Division della stessa Conference (12 partite)
- 2 sfide con le 15 squadre dell'altra Conference (30 partite)

Per quanto riguarda la stesura della classifica delle franchigie durante il campionato, a differenza di altri campionati in cui è presente un sistema di assegnazione di punti in relazione al risultato dei match (ad esempio nel calcio alla squadra vincente viene attribuito un punteggio in classifica di 3 punti mentre di 0 punti alla squadra perdente, mentre in caso di pareggio viene assegnato 1 punto ad entrambe le squadre), l'NBA si basa sulle percentuali di vittoria di ogni squadra (semplice operazione matematica che va a dividere il numero di partite vinte per il numero di partite giocate ed infine moltiplicando per cento) che vanno a definire il piazzamento in classifica durante la Regular Season.

Una volta conclusa la Regular Season, a cavallo dei mesi di aprile e maggio, si disputano i Play-Off, ovvero una sorta di torneo in cui partecipano, divise per Conference, le prime 16 franchigie del campionato, allo scopo di incoronare il vincitore di ogni Conference.

Le squadre verranno accoppiate in relazione alla posizione in classifica durante la Regular Season (la prima con l'ottava, la seconda con la settima, la terza con la sesta e la quarta con la quinta) e si sfideranno in un confronto al meglio di sette partite (supera il turno chi vince almeno 4 partite) con il fattore campo inizialmente a favore della squadra con il miglior piazzamento in campionato (ovvero 2-2-1-1-1 i primi due match le gioca in casa il team con il record migliore, solo successivamente l'altra squadra giocherà 2 partite in casa ed infine una gara a testa in casa fino alla cosiddetta gara 7.

3.2 Preparazione del modello

Come ricordato nei capitoli precedenti, l'obbiettivo dell'analisi statistica applicata all'universo sportivo ha come scopo finale l'ottenimento di informazioni utili a supportare le decisioni prese da allenatori, staff tecnici e sfere dirigenziali. Per questo il seguente progetto verte sull'analisi della stagione regolare NBA, considerando i quattro fattori critici di successo teorizzati da Dean Oliver, andando ad individuare le variabili che hanno contribuito maggiormente al successo o all'insuccesso finale di una squadra, così da poter riproporre un comportamento di gioco se questo ha prodotto esiti positivi oppure evitarlo nel caso abbia prodotto esiti negativi. In più, tramite modelli di regressione lineare è

possibile studiare la correlazione tra le performance registrate prima degli All Star Game e dopo, al fine di capire se la prima parte di stagione possa influenzare la seconda parte ed in seconda battuta il risultato finale della competizione.

L'elaborato è stato sviluppato interamente tramite i fogli di calcolo di Microsoft Excel, utilizzando, come punto di partenza dell'analisi, i box-scores resi noti nella sezione statistiche dal sito ufficiale della National Basketball Association.

Il primo step dell'analisi è stato il reperimento e l'organizzazione dei dati. Infatti, interrogando il database fornito da NBA.com ed utilizzando le specifiche query, è stato possibile ottenere i dati grezzi descritti nel paragrafo 2.4, relativi alle performance di ogni franchigia durante le 82 gare di campionato, rispettivamente rilevati in modalità "Totali", distinti per i segmenti Pre e Post All Star Game, relativi alla Regular Season 2021/2022.

In particolare, la completezza del database messo a disposizione ha permesso di estrarre, oltre alle statistiche relative ad ogni singolo roster, anche quelle relative alle cosiddette "opponent", ovvero le statistiche fatte registrare dalle squadre avversarie durante le partite con la specifica franchigia, utili per calcolare tutti gli indicatori di performance descritti nel paragrafo 2.5.

Così facendo è stato realizzato un foglio di calcolo composto da 31 righe e 117 colonne per un totale di 3627 celle occupate.
Il secondo step del processo è stato il calcolo sistematico di tutti gli indicatori di performance per ogni franchigia, compresa la stima dei valori di benchmark (per ogni indicatore è stata calcolata la media aritmetica dei risultati legati ai vari team), così da poter discriminare se le performance di una squadra erano superiori o inferiori rispetto a quelle medie fatte registrare dalla lega in generale (utilizzando la funzione Formattazione Condizionale presente su Excel).
Inoltre, sfruttando il concetto dei quartili, è stato possibile suddividere il campo di variazione delle percentuali di vittoria in intervalli (utilizzando la funzione SE e la funzione QUARTILE), classificando così le franchigie in 4 diverse categorie: partendo dalla prima, con una percentuale di successo più alta (ospita quindi le squadre più "forti" che hanno raccolto più vittorie), scalando fino alla quarta ed ultima categoria dove si trovano le squadre con una vittoria percentuale più bassa e quindi più "deboli".

Infine, l'ultimo passaggio del lavoro si incentrato sulla visualizzazione dei risultati tramite la costruzione di un diagramma a dispersione degli indicatori di performance come argomentato nel capitolo 1, occupandosi poi della costruzione della retta di regressione.

3.3 Analisi dei Four Factors

Nella tabella (3) sono stati riportati i *Four Factor* di ogni franchigia per la Regular Season 2021-2022, calcolati attraverso le formule riportate nel paragrafo 2.5.

Analizzando i risultati, la prima cosa che si può notare è che, in media, l'intera lega ha migliorato le proprie performance con l'avvicinarsi dei Play Off e delle Finals.
Infatti, calcolando i valori di benchmark per ogni indicatore nella finestra temporale PRE All Star Game e POST All Star Game e facendo poi una semplice differenza delle medie, risulta che gli indicatori sono complessivamente aumentati o rimasti costanti.
Quelli che hanno avuto una crescita maggiore sono quelli legati alle performance di tiro. Più precisamente l' Effective Field Goal Percentage è migliorata in media di 1.96 punti percentuali; per l'appunto, solamente quattro squadre su trenta hanno peggiorato le loro statistiche(in termini di punti percentuale i Portland Trail Blazers di circa 3.88, Chicago Bulls di 2.17, gli Utah Jazz di 0.78, Toronto Raptors 0.07), mentre il tasso dei tiri liberi (FTM/FGA) è aumentato di circa 2,12 punti percentuali, con picchi di 7.26 per i Philadelphia 76ers, 6.81 per i Denver Nuggets e 6.43 per San Antonio Spurs.
Per quanto riguarda la percentuale media di palle perse per possesso è diminuita di 0.55 punti percentuali, evidenziando una diminuzione complessiva del tasso di errore, con un massimo di 1.81 punti percentuali, fatto registrare dagli Atlanta Hawks. È invece interessante osservare come mediamente la percentuale di rimbalzi offensivi della lega è praticamente rimasta invariata per tutta la stagione.
Andando ad approfondire le performance della prima categoria, si noti come nessuna franchigia abbia registrato dei valori superiori alla media in tutti gli indicatori, soprattutto per entrambi i periodi analizzati, nonostante siano le squadre con una percentuale di vittorie superiore alle altre. Solamente i Milwaukee Bucks, campioni in carica della competizione, si sono avvicinati a questo risultato all'interno della prima categoria.
Un aspetto che emerge chiaramente analizzando le prestazioni della prima categoria è che, ad eccezione dei Memphis Grizzlies, tutte le squadre hanno avuto un'efficienza di tiro

superiore alla media della lega nella seconda parte di stagione, questo a riprova delle affermazioni di Dean Oliver sull'importanza dell'eFG% rispetto agli altri indicatori. In questo frangente è sicuramente da evidenziare l'exploit dei Boston Celtics con un record stagionale di 58.72% nel periodo Post All Star Game.
Alla luce di ciò, si può intuire ancora una volta, che la fase cruciale della stagione, in cui le franchigie più competitive concentrano maggiormente gli sforzi, non sia sicuramente l'inizio del campionato, bensì la fase finale.

Per quanto riguarda la seconda categoria è interessante notare come, nonostante la percentuale di vittorie dei Denver Nuggets ed i Toronto Raptors siano uguali, le rispettive vitorie derivino da un gioco differente, infatti, se dal punto di vista statistico i Denver hanno avuto una eFG% nettamente superiore per tutta la stagione ed un'esplosione nei free throw rate nel Post All Star Game, la franchigia di Toronto ha sfoggiato una percentuale di palle perse per possesso tra le più basse della lega per l'intera stagione ed una percentuale di rimbalzi offensivi migliore di molte franchigie di prima categoria. Considerando il peso attribuito da Oliver ai diversi indicatori, ciò risulta plausibile.
Un'ulteriore evidenza da menzionare tra le squadre presenti nella seconda categoria è sicuramente la brillante stagione in termini statistici dei Brooklyn Nets, trainati da campioni del calibro di James Harden e Kevin Durant. Infatti, la squadra di Brooklyn ha avuto delle percentuali superiori alla media in tutti gli indicatori per il periodo Pre All Star Game, facendo poi registrare una flessione dello 0.68% solamente nelle palle perse per possesso nella parte conclusiva della stagione, venendo infine eliminati al primo turno dei Play Off dai Celtics, poi finalisti.
Come preventivabile, analizzando gli indicatori della terza categoria, si può assistere ad un progressivo abbassamento delle performance rilevate dagli indicatori, questo in quasi tutte le franchigie in esame, ad eccezione degli Atlanta Hawks e degli Charlotte Hornets.
Le due franchigie hanno viaggiato durante la stagione sulla stessa percentuale di vittorie, affrontandosi poi ai Play In (mini-torneo che si svolge per entrambe le Conference tra le squadre dalla settima alla decima posizione, al fine di stabilire le ultime due forze che potranno accedere ai Play Off), dove la squadra di Atlanta ha avuto la meglio, giustificando sul campo le statistiche fatte registrare nei box score durante la Regular Season, le quali non sempre venivano tradotte in vittorie.
Invece, come si poteva supporre, le franchigie che compongono la quarta ed ultima categoria risultano quelle con gli indicatori di efficienza peggiori, quasi costantemente sotto la media

di lega. Quello che sicuramente non era atteso era la militanza dei Los Angeles Lakers in questa zona di classifica. Il dream team composto da Lebron James, Anthony Davis, Russell Westbrook, Carmelo Anthony, Isaiah Thomas, giusto per menzionare qualche giocatore, era partito ad inizio stagione con il favore dei pronostici, per poi fermarsi all'undicesima piazza della Western Conference.

La peggior squadra della competizione, in termini di Four Factors, sono stati gli Orlando Magic, con nessun indicatore sopra i valori di benchmark, meritando di fatto l'ultima posizione nella Eastern Conference.

TEAM	W%	CAT	eFG%-PRE	eFG%-POST	DIFF	FTM/FGA-PRE	FTM/FGA POST	DIFF	TOV%-PRE	TOV%-POST	DIFF	OREB%-PRE	OREB%-POST	DIFF
Phoenix Suns	78,05	1	54,44%	56,00%	1,56%	17,79%	17,25%	-0,54%	12,77%	12,44%	-0,33%	22,93%	20,90%	-2,03%
Memphis Grizzlies	68,29	1	51,86%	53,26%	1,40%	18,05%	17,87%	-0,18%	12,94%	12,18%	-0,75%	29,71%	30,71%	1,00%
Golden State Warriors	64,63	1	55,03%	55,58%	0,55%	17,97%	18,29%	0,32%	15,08%	14,14%	-0,94%	22,74%	23,05%	0,31%
Miami Heat	64,63	1	53,80%	57,04%	3,24%	19,62%	22,33%	2,71%	15,07%	13,83%	-1,24%	25,30%	18,36%	-6,94%
Dallas Mavericks	63,41	1	52,86%	56,40%	3,54%	18,43%	21,26%	2,84%	12,79%	12,80%	0,01%	21,39%	21,16%	-0,23%
Milwaukee Bucks	62,20	1	53,99%	56,39%	2,40%	19,11%	22,03%	2,92%	13,31%	12,42%	-0,88%	22,81%	23,70%	0,88%
Philadelphia 76ers	62,20	1	52,62%	55,49%	2,87%	21,07%	28,32%	7,26%	12,62%	12,71%	0,08%	19,80%	20,84%	1,04%
Boston Celtics	62,20	1	52,53%	58,72%	6,20%	19,42%	19,69%	0,27%	14,08%	12,55%	-1,53%	23,84%	24,64%	0,80%
Utah Jazz	59,76	2	55,74%	54,96%	-0,78%	20,72%	20,99%	0,27%	14,20%	13,66%	-0,54%	24,26%	28,01%	3,74%
Denver Nuggets	58,54	2	54,75%	57,66%	2,91%	17,40%	24,21%	6,81%	14,34%	14,44%	0,10%	21,73%	22,37%	0,64%
Toronto Raptors	58,54	2	51,05%	50,98%	-0,07%	17,34%	18,58%	1,25%	12,80%	11,73%	-1,07%	27,98%	29,27%	1,29%
Chicago Bulls	56,10	2	54,74%	52,57%	-2,17%	19,83%	20,98%	1,15%	12,97%	12,27%	-0,70%	20,86%	19,15%	-1,71%
Minnesota Timberwolves	56,10	2	52,57%	57,25%	4,68%	18,30%	23,62%	5,32%	13,93%	13,30%	-0,63%	25,22%	22,01%	-3,21%
Cleveland Cavaliers	53,66	2	53,54%	54,32%	0,77%	18,52%	23,03%	4,51%	15,06%	13,55%	-1,51%	24,14%	23,49%	-0,65%
Brooklyn Nets	53,66	2	52,99%	56,44%	3,44%	18,96%	21,88%	2,92%	13,63%	14,31%	0,68%	23,40%	25,21%	1,81%
Atlanta Hawks	52,44	3	53,96%	54,98%	1,02%	19,91%	21,93%	2,02%	12,39%	10,59%	-1,81%	22,82%	23,38%	0,57%
Charlotte Hornets	52,44	3	53,05%	58,27%	5,22%	17,19%	17,67%	0,48%	12,83%	12,87%	0,05%	23,13%	23,91%	0,79%
LA Clippers	51,22	3	52,56%	54,73%	2,17%	17,12%	19,59%	2,46%	14,03%	12,49%	-1,54%	20,10%	22,03%	1,93%
New York Knicks	45,12	3	50,95%	52,34%	1,39%	19,56%	24,16%	4,60%	13,77%	12,61%	-1,16%	24,94%	25,43%	0,50%
New Orleans Pelicans	43,90	3	50,89%	53,81%	2,93%	20,29%	22,10%	1,81%	14,43%	13,01%	-1,42%	25,80%	29,70%	3,90%
Washington Wizards	42,68	3	52,33%	55,47%	3,14%	19,80%	19,59%	-0,21%	13,32%	12,81%	-0,51%	21,24%	20,12%	-1,12%
San Antonio Spurs	41,46	3	52,57%	53,16%	0,59%	14,82%	21,26%	6,43%	12,19%	12,59%	0,40%	23,63%	23,86%	0,24%
Los Angeles Lakers	40,24	4	53,36%	54,40%	1,04%	18,10%	21,04%	2,93%	14,16%	13,57%	-0,59%	21,09%	21,21%	0,12%
Sacramento Kings	36,59	4	52,19%	53,23%	1,03%	19,24%	23,27%	4,03%	13,79%	13,47%	-0,33%	22,54%	17,96%	-4,59%
Portland Trail Blazers	32,93	4	52,64%	48,75%	-3,88%	18,91%	18,65%	-0,26%	13,82%	15,77%	1,95%	22,93%	22,98%	0,05%
Indiana Pacers	30,49	4	52,35%	55,11%	2,76%	18,04%	19,20%	1,16%	14,11%	14,21%	0,10%	25,06%	26,73%	1,67%
Oklahoma City Thunder	29,27	4	48,34%	53,18%	4,84%	16,46%	17,86%	1,40%	13,37%	14,70%	1,33%	21,89%	21,64%	-0,25%
Detroit Pistons	28,05	4	48,36%	52,00%	3,64%	19,08%	20,33%	1,25%	14,32%	13,11%	-1,21%	22,93%	24,59%	1,66%
Orlando Magic	26,83	4	50,10%	50,66%	0,55%	17,66%	17,21%	-0,45%	14,39%	13,52%	-0,87%	19,97%	19,50%	-0,47%
Houston Rockets	24,39	4	52,80%	54,77%	1,97%	20,84%	18,84%	-2,00%	16,33%	14,63%	-1,71%	22,21%	20,57%	-1,65%
VALORI DI BENCHMARK			**52,63%**	**54,60%**	**1,96%**	**18,65%**	**20,77%**	**2,12%**	**13,76%**	**13,21%**	**-0,55%**	**23,21%**	**23,22%**	**0,00%**
COEFFICIENTE DI CORRELAZIONE				**0,473747231**			**0,404624709**			**0,59901826**			**0,740884193**	

Tabella 3: Four Factors Regular Season 2021/2022, fonte: Nostre elaborazioni

3.4 Modello di regressione lineare inerente all'Effective Field Goal Percentage

Plottando all'interno della figura (7) i risultati descritti nel paragrafo 3.4, dove l'asse delle X rappresenta l'eFG% Pre All Star Game e l'asse delle Y l'eFG% Post All Star Game, emergono nuove sfaccettature che attraverso la semplice rappresentazione tabellare potrebbero essere difficili da cogliere, agevolando il lavoro di analisi.

Infatti, il coefficiente di correlazione lineare di Peason raggiunge circa il 47% del suo massimo teorico, per cui vi è una moderata correlazione positiva tra l'Effective Field Goal Percentage Pre e Post All Star Game. Ciò significa che quando aumenterà il valore della variabile indipendente ci si aspetta un aumento quasi proporzionale anche della variabile dipendente (con proporzioni che possono essere inferiori o superiori a 1) e questo è chiaramente illustrato dalla retta di regressione, avente un'inclinazione crescente con coefficiente angolare un valore di circa 0,6334, per cui aumentando di un punto percentuale l'eFG% Pre All Star Game, la variabile Post aumenterà in media di 0,63 punti percentuali.

È vero però che l'indice di determinazione lineare R^2 raggiunge solamente il 22% del suo massimo teorico per cui la retta interpolante spiega solamente il 22% della variabilità dei dati, ovvero solamente il 22% circa dell'aumento dell'Effective Field Goal Percentage Post All Game può essere spiegato da quella Pre All Star Game.

Questo significa che la nube di punti, come si nota chiaramente anche dal grafico di Figura XXX, è moderatamente dispersa attorno alla retta usata per rappresentarla. Pur tenendo conto di questo limite, si tratta comunque di un valore che indica un modello accettabile dal punto di vista di una rappresentazione a grandi linee dei dati. Il fatto che ci sia un elevata variabilità, difficile da spiegare in modo compiuto attraverso la retta, può essere considerato un fattore interessante: la relazione tra l'eFG% di una squadra nella prima e nella seconda parte della stagione non è così intensa come si potrebbe supporre a priori. Questo porta con sé due osservazioni importanti: (1) esistono fattori che comportano rilevanti variazioni nel tempo della performance di una squadra (tra questi possiamo annoverare anche gli infortuni, ad esempio) e quindi (2) la classifica provvisoria Pre All Stat Game potrebbe variare considerevolmente prima della fine della stagione regolare.

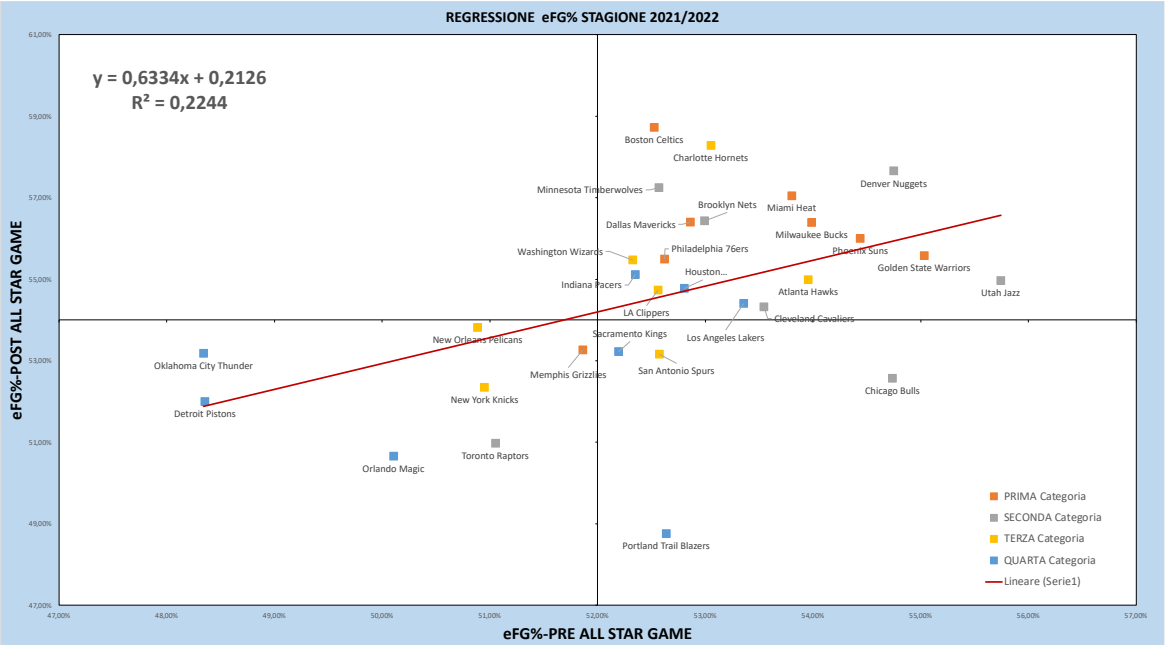


Figura 7: Regressione eFG% stagione 2021/2022 (fonte: Nostre elaborazioni)

3.5 Modello di regressione lineare inerente al Free Throw Rate

Andando a reiterare il ragionamento illustrato per l'eFG%, si noti come, anche nel diagramma di figura (8) sviluppato per il tasso di tiri liberi, il risultato sia approssimativamente uguale.

Infatti, il coefficiente di correlazione lineare r raggiunge circa il 40% del suo massimo teorico, sottolineando una modesta correlazione positiva tra le due variabili, per cui quando aumenterà la variabile esplicativa X, anche quella di risposta Y tenderà ad aumentare.

Ciò è più immediato andando a studiare la retta di regressione lineare, dove il coefficiente angolare risultante è di circa 0.74, ciò comporta che in linea di massima aumentando di un punto percentuale il tasso di tiri liberi Pre All Star Game, ci si aspetta un aumento anche dell'indicatore Post All Star Game, con una proporzione media di 0.74.

La dispersione dei punti dello scatter plot lontano dalla retta di regressione lascia presagire ancora una volta una scarsa aderenza del modello alla realtà; infatti, il coefficiente di determinazione lineare R^2 raggiunge solamente 16% del suo massimo teorico, per cui la

retta di regressione spiega solamente il 16% della variabilità dei dati del fenomeno, non rappresentando una buona interpolante. Riguardo questa circostanza valgono le osservazioni sottolineate in precedenza.

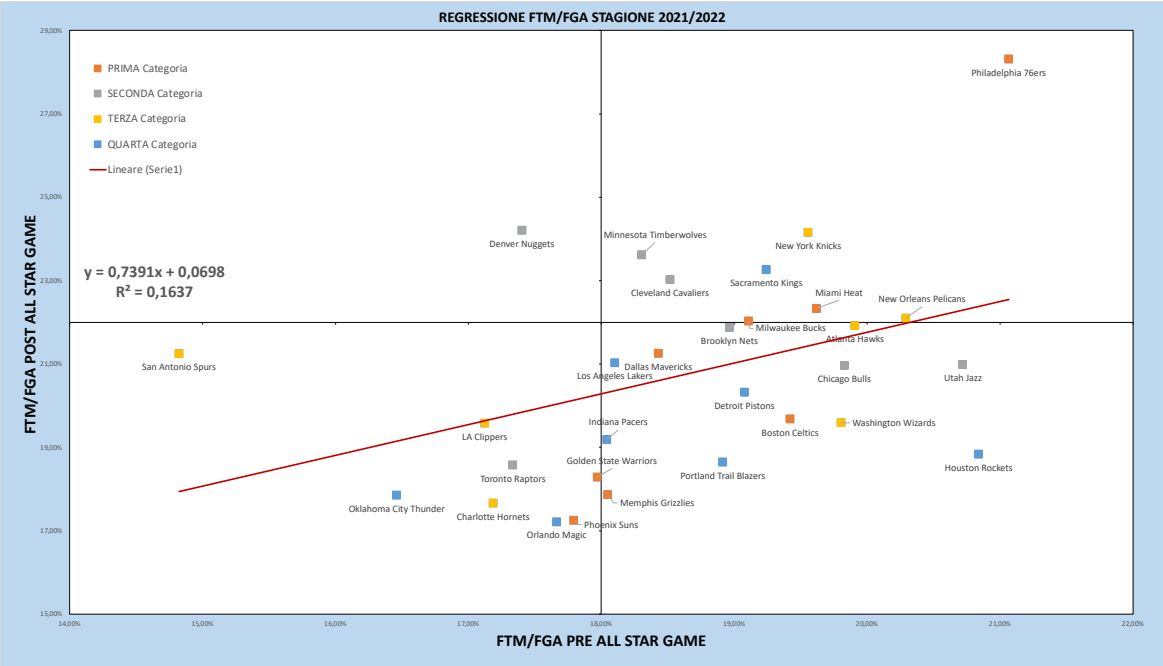


Figura 8: Regressione FTM/FGA stagione 2021/2022 (fonte: Nostre elaborazioni)

3.6 Modello di regressione lineare inerente ai Turnovers per Possession

Il diagramma a dispersione di Figura (9) si riferisce agli indicatori relativi alle palle perse per possesso e si può intuire subito una migliore correlazione lineare tra le variabili prese in considerazione rispetto ad i modelli precedenti.

Infatti, la nuvola di punti è più aderente alla retta di regressione e il coefficiente di correlazione lineare di Pearson raggiunge circa il 60% del suo massimo teorico, facendo emergere una correlazione lineare positiva piuttosto elevata tra le palle perse per possesso Pre All Star Game e quelle Post All Star Game.

La retta di regressione con un coefficiente angolare di 0,67 circa, fa emergere il fatto che andando ad aumentare di 1 punto percentuale la variabile indipendente X, la variabile dipendente Y registrerà in media un aumento di 0,67% punti percentuali circa.

Inoltre, l'indice di determinazione R^2 raggiunge il 36% circa del suo massimo teorico, sicuramente meglio dei precedenti risultati, poiché la retta interpolate spiega già il 36% della variabilità dei dati e infatti la nube di punti risulta anche visivamente meno dispersa attorno alla retta interpolante.

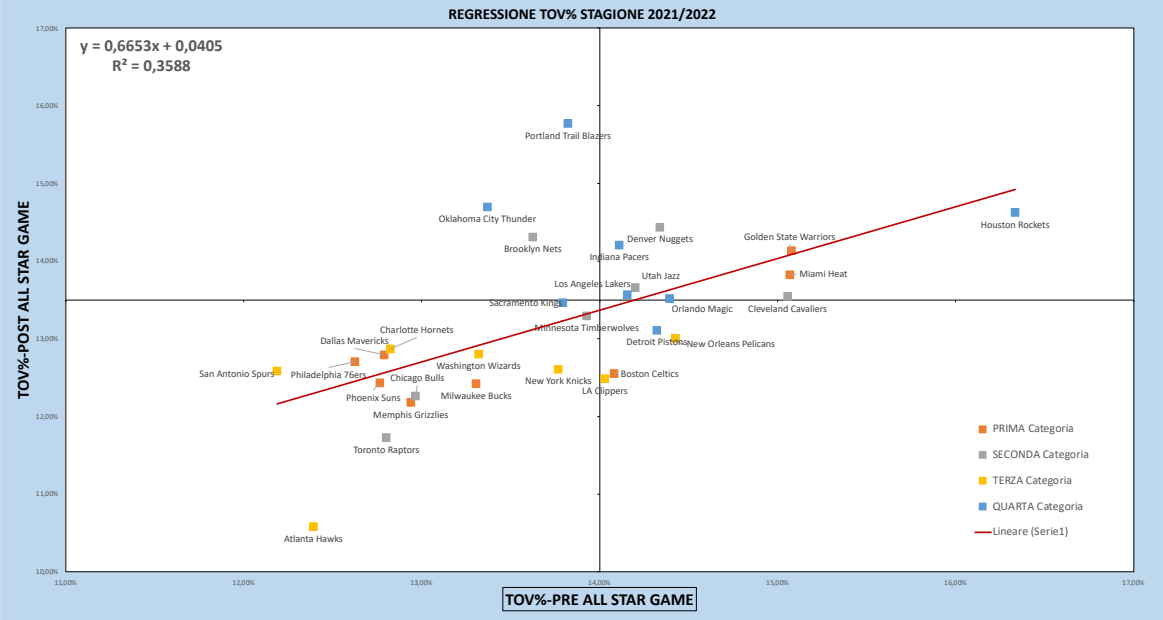


Figura 9: Regressione Tov% stagione 2021/2022 (fonte: Nostre elaborazioni)

3.6 Modello di regressione lineare inerente all'Offensive Rebounding Percentage

L'ultimo dei Four Factors preso in considerazione è quello relativo ai rimbalzi offensivi.

Come si può notare dal diagramma di Figura (10) questo indicatore è quello che si presta meglio all'implementazione di un modello regressivo.

Infatti, vi è un'elevata correlazione lineare tra i rimbalzi offensivi Pre e Post All Star Game, e l'osservazione della nube consente di osservare che molti dei punti in esame si posizionano a ridosso dell'interpolante.

Dal punto di vista algebrico tutto ciò è dimostrato dal fatto che coefficiente di correlazione lineare raggiunge il 74% del suo massimo teorico, sottolineando una forte correlazione lineare positiva tra le variabili di rimbalzo, per cui ancora una volta all'aumentare di X aumenterà anche la Y.

In più, la retta di regressione gode di un coefficiente angolare di circa 1.08, per cui ad un aumento di un punto percentuale dell'indicatore Pre All Star Game, seguirà un aumento di 1.08 punti percentuali dell'indicatore Post.

In questo caso, l'indice di determinazione raggiunge il 55% del suo massimo teorico, per cui la retta di regressione spiega il 55% della variabilità dei dati, che si può considerare un risultato decisamente buono. Il valore sarebbe anche più elevato, se non fosse per alcuni punti outlier, che si posizionano lontano dalla retta ed evidenziano delle situazioni anomale che sarebbe interessante investigare più in profondità (Sacramento Kings e Miami Heat).

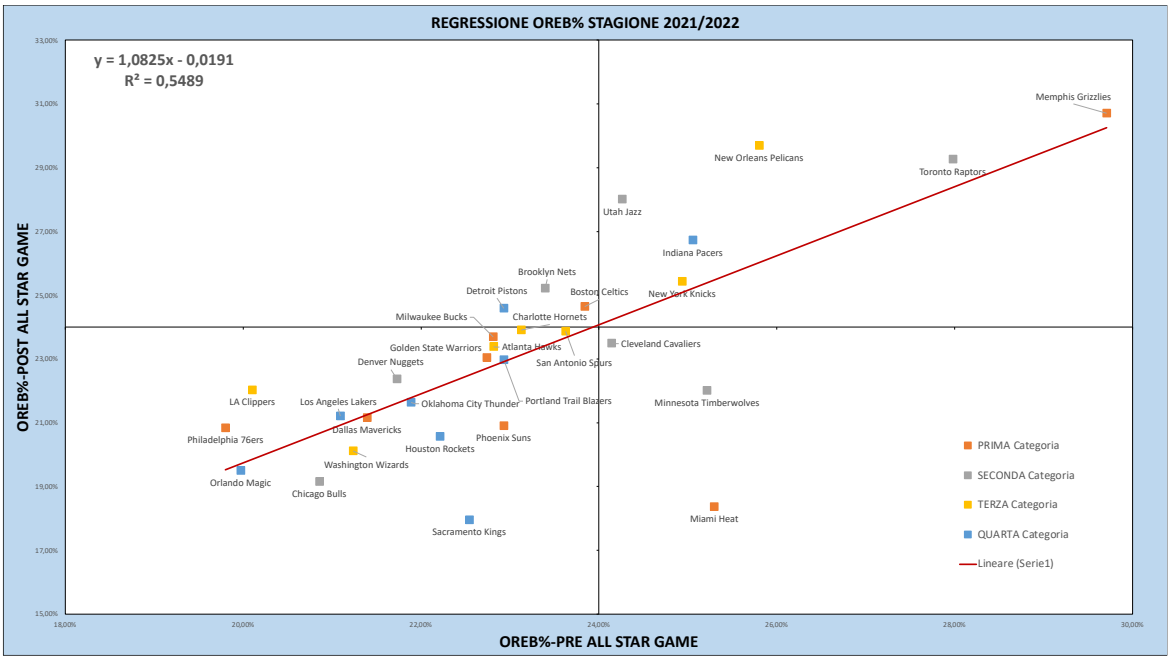


Figura10: Regressione OREB% stagione 2021/2022 (fonte: Nostre elaborazioni)

Conclusioni

Nel corso dei precedenti capitoli sono state messe in evidenza le tematiche relative ai Big Data, in particolar modo sottolineando la tendenza sempre più pervasiva, in qualsiasi campo, di approcci data-driven alle decisioni più nevralgiche e delicate, ovvero utilizzando in modo razionale un metodo prettamente scientifico basato sui dati, piuttosto che sull'intuito e sull'esperienza del singolo decisore.

Dopo aver descritto le possibili aree di utilizzo e i risvolti derivanti dall'applicazione di tecniche di Data Science, si è preferito concentrarsi sul mondo degli Sport Analytics, in particolar modo sulla pallacanestro, attività da sempre legata a doppio filo con la statistica e i numeri.

Se in prima battuta i risultati forniti dai box scores e dagli indicatori di performance possono sembrare confusionari, vaghi e poco utili, effettuando una riorganizzazione e una manipolazione degli stessi, emerge invece come rappresentino un punto di partenza ottimo per approfondire qualsiasi aspetto del gioco. In particolare, è stato possibile evidenziare quali sono gli aspetti cruciali che hanno portato alcune franchigie al successo ed all'insuccesso di altre.

Inoltre, i modelli di regressione lineare implementati utilizzando come variabili i Four Factors inerenti alla stagione NBA 2021/2022, hanno permesso di dimostrare la presenza di una correlazione positiva, ma comunque non sempre elevata, tra le performance della prima parte di stagione, quella Pre All Star Game, e la seconda, quella Post All Star Game. Il fatto che la correlazione esista si può considerare un risultato atteso, meno atteso è il fatto che comunque ci sia una variabilità piuttosto elevata tra la prima e la seconda parte della stagione, specialmente sugli indicatori riguardanti l'efficienza di tiro, i tiri liberi e, in misura inferiore, le palle perse.

Inoltre, valutando i risultati della prima parte di stagione, si è osservato un generalizzato miglioramento nella seconda parte, oppure alle medesime performance, ma raramente ad un peggioramento.

Infatti, anche in funzione del successo finale dei Golden State Warriors, si può notare che i risultati della seconda parte della stagione sono risultati maggiormente determinanti sull'esito finale del campionato. Un'ipotesi è che i team di vertice dosino le forze nel corso delle 82 partite, giocandosi poi il tutto e per tutto nelle fasi finali.

Si ricordi che tali modelli non possono sicuramente essere perfettamente attinenti alla realtà e alcuni di essi presentano evidenti problemi di underfitting. Per l'appunto, riassumere in maniera calzante la relazione che intercorre tra variabili riguardanti un'intera stagione,

sfruttando solamente la funzione matematica di una retta è piuttosto ambizioso. Ciò è dovuto a molteplici cause, ma soprattutto all'incapacità di un modello basico come la regressione lineare semplice di riassumere relazioni complesse su cui influiscono una molteplicità di fattori diversi. Nonostante questo, pur non padroneggiando modelli statistici più sofisticati, è stato possibile effettuare un'indagine statistica in grado di monitorare le prestazioni dei vari attori coinvolti nel campionato NBA, fornendo delle indicazioni generali sugli aspetti fondamentali da tenere in considerazione nella valutazione delle performance di una squadra e cosa può impattare maggiormente sul successo o l'insuccesso di essa, con un fuoco sul modo in cui le prestazioni si distribuiscono nel tempo.

Sicuramente è possibile approfondire ulteriormente l'indagine, per esempio, andando a valutare le singole performance di ogni giocatore, o utilizzando più parametri di benchmark. Il tema conclusivo della relazione è proprio questo, comprendere la complessità delle problematiche che un data scientist è chiamato ad affrontare per risolvere questioni che apparentemente possono sembrare solamente chiacchere da bar che possono essere risolte con qualche vaga supposizione. In realtà l'implementazione di qualsiasi modello richiede delle conoscenze statistiche, matematiche, informatiche, di machine learning e IA così approfondite da richiedere anni di studi e specializzazioni, per questo ad oggi i data scientist sono tra le figure più ricercate e retribuite nel mercato del lavoro.

Bibliografia

Zuccolotto & Manisera (2020) Basketball Data Science with Applications in R

Thomas A. Severini (2015). Analytic Methods in Sports

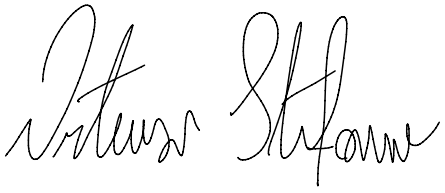
Chapman & Hall (2017). Handbook of Statistical Methods and Analyses in Sports

Thomas W. Miller (2016). Sports Analytics and Data Science

Robert P. Schumaker, Osama K. Solieman, Hsinchun Chen, Sports Data Mining

Dean Oliver, Justin Kubatko, Kevin Pelton, Dan T. Rosenbaum (2007). A Starting Point for Analyzing Basketball Statistics

Sitografia

- https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century
- https://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/
- https://www.statista.com/chart/17727/global-data-creation-forecasts/
- https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/
- https://en.wikibooks.org/wiki/Lentis/%22Data_is_the_new_oil%22
- https://towardsdatascience.com/5-steps-of-a-data-science-project-lifecycle-26c50372b492
- https://it.mathworks.com/discovery/overfitting.html
- //i.stack.imgur.com/cIDB3.png
- https://www.coni.it/images/documenti/Carta_europea_dello_Sport.pdf
- https://www.ilsole24ore.com/art/lo-sport-italia-vale-788-miliardi-pari-3percento-pil-AEG3V5NB

- https://it.wikipedia.org/wiki/Storia_della_pallacanestro
- https://www.museodelbasket-milano.it/leggi.php?s=&idcontenuti=266
- https://www.nba.com/stats/players/traditional?PerMode=PerGame&sort=PTS&dir=-1&Season=2021-22&SeasonType=Regular+Season&TeamID=0
- https://www.dailybasket.it/nba/nba-stats-welcome-to-moreyball/
- https://masterx.iulm.it/uncategorized/three-is-the-magic-number-levoluzione-recente-del-tiro-da-tre/
- https://www.basketball-reference.com/about/factors.html