LUISS

Dipartimento di Impresa e Management
Corso di Laurea Magistrale in Marketing

Cattedra in Business & Marketing Analytics

Machine Learning: l'integrazione delle leve di trade marketing per la previsione delle vendite

Prof. Frontoni Emanuele

RELATORE

Prof. Laura Luigi

CORRELATORE

Caso Fabio 764581

CANDIDATO

ANNO ACCADEMICO 2023/2024

INDICE

1 INTRODUZIONE 3

1.1 CONTESTO GENERALE 3

1.1.1 Il trade marketing 6

1.1.2 Il machine learning 7

1.2 TRADE MARKETING IN BARILLA: LA STIMA DEI VOLUMI 12

1.2.1 Dinamiche di sell-in e sell-out nel fmcg 12

1.2.2 Metodologia di stima 14

1.2.3 Limiti del processo 16

1.3 OBIETTIVI DI RICERCA 17

1.4 STRUTTURA DELLA TESI 22

2 REVISIONE DELLA LETTERATURA 25

2.1 APPROCCI TRADIZIONALI ALLA PREVISIONE DELLE VENDITE 25

2.2 UTILIZZO DEL MACHINE LEARNING PER LA PREVISIONE DELLE VENDITE 27

2.2.1 Gli algoritmi utilizzati 29

2.3 IL TRADE MARKETING E L'INTEGRAZIONE CON IL ML 33

2.4 IMPLICAZIONI DALLA LITERATURE REVIEW 36

3 METODOLOGIA 38

3.1 DESIGN DELLA RICERCA 38

3.1.1 Metodologia di raccolta e analisi 40

3.1.2 Linguaggio di programmazione: python 42

3.2 DATASET 44

3.2.1 Feature 46

3.3 DATA PRE-PROCESSING 49

3.4 ALGORITMI UTILIZZATI 54

4 RISULTATI 57

4.1 DESCRIZIONE DEI DATI 57

4.2 RELAZIONE TRA I DATI 62

4.3 PERFORMANCE DEI MODELLI 65
4.4 IMPLICAZIONI MANAGERIALI 66
4.5 LIMITI E RICERCHE FUTURE 68
5 CONCLUSIONI 70

BIBLIOGRAFIA 72

1 INTRODUZIONE

1.1 CONTESTO GENERALE

Il mercato dei beni di largo consumo (FMCG) rappresenta uno dei settori più dinamici dell'economia globale. Questa industria svolge un ruolo cruciale nel soddisfare le esigenze quotidiane dei consumatori: secondo il Barometro dei consumi di Nielsen (NIQ), nel contesto italiano, il FMCG nel 2023 ha registrato un notevole fatturato di 134 miliardi di euro.

Il mondo del largo comsumo è estremamente diversificato, e include una vasta gamma di prodotti come cibo, bevande, prodotti per la pulizia, igiene personale, cosmetici, farmaci da banco, articoli per la casa, prodotti per l'infanzia e articoli per animali domestici. Sempre secondo dati NIQ, nel panorama italiano la categoria che emerge come la più rilevante è quella alimentare, la quale ha contribuito ad oltre il 60% del fatturato totale nel 2023.

Tra i prodotti alimentari italiani la pasta rappresenta un elemento iconico e centralissimo (sia per i consumatori che per i produttori e rivenditori), in quanto non è solo un elemento essenziale della dieta italiana, ma è anche un bene di largo consumo che genera milioni di tonnellate vendute ogni anno. Il mercato della pasta è caratterizzato da una forte competizione sia a livello nazionale che internazionale. Le aziende italiane sono leader nel settore ed esportano pasta in tutto il mondo. Da questo punto di vista, la reputazione del prodotto made in Italy è un grande vantaggio competitivo, spesso utilizzato per differenziare i prodotti sul mercato globale.

Da anni ormai, uno dei principali attori nel mercato internazionale della pasta è *Barilla,* con una quota di mercato dominante e una presenza consolidata non solo in Italia ma in tutto il mondo. L'azienda, negli anni, è stata in grado di costruirsi un'elevata brand equity, posizionandosi come il top-of-mind brand tra i consumatori. Ad oggi, grazie ad una strategia efficace di brand positioning e brand loyalty, ha creato un'immagine di marca associata a qualità, tradizione e affidabilità, consolidando la sua brand salience.

Una delle caratteristiche distintive dei beni di largo consumo, come ad esempio la pasta, è l’alto tasso di rotazione. Questi prodotti, infatti, sono acquistati frequentemente dai consumatori, garantendo un flusso costante di domanda e vendite. Inoltre, la loro ampia disponibilità geografica e la relativa economicità li rendono accessibili a un vasto segmento della popolazione.

In questo contesto, diventa quindi fondamentale per le aziende comprendere e anticipare la domanda del consumatore.

Una previsione accurata dei volumi di vendita è di cruciale importanza non solo per ottimizzare la produzione o gestire efficientemente lo stock, ma consente anche di ottenere un importante vantaggio competitivo, dato dalla capacità di rispondere prontamente ai cambiamenti del mercato, e di ridurre al minimo i casi di out of stock per far sì che i prodotti siano sempre presenti sugli scaffali dei retailers.

L'implementazione di tecniche di demand planning o sales forecasting consente alle aziende di ottimizzare le proprie risorse sotto numerosi punti di vista.

Secondo un articolo pubblicato su *International Journal of Production Economics*, una stima accurata delle vendite permette di migliorare l'efficienza della supply chain, riducendo gli sprechi e i costi associati all'inventario e alla logistica (W.K. Wong n & Z.X. Guo, 2010). Questa ottimizzazione è fondamentale per mantenere margini di profitto adeguati in un settore caratterizzato da basse marginalità.

Anticipare correttamente la domanda dei consumatori permette alle aziende di mantenere livelli di inventario adeguati, assicurando la disponibilità del prodotto nei punti vendita. Rendere i prodotti disponibili quando e dove gli shopper li cercano, come evidenziato da Christopher (2016), porta ad una maggiore soddisfazione del cliente, incentiva l'acquisto da parte dello shopper e migliora la fidelizzazione del consumatore.

Un'asset fondamentale per costruire e mantenere relazioni solide con i clienti è la capacità di rispondere tempestivamente alle esigenze del mercato. Le aziende che utilizzano modelli avanzati di previsione sono in grado di rispondere più rapidamente alle variazioni del mercato, adattandosi ai cambiamenti nelle preferenze dei consumatori e alle condizioni economiche. Un articolo di Fisher, Raman e McClelland (2019) su *Harvard Business Review* sottolinea come la flessibilità operativa derivante da un'efficace pianificazione delle vendite possa rappresentare un vantaggio competitivo critico che consente alle aziende di sfruttare nuove opportunità di mercato.

Inoltre, la conoscenza approfondita delle tendenze della domanda può ottimizzare le strategie di marketing ed aiutare nello sviluppo di nuovi prodotti. A conferma di questo assunto, Kotler e Keller (2020), in *Marketing Management*, affermano che le aziende in grado di comprendere meglio i comportamenti e le preferenze dei consumatori possono sviluppare campagne di marketing più

mirate e lanciare prodotti innovativi che rispondano effettivamente alle esigenze del mercato. Questo approccio basato sui dati contribuisce a ridurre il rischio associato all'introduzione di nuovi prodotti e a massimizzare il ritorno sugli investimenti di marketing.

Una accurata previsione dei volumi di vendita offre quindi molteplici vantaggi operativi, finanziari e strategici. Tuttavia, questa attività è diventata una sfida estremamente complessa a causa della moltitudine di variabili, esogene ed endogene, che influenzano il processo di previsione.

Tra i vari fattori esogeni in grado di influenzare le stime previsionali, dobbiamo considerare, in primis, le dinamiche competitive di un mercato in continua evoluzione. Altro elemento cruciale è il ruolo delle innovazioni tecnologiche, che possono modificare drasticamente i comportamenti dei consumatori e le strategie aziendali. Ulteriori variabili da monitorare sono anche le nuove leggi e regolamentazioni oppure i trend sociali e culturali, in costante cambiamento, che influenzano le preferenze e le abitudini di consumo, richiedendo un continuo adattamento delle strategie di marketing e di prodotto. Infine, un fattore determinante che può influire sulla capacità di spesa dello shopper è anche l’impatto delle dinamiche economiche globali e locali.

I fenomeni esterni, nella gran parte dei casi, hanno la caratteristica distintiva di essere difficilmente prevedibili dalle aziende. Questi fattori agiscono al di fuori della sfera di influenza aziendale, rendendo complesso anticiparne l'evoluzione e l'impatto. Pertanto, le imprese devono sviluppare tecnologie flessibili in grado di adattarsi ai cambiamenti esterni, gestire le incertezze e mitigare i rischi.

Differentemente da quest’ultimi, i fattori endogeni possono invece essere regolati e gestiti dalle aziende in quanto sono elementi interni, come la cultura organizzativa, le capacità tecnologiche e le strategie di gestione, che influenzano le sue performance e il suo sviluppo competitivo. Queste variabili rappresentano una solida base per effettuare proiezioni dettagliate dei volumi di vendita.

In particolare, nel presente studio, si utilizzano le leve endogene della disciplina aziendale del trade marketing per prevedere le vendite di un’azienda nel mercato della pasta in Italia.

1.1.1 IL TRADE MARKETING

"In quanto processo volto ad integrare gli obiettivi e le strategie di vendita e marketing, il trade marketing è progettato per garantire che le esigenze del retailer (ad esempio in termini promozionali), siano soddisfatte dal marketing mix del marchio” - definizione di Dewsnap e Jobber (2009, p. 989).

In altri termini, possiamo dire che il Trade Marketing è una disciplina che si concentra sulla creazione e implementazione di strategie di marketing rivolte ai canali di distribuzione e ai partner commerciali, piuttosto che direttamente ai consumatori finali. Il suo lavoro è essenziale per creare sinergie tra produttori e distributori, migliorando la visibilità dei prodotti e ottimizzando la loro presenza sul mercato.

Le principali leve operative di questa funzione aziendale sono:

1. Prezzo
2. Promozioni
3. Distribuzione & Assortimento

Nello specifico le varie attività che deve affrontare un trade marketeer sono:

- Definire gli action plans in cui vengono sviluppati programmi promozionali in grado di attirare l'attenzione dei clienti, aumentando così l'engagement e la conversion (trade promotions)
- Sviluppo di attività volte ad ottimizzare la visibilità dei prodotti nei punti vendita e sugli scaffali per stimolare l’acquisto da parte dello shopper
- Coordinare gli obiettivi di sales e marketing per assicurare che le attività sul campo siano allineate con le strategie aziendali complessive
- Assicurare una distribuzione efficiente delle risorse di vendita, garantendo che gli sforzi siano concentrati sui clienti e sui mercati con il maggior potenziale di crescita
- Collaborare con il KAM (Key Account Management) per creare strategie di vendita e promozioni tailor-made per i principali clienti e canali di distribuzione
- Assistere il KAM nelle negoziazioni con i clienti chiave, fornendo dati e analisi che supportino le proposte di valore
- Analisi dei dati di vendita (analisi di sell-out) al fine di fornire al team di KAM e alla forza vendita sul campo insights dettagliati basati su dati di mercato.

- Creare campagne di co-marketing con i rivenditori. Queste collaborazioni possono includere iniziative di branding condivise, eventi promozionali congiunti e campagne pubblicitarie coordinate. Tali attività non solo amplificano l’efficacia delle promozioni, ma rafforzano anche le alleanze strategiche tra produttori e distributori, creando un ecosistema di vendita più coeso e reattivo.

(Cespedes 1993, 1995; Davies 1993, 1994; Dewsnap and Jobber 2003, 2009; Piercy 1985; Promotion Optimization Institute 20 March 2013; Randall 1994; Thain and Bradley 2012; Zentes 1989; Kotler, P., & Keller, K. L. 2016).

Uno studio pubblicato nel *Journal of Retailing and Consumer Service* sottolinea che il prezzo è uno dei fattori più influenti nelle decisioni di acquisto dei consumatori di beni di largo consumo (Fornari, E., Fornari, D., Grandi, S., & Menegatti, M. - 2013). Il trade marketing, attraverso attività volte alla definizione dei programmi promozionali, influenza direttamente la leva del prezzo e la percezione del valore da parte dello shopper, incentivando così la conversione all’acquisto.

Precedentemente, Shankar e Bolton (2004), nel *Journal of Marketing*, hanno affermato che le decisioni di prezzo sono influenzate dalla collaborazione tra produttori e rivenditori, facilitata dalle strategie di trade marketing. La collaborazione permette di creare strategie di pricing più efficaci, che tengono conto sia degli obiettivi di vendita dei produttori sia delle esigenze di profitto dei rivenditori.

Pertanto, questa disciplina ha il potenziale di influenzare significativamente le performance di vendita di un'azienda. Le sue leve operative, utilizzate in modo strategico, possono migliorare la visibilità del prodotto, incentivare i distributori, stimolare la domanda e ottimizzare la penetrazione del mercato. Tuttavia, è fondamentale gestire attentamente ogni attività per evitare impatti negativi sui margini di profitto e preservare la percezione del valore del marchio.

1.1.2 IL MACHINE LEARNING

Una delle tecniche più diffuse e popolari volta ad affrontare la grande sfida del sales forecasting è il machine learning (ML o apprendimento automatico). Quest’ultimo è un campo dell'intelligenza artificiale che si concentra sullo sviluppo di algoritmi e tecniche che permettono ai computer di

imparare dai dati. Gli algoritmi di apprendimento automatico, infatti, si basano sulla capacità di trovare pattern informativi nei dati passati, con l’obiettivo di migliorare le prestazioni sui dati futuri.

Gli esseri umani imparano attraverso l'esperienza, utilizzando un processo di tentativi ed errori per capire quali azioni dovrebbero essere intraprese in determinate circostanze. Questo ci permette di fare astrazioni e costruire conoscenza. Il ML è in qualche modo simile, in quanto include tecniche che permettono alle macchine di migliorare nei tasks grazie all'esperienza (Raschka et al, 2017, e Mitchell, 1997).

Tale tecnologia sta trasformando il modo in cui le aziende analizzano i dati e fanno previsioni. Secondo un articolo di Domingos (2012), il machine learning offre strumenti potenti per identificare pattern e tendenze che sarebbero impossibili da rilevare attraverso le tradizionali tecniche analitiche. Gli algoritmi di apprendimento automatico sono utilizzati ogni giorno per prendere decisioni critiche nella diagnosi medica, nel commercio azionario, nella previsione del carico energetico e nella manutenzione predittiva.

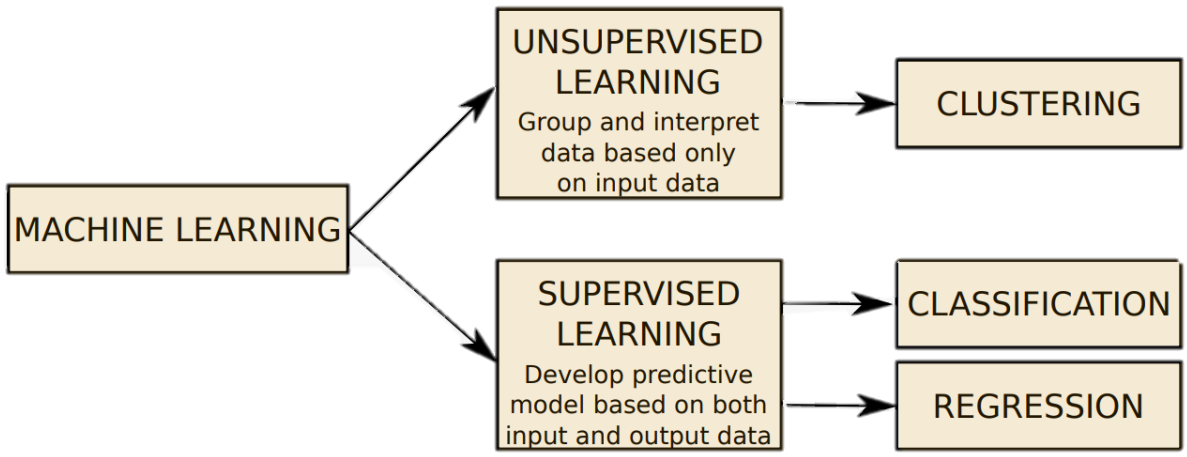
In particolare, nel contesto delle previsioni di vendita, permette di analizzare enormi quantità di dati storici e di mercato per generare modelli predittivi altamente accurati. Ad esempio, Bajari et al. (2015) hanno dimostrato come le tecniche di machine learning possano migliorare significativamente le previsioni di domanda e le decisioni di pricing, permettendo alle aziende di ottimizzare le loro strategie commerciali. Questo è particolarmente importante in un mercato dinamico come quello dei beni di largo consumo, dove le preferenze dei consumatori e le condizioni di mercato possono cambiare rapidamente.

Esistono 2 tipologie di tecniche di apprendimento:

- Apprendimento supervisionato
- Apprendimento non supervisionato

Ciascuna di queste due metodologie presenta caratteristiche e applicazioni diverse dall’altra. La differenza sostanziale tra i due approcci risiede nel tipo di dati utilizzati e negli obiettivi perseguiti: l'apprendimento supervisionato richiede dati etichettati e si concentra sulla predizione accurata di output specifici, mentre l'apprendimento non supervisionato esplora i dati non etichettati per scoprire pattern nascosti e nuove conoscenze.

Figura 1: Composizione del machine learning



Tenure Track Assistant Professor @SPOCRI UNIMC

Dato l'obiettivo di questo studio e le caratteristiche del dataset a disposizione, l'approccio metodologico più appropriato e scientificamente valido è rappresentato dall'apprendimento supervisionato. In particolare, si considerano le tecniche di regressione come strumento primario, in quanto sono specificamente progettate per prevedere output continui, come i volumi di vendita. I regressori, grazie alla loro capacità di modellare e analizzare le relazioni tra variabili dipendenti e indipendenti, risultano particolarmente efficaci nel fornire previsioni accurate e affidabili.

La maggior parte delle sfide nel machine learning riguardano la gestione dei dati e la scelta del modello giusto. I dati possono presentarsi in diverse forme e dimensioni. I dataset del mondo reale possono essere disordinati, incompleti e in vari formati. Per far fronte alla prima sfida è necessario avere a propria disposizione conoscenze e strumenti specializzati al fine di essere in grado di pre-processare i dati. Ad esempio, per selezionare le caratteristiche necessarie ad addestrare un algoritmo di rilevamento oggetti, è necessario avere competenze specifiche nell'elaborazione delle immagini. Diversi tipi di dati richiedono approcci differenti per il pre-processamento.

Invece, per quanto riguarda la seconda sfida, trovare il modello (o algoritmo) migliore per adattarsi ai dati richiede tempo. La selezione del modello appropriato rappresenta un equilibrio complesso tra velocità, accuratezza e complessità computazionale. Non esiste un algoritmo che possa essere considerato universalmente superiore rispetto agli altri; la sua efficacia dipende strettamente dalle caratteristiche intrinseche dei dati utilizzati per la previsione. In altre parole, le performance dei

modelli sono in gran parte determinate dalla natura e dalla qualità dei dati che alimentano il processo predittivo. Ognuno ha punti di forza e limiti specifici, e la scelta ottimale deve essere orientata dalle peculiarità del dataset e dagli obiettivi analitici predefiniti.

La letteratura esistente sull'uso del Machine Learning per la previsione delle vendite è robusta e in rapida evoluzione. Attualmente, la ricerca ha esplorato una vasta gamma di algoritmi, incluse tecniche di regressione, alberi decisionali, metodi ensemble, reti neurali profonde, support vector machines e metodi non supervisionati. Questa diversità di approcci riflette un tentativo di adattare il modello migliore ai vari contesti specifici delle vendite, come il retail, l'e-commerce e le vendite stagionali.

Pedro Domingos, nel suo articolo "A Few Useful Things to Know About Machine Learning", discute vari principi fondamentali che devono essere considerati quando si lavora con il machine learning. Uno dei punti chiave del paper riguarda l'importanza della scelta dell'algoritmo corretto per un dato problema. In tale studio l’autore conferma quanto detto precedentemente. Infatti, sostiene che non esiste un modello che funzioni meglio per tutti i tipi di problemi perché la performance dipende fortemente dal tipo di dati e dal problema specifico che si sta cercando di risolvere. Pertanto, la scelta deve essere fatta in base alle caratteristiche specifiche dei dati e agli obiettivi del progetto. Inoltre, il paper raccomanda un approccio empirico durante la fase di selezione che implica di testare diversi algoritmi sui dati specifici e confrontare le loro prestazioni.

In particolare, i modelli di regressione presi in esame in questo studio sono stati selezionati seguendo le indicazioni di Pedro Domingos e comprendono una vasta gamma di tecniche avanzate: algoritmi basati sugli alberi decisionali (Decision Tree Regressor), metodi ensemble (Random Forest Regression e Extra Tree Regression), modelli lineari (Linear Regression), Gradient Boosting, Support Vector Machine, K-NN Regression, e Stepwise Regression.

Gli studi recenti, presenti in letteratura, sulle previsioni di vendita attraverso l’apprendimento automatico, tendono a integrare non solo dati storici di vendita, ma anche dati esterni come condizioni meteorologiche, festività, dati economici e social media analytics. Questa integrazione permette di catturare meglio le complessità e le variabili esterne che influenzano le vendite. La maggior parte delle ricerche si concentrano sull'aumento dell'accuratezza delle previsioni, dimostrando che i modelli di Machine Learning spesso superano i metodi statistici tradizionali, specialmente nella gestione di grandi quantità di dati e variabili complesse.

Emerson Martins & Napoleão Verardi Galegale nel 2023, attraverso la valutazione e comparazione dei vari metodi statistici presenti in letteratura, sono andati ad identificare quello che ci assicura la maggiore accuratezza nello svolgimento dell’attività di sales forecasting. Il confronto è stato fatto tra 3 principali metodi: i metodi di previsione tradizionali (che si basano sulle serie temporali); le reti neurali artificiali e gli algoritmi di apprendimento supervisionato. Il risultato di questo studio mostra come gli algoritmi di apprendimento supervisionato rappresentano la miglior metodologia da utilizzare. Come anticipato precedentemente quando si guarda nel lungo periodo, svolgono un ruolo fondamentale nella stima di tutti i fenomeni esterni difficilmente prevedibili. La caratteristica che rende questi modelli più efficienti rispetto agli altri metodi, è la loro capacità di comprendere le logiche operative sia delle variabili endogene che di quelle esogene. Pertanto, il ML offre una comprensione più approfondita dei trend di mercato e una maggiore accuratezza nelle previsioni finali.

Tale miglioramento delle previsioni è evidente anche nell’industria del FMCG. Jakob Huber e Heiner Stuckenschmidt, nel loro studio pubblicato sull'*International Journal of Forecasting (2020),* dimostrano attraverso prove scientifiche ed empiriche che i metodi di apprendimento automatico non solo offrono previsioni più accurate, ma si rivelano anche particolarmente adatti per applicazioni su larga scala (inclusa la vendita al dettaglio).

Stuti Raizada & Jatinderkumar R. Saini nel 2021 hanno utilizzato le tecniche di machine learning per stimare le possibili vendite di 45 Walmart store allocati in differenti posizioni geografiche. Lo studio dimostra che anche i fattori esterni, come il tasso di disoccupazione, il costo della benzina o la settimana di vacanza svolgono un ruolo fondamentale nella previsione delle vendite di qualsiasi negozio al dettaglio. A tal proposito, i modelli di regressione utilizzati dagli autori hanno mostrato come, con questo set di variabili, l’Extra Tree Regressor è in grado di svolgere previsioni di vendite con un’accuratezza superiore al 98%.

Stephen P. Ryan and Miaoyu Yang et al. nel 2015 hanno invece contribuito ad approfondire la conoscenza esistente entrando a fondo nella categoria del food retail e rispondendo alla necessità di prevedere le vendite di una catena di negozi alimentari attraverso le tecniche di apprendimento automatico. In linea con tali contributi, successivamente, E. Tarallo et al. (2019), hanno ulteriormente esplorato questa tematica, confermando come l'adozione del machine learning comporti vantaggi previsionali significativi nel segmento del food & beverage. Inoltre, la ricerca dimostra che le analisi complesse, come l'impatto delle promozioni, sono gestite in modo più

efficace dalle soluzioni di machine learning rispetto ai metodi tradizionali, migliorando l’accuratezza delle stime anche in mercati con domanda soggetta a minori interferenze esterne.

1.2 TRADE MARKETING IN BARILLA: LA STIMA DEI VOLUMI

Nell'organizzazione aziendale di Barilla, il Trade Marketing è l’unità responsabile anche della definizione dei volumi annuali di vendita e delle successive revisioni. In altre parole, si occupa di prevedere quanto l’azienda venderà in un anno e di revisionare periodicamente (ogni 4 mesi) le proprie stime, al fine di monitorare l'andamento delle vendite ed intervenire, se necessario, con azioni correttive. L'obiettivo è garantire il raggiungimento dei target annuali prefissati, monitorando costantemente le previsioni e i risultati effettivi, in modo da ottimizzare la performance commerciale dell'azienda.

La responsabilità di questa task è affidata al team di Trade Marketing in quanto fungendo da ponte tra il Marketing, responsabile delle strategie di marca, e il team di Vendite, che implementa le strategie commerciali, è in grado di sintetizzare tutte le informazioni disponibili e formulare previsioni accurate. Il Trade Marketing, inoltre, è una funzione che possiede una profonda conoscenza delle dinamiche di consumo, sia all'interno della categoria di prodotto che della rete distributiva. Questa expertise è fondamentale per la definizione dei volumi di vendita. Comprendere le tendenze di acquisto e i comportamenti dei consumatori, insieme alla conoscenza delle caratteristiche della rete distributiva, consente di avere la sensibilità necessaria per sviluppare stime più precise e realistiche.

In Barilla l’attività di forecast si concentra sui volumi di vendita al sell-in (in questo studio, ci occupiamo invece di stimare il sell-out).

1.2.1 DINAMICHE DI SELL-IN E SELL-OUT NEL FMCG

Il concetto di *sell-in* e *sell-out* è cruciale per comprendere le dinamiche di vendita e l'efficacia delle strategie commerciali. Sebbene entrambi i termini siano strettamente legati al ciclo di vita di un prodotto, si riferiscono a fasi diverse della catena di distribuzione.

Il *sell-in* fa riferimento al volume di prodotti venduti dall'azienda produttrice ai suoi clienti diretti (tipicamente distributori, grossisti o rivenditori). In altre parole, è la quantità di merce che esce dai magazzini del produttore ed entra nei canali di distribuzione dei clienti. Questa fase offre informazioni fondamentali per il produttore, in quanto riflettendo l'accettazione del prodotto da parte dei partner commerciali fornisce una misura preliminare del successo commerciale. Un alto volume di *sell-in* suggerisce che il prodotto ha ottenuto un buon livello di accettazione e interesse da parte dei distributori e rivenditori. Ciò può essere il risultato di strategie di marketing efficaci, buone relazioni commerciali o una percezione positiva del prodotto. Inoltre, questo livello della catena distributiva è fondamentale per pianificare e gestire la produzione e la logistica, in quanto permette di coordinare le scorte e le consegne per soddisfare la domanda. Tuttavia, un forte sell-in non garantisce una visione completa del successo del prodotto, poiché non tiene conto del comportamento del consumatore finale.

Il *sell-out*, invece, si riferisce al volume di prodotti effettivamente venduti dai retailers al consumatore finale. Questo ci consente di valutare la domanda reale del mercato e la capacità del prodotto di attrarre i consumatori. Il *sell-out*, quindi, offre una visione più accurata dell'efficacia delle strategie di marketing, promozionali e di pricing, e fornisce una visione più completa e reale delle prestazioni del prodotto sul mercato.

La comprensione completa della performance di un prodotto richiede un'analisi integrata del *sell-in* e del *sell-out*. Sebbene il *sell-in* fornisca una misura dell'accettazione da parte dei distributori, è il *sell-out* che rivela come il prodotto stia effettivamente performando sul mercato finale. Se il *sell-in* è elevato ma il *sell-out* è basso, può indicare problemi come scelte di prodotto non allineate con le preferenze dei consumatori, inefficacia delle strategie di promozione o problemi di visibilità nei punti vendita. Al contrario, un basso *sell-in* ma un alto *sell-out* potrebbe indicare un problema di scorte insufficienti, con il rischio di perdere opportunità di vendita.

In questo contesto, la situazione più favorevole per un'azienda produttrice si concretizza quando c'è un allineamento ottimale tra le due fasi della catena di distribuzione. Questo equilibrio è cruciale per garantire un funzionamento efficiente dell'intera filiera. Quando i volumi di prodotto venduti al *sell-in* sono in sintonia con le vendite effettive al consumatore finale (*sell-out*), l'azienda è in grado, in primis, di ridurre significativamente i rischi legati all'overstocking (accumulo eccessivo di inventario) e stock-out (mancanza di prodotti disponibili). In secundis, l’azienda produttrice riesce a consolidare le relazioni commerciali (poiché garantisce una fornitura costante e adeguata ai punti vendita), a migliorare la redditività, ed a minimizzare le speculazioni dei retailers.

La speculazione, nelle relazioni tra produttori e rivenditori, si manifesta quando i retailers acquistano volumi maggiori di prodotto rispetto a quanto prevedano effettivamente di vendere al consumatore finale. Questa pratica è particolarmente diffusa nel settore dei beni di largo consumo, dove ad esempio i rivenditori trovano vantaggioso approfittare delle promozioni sul sell-in per acquistare grandi quantità di prodotto a prezzi scontati. L'obiettivo è accumulare scorte che potranno poi essere vendute successivamente a prezzi più elevati, una volta terminato il periodo di promozione al sell-out. In questo modo, i rivenditori riescono a massimizzare i margini di profitto, sfruttando le differenze di prezzo tra il periodo promozionale e il periodo successivo.

1.2.2 METODOLOGIA DI STIMA

L’attività di forecast in Barilla si compone di 4 momenti diversi, riportati di seguito nella loro nomenclatura aziendale:

1. *Budget forecast*: circa tre mesi prima della fine dell’anno, nel mese di ottobre, inizia il processo di pianificazione della domanda per l'anno successivo. Il budget forecast rappresenta il primo passo nella definizione del budget annuale e richiede una stretta collaborazione del Trade Marketing con i team di Vendite e Demand Planning. L'obiettivo è stimare con precisione la domanda totale dell'anno successivo, tenendo conto delle previsioni di chiusura dell'anno corrente e delle proiezioni di crescita o contrazione del mercato. Questo processo di pianificazione anticipata è fondamentale per garantire che l'azienda possa prepararsi adeguatamente alle sfide e alle opportunità del nuovo anno, allineando le risorse e le strategie alle aspettative di mercato.
2. *Rolling forecast*: si svolge tipicamente a febbraio, e rappresenta la prima revisione critica del nuovo anno. In questa fase, l'attenzione si concentra sull'analisi dei primi dati disponibili per confrontarli con le aspettative delineate nella pianificazione iniziale. L'importanza del *rolling forecast* risiede nella sua capacità di far mantenere il controllo sulle performance aziendali sin dalle prime fasi dell'anno. Attraverso questa revisione tempestiva e accurata, si ha l'opportunità di anticipare eventuali problematiche e di agire proattivamente (con manovre che possono includere l'ottimizzazione delle attività promozionali, l'adeguamento delle strategie di pricing o una rivisitazione delle priorità distributive), invece di aspettare che i risultati finali si discostino significativamente dagli obiettivi.

3. *Mid-year forecast*: è la revisione di metà anno, attraverso cui vengono valutati eventuali scostamenti dall'obiettivo di budget e, soprattutto, vengono identificate le aree su cui intervenire per migliorare. A metà anno, grazie allo storico già consolidato, c’è una base solida su cui fondare le analisi e le eventuali azioni correttive. Questo è il momento ideale per implementare piani di contingency (se necessario) al fine di garantire che l’obiettivo annuale sia ancora raggiungibile. La *mid-year forecast*, dunque, rappresenta un’opportunità strategica per affinare le strategie e massimizzare le performance nella seconda metà dell’anno.
4. *Last estimate (LE)*: rappresenta l'ultima revisione del budget, solitamente condotta nel mese di settembre. In questa fase ci avviciniamo alla fine dell'anno e si ha ormai una chiara visione delle prestazioni realizzate fino a quel punto. *LE* ci permette di fare un bilancio accurato rispetto agli obiettivi annuali, valutando se è necessario mettere in atto le ultime azioni migliorative. Inoltre, questa revisione è di cruciale importanza anche per la definizione del nuovo *budget*. Infatti, il *budget* per l'anno successivo, viene elaborato sulla base delle previsioni di chiusura dell'anno corrente, quindi, utilizzando le stime aggiornate di *LE*. Questo processo ci permette di costruire su una base informata e realistica la pianificazione del nuovo anno. D'altro canto, tale approccio comporta un certo grado di incertezza, poiché il *budget* per l'anno successivo si basa su una stima di chiusura dell'anno corrente che potrebbe non riflettere pienamente l'andamento reale degli ultimi mesi. In questa fase l'accuratezza delle previsioni diventa cruciale: per minimizzare il margine di errore, sarebbe opportuno disporre di una stima precisa del *sell-in* e del *sell-out* per i mesi finali dell'anno.

Durante ogni fase del processo di previsione, è essenziale una stretta collaborazione tra il Trade Marketing ed i team di Vendite e Demand Planning per ottenere stime più accurate.

Il Demand Planning, ha il compito di pianificare la domanda e fornire agli stabilimenti produttivi una previsione settimanale dei volumi da caricare e spedire. Nonostante si tratti di una funzione tradizionalmente legata alla supply chain, nella *Region Italy* di Barilla, ha assunto un ruolo più commerciale a causa della natura altamente frammentata della grande distribuzione italiana, che richiede una continua interazione della supply chain con i clienti diretti dell’azienda.

In mercati come Francia, Germania e Spagna, dove la distribuzione è centralizzata nelle mani di pochi grandi retailer, il Demand Planning non necessita di un'integrazione così stretta con il team commerciale. In Italia, invece, questa sinergia è indispensabile.

Durante l'attività di pianificazione, il Demand Planning è responsabile delle previsioni dei primi tre mesi. Nel breve periodo ha piena visibilità delle promozioni che sono state pianificate in maniera dettagliata. Perciò, grazie a tecniche e capacità previsionali avanzate, è in grado di fornire stime realistiche per questo periodo. Nel medio e lungo termine, invece, non possiede una comprensione approfondita delle logiche commerciali, in quanto, non dispone di informazioni dettagliate sui piani promozionali dei clienti.

A questo punto, l'analisi previsionale passa al Trade Marketing, il quale, essendo responsabile dell'allocazione degli investimenti promozionali e dello sviluppo di attività out-of-shelf (che impattano direttamente i volumi di vendita al sell-out), ha una visione più strategica ed integrata delle attività pianificate nel BOY (Balance Of the Year).

Per la stima del medio/lungo periodo, è fondamentale la collaborazione con il key account management, al fine di raccogliere informazioni dettagliate relative agli accordi commerciali presi con i clienti. In questa fase, ci si concentra principalmente sui grandi eventi e sulle grandi iniziative promozionali ricorrenti che vengono svolte annualmente con i clienti chiave. Si inizia da un'analisi dettagliata dello storico (le performance passate del prodotto) e, sulla base di questo, si sviluppa una proiezione delle vendite future, che prende in considerazione le nuove condizioni commerciali pianificate per l'anno in corso e gli ultimi trend del mercato. Ad esempio, un trend significativo degli ultimi anni è la crescita di quota dei discounts (in cui Barilla ha una presenza meno consolidata). Di conseguenza, ogni incremento della market share dei discounts rappresenta un potenziale rischio di perdita di volumi che il trade marketeer deve considerare quando elabora la propria stima.

1.2.2 LIMITI DEL PROCESSO

La maggiore difficoltà di questa attività di forecast è capire la reale resa promozionale al *sell-out*. Il *sell-out* è influenzato da tanti fattori che incidono direttamente sulle performance annuali di un prodotto.

Tra questi ci sono certamente alcuni fattori esterni, come le condizioni climatiche, che possono avere un impatto significativo sulle vendite. Per esempio, durante stagioni particolarmente calde, i consumi di merendine possono diminuire a favore di gelati. Oppure, se settembre e ottobre sono estremamente caldi e i consumatori continuano a frequentare le spiagge, le vendite di prodotti

destinati al consumo domestico (come la pasta) possono essere inferiori alle previsioni. Inoltre, anche fattori economici più ampi, come l’inflazione o una crisi economica, influenzano notevolmente il comportamento dei consumatori.

Un altro rischio fondamentale è poi dettato dal fatto che le performance di un prodotto possono essere significativamente influenzate dalle precedenti attività effettuate dai concorrenti. Ad esempio, implementare una promozione che segue un’importante iniziativa promozionale di un concorrente forte, può risultare meno efficace, poiché i consumatori potrebbero aver già accumulato grandi quantità di pasta, riducendo così l'impatto della promo successiva.

Il processo di definizione delle vendite al *sell-in*, è caratterizzato da una relazione diretta con le stime dei volumi che saranno venduti ai consumatori finali (*sell-out*). L'elevata aleatorietà delle dinamiche di *sell-out* sottolinea la necessità di adottare modelli previsionali avanzati, capaci di analizzare i dati storici di vendita al consumatore finale. Questi modelli, attraverso algoritmi sofisticati, sono in grado di identificare pattern e tendenze, permettendo previsioni più accurate delle performance future. Un approccio basato sui dati storici consente di ridurre l'incertezza e migliorare l'affidabilità delle previsioni. Quindi, la capacità di prevedere con precisione il *sell-out* non solo andrebbe a colmare un gap significativo, ma consentirebbe anche di ottimizzare le strategie commerciali, di marketing e la gestione dell’inventario.

1.3 OBIETTIVI DI RICERCA

Nel contesto attuale delle previsioni di vendita nel settore dei beni di largo consumo (FMCG), si rileva un significativo gap di ricerca riguardante l'integrazione delle attività di trade marketing come variabili predittive fondamentali nei modelli di machine learning. La maggior parte degli studi accademici si concentrano su modelli predittivi basati su dati storici di vendita, tendenze di mercato generali e fattori esterni, come le condizioni economiche o le promozioni stagionali. Tuttavia, c'è una carenza significativa di ricerche che esplorano in profondità come le attività di trade marketing possano essere utilizzate per migliorare l'accuratezza e la robustezza delle stime.

La disciplina del Trade Marketing, gioca un ruolo determinante sul comportamento d'acquisto degli shopper e la mancanza di studi specifici su come queste variabili possano essere utilizzate per affinare le previsioni delle vendite rappresenta una lacuna significativa.

Inoltre, come spiegato nel paragrafo 1.2 (*Il Trade Marketing in Barilla: La stima dei volumi*) l'azienda leader del mercato della pasta in Italia (Barilla) si impegna a stimare i volumi di vendita al fine di adeguare ed ottimizzare i propri processi produttivi e commerciali. In tal caso però, viene elaborata una previsione al *sell-in*, in quanto, in effetti, il flusso di cassa positivo per un produttore come la Barilla è generato dalla vendita ai retailers e non ai consumatori finali.

Questo studio si propone invece di sviluppare un modello di previsione avanzato che permetta alle aziende di ottenere stime precise dei volumi futuri di *sell-out*. In questo modo si va a colmare un gap cruciale, fornendo uno strumento che integra e ottimizza le attività annuali di previsione del *sell-in*. Con i risultati, le aziende potranno migliorare la loro capacità di pianificazione, assicurando stime più accurate e allineate con la domanda effettiva del mercato.

L'obiettivo di ricerca di questa tesi viene perseguito ampliando la conoscenza esistente sui metodi di previsione dei volumi di vendita dei beni di largo consumo, andando ad integrare al machine learning le attività di trade marketing. In particolare, al fine di migliorare la precisione del processo predittivo, viene utilizzato un dataset composto da informazioni di in-store marketing. Attraverso una rigorosa revisione della letteratura e l'analisi critica delle pratiche esistenti, la ricerca intende colmare le attuali lacune conoscitive presenti e proporre nuove prospettive per l'integrazione di questi due ambiti, fornendo così un contributo significativo alla teoria e alla pratica del marketing e della gestione delle vendite.

Il raggiungimento dell'obiettivo permette alle aziende di identificare con precisione l'algoritmo predittivo ottimale da utilizzare quando si analizzano dati con caratteristiche simili a quelle del dataset di questo studio. Inoltre, la metodologia adottata per l'estrazione e l'analisi è altamente scalabile e consente una facile replicabilità del processo da parte di altre organizzazioni.

In questo caso affrontiamo come caso studio il mercato della pasta, ma il processo di predizione è facilmente condivisibile con altri mercati e categorie del FMCG. Come è possibile? Perché il dataset impiegato per l'addestramento degli algoritmi applicati è stato estratto da Nielsen IQ, un'azienda di riferimento nel settore dell'intelligence dei dati di mercato. NIQ traccia tutte le informazioni dei punti vendita utilizzando una combinazione di metodologie avanzate, tecnologie sofisticate e collaborazioni strategiche con rivenditori. In questo modo Nielsen riesce ad avere una conoscenza a 360° sull'industria dei beni di largo consumo. Conoscenza che poi rivende ai suoi clienti sottoforma di dati di mercato tramite le sue piattaforme online. Le informazioni utilizzate in

questo studio, quindi, possono essere facilmente estratte anche dagli altri produttori che operano nel settore del fast moving consumer goods.

La collaborazione con Nielsen IQ ci consente di sfruttare un dataset ricco e dettagliato e di analizzare le dinamiche del mercato della pasta con una profondità senza precedenti. In particolare, il dataset comprende dati di mercato dal 2022 ad oggi.

L'analisi si focalizzerà esclusivamente sui Key Performance Indicators (KPI) che definiscono le attività in-store implementate dai principali player del mercato della pasta in Italia:

1. Distribuzione ponderata
2. Numero medio di referenze per punto di vendita
3. Numero medio punti di vendita vendenti ad un prezzo non promozionato
4. P.E.I. (indicatore che misura l’efficacia delle promozioni implementate)
5. Sconto % medio
6. Prezzo medio per confezione di prodotto
7. Numero volantini
8. % inserimenti in volantini “Sottocosto”
9. % inserimenti in volantini “Sconto – quantità”
10. % inserimenti in volantini “Taglio prezzo”
11. % inserimenti in volantini “Concorso”

Nel contesto del trade marketing, questo studio rappresenta un avanzamento significativo nella comprensione dell'impatto che hanno le risorse allocate negli store. Attraverso l'implementazione e la validazione di modelli predittivi che integrano queste variabili, le aziende possono acquisire una visione più dettagliata e accurata delle dinamiche commerciali direttamente legate agli sforzi di trade marketing. Questo approccio consente di identificare con precisione quali risorse e iniziative hanno il maggiore impatto sulle vendite.

La ricerca contribuisce a migliorare la precisione delle previsioni ed a far sì che le imprese possano assicurarsi di avere sempre a portata di mano informazioni critiche necessarie per l’ottimizzazione delle strategie commerciali e dei piani di marketing che gli consentono di costruire una crescita sostenibile a lungo termine. Infine, l'integrazione delle tecniche di machine learning e dei big data analytics nei processi di trade marketing rappresenterebbe una svolta innovativa, che permette di superare le tradizionali limitazioni delle analisi di mercato. Le aziende possono così ottenere un

vantaggio competitivo significativo, posizionandosi come leader nel settore grazie alla capacità di prendere decisioni informate e strategiche basate su dati approfonditi e analisi predittive.

Come riportato nella Figura 1, le ipotesi che delineano il percorso di ricerca sono 4.

Figura 2: Ipotesi di ricerca e la letteratura a supporto

Ipotesi di ricerca	Letteratura a supporto
H1: Le attività di Trade Marketing hanno un impatto significativo sulle vendite di beni di largo consumo.	Fornari, D. (2009). Trade marketing: Relazioni di filiera e strategie commerciali. Egea.
H2: Le tecniche di machine learning possono ottimizzare significativamente l'impatto delle attività di trade marketing sulle vendite, migliorando la previsione della domanda e l'efficacia delle strategie promozionali.	Tarallo, E., Akabane, G. K., Shimabukuro, C. I., Mello, J., & Amancio, D. (2019). Machine learning in predicting demand for fast-moving consumer goods: An exploratory research. IFAC-PapersOnLine, 52(13), 737-742.
H3: I modelli di Machine Learning superano i modelli tradizionali in termini di accuratezza predittiva delle vendite.	Emerson, M., Napoleão, V. G. (2023). Sales forecasting using machine learning algorithms. Revista Gestão e Secretariado (GeSec), São Paulo, SP, v. 14, n. 7, 2023, p. 11294-11308.
H4: Un modello previsionale che stima con elevata accuratezza le performance dei prodotti al sell-out rappresenta un asset fondamentale che consente alle aziende di ottimizzare le proprie decisioni strategiche e operative.	Vishal Das, Tianyi Mao, Zhicheng Geng, Carmen Flores, Diego Pelloso, Fang Wang (2024). Enhancing sell-in and sell-out forecasting using ensemble machine learning method.

La prima ipotesi (H1) postula che le attività di trade marketing influenzino in modo significativo le vendite di prodotti FMCG. A supporto di questa ipotesi, Fornari, Daniele (2009) nel libro "Trade Marketing (Relazioni di filiera e strategie commerciali)" approfondisce le relazioni di filiera e le

strategie commerciali, sottolineando come il trade marketing possa migliorare le performance di vendita.

La seconda ipotesi (H2), invece, suggerisce che l'uso di tecniche di machine learning può ottimizzare l'impatto delle attività di trade marketing sulle vendite. In altre parole, l'integrazione di machine learning nelle strategie di trade marketing può migliorare la loro efficacia e precisione. Tarallo, Akabane, Shimabukuro, Mello e Amancio (2019), in "Machine learning in predicting demand for fast-moving consumer goods: An exploratory research", dimostrano chiaramente come l'uso del machine learning migliorando la precisione nelle previsioni della domanda, fornisce ai produttori un vantaggio competitivo significativo, consentendo loro di ottimizzare i piani di vendita e le attività commerciali, inclusa la pianificazione delle promozioni e le strategie di visibilità del prodotto.

La terza ipotesi (H3) afferma che i modelli di machine learning che incorporano variabili di trade marketing offrono previsioni più accurate rispetto ai modelli tradizionali che si basano esclusivamente sui dati storici di vendita. Questo implica che l'inclusione di variabili specifiche del trade marketing nei modelli predittivi migliora significativamente la loro accuratezza. La letteratura a supporto include: Emerson Martins & Napoleão Verardi Galegale (2023) con l'articolo "*Sales forecasting using machine learning algorithms*" pubblicato sul *Journal of Forecasting*. Questo studio mostra come l'utilizzo di algoritmi di machine learning per le previsioni di vendita, che includono variabili di trade marketing, possa superare i modelli tradizionali in termini di accuratezza predittiva.

Infine, la quarta ipotesi (H4) sostiene che un modello previsionale in grado di stimare con elevata accuratezza le performance dei prodotti al sell-out rappresenta un asset fondamentale per le aziende, poiché, oltre che essere un’informazione cruciale per supportare il processo previsionale dei volumi di vendita al sell-in, permette di ottimizzare anche le strategie commerciali, di marketing e la gestione dell’inventario. Questo implica che l’accuratezza nelle previsioni di sell-out consente di migliorare la gestione complessiva delle attività aziendali, influenzando positivamente direttamente anche la redditività. La letteratura di riferimento include lo studio di Vishal Das et al. (2021) "*Enhancing sell-in and sell-out forecasting using ensemble machine learning methods*" pubblicato su "*Journal of Business Research*". Il loro studio evidenzia come l'integrazione di diverse fonti di dati e tecniche avanzate di machine learning possa ottimizzare le previsioni di vendita (al sell-in e al sell-out), migliorando la capacità delle aziende di pianificare e gestire le scorte, nonché di ottimizzare le strategie di marketing e distribuzione.

1.4 STRUTTURA DELLA TESI

La tesi esamina in dettaglio lo stato attuale delle conoscenze in letteratura sull'integrazione della disciplina del trade marketing alle tecniche di machine learning per prevedere le vendite delle aziende operanti nel settore dei beni di largo consumo (FMCG), ed identifica le aree ancora inesplorate e le opportunità di innovazione. La ricerca combina un'analisi approfondita della letteratura con un'analisi statistica e quantitativa, al fine di sviluppare proposte concrete su un fenomeno rilevante nell'organizzazione delle attività di marketing e vendita.

L'obiettivo è quello di fornire una comprensione più completa di come le metodologie avanzate di machine learning possano essere applicate alle strategie di trade marketing, migliorando così la precisione delle previsioni di vendita e ottimizzando le performance commerciali nel settore FMCG.

Questo capitolo funge da introduzione all'approccio di ricerca e da schema della struttura della tesi che ne consegue. L'approccio di ricerca si basa su tre attività chiave:

- Valutazione della letteratura

La revisione della letteratura è suddivisa in 2 parti. La prima parte esamina gli approcci tradizionali alla previsione delle vendite e perché è sorta negli anni l'esigenza di passare all'applicazione dei modelli di machine learning. La seconda parte invece esplora cos'è il Trade Marketing e perché l'integrazione di questa disciplina può migliorare l'accuratezza delle previsioni.

- Ricerca empirica quantitativa

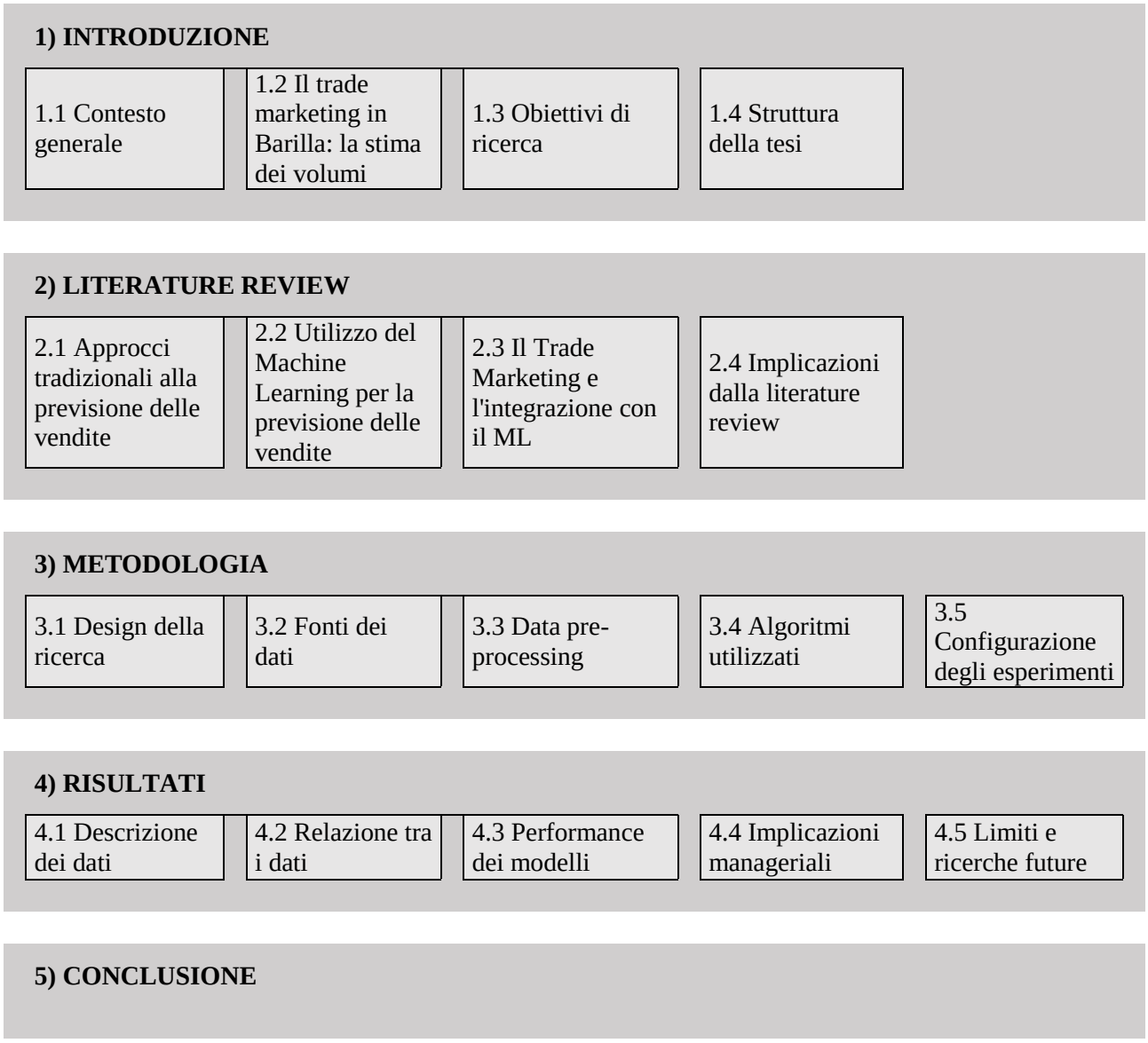
La ricerca quantitativa vuole descrivere l'approccio metodologico di ricerca condotto per prevedere le vendite dell'azienda *"Barilla G. e R. Fratelli S.p.A"* utilizzando modelli di machine learning basati sui dati di trade marketing. L'approccio viene descritto in tutte le sue diverse fasi chiave: raccolta di dati, pre-processing dei dati, feature selection, scelta degli algoritmi da applicare, definizione del modello più performante.

- Sviluppo degli insights strategici

Nell'ultima fase, viene effettuata un'analisi approfondita e una diagnosi delle informazioni emerse dalla ricerca. Gli insights strategici sono sviluppati concentrando l'attenzione sulle opportunità di

crescita e di vantaggio competitivo. Questo processo mira a trasformare i dati grezzi in conoscenze strategiche, offrendo alle aziende una guida precisa per ottimizzare le loro operazioni di marketing e migliorare le performance di vendita. Il risultato finale è quello di fornire un framework chiaro e dettagliato in grado di supportare le decisioni aziendali con evidenze empiriche solide e promuovendo un approccio innovativo e data-driven al trade marketing.

Figura 3: Overview della struttura della tesi



La struttura a capitoli della tesi rispecchia l'approccio di ricerca (*figura 3*). I capitoli della tesi sono strutturati, oltre l'introduzione, in altre quattro parti principali. La prima parte tratta della literature

review (capitolo 2). La seconda parte descrive la metodologia di ricerca orientata alla scoperta e i metodi empirici utilizzati. (capitolo 3). La terza parte infine descrive i risultati della ricerca e le vie per gli sviluppi futuri (capitolo 4) e conclude con le conseguenze pratiche e teoriche che derivano dai risultati dello studio (capitolo 5).

2 REVISIONE DELLA LETTERATURA

La revisione della letteratura offre un'analisi approfondita dei metodi di previsione delle vendite, partendo dalle tecniche tradizionali (capitolo 2.1) e mettendo in luce le ragioni che hanno portato all'adozione dei modelli di machine learning (capitolo 2.2). Infine, viene esaminato il ruolo del trade marketing e il suo impatto sull'affinamento delle previsioni di vendita, rendendo i modelli di apprendimento automatico più robusti e adattati alle reali dinamiche di mercato (capitolo 2.3).

2.1 APPROCCI TRADIZIONALI ALLA PREVISIONE DELLE VENDITE

La revisione della letteratura offre un'analisi approfondita dei metodi di previsione delle vendite, partendo dalle tecniche tradizionali e mettendo in luce le ragioni che hanno portato all'adozione dei modelli di machine learning. I metodi di previsione tradizionali si basano su serie temporali, ciò significa che vengono applicati nell'ipotesi che la domanda passata possa stimare statisticamente la domanda futura. Questo approccio si compone di un'analisi dei dati storici al fine di identificare pattern e tendenze che possono essere proiettati nel futuro. In genere, questi metodi sono facili da applicare e funzionano bene nei mercati in cui la domanda è per lo più stabile ed indicativa delle tendenze future (CHOPRA et al, 2013).

La letteratura sui metodi tradizionali di previsione delle vendite basati sulle serie temporali è ampia e consolidata. Makridakis e Hibon (2000), hanno evidenziato l'efficacia dei metodi di serie temporali, come ARIMA, nel fornire previsioni accurate in vari contesti aziendali. Precedentemente, Box e Jenkins (1970) hanno introdotto una metodologia sistematica per l'analisi delle serie temporali, noto come Box-Jenkins ARIMA, che è diventata una pietra miliare nella previsione statistica. Taylor (2003) ha dimostrato l'efficacia del livellamento esponenziale stagionale per la previsione della domanda in settori con forti componenti stagionali, metodo che può essere applicato efficacemente anche alle vendite al dettaglio. Infine, Fildes e Goodwin (2007) hanno discusso l'applicazione pratica delle tecniche di previsione tradizionali in aziende reali, sottolineando come metodi come ARIMA e il livellamento esponenziale possano essere utilizzati per ottenere previsioni di vendita affidabili, ma anche come questi necessitino di un'attenta gestione per evitare errori di giudizio.

Nonostante la loro utilità, i metodi tradizionali di previsione delle vendite presentano significative limitazioni, soprattutto in contesti caratterizzati da alta volatilità e complessità. Questi metodi tendono a basarsi su assunti di linearità e stabilità, ovvero presuppongono che il comportamento passato delle vendite possa essere utilizzato per prevedere in modo accurato il loro andamento futuro. Tuttavia, questa visione risulta spesso riduttiva, in quanto non tiene conto della natura dinamica e interconnessa dei mercati moderni. In ambienti in cui le condizioni possono cambiare rapidamente e in modo imprevedibile, come nel caso di crisi economiche, innovazioni tecnologiche o cambiamenti normativi, i modelli basati su serie storiche potrebbero non essere in grado di fornire previsioni accurate. I modelli di serie temporali, cercano di identificare pattern ricorrenti nei dati storici, come trend o stagionalità, per estrapolarli nel futuro. Tuttavia, spesso falliscono in quanto non riescono a catturare eventi imprevisti o cambiamenti repentini nelle condizioni di mercato (Chatfield, 2001). Un esempio evidente di questa limitazione è rappresentato dai servizi di trasporto on-demand, come Uber. La domanda per questi servizi non può essere stimata in modo accurato basandosi esclusivamente su dati storici. Fattori come le condizioni meteorologiche, l'ora del giorno, il giorno della settimana o persino eventi locali possono influenzare significativamente la richiesta di corse, rendendo i modelli tradizionali inadeguati per cogliere la complessità del mercato (Ke et al., 2017).

Per affrontare queste sfide, che si trovano nelle tecniche tradizionali, sono stati sviluppati modelli di previsione più avanzati, noti come modelli causali. Questi si distinguono per la loro capacità di includere variabili esogene, ossia fattori esterni che possono influenzare le vendite, come le condizioni macroeconomiche, le condizioni meteorologiche, le strategie di marketing o altre dinamiche del mercato (CHOPRA et al, 2013). A differenza dei modelli basati su serie storiche, le tecniche previsionali avanzate non si limitano a osservare il passato, ma cercano di comprendere e quantificare le relazioni tra diversi fattori e il loro impatto sulle vendite future. L'integrazione di modelli causali, che tengono conto di variabili esogene, rappresenta un passo avanti significativo, offrendo alle aziende uno strumento più flessibile e accurato per affrontare le sfide dei mercati moderni.

2.2 UTILIZZO DEL MACHINE LEARNING PER LA PREVISIONE DELLE VENDITE

In questo senso, il ML stesso potrebbe essere considerato come un fornitore di modellazione causale, che si distingue per la sua capacità di elaborare previsioni più accurate gestendo e analizzando enormi quantità di dati complessi. In contesti in cui i metodi tradizionali di previsione possono risultare inadeguati, il ML consente di identificare e generalizzare pattern complessi che altrimenti potrebbero sfuggire a una semplice analisi statistica. Grazie a tecniche avanzate come le reti neurali, gli alberi decisionali e i modelli di apprendimento profondo, gli algoritmi di ML sono in grado di migliorare progressivamente la loro capacità predittiva, adattandosi ai cambiamenti ambientali e cogliendo le sottili variazioni nei dati.

Un altro vantaggio chiave di varie tecniche di apprendimento automatico risiede nella loro natura non parametrica. Ciò implica che, a differenza degli approcci statistici classici, come la regressione lineare, non richiedono che le caratteristiche del set di dati siano normalmente distribuite (Kolkman & Van Witteloostuijn, 2019; Van Liebergen, 2017). Questo aspetto è particolarmente rilevante perché, i dati raccolti nel mondo reale raramente seguono distribuzioni perfettamente normali. Le tecniche di ML sono progettate per lavorare efficacemente con dati eterogenei e non strutturati, permettendo agli algoritmi di adattarsi a una vasta gamma di situazioni senza la necessità di trasformazioni complesse dei dati per conformarsi ai requisiti statistici. Questa flessibilità consente agli algoritmi di apprendere, adattarsi e, nel processo, scoprire sottili intuizioni nei dati (Leo et al., 2019).

Secondo Ryll e Seidens (2019), il concetto di apprendimento automatico, nonostante la sua crescente popolarità, rimane mal definito nella letteratura esistente. Gli autori lo definiscono come un processo attraverso il quale un sistema interagisce con il suo ambiente in modo tale che la struttura del sistema cambi e, a causa di alterazioni strutturali, cambi anche il processo di interazione. Shalev-Shwartz e Ben-David (2014) affermano che l'apprendimento automatico è il rilevamento di modelli di dati significativi da parte di algoritmi in modo automatizzato, indicando essenzialmente che le tecniche di apprendimento automatico dotano i programmi della capacità di "imparare" e adattarsi di conseguenza. Questa concezione è in linea con quella di Goodfellow et al. (2016), che definiscono l'apprendimento automatico come la capacità dei sistemi di intelligenza artificiale (AI) di acquisire conoscenza raccogliendo modelli da set di dati grezzi. Lantz (2019) concepisce l'apprendimento automatico come una tecnologia che elabora e trasforma i dati in

informazioni fruibili. Mohammed et al. (2016) descrivono l'apprendimento automatico come l'abilitazione delle macchine ad apprendere senza programmazione esplicita.

L'apprendimento automatico è stato ampiamente adottato per una varietà di scopi, tra cui la modellazione finanziaria, l'analisi della salute e della sicurezza, la diagnosi medica e il rilevamento delle frodi (Crane-Droesch, 2017; Enkono & Suresh, 2020; Gholizadeh et al., 2018; Mohammed et al., 2016). Le tecniche di apprendimento automatico sono state adottate anche per prevedere la domanda di mercato, i volumi di vendita e il comportamento dei consumatori (Bajari et al., 2015; Sekban, 2019; Tsoumakas, 2019; Venishetty, 2019). Una delle principali aree di utilizzo del machine learning è la modellazione predittiva delle performance di vendita (Sekban, 2019; Tsoumakas, 2019; Venishetty, 2019).

Negli ultimi anni, i modelli di machine learning hanno guadagnato terreno come strumenti potenti per la previsione delle vendite e la letteratura presenta numerosi paper scientifici sulla tematica. Raizada e Saini (2021) nel loro studio dimostrano come algoritmi di apprendimento automatico come la regressione lineare, le foreste casuali e le reti neurali superino le performance dei modelli tradizionali in termini di accuratezza e adattabilità. G. Sujatha e R. Sangeetha (2021) hanno esplorato l'applicazione del machine learning anche nel contesto delle promozioni di vendita nei supermercati, evidenziando come tecniche come gli alberi decisionali e KNN possano prevedere con precisione l'impatto delle promozioni sulle vendite. Questo studio mostra la capacità di integrare variabili esogene e di migliorare le decisioni strategiche di marketing.

Ancora, un esempio applicativo rilevante è fornito da Rohaan et al. (2019), che illustrano come i modelli di machine learning possano gestire efficacemente la variabilità e le complessità delle vendite B2B, dimostrando un miglioramento significativo nella precisione delle previsioni rispetto ai modelli tradizionali.

In "Machine Learning Methods for Data-Driven Demand Estimation and Assortment Planning Considering Cross Selling and Substitutions" Chen, Fan e Sun (2019) evidenziano l'abilità dei modelli di apprendimento automatico di adattarsi a complesse interazioni di mercato. In particolare, discutono come il machine learning possa essere utilizzato per la stima della domanda e la pianificazione dell'assortimento, tenendo conto di variabili di stampo strettamente commerciale come le vendite incrociate e le sostituzioni.

Emerson Martins e Napoleão Verardi Galegale (2019) nel loro lavoro "Sales forecasting using machine learning algorithms" sottolineano come le tecniche di apprendimento automatico, incluse le reti neurali, possano offrire previsioni di vendita precise, facilitando decisioni strategiche aziendali. Altri studi analizzati per una completa revisione della letteratura esistente sono ad esempio quelli di Cheriyan et al. (2019) , P. Bajari et al. (2015) e Usuga Cadavid et al. (2020), che forniscono una panoramica delle tendenze attuali nell'applicazione del machine learning alle previsioni di vendita, dimostrando come queste tecniche stiano trasformando il panorama delle previsioni aziendali con maggiore precisione e flessibilità.

Infine, Helper Zhou e Victor Gumbo (2020) esplorano l'efficacia di diversi algoritmi di apprendimento automatico per la previsione delle vendite nelle piccole e medie imprese, dimostrando quindi la scalabilità di questi modelli, che possono essere applicati anche in realtà più piccole purché venga fornito alla macchina un grande quantitativo di dati di input. Inoltre i due autori sottolineano l'importanza della scelta dell'algoritmo appropriato per ottenere risultati ottimali.

Pertanto, dopo un'analisi approfondita della letteratura scientifica, è emerso che esiste un ampio ventaglio di algoritmi di machine learning applicabili per la previsione delle vendite, ciascuno con specifiche caratteristiche e modalità operative. La performance di questi algoritmi è strettamente legata alle peculiarità del dataset su cui vengono applicati. Questo implica che la scelta dell'algoritmo appropriato è una decisione strategica di primaria importanza e deve essere fatta con attenzione.

2.2.1 GLI ALGORITMI UTILIZZATI

La selezione dell'algoritmo più idoneo può risultare complessa, considerando che il panorama degli algoritmi di apprendimento automatico comprende sia approcci supervisionati che non supervisionati, ognuno con metodologie diverse e punti di forza distintivi. Non esiste un algoritmo universale che possa garantire risultati ottimali in ogni contesto; piuttosto, la scelta deve essere basata su una valutazione dettagliata delle specifiche caratteristiche dei dati di input, inclusi la dimensione, la qualità, la distribuzione e la struttura dei dati.

Per esempio, dataset di grande dimensione e alta complessità spesso richiedono algoritmi avanzati come le Deep Neural Networks (DNNs)o le Support Vector Machines, che sono capaci di estrarre e modellare pattern complessi (Ian Goodfellow, Yoshua Bengio e Aaron Courville, 2016). Al

contrario, per dataset più piccoli o meno complessi, modelli più semplici come la Random Forest possono essere sufficienti (Aurélien Géron, 2019). Inoltre, la decisione di quale algoritmo utilizzare non dipende solo dalla dimensione e dalla complessità del dataset, ma deve considerare anche le informazioni che si intendono ottenere dai dati e come queste informazioni verranno impiegate per supportare le decisioni strategiche aziendali.

Ad ogni modo non c’è modo di garantire a priori quale algoritmo fornirà le migliori performance senza una fase di sperimentazione e ottimizzazione. Ogni scenario di previsione richiede una calibrazione e una validazione specifica per assicurare che l'algoritmo selezionato sia il più adatto a generare previsioni precise e utili per le esigenze aziendali. Questo processo di selezione e ottimizzazione è un processo di prove ed errori, ma è cruciale per massimizzare l'efficacia delle previsioni di vendita e, conseguentemente, migliorare la competitività e la sostenibilità delle strategie di business.

Data l’elevata importanza della scelta dell’algoritmo, il presente studio prende in analisi 8 regressori che secondo la letteratura possono essere considerati appropriati alle caratteristiche dei dati che vengono passati alla macchina. Di seguito vengono presentati i principi di funzionamento, i vantaggi e svantaggi degli algoritmi applicati per svolgere l’analisi predittiva.

Il Decision Tree Regressor è un algoritmo di apprendimento supervisionato utilizzato per la regressione, ossia per prevedere valori continui in base a un set di dati di input. Questo algoritmo crea un modello che suddivide il dataset in sottoinsiemi più piccoli, in base a variabili esplicative, fino a quando non si raggiungono criteri di arresto predeterminati, come la profondità massima dell'albero o il numero minimo di campioni per foglia. Durante la costruzione dell'albero, l'algoritmo seleziona la variabile e la soglia che meglio dividono i dati per minimizzare l'errore nella previsione dei valori target. Una volta costruito, l'albero può essere utilizzato per fare previsioni su nuovi dati, seguendo il percorso di decisione fino a una foglia, che contiene il valore previsto (Breiman, L., Friedman, J., Olshen, R., & Stone, 1984). Tra i suoi vantaggi ritroviamo la sua semplicità e facilità di interpretazione. Le decisioni prese in ogni nodo sono facili da comprendere e spiegare, il che rende l'algoritmo particolarmente utile in contesti dove l'interpretabilità è importante. Inoltre è capace di gestire sia variabili numeriche che categoriche senza richiedere alcuna trasformazione preliminare dei dati. Questo lo rende flessibile e applicabile in vari contesti. D’altro canto invece, uno dei suoi limiti, è l’overfitting. Questo significa che l'albero può adattarsi troppo strettamente ai dati di training, perdendo la capacità di generalizzare su nuovi dati. Inoltre, il Decision Tree Regressor è particolarmente sensibile a piccole variazioni nei dati. Un piccolo

cambiamento nei dati di training può portare a un albero completamente diverso. Questo problema può essere mitigato attraverso l'uso di metodi ensemble come Random Forest, che aggregano le previsioni di molti alberi per migliorare la robustezza e l'accuratezza del modello (Loh, 2011).

Il Random Forest Regression è una versione estesa del metodo degli alberi decisionali (Decision Trees) e si basa su una tecnica di ensemble che combina le previsioni di molti alberi decisionali per ottenere risultati più accurati e robusti. A differenza dei singoli alberi decisionali, oltre ad aver una maggiore accuratezza e robustezza, è anche meno incline all'overfitting, poiché aggrega le previsioni di molti alberi, ciascuno dei quali è addestrato su un diverso campione del dataset. Dall’altro lato però, un limite del Random Forest è la sua scarsa interpretabilità rispetto ai modelli di regressione più semplici come la regressione lineare o i singoli alberi decisionali. Le previsioni sono basate su una combinazione di molti alberi, rendendo difficile spiegare come una singola decisione è stata presa. Inoltre, a causa della natura ensemble, l'addestramento e l'uso di un Random Forest possono richiedere un notevole tempo di calcolo e memoria, soprattutto con dataset di grandi dimensioni e con molti alberi. (Breiman, 2001; Biau & Scornet 2016; Friedman, J., Hastie, T., & Tibshirani 2001)

Il Gradient Boosting è un altro metodo di ensemble. Questo modello costruisce una serie di modelli (generalmente alberi decisionali) in modo sequenziale, dove ogni nuovo modello cerca di correggere gli errori dei modelli precedenti. In pratica, vengono aggiunti alberi successivi fino a quando l'errore non può essere ulteriormente ridotto o si raggiunge un numero predefinito di alberi (Friedman, 2001). Sebbene sia noto per la sua elevata accuratezza predittiva grazie alla sua capacità di ridurre l’errore residuo attraverso l’aggiunta sequenziale di modelli, può essere incline all'overfitting, specialmente se il numero di alberi è troppo elevato o se non vengono applicati adeguati meccanismi di regolarizzazione. Inoltre, come il random forest, può risultare meno interpretabile rispetto ai singoli modelli e può richiedere un notevole tempo di calcolo e risorse (Liaw & Wiener 2002).

Altro algoritmo di ensemble è l’Extra Tree Regression. Questo è una variante della Random Forest, ma costruisce gli alberi in modo più casuale. Una delle differenze tra le due tecniche è la selezione del punto di taglio. Le tecniche Extra Tree scelgono i punti di taglio in modo casuale; tuttavia, Random Forest seleziona la divisione ottimale. Extra Trees è noto per la sua velocità di addestramento e alta accuratezza grazie alla sua randomizzazione, che riduce il rischio di overfitting. Tuttavia, come gli altri metodi di ensemble, può essere meno interpretabile rispetto ai singoli alberi decisionali (Geurts et al., 2006; Stuti Raizada e Jatinderkumar R. Saini, 2021).

Il Support Vector Regressor (SVR) è una variante del Support Vector Machine (SVM) progettata per compiti di regressione. L'algoritmo cerca di trovare una funzione che approssima i dati in modo che l'errore di previsione rientri entro un intervallo epsilon (tolleranza accettabile per l'errore della predizione) per la maggior parte dei dati. Il Support Vector Regressor è meno sensibile ai valori anomali rispetto ad altri metodi di regressione, poiché minimizza l'influenza degli errori entro un margine predefinito (epsilon). Tuttavia, la performance di SVR dipende fortemente dalla scelta dei parametri come il margine epsilon e i parametri del kernel. Una scelta subottimale può portare a modelli che non generalizzano bene sui dati di test. Inoltre, è computazionalmente intensivo, specialmente con dataset di grandi dimensioni, e la sua interpretabilità può risultare complessa, rendendo difficile capire come vengono fatte le previsioni. (Cortes & Vapnik, 1995).

La regressione lineare è uno dei modelli fondamentali utilizzati in Machine Learning (ML) per predire un valore continuo. Questo modello si basa sull'assunzione che esista una relazione lineare tra la variabile target e una o più variabili indipendenti (Stuti Raizada e Jatinderkumar R. Saini – 2021). Il principale vantaggio della regressione lineare è la sua semplicità e interpretabilità. È particolarmente efficace quando la relazione tra le variabili è lineare e i dati non presentano significative deviazioni dalla linearità. Tuttavia, può essere poco efficace se la relazione tra le variabili è non lineare, e i risultati possono essere influenzati da outliers. Inoltre, un limite fondamentale di questo modello è che non gestisce bene le interazioni complesse tra le variabili (Shai Shalev-Shwartz e Shai Ben-David, 2014).

Il K-Nearest Neighbors (K-NN) Regression, invece, è un algoritmo di apprendimento supervisionato che predice il valore di una variabile target trovando i K esempi più vicini nel dataset di addestramento e calcolandone la media. La "vicinanza" viene solitamente misurata utilizzando la distanza euclidea, ma possono essere usate altre metriche a seconda della specificità del problema. Il parametro K è cruciale: un valore troppo piccolo può rendere il modello eccessivamente sensibile al rumore nei dati, mentre un valore troppo grande può portare a previsioni eccessivamente generalizzate (Zhang, 2016). Nonostante la sua semplicità, il K-NN presenta svantaggi rilevanti, soprattutto in termini di costo computazionale, poiché ogni previsione richiede il calcolo delle distanze tra il nuovo punto e tutti i punti del dataset di addestramento. Questo può diventare proibitivo con dataset di grandi dimensioni. Inoltre, il modello è sensibile alla scelta del valore di K e alla metrica di distanza utilizzata (Bhatia, N., & Vandana, 2010).

Infine, un'ultimo regressore applicato durante l'analisi previsionale è la Stepwise Regression, tecnica di regressione lineare utilizzata per selezionare le variabili più rilevanti in un modello

predittivo. L'obiettivo è costruire un modello che con il minor numero di variabili che ancora spiegano una porzione significativa della varianza del target (ad esempio, i volumi di vendita). Il suo principio di funzionamento permette di semplificare il modello, riducendo il numero di variabili incluse e migliorando così l'interpretabilità. Nonostante la sua utilità, la Stepwise Regression ha alcuni limiti. In primis, può portare a modelli instabili, specialmente se il dataset è piccolo o se le variabili sono fortemente correlate tra loro (multicollinearità). A tal proposito, in presenza di multicollinearità, potrebbe eliminare variabili che sono rilevanti ma appaiono ridondanti a causa delle correlazioni con altre variabili. Inoltre, si adatta troppo bene ai dati storici (overfitting), perdendo la capacità predittiva sui nuovi dati (Smith, 2018; P Bajari, D Nekipelov, SP Ryan, M Yang, 2015).

2.3 IL TRADE MARKETING E L’INTEGRAZIONE CON IL ML

Negli anni '80 e '90 molti produttori di beni di consumo hanno implementato un'unità funzionale di trade marketing nella loro organizzazione di marketing e vendita (Dewsnap e Jobber 2004a, 2009; Piercy 1985; Randall 1994). La letteratura e la pratica commerciale utilizzano i termini marketing del cliente, pianificazione del cliente e sviluppo delle vendite come sinonimi di trade marketing (Dewsnap e Jobber 2003).

In letteratura il trade marketing non è definito in modo coerente. La definizione di Trade Marketing di Dewsnap e Jobber (2009, p. 989) è una delle poche definizioni nei documenti accademici: "In quanto processo volto ad integrare gli obiettivi e le strategie di vendita e marketing, il trade marketing è progettato per garantire che le esigenze del retailer (ad esempio in termini promozionali), comunicate internamente dal personale di vendita, siano soddisfatte dal marketing mix del marchio coordinato dal marketing (Cespedes, 1993)”. Nel libro manageriale di Corstjens e Corstjens (1999, pp. 222-223), successivamente aggiornato da Thain e Bradley (2012), il concetto di trade marketing viene invece definito come una forma di marketing industriale, specificamente orientato alle dinamiche del business-to-business. Questo tipo di marketing si distingue nettamente dal marketing rivolto ai consumatori finali, in quanto focalizzato sulle relazioni e le strategie tra le imprese.

La definizione di Dewsnap e Jobber (2009) si concentra maggiormente sul trade marketing come ruolo di integratore. Corstjens e Corstjens (1999) vedono invece il trade marketing come un'unità funzionale separata focalizzata sul marketing business-to-business per i rivenditori.

Le ragioni per implementare il trade marketing sono molteplici (Dewsnap e Jobber 2004b). Ma gli autori evidenziano due delle ragioni.

In primis, sottolineano la crescente pressione per integrare il marketing e le vendite in risposta ai rivenditori più consolidati, centralizzati e sofisticati (Davies 1993, Dewsnap e Jobber 2004a, 2004b; Piercy 1985; Randall 1994). In secundis, sottolineano l'aumento vertiginoso dei costi di promozione e la necessità di un'unità funzionale specializzata al marketing business-to-business rivolta al dettagliante, che sia responsabile della gestione della promozione commerciale e dei materiali di comunicazione (Davies 1994; Thain e Bradley 2012). Sotto questo punto di vista, il trade marketing aiuta i rivenditori a differenziarsi dalla concorrenza offrendo loro promozioni su misura (Zentes 1989).

Questa disciplina non si limita semplicemente alla promozione dei prodotti nei punti vendita, ma integra una serie di attività mirate a ottimizzare la disponibilità dei prodotti, migliorare la visibilità sugli scaffali, e stimolare la domanda attraverso la collaborazione con i partner commerciali. In questo contesto, l'unità aziendale attraverso le proprie attività impatta key performance indicators (KPIs) strategici che influenzano significativamente le performance di vendita finali. Attraverso la raccolta e l'analisi di dati relativi alla disponibilità e alle performance dei prodotti, il trade marketing può fornire raccomandazioni mirate, in primis ai colleghi di sales, e in secondo luogo ai clienti, per migliorare la presenza e l'esposizione dell'assortimento, aumentando così la probabilità di vendita. Ad esempio, indicatori come il numero di facing, il numero di punti vendita che aderiscono alle campagne promozionale, la quota a scaffale o la distribuzione ponderata forniscono dati cruciali sul successo delle iniziative commerciali. La costante monitorizzazione di tali dati consente alle aziende di adattare e affinare continuamente le loro strategie, al fine di massimizzare le vendite (P. J. Johnson e H. R. Thompson, 2019; Philip Kotler e Gary Armstrong, 2017).

In sintesi Philip Kotler e Gary Armstrong affermano che il trade marketing si compone di tre leve operative: prezzo, promozione e distribuzione. Questa affermazione è supportata dal loro testo ampiamente utilizzato nel campo del marketing: Principles of Marketing (17th ed.) (2018).

La leva del prezzo riveste un ruolo cruciale nell'ambito del trade marketing, influenzando direttamente le dinamiche di acquisto dei rivenditori. Kotler e Armstrong evidenziano che una strategia di pricing ben strutturata può includere sconti quantitativi, incentivi per acquisti anticipati e condizioni di pagamento preferenziali. Questi strumenti di pricing sono progettati per rendere il prodotto più attraente e competitivo, favorendo la decisione dei rivenditori di inserire e promuovere il prodotto all'interno del loro assortimento. Inoltre, una politica di prezzo efficace contribuisce a ottimizzare il margine di profitto e a posizionare il prodotto in maniera vantaggiosa rispetto alla concorrenza.

La leva della promozione è strategicamente utilizzata per stimolare l'interesse e il supporto dei rivenditori. Le attività promozionali possono comprendere sconti temporanei, esposizioni speciali in punto vendita e campagne pubblicitarie mirate. Kotler e Armstrong sottolineano che queste iniziative hanno l'obiettivo di incrementare la visibilità del prodotto e migliorare il suo posizionamento sugli scaffali. Attraverso una promozione ben orchestrata, è possibile incentivare i rivenditori a mettere in risalto il prodotto, facilitare l'acquisizione di nuovi clienti e rafforzare la brand loyalty. Inoltre, le promozioni possono agire come leva per spingere il prodotto in cima alle preferenze dei consumatori.

Infine, la leva della distribuzione è fondamentale per garantire che i prodotti siano disponibili nei punti vendita giusti e nelle giuste quantità. Il trade marketing attraverso le proprie analisi sui singoli item aiuta i colleghi delle sales assicurare che l’assortimento di prodotti nei punti vendita sia adeguato alle esigenze dei consumatori. Questo processo include la pianificazione dell'allocazione dei prodotti basata su dati di vendite passate e previsioni di domanda, con l'obiettivo di massimizzare la disponibilità e l'efficacia del prodotto (Ailawadi et al., 2014).

La letteratura accademica esistente sull'integrazione tra machine learning (ML) e trade marketing per la previsione dei volumi di vendita evidenzia l'importanza e il potenziale di questa sinergia. L'applicazione del machine learning nel contesto del trade marketing è stata analizzata attraverso vari studi, che mostrano come l'adozione di tecniche avanzate di ML possa migliorare significativamente l'accuratezza delle previsioni di vendita e l'efficacia delle strategie di trade marketing.

Sujatha e Sangeetha (2017) hanno dimostrato come l'uso di algoritmi di ML possa aiutare a identificare le variabili chiave che influenzano le vendite in un contesto di trade marketing. Attraverso tecniche come il clustering e l'analisi delle correlazioni, i modelli di ML possono rivelare

l'impatto delle promozioni, delle strategie di pricing e della distribuzione dei prodotti sui volumi di vendita. Continuando, altri studi hanno poi preso in analisi l’efficacia delle attività di promozione. Cadavid, Lamouri e Grabot (2019) hanno analizzato l'impatto delle campagne promozionali utilizzando algoritmi di machine learning. I loro studi suggeriscono che i modelli di ML possono prevedere l'efficacia delle campagne promozionali e suggerire modifiche per massimizzare il ritorno sugli investimenti. Altri studi evidenziano l'importanza delle previsioni della domanda in tempo reale nel contesto del trade marketing. Bajari et al. (2015) attraverso i loro studi hanno evidenziato come i modelli di ML possano fornire previsioni in tempo reale, permettendo alle aziende di adattare rapidamente le loro strategie di marketing e migliorare la gestione delle scorte.

In conclusione, il ruolo del Trade Marketing e il suo potenziale nel migliorare l'accuratezza delle previsioni di vendita è sottolineato in più studi in letteratura. Infatti, il Trade Marketing, focalizzato sull'ottimizzazione delle attività promozionali e distributive nei punti vendita, può fornire dati preziosi che, integrati nei modelli di Machine Learning, permettono di affinare ulteriormente le previsioni. L'integrazione di queste strategie con i modelli predittivi consente di ottenere una visione più completa e accurata delle dinamiche di mercato, fornendo alle aziende strumenti avanzati per la pianificazione commerciale e l'ottimizzazione delle vendite.

2.4 IMPLICAZIONI DALLA LITERATURE REVIEW

Questo paragrafo riassume i risultati della revisione della letteratura e getta le basi per la ricerca empirica sul ruolo del trade marketing nel contesto del machine learning pre svolgere predizioni dei volumi di vendita.

La revisione della letteratura ha messo in luce la crescente necessità per le aziende di integrare metodi di previsione avanzati, come gli algoritmi di apprendimento supervisionato, nei propri processi di stima dei volumi di vendita. Questo approccio permette di superare i limiti dei modelli tradizionali, offrendo soluzioni più flessibili e precise, capaci di rispondere alle sfide imposte dai mercati contemporanei, caratterizzati da volatilità e complessità crescenti.

Inoltre, numerosi studi hanno dimostrato come l'integrazione tra machine learning (ML) e strategie di trade marketing rappresenta un approccio innovativo e potente per migliorare la precisione delle previsioni di vendita. Questa combinazione, non solo migliora l'affidabilità delle previsioni, ma

consente anche di affinare le strategie di trade marketing, rendendole più efficaci nel raggiungere gli obiettivi aziendali.

In questo contesto, la ricerca empirica ha evidenziato l'importanza di perfezionare i modelli predittivi includendo dimensioni specifiche del trade marketing. Tali dimensioni, approfondite nel capitolo 3, rappresentano variabili fondamentali che influenzano direttamente i volumi di vendita. Dalle analisi condotte emergono proposizioni che esplorano le relazioni tra le variabili indipendenti e la variabile target, offrendo una visione più completa di come i diversi fattori interagiscono e impattano sulle performance commerciali.

Questi presupposti costituiscono una base solida per avviare un'analisi approfondita volta all'individuazione dell'algoritmo di apprendimento supervisionato più adatto a prevedere le vendite future delle aziende operanti nel FMCG. L'obiettivo è sfruttare al massimo il potenziale dei dati storici aziendali, fornendo alla macchina le informazioni relative alle attività in-store implementate in tutti gli ipermercati e supermercati italiani dal 2022 ad oggi dai principali produttori di pasta che operano nel settore dei beni di largo consumo in Italia.

3 METODOLOGIA

Questo capitolo si propone di delineare in maniera dettagliata la metodologia adottata per condurre il processo di raccolta e analisi del dataset utilizzato per l’implementazione del modello predittivo. L'obiettivo è fornire una chiara descrizione delle fasi operative che hanno guidato l'intera ricerca, dalla selezione e acquisizione dei dati fino all'applicazione dei modelli di machine learning per l'elaborazione delle previsioni.

In questo studio, è stata dedicata particolare attenzione alla struttura del processo analitico, concepito per massimizzare la precisione delle previsioni e minimizzare i margini di errore. Inoltre, la robustezza metodologica è stata assicurata attraverso l'adozione di tecniche avanzate di data mining e machine learning, integrate con una scrupolosa fase di pre-processing e di esplorazione dei dati. Ciò ha permesso di affrontare le sfide poste dalla complessità e dalla variabilità dei dati storici presi in considerazione.

Il capitolo, non solo illustra le tecniche e gli strumenti utilizzati, ma ne discute anche la giustificazione teorica e pratica, dimostrando come l'approccio metodologico scelto sia stato calibrato per rispondere efficacemente alle specifiche esigenze del caso di studio. Pertanto, la scelta della metodologia e degli strumenti analitici è stata attentamente ponderata in funzione degli obiettivi dello studio, assicurando che le conclusioni tratte siano sostenute da un robusto processo analitico.

3.1 DESIGN DELLA RICERCA

Come delineato in precedenza, la presente tesi adotta un approccio di ricerca mirato ad identificare e colmare un vuoto esistente nella letteratura scientifica relativo all'integrazione tra le tecniche di machine learning ed il trade marketing, al fine di fornire nuova conoscenza utile per ottimizzare le strategie e la pianificazione aziendale. Per fare ciò, lo studio si propone di sviluppare un modello di previsione avanzato che permetta alle aziende di ottenere stime precise dei volumi futuri di sell-out.

RESEARCH QUESTION	*Qual è il modello di machine learning più efficace per prevedere, sulla base di dati di trade marketing, i volumi di vendita di un'azienda operante nell’industria dei beni di largo consumo?*

Le ipotesi che guidano l’obiettivo della ricerca sono state precedentemente testate e validate in altri studi scientifici. Tuttavia, è fondamentale sottolineare che, nonostante queste ipotesi siano già state esplorate, esse non sono mai state applicate in modo specifico ad uno scenario simile a quello sviluppato in questo studio. In altre parole, non sono mai state utilizzate per l'identificazione di un modello predittivo basato su dati commerciali relativi ad un'intera categoria di mercato (la pasta di semola) all'interno del settore dei beni di largo consumo.

Il contesto in cui vengono applicate queste ipotesi è quindi del tutto nuovo e rappresenta un contributo significativo alla letteratura esistente. Il settore del FMCG presenta dinamiche di mercato uniche, caratterizzate da un'elevata frequenza di acquisto, prezzi relativamente bassi e margini di profitto ridotti. Queste caratteristiche rendono cruciale l'accuratezza e la tempestività delle previsioni di vendita, e giustificano la necessità di esplorare l'applicabilità delle ipotesi esistenti in questo specifico contesto.

H1	*Le attività di Trade Marketing hanno un impatto significativo sulle vendite dei beni di largo consumo.*
H2	*Le tecniche di machine learning possono ottimizzare significativamente l'impatto delle attività di trade marketing sulle vendite, migliorando la previsione della domanda e l'efficacia delle strategie promozionali.*
H3	*I modelli di machine learning superano i modelli tradizionali in termini di accuratezza predittiva delle vendite.*
H4	*Un modello previsionale che stima con elevata accuratezza le performance dei prodotti al sell-out rappresenta un asset fondamentale che consente alle aziende di ottimizzare le proprie decisioni strategiche e operative.*

In questo studio, le ipotesi vengono testate non solo per verificarne la validità, ma anche per costruire un modello predittivo avanzato che possa supportare le decisioni strategiche nel mondo FMCG. Questo approccio innovativo consente di esplorare nuove possibilità di applicazione delle

tecniche di machine learning in un settore in cui l'ottimizzazione delle vendite e della supply chain è di fondamentale importanza.

La ricerca è stata svolta con un approccio empirico, in quanto si basa su dati concreti e osservabili relativi al mercato della pasta di semola in Italia. Questo approccio consente di analizzare e interpretare le informazioni storiche per ottenere insights pratici e realistici. L'analisi si concentra sull'osservazione di trend e pattern passati per formulare previsioni e inferenze riguardanti il futuro andamento del mercato.

3.1.1 METODOLOGIA DI RACCOLTA E ANALISI

I dati utilizzati sono stati forniti dalla *Barilla G. e R. Fratelli S.p.A* (azienda con un posizionamento da leader di settore nella categoria della pasta di semola in Italia). Tuttavia, è importante sottolineare che i dati di mercato condivisi da *Barilla* sono acquisiti da un provider esterno: *NielsenIQ* (*NIQ*). La collaborazione tra queste due società assicura un elevato grado di accuratezza e affidabilità.

NielsenIQ è una delle principali società di ricerca di mercato a livello globale, riconosciuta per la sua leadership nel settore dei beni di largo consumo. Con una lunga storia di expertise e innovazione, *NIQ* fornisce analisi e insights approfonditi che sono fondamentali per le strategie aziendali e le decisioni di marketing nel panorama globale del FMCG. La sua autorevolezza è supportata da metodi di ricerca avanzati e da una vasta rete di dati, che la rendono un punto di riferimento imprescindibile per comprendere le dinamiche di mercato e le tendenze dei consumatori. Attraverso le piattaforme *Nielsen*, i produttori (come in questo caso la *Barilla*) di beni di largo consumo possono consultare in tempo reale le informazioni di mercato di cui necessitano, ed ottenere insights in grado di facilitare risposte rapide e informate.

Successivamente all'estrazione del dataset dalla piattaforma del provider, è iniziato il processo di analisi dei dati attraverso l'utilizzo del popolare linguaggio di programmazione Python e dei modelli di machine learning.

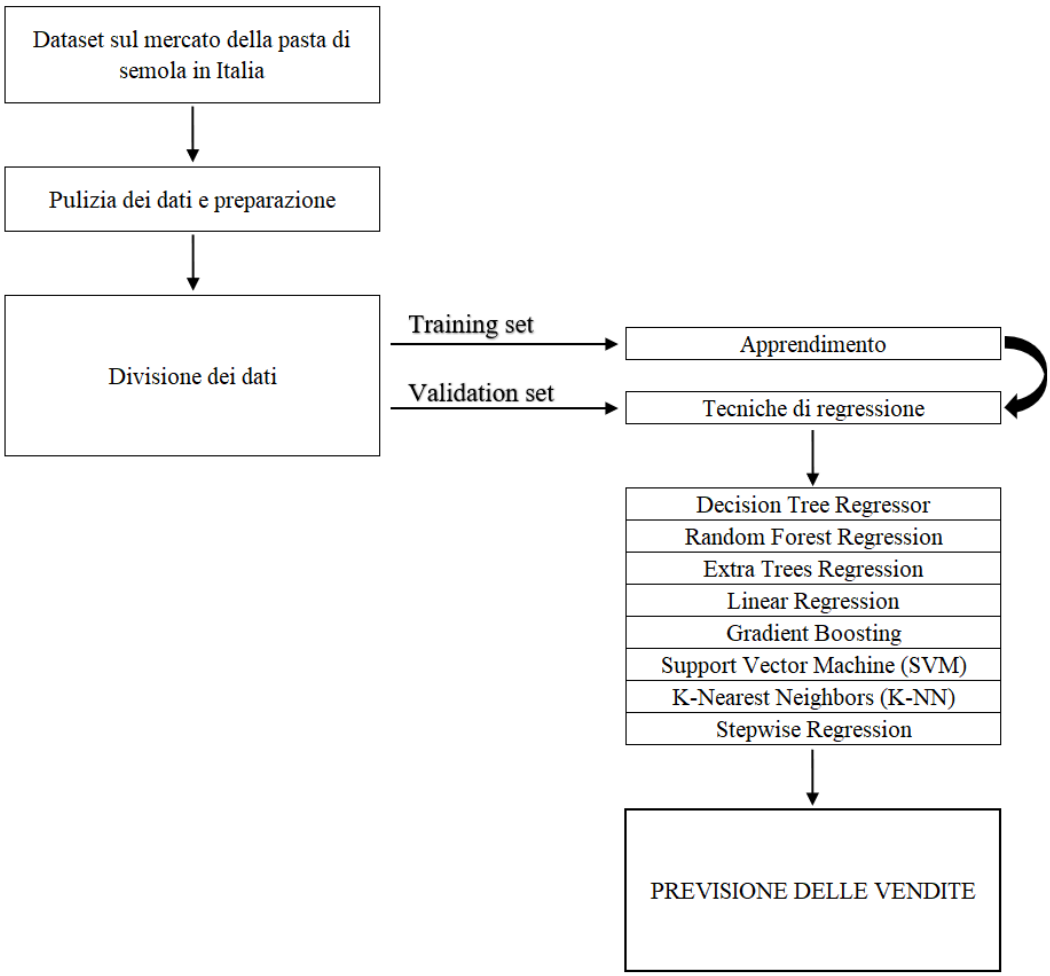
Come spiegato nel capitolo 2.2 *"Utilizzo del machine learning nella previsione delle vendite"*, l'apprendimento automatico, nell'ambito dell'analisi statistica, rappresenta un avanzamento significativo rispetto ai metodi tradizionali, poiché permette di identificare pattern complessi e fare

previsioni basate su grandi volumi di dati. Diversamente dall'analisi statistica classica, che si basa su modelli predeterminati e assunzioni specifiche, il machine learning è in grado di apprendere autonomamente dalle informazioni fornite, migliorando continuamente la propria performance man mano che vengono introdotti nuovi dati.

Il processo di analisi è stato strutturato con l'obiettivo di garantire una robustezza metodologica tale da assicurare risultati validi e affidabili. A tal fine, il processo ha seguito diverse fasi chiave, ciascuna con un ruolo cruciale nell'assicurare l'integrità e la qualità delle conclusioni derivanti dallo studio:

1. *Data Pre-Processing*: Questa fase iniziale è fondamentale per preparare i dati grezzi in modo che siano adatti all'analisi successiva. Include operazioni come la pulizia dei dati, la gestione dei valori mancanti, la normalizzazione, la trasformazione delle variabili. L'obiettivo è ridurre gli errori che potrebbero compromettere l'affidabilità dei risultati, garantendo che i dati siano rappresentativi e coerenti con le esigenze dell'analisi.
2. *Data Exploration*: Una volta che i dati sono stati pre-processati, si procede con l'esplorazione dei dati. Questa fase prevede l'analisi esplorativa, utilizzando strumenti statistici e visualizzazioni per comprendere le caratteristiche principali del dataset. Vengono identificate tendenze, pattern e anomalie, oltre a testare ipotesi preliminari. L'esplorazione dei dati è essenziale per guidare le scelte metodologiche successive e per individuare eventuali problematiche che potrebbero influenzare la validità dell'analisi.
3. *Data Mining*: Il cuore del processo analitico risiede nel data mining, la fase in cui vengono utilizzati algoritmi avanzati per estrarre conoscenze significative dai dati. In questa fase dell'analisi, vengono implementate le tecniche di regressione, le quali sono fondamentali per modellare e quantificare le relazioni esistenti tra le variabili. Attraverso l'uso della regressione, è possibile prevedere con precisione i valori continui, permettendo di ottenere stime accurate basate sui dati osservati. Oltre alla regressione, possono essere impiegate altre metodologie come il clustering o la classificazione, anch’esse finalizzate all'individuazione di relazioni nascoste e alla costruzione di modelli predittivi. Il data mining non solo consente di scoprire insight che non sarebbero immediatamente evidenti, ma fornisce anche una base empirica solida su cui costruire le conclusioni della ricerca. È essenziale che questa fase venga eseguita con precisione, selezionando i modelli più appropriati per il tipo di dati e il contesto specifico dell'analisi.

Figura 4: Metodologia di analisi



3.1.2 LINGUAGGIO DI PROGRAMMAZIONE: PYTHON

Il linguaggio di programmazione utilizzato per l'analisi dei dati è Python. Quest'ultimo è stato selezionato per la sua ampia disponibilità di librerie e framework specializzati, che lo rendono uno strumento particolarmente efficace per affrontare le sfide complesse legate all'analisi dei dati.

Negli ultimi 10 anni, Python è diventato uno dei linguaggi più importanti per la scienza dei dati, l'apprendimento automatico e lo sviluppo generale del software nel mondo accademico e nell'industria. Il suo supporto avanzato per librerie come Pandas e Scikit-learn ha consolidato la sua popolarità, rendendolo la scelta preferita per numerose attività di analisi dati, grazie alla sua capacità di gestire e manipolare grandi quantità di informazioni in modo efficiente e scalabile (McKinney, W. - 2012).

Nel presente studio sono state utilizzate varie librerie di Python. Di seguito vediamo i loro contributi specifici al processo.

NUMPY: Abbreviazione di "*Numerical Python*", è fondamentale per la composizione numerica in Python. Questa libreria fornisce supporto per l'elaborazione numerica avanzata, offrendo array multidimensionali e una vasta gamma di funzioni matematiche. NumPy è essenziale per il trattamento di grandi volumi di dati numerici, ottimizzando le operazioni computazionali e garantendo prestazioni elevate anche su dataset di notevoli dimensioni.

PANDAS: Abbreviazione di "*Panel Data System*". Si tratta di una libreria fondamentale per la gestione e la manipolazione dei dati, costruita sulla base di Numpy. NumPy fornisce array multidimensionali e funzioni matematiche avanzate, che costituiscono il fondamento su cui Pandas sviluppa le sue strutture dati principali: Series e DataFrame. Il DataFrame è una struttura dati tabulare orientata alle colonne con etichette sia di riga che di colonna; mentre la Serie è un oggetto array unidimensionale etichettato. In questo modo, Pandas offre strutture dati che facilitano la gestione e la pulizia dei dataset. Grazie ad essa, è possibile eseguire operazioni complesse sui dati con poche righe di codice, semplificando la fase di pre-processamento e rendendo il flusso di lavoro più efficiente. Inoltre, consente l'elaborazione sia di fogli Excel che di dati SQL.

MATPLOTLIB: La visualizzazione dei dati è un aspetto cruciale dell'analisi, poiché permette di comprendere meglio le caratteristiche e le relazioni presenti nei dataset. Matplotlib è una libreria altamente versatile per la generazione di grafici bidimensionali e tridimensionali. Nel processo di analisi, è stata fondamentale per la fase di Data Exploration (vedi cap. 4 "*Risultati*"), in quanto permette la creazione di visualizzazioni personalizzate, facilitando l'identificazione intuitiva di pattern, outlier e altre strutture rilevanti nei dati, rendendo tali informazioni immediatamente comprensibili e interpretabili.

SHIKIT-LEARN: Rappresenta lo strumento chiave per l'implementazione di algoritmi di machine learning e tecniche di data mining. Questa libreria include una vasta gamma di algoritmi per la classificazione, il clustering, la regressione e la riduzione della dimensionalità, tutti facilmente integrabili nel workflow analitico. La sua semplicità d'uso consente di sperimentare rapidamente diversi modelli, selezionando quello più adatto alle specificità dei dati analizzati.

L'87.5% degli algoritmi implementati nella nostra analisi sono modelli di machine learning presenti nella libreria Scikit-learn:

1. Decision Tree Regressor
2. Random Forest Regression
3. Extra Trees Regression
4. Linear Regression
5. Gradient Boosting
6. Support Vector Machine (SVM)
7. K-Nearest Neighbors (K-NN) Regression

(McKinney, W. (2012). Python for Data Analysis: data wrangling with pandas, numpy, and ipython. O'Reilly)

3.2 DATASET

Per la conduzione dell'analisi, sono stati utilizzati dati storici di vendita e di mercato, nonché informazioni dettagliate sulle attività di marketing in-store.

I dati estratti comprendono: trend storici di mercato, vendite effettive, le metriche associate alle campagne promozionali e alle strategie di trade marketing implementate, ed altre informazioni specifiche sulle dinamiche di marketing in-store, come la presenza nei volantini dei distributori o le esposizioni. Questi elementi sono stati integrati per fornire una visione comprensiva e dettagliata dell'impatto delle strategie di trade marketing sui volumi di vendita e sul comportamento dei consumatori.

I dati utilizzati sono stati ottenuti attraverso la piattaforma NIQ, la quale assicura una precisione elevata e una rappresentatività globale delle tendenze di mercato, grazie ad avanzati metodi di raccolta e monitoraggio dati, inclusi panel di consumatori e scanner data. Questo assicura una base solida per la modellizzazione e la previsione, permettendo di ottenere risultati empirici e affidabili.

Il database presenta informazioni dettagliate inerenti allo scenario competitivo in Italia nel mercato della pasta di semola. In particolare vengono presi in analisi i maggiori player del marcato, i quali rappresentano circa il 75% delle vendite del mercato. I brands in questione sono:

- *Barilla*
- *Divella*

- *De Cecco*
- *La Molisana*
- *Garofalo*
- *Rummo*
- *Private Label (o “marca del distributore”)*

Per fornire una panoramica esaustiva della situazione competitiva, i dati analizzati non solo comprendono l'intero mercato italiano, ma offrono anche un livello di dettaglio regionale approfondito. L'integrazione di dati a livello nazionale e regionale consente di ottenere una visione dettagliata e stratificata del mercato. Infatti, i dati aggregati a livello nazionale forniscono una visione completa del mercato, e delle tendenze generali e performance complessive del settore. Questa visione d'insieme ci consente di identificare le principali forze competitive e i cambiamenti nel mercato a livello nazionale. Invece, i dati disaggregati per regione offrono un'analisi più profonda delle performance di un determinato territorio, evidenziando le differenze significative tra le varie aree geografiche. Questo livello di dettaglio permette di analizzare le peculiarità locali, come le preferenze dei consumatori, le dinamiche della domanda e l'efficacia delle strategie di marketing specifiche per ciascuna regione.

Il dataset fornisce informazioni con riferimento al canale di distribuzione “*Hyper + Super”*, che tiene conto solo degli ipermercati e supermercati presenti sul territorio nazionale.

Il periodo di analisi considerato in questo studio va dal 1/01/2022 fino al 31/03/2024 (primo trimestre dell’anno in corso). La scelta di iniziare dal 2022 è dovuta dalla necessità di escludere gli effetti distorsivi causati dalla pandemia di COVID-19. Durante la pandemia, infatti, il mercato della pasta ha subito oscillazioni anomale nelle vendite a causa di fattori straordinari, come l'aumento della domanda dovuto al lockdown e alle restrizioni imposte sui movimenti, che hanno spinto i consumatori a fare scorte alimentari e a modificare le proprie abitudini di consumo.

Questa situazione ha creato un contesto atipico, con picchi di vendite non rappresentativi delle tendenze a lungo termine. L'inclusione di tali dati nella nostra analisi avrebbe potuto compromettere l'affidabilità dei modelli predittivi, poiché le fluttuazioni indotte dal COVID-19 non riflettono le condizioni di mercato attuali e future in un contesto di "normalità" economica. Pertanto, limitare l'analisi ai dati successivi al 2021 permette di ottenere un quadro più rappresentativo delle dinamiche post-pandemiche, consentendo di sviluppare modelli predittivi più robusti e applicabili alle condizioni operative contemporanee e future.

3.2.1 FEATURE

Per garantire l'accuratezza e la validità di una regressione statistica, è essenziale disporre di un insieme esaustivo di feature che esercitano un'influenza significativa sulla variabile dipendente.

Il dataset in questione comprende 4323 righe e 30 variabili, di cui 29 sono indipendenti mentre una sola è dipendente. Le feature sono state selezionate per il loro potenziale impatto sulla variabile target, che, in questo caso specifico, è rappresentata dai volumi di vendita. Questa variabile è codificata nel dataset con la nomenclatura: V. (KGS).

La composizione dei dati è stata progettata per supportare l'analisi e l'identificazione delle relazioni statistiche tra le variabili indipendenti e la variabile target, garantendo una base solida per la regressione statistica e per l'elaborazione di modelli predittivi accurati.

Le feature selezionate per la presente analisi sono state identificate come fondamentali per una comprensione approfondita della competitività del mercato della pasta di semola dal 2022 a oggi. In particolare, le variabili, descrivono, in maniera dettagliata, le tre leve strategiche gestite dal trade marketing: Distribuzione e Assortimento, Prezzo e Promozioni.

Per analizzare i trend relativi alla distribuzione, vengono presi in considerazione indicatori chiave come: la *distribuzione numerica*, che indica la percentuale di punti vendita in cui un brand è presente rispetto al totale dei punti vendita disponibili, offrendo una visione della penetrazione del prodotto sul mercato; e la *distribuzione ponderata* (o *weighted distribution, WD*), che, invece, non si limita a misurare la presenza quantitativa, ma fornisce una valutazione qualitativa della distribuzione, ponderando la presenza del brand in base al volume di vendite generato dagli store in cui la marca è presente. Questo ci permette di capire non solo dove il prodotto è disponibile, ma anche l'importanza dei punti di vendita in cui è inserito. Inoltre, ritroviamo anche indicatori che ci aiutano a comprendere come varia la distribuzione in base alla strategia promozionale e di prezzo del brand: come il *numero di punti vendita che offrono un prodotto in promozione* ed il *numero di punti vendita con il prezzo in normalità*.

Infine, per una visione completa della qualità distributiva è fondamentale valutare le performance assortimentali (di un produttore). A tal proposito vengono presi in analisi il *numero medio di referenze per punto vendita*, che indica il numero medio di sku (stock keeping unit) di un brand presenti nei punti di vendita; e la *quota assortimentale*, che rappresenta la percentuale di spazio a

scaffale occupata da una marca rispetto all'intero assortimento di una determinata categoria merceologica.

Per quanto riguarda il prezzo, sono fornite informazioni che fanno riferimento ad una serie di dati strategici, tra cui i dati relativi al *prezzo medio per confezione di prodotto* o lo *sconto medio* %. Quest’ultimo rappresenta la percentuale media di sconto applicata durante le promozioni rispetto al prezzo normale di listino. Tali indicatori forniscono una chiara indicazione del posizionamento competitivo del brand in termini di prezzo percepito.

Infine, nell'ambito della leva delle promozioni, il dataset offre una serie di variabili chiave che consentono di monitorare e valutare in modo approfondito l'efficacia delle strategie promozionali adottate nel tempo. Queste variabili includono metriche dettagliate che misurano l'impatto delle promozioni sui volumi di vendita. Per l'analisi, sono stati presi in considerazione i KPI più significativi ed esplicativi, che meglio rappresentano la performance delle iniziative promozionali e il loro contributo al raggiungimento degli obiettivi aziendali: *Intensity Index* e *Promo Efficiency Index* (*P.E.I.*) Questi indicatori offrono una visione chiara e quantitativa dell'efficienza delle promozioni, consentendo di identificare le strategie più efficaci e di ottimizzare le future campagne promozionali. Il primo, misura l'intensità e la frequenza con cui le promozioni vengono implementate, offrendo una visione dettagliata dell'impegno promozionale di un brand. Il secondo, invece, valuta l'efficacia delle promozioni nel generare incrementi nelle vendite: maggiore è il valore di questo indicatore, maggiore è stato l’impatto della promo sulle vendite effettuate. Infine, per completare l’analisi promozionale, sono state incluse variabili relative alle promozioni presenti sui volantini dei clienti, che forniscono un'ulteriore dimensione di comprensione delle strategie promozionali e del loro impatto sui comportamenti d'acquisto nei supermercati e ipermercati.

L’analisi integrata di questi KPIs fornisce un quadro dettagliato della performance di mercato, di ogni singolo player e dell'efficacia delle strategie di trade marketing. L'integrazione e l'analisi di tali indicatori permettono una comprensione approfondita delle dinamiche competitive, facilitando l'ottimizzazione delle strategie di marketing e migliorando la pianificazione e l'esecuzione delle attività commerciali.

Di seguito vengono elencate tutte le variabili presenti nel dataset estratto:

- Market Share
- Distribuzione Ponderata

- Distribuzione Ponderata Any promo
- Distribuzione Ponderata No promo
- Distribuzione Numerica
- Numero medio di referenze per pdv
- Quota Assortimentale
- N.Medio pdv vendenti No Promo
- N.Medio pdv vendenti Any Promo
- Total TPR Int.Idx (KGS) → *Intensity Index*
- Price Reduction 10<20% Int.Idx (KGS)
- Price Reduction 20<30% Int.Idx (KGS)
- Price Reduction 30<40% Int.Idx (KGS)
- Price Reduction >40% Int.Idx (KGS)
- Total TPR (KGS) % Lift → *P.E.I*
- Price Reduction 10<20% (KGS) % Lift
- Price Reduction 20<30% (KGS) % Lift
- Price Reduction 30<40% (KGS) % Lift
- Price Reduction >40% (KGS) % Lift
- Sconto % Total TPR vs No Promo (KGS) → *Sconto medio*
- Prezzo Medio Conf.
- Prezzo Medio No Promo Conf.
- Prezzo Medio Any Promo Conf.
- V. (KGS)
- N° Folder → *Numero di volantini in cui il brand x è atterrato nel periodo preso in considerazione*
- Prezzo promo medio folder
- % ins sottocosto
- % ins sconto quantità
- % ins taglio prezzo
- % ins concorso

3.3 DATA PRE-PROCESSING

La prima fase del processo di analisi prevede il pre-processamento dei dati, il cui obiettivo è quello di trasformare i dati grezzi, caratterizzati da incompletezza e disomogeneità, in un dataset strutturato e pulito, idoneo per una corretta analisi statistica e per l’addestramento di modelli di machine learning. Il pre processamento rappresenta uno step cruciale nel workflow dell’analisi, poiché l’accuratezza, l’affidabilità e la robustezza delle previsioni dipendono direttamente dalla qualità e dalla preparazione dei dati utilizzati. Non è soltanto una pratica di routine, ma un passaggio imprescindibile per garantire la validità dei risultati ottenuti. Senza un'adeguata preparazione, anche i modelli più sofisticati rischiano di produrre previsioni distorte o inaccurate. Inoltre, un dataset ben pre-processato riduce il rischio di overfitting, consentendo di ottenere performance migliori in fase di test sui dati non visti.

Ci sono varie attività che possono essere applicate durante il processo di data pre-processing:

1. DATA CLEANING: Il dataset estratto da Nielsen conteneva un elevato numero di valori mancanti e di outliers (*figura 6*). Per poter svolgere un’analisi predittiva puntuale, i dati incompleti, sono stati trattati con la tecnica dell’imputazione della mediana (“*median imputation*”). Nello specifico, l’incompletezza del dataset è stata gestita adottando una tecnica di inserimento del valore centrale della distribuzione (cioè la mediana) nelle celle vuote.

   La scelta della tecnica di imputazione da implementare è intrinsecamente determinata dalle specifiche caratteristiche del dataset in questione, tenendo conto delle sue peculiarità strutturali e della natura delle variabili coinvolte. Un'analisi comparativa dei metodi di imputazione, condotta da Khan e Hoque (2020) e pubblicata nel *Journal of Big Data*, evidenzia che l'imputazione mediante mediana risulta particolarmente efficace in contesti caratterizzati da distribuzioni non normali, come quelle asimmetriche, ed in presenza di valori anomali.

   In maniera analoga, uno studio di Joel, Doorsamyha e Paul, ha approfondito l'efficacia dei diversi metodi di imputazione con un focus specifico sul contesto sanitario. Questo studio ha fornito un'analisi dettagliata delle tecniche di imputazione e della loro efficacia in scenari di dati mancanti o incompleti, rivelando nuovamente che l’imputazione con la mediana è particolarmente vantaggiosa. Anche in questo caso, i risultati hanno dimostrato che l’uso

della mediana può ridurre significativamente l'errore medio assoluto (MAE) e l'errore quadratico medio (RMSE), soprattutto in presenza di distribuzioni asimmetriche o quando i dati presentano outlier significativi.

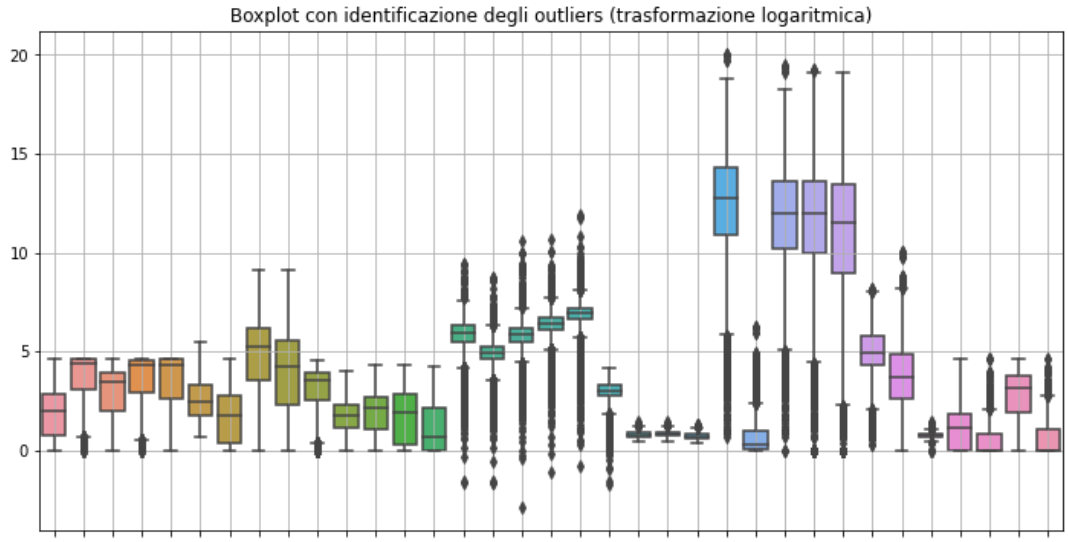
Sulla base di queste evidenze, prima di procedere con l'imputazione, è stato calcolato il coefficiente di skewness (*figura 5*), che misura il grado di asimmetria di una distribuzione. L'analisi ha rivelato che tutti i coefficienti risultano differenti da zero, indicando l'assenza di simmetria nelle distribuzioni delle variabili del dataset. Di conseguenza, l'imputazione con la mediana emerge come una metodologia di data cleaning particolarmente efficace per questa ricerca, in grado di preservare la precisione delle stime e migliorare l'affidabilità dei risultati.

Figura 5: Coefficienti di Skewness

```
I coefficienti di asimmetria per le colonne sono:
Quota in (KGS) - Product                  1.719031
Distr. Pond.                             -0.532850
Distr. Pond. Any Promo                    0.801819
Distr. Pond. No Promo                    -0.424609
Distr. Num.                              -0.375667
N. Medio Ref. per pdv                     1.982083
Quota Assortimentale - Product            1.786066
N.Medio pdv vendenti No Promo             5.286807
N.Medio pdv vendenti Any Promo            7.403164
Total TPR Int.Idx (KGS)                   0.033954
Price Reduction 10<20% Int.Idx (KGS)      1.964272
Price Reduction 20<30% Int.Idx (KGS)      1.347025
Price Reduction 30<40% Int.Idx (KGS)      1.242469
Price Reduction >40% Int.Idx (KGS)        2.494439
Total TPR (KGS) % Lift                   10.936707
Price Reduction 10<20% (KGS) % Lift      14.694433
Price Reduction 20<30% (KGS) % Lift      21.085664
Price Reduction 30<40% (KGS) % Lift      19.183853
Price Reduction >40% (KGS) % Lift        27.610599
Sconto % Total TPR vs No Promo (KGS)     -1.344470
Prezzo Medio Conf.                        0.878819
Prezzo Medio No Promo Conf.               0.566749
Prezzo Medio Any Promo Conf.              0.939391
V. (KGS)                                 17.052642
V. (KGS) No Promo                        17.148956
V. (KGS) Any Promo                       16.768073
V. (KGS) Total TPR                       16.729027
V. Conf. per pdv ponderate                2.138363
N° Folder                                14.999838
Prezzo promo medio folder                 0.855892
% ins sottocosto                          4.470491
% ins sconto quantità                    12.314587
% ins taglio prezzo                       0.952316
% ins concorso                            7.514160
```

Inoltre, come evidenziato nella *figura 6*, il dataset presenta un numero significativo di outliers. Tuttavia, questi valori non influiscono negativamente sui risultati del nostro studio. Al contrario, dopo diverse analisi, è emerso che i modelli predittivi mostrano performance migliori in presenza dei valori anomali. Pertanto, si è deciso di mantenere questi valori anziché sostituirli.

Figura 6: Outliers



2. DATA TRASFORMATION: Un altro aspetto cruciale, della preparazione dei dati, è stata la trasformazione della scala dei valori della variabile dipendente. Originariamente, i volumi di vendita (variabile target) erano espressi in *chilogrammi*, un'unità di misura che comportava valori di ampia portata, influenzando negativamente le prestazioni dei modelli di machine learning. In particolare, l'unità di misura aveva un forte impatto su MAE (Mean Absolute Error), MSE (Mean Squared Error) e RMSE (Root Mean Squared Error), le quali sono fortemente dipendenti dalla scala dei dati. Valori molto grandi (come in questo caso) portavano a errori eccessivamente elevati, distorcendo le metriche di valutazione.

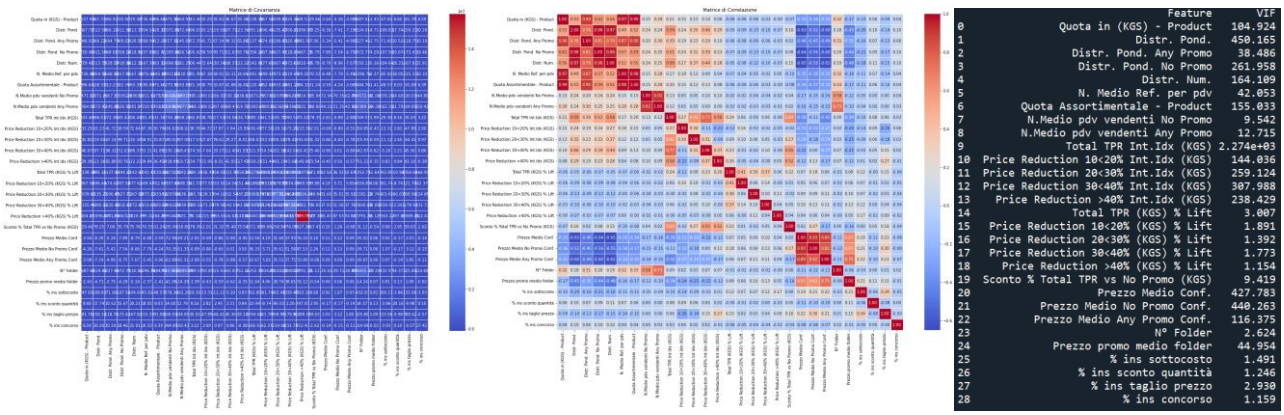
   Per mitigare questo problema, è stato necessario trasformare l'ordine di grandezza, convertendo i valori da *chilogrammi* a *chilotonnellate* al fine di facilitare l'analisi e l'interpretazione.

3. FEATURE SELECTION: Come già anticipato, originariamente il dataset era composto da 29 variabili indipendenti, selezionate ed estratte da Nielsen perché in grado di influenzare significativamente la variabile dipendente. Il processo di identificazione e selezione delle variabili più rilevanti (o feature selection) ci aiuta a capire quali sono le informazioni riportate nel dataset, che effettivamente contribuiscono alla previsione. Spesso, le informazioni presenti in un dataset, anche se sono concettualmente utili, possono risultare statisticamente superflue o già spiegate da un'altra variabile. È importante trovare un equilibrio tra il numero di caratteristiche e la performance del modello. Troppe variabili possono introdurre rumore, mentre troppo poche potrebbero non catturare sufficientemente i pattern nei dati.

   Ci sono vari metodi per svolgere una corretta selezione delle caratteristiche. In questo studio le tecniche applicate si basano sull'analisi statistica delle caratteristiche per determinare la loro rilevanza rispetto alla variabile target. Il principio alla base è quello di applicare test statistici o misure quantitative per valutare e selezionare le variabili più significative prima dell'addestramento del modello. In dettaglio, le variabili sono state testate mediante l'analisi della covarianza, della correlazione e della multicollinearità.

   Le 29 feature oggetto della prima analisi, hanno riportato risultati che hanno evidenziato la presenza di variabili con una correlazione molto elevata, il che indica ridondanza nel modello. Inoltre, le variabili con valori di VIF elevati suggeriscono multicollinearità, il che potrebbe influenzare negativamente la stabilità e l'interpretabilità del modello.

Figura 7: Matrice di covarianza, correlazione e calcolo del VIF



	Feature	VIF
0	Quota in (KGS) - Product	104.924
1	Distr. Pond.	450.165
2	Distr. Pond. Any Promo	38.486
3	Distr. Pond. No Promo	261.958
4	Distr. Num.	164.109
5	N. Medio Ref. per pdv	42.053
6	Quota Assortimentale - Product	155.033
7	N.Medio pdv vendenti No Promo	9.542
8	N.Medio pdv vendenti Any Promo	12.715
9	Total TPR Int.Idx (KGS)	2.274e+03
10	Price Reduction 10<20% Int.Idx (KGS)	144.036
11	Price Reduction 20<30% Int.Idx (KGS)	259.124
12	Price Reduction 30<40% Int.Idx (KGS)	307.988
13	Price Reduction >40% Int.Idx (KGS)	238.429
14	Total TPR (KGS) % Lift	3.007
15	Price Reduction 10<20% (KGS) % Lift	1.891
16	Price Reduction 20<30% (KGS) % Lift	1.392
17	Price Reduction 30<40% (KGS) % Lift	1.773
18	Price Reduction >40% (KGS) % Lift	1.154
19	Sconto % Total TPR vs No Promo (KGS)	9.419
20	Prezzo Medio Conf.	427.783
21	Prezzo Medio No Promo Conf.	440.263
22	Prezzo Medio Any Promo Conf.	116.375
23	N° Folder	2.624
24	Prezzo promo medio folder	44.954
25	% ins sottocosto	1.491
26	% ins sconto quantità	1.246
27	% ins taglio prezzo	2.930
28	% ins concorso	1.159

A questo punto, il processo di feature selection è stato eseguito con un approccio di "prove ed errori", cioè filtrando le caratteristiche in base alla loro capacità di mantenere una covarianza, correlazione e multicollinearità accettabile.

Dopo un attento lavoro di pre-processamento dei dati, l'output finale presenta 11 variabili indipendenti, tutte numeriche (*figura 8*).

Figura 8: Variabili del dataset

KPI	Leva strategica	Spiegazione
Distribuzione Ponderata	DISTRIBUZIONE & ASSORTIMENTO	Esprime la percentuale di giro d'affari della categoria sviluppata dai punti di vendita nei quali un prodotto è venduto
Numero medio di referenze per pdv	DISTRIBUZIONE & ASSORTIMENTO	Numero medio di sku di un brand presenti nei punti di vendita
N.Medio pdv vendenti No Promo	DISTRIBUZIONE & ASSORTIMENTO	Numero di punti di vendita che offrono almeno un prodotto del brand al prezzo di listino
Total TPR (KGS) % Lift	PROMOZIONE	Indice di efficacia promozionale - P.E.I. (Promotion Effectiveness Index). Fornisce una valutazione sull'efficacia dell'attività promozionale, valutando il rapporto tra le vendite effettivamente realizzate e le vendite di baseline
Sconto % Total TPR vs No Promo (KGS)	PROMOZIONE	Sconto medio
Prezzo Medio Conf.	PREZZO	Prezzo medio per singola confezione. Questo indicatore è soggetto a variazioni dovute ad attività promozionali sul prezzo
N° Folder	PROMOZIONE	Numero di volantini in cui è presente un prodotto del brand
% ins sottocosto	PROMOZIONE	% di volantini in cui è presente una promozione "sottocosto"
% ins sconto quantità	PROMOZIONE	% di volantini in cui è presente una promozione "sconto quantità"
% ins taglio prezzo	PROMOZIONE	% di volantini in cui è presente una promozione "taglio prezzo"
% ins concorso	PROMOZIONE	% di volantini in cui è presente una meccanica di partecipazione ad un concorso

3.4 ALGORITMI UTILIZZATI

Data l'elevata importanza della scelta dell'algoritmo, il presente studio prende in analisi 8 regressori che secondo la letteratura possono essere considerati appropriati alle caratteristiche dei dati che vengono passati alla macchina e ne identifica il migliore in termini di accuratezza delle previsioni. In questo modo mira ad aiutare le aziende a decidere quale algoritmo di apprendimento supervisionato applicare senza incorrere nella prima complessità della scelta dell'algoritmo.

- Decision Tree Regressor: spesso utilizzato in letteratura, è un modello adatto per dataset con relazioni non lineari e interazioni complesse tra le variabili. Funziona bene con dati con molte feature categoriche o misti (categorici e continui). Inoltre un problema di questo modello è l'overfitting (Asma' Amro, Mousa Al-Akhras, Khalil El Hindi, Mohamed Habib, and Bayan Abu Shawar – 2021);

- Random Forest Regression: funziona costruendo diversi alberi decisionali durante il corso dell'addestramento e utilizzando la media delle classi come output per la stima di tutti gli alberi. In particolare, la previsione viene generata prendendo una media di tutte le stime prodotte dagli alberi nella foresta, che aumenterà l'accuratezza della previsione (Stuti Raizada, Jatinderkumar R. Saini - 2021). Tale modello eccelle in dataset con molte feature, sia categoriche che continue. È una tecnica che combina i risultati di diversi alberi decisionali per migliorare la precisione complessiva del modello. La random forest è generalmente robusta e meno suscettibile all'overfitting rispetto a un singolo albero decisionale (Leo Breiman – 2001);

- Gradient Boosting: ideale per dataset complessi con relazioni non lineari. Se il problema che si affronta ha relazioni complesse tra var. dipendenti e var. indipendente, allora questo metodo può catturare queste relazioni in modo più efficace rispetto ad altri modelli più semplici. Nonostante ciò è molto sensibile agli outliers (G. Sujatha, R. Sangeetha – 2021);

- Support Vector Machine (SVM): il paper "A Novel Approach to Identify the Promotion of Sales in Supermarkets using Machine Learning Algorithms" descrive come la SVM sia utilizzata per analizzare il comportamento dei clienti e prevedere le vendite. Si evidenzia che la SVM è efficace nel gestire dataset complessi, trovando l'iperpiano che meglio separa le diverse classi di dati. Questo modello è particolarmente utile per problemi di classificazione con margini chiari tra le classi e può essere applicato anche per problemi di regressione;

- Linear Regression: ottimale per dataset con relazioni lineari tra le feature e l'output. È semplice e interpretabile, ma non adatto per dati con relazioni non lineari complesse. Ed è sensibile agli outlier, che possono influenzare significativamente il modello e le sue previsioni (Stuti Raizada e Jatinderkumar R. Saini – 2021);

- K-NN Regression: L a KNN Regression predice il valore di una variabile dipendente (come le vendite) basandosi sui valori delle variabili indipendenti dei "k" punti dati più vicini nel dataset di addestramento. Metodo efficace per la previsione delle vendite quando le relazioni tra i dati sono complesse e non lineari, ma richiede attenzione particolare all'efficienza computazionale (riduzione delle dimensioni) e alla preparazione dei dati (normalizzazione, sensibilità outliers) (Stuti Raizada e Jatinderkumar R. Saini – 2021);

- Extra Tree Regression: è una variante della Random Forest, ma costruisce gli alberi in modo più casuale. Una delle differenze tra le due tecniche è la selezione del punto di taglio. Le tecniche Extra Tree scelgono i punti di taglio in modo casuale; tuttavia, Random Forest seleziona la divisione ottimale. Questo porta a una maggiore diversificazione degli alberi e tende a ridurre l'overfitting. Poiché i punti di divisione vengono scelti casualmente, l'algoritmo può essere più veloce rispetto alla Random Forest, che cerca i punti di divisione ottimali. L'Extra Tree Regression è robusta rispetto alla variazione nei dati e può gestire bene dataset complessi con molte caratteristiche (Stuti Raizada e Jatinderkumar R. Saini – 2021).

- Stepwise Regression: utile per selezionare feature rilevanti in dataset con molte variabili. Funziona meglio quando c'è una combinazione di variabili irrilevanti e rilevanti, aiutando a costruire modelli più interpretabili (P Bajari, D Nekipelov, SP Ryan, M Yang – 2015).

Per quanto riguarda la configurazione degli esperimenti, per ognuno dei modelli testati è stata effettuata una *K-fold Cross Validation*, suddividendo il dataset di training in cinque parti per garantire una valutazione accurata e prevenire l'overfitting. In particolare, i dati sono stati suddivisi in un set di training e un set di testing, con una proporzione dell'80% per l'addestramento e del 20% per la verifica delle prestazioni. Questa divisione ha permesso di addestrare i modelli su una porzione ampia dei dati e di testarlo su dati mai visti.

Infine, le prestazioni sono state valutate utilizzando metriche standard di regressione, inclusi l'errore quadratico medio (MSE), l'errore assoluto medio (MAE) e il coefficiente di determinazione (R^2).

Inoltre, è stato calcolato l'errore quadratico medio radice (RMSE) per fornire una misura interpretabile della deviazione media delle predizioni.

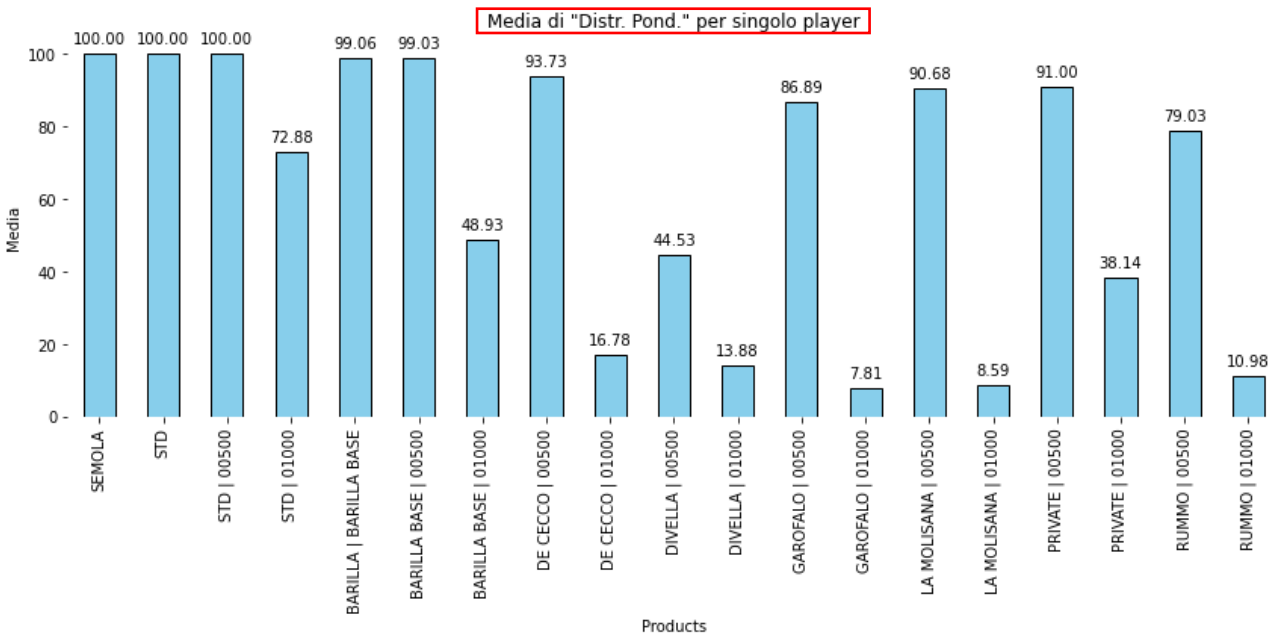
4 RISULTATI

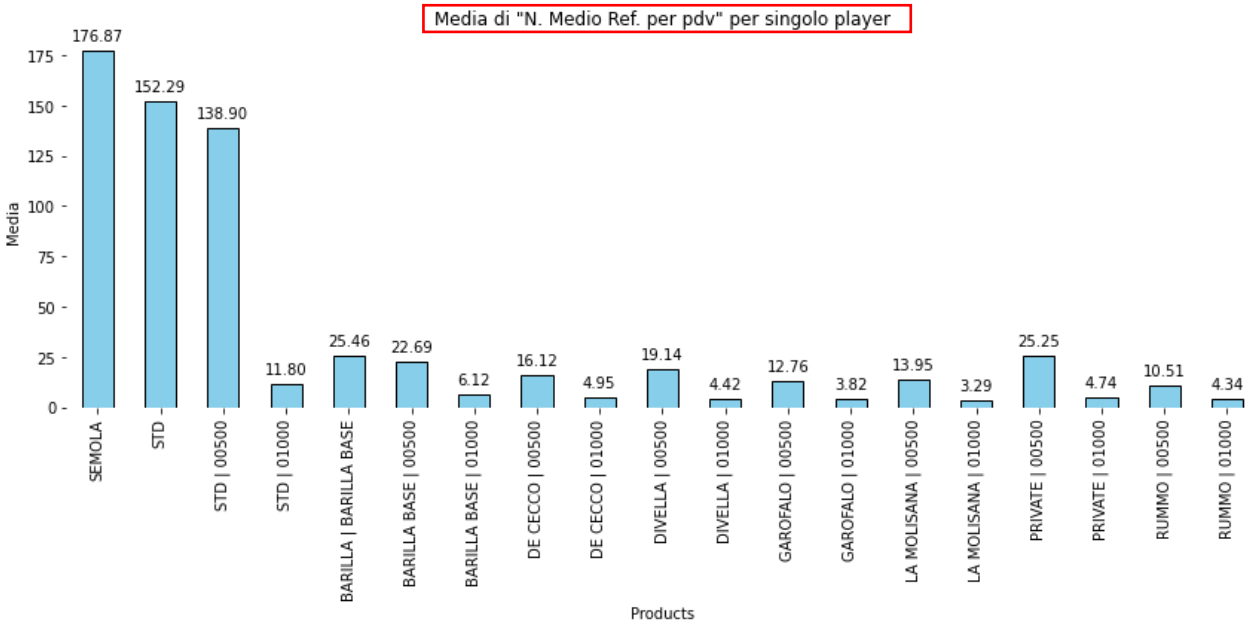
4.1 DESCRIZIONE DEI DATI

Come visto nel paragrafo 3.3, durante la fase di pre-processamento dei dati è stato svolto un lavoro di analisi statistica sul dataset al fine di conoscere le caratteristiche specifiche dei dati a disposizione. In seguito, il dataset che è stato utilizzato per l'addestramento dei vari algoritmi di machine learning, conta 11 variabili indipendenti (vedi *figura 8*), una variabile dipendente e ben 51864 valori numerici, che spiegano le strategie adottate dai vari players del mercato della pasta di semola negli ultimi anni.

I grafici riportati di seguito, mostrano i trend distributivi, promozionali e di prezzo dei vari produttori (con un focus sulla grammatura delle confezioni di pasta). Questa visualizzazione consente di identificare chiaramente le strategie principali adottate dai diversi attori di mercato negli ultimi anni, offrendo una panoramica sull'evoluzione delle politiche commerciali e delle dinamiche competitive del settore.

Figura 9: Trend distributivi e assortimentali





Per quanto riguarda i KPI distributivi (*figura 9*), emergono diverse considerazioni. In primo luogo, la distribuzione ponderata evidenzia l'impegno costante di *Barilla* nel garantire una forte presenza in tutti i punti vendita del territorio nazionale. La confezione da 500g, infatti, presenta una qualità distributiva del 99,03%, un risultato che nessun altro competitor è riuscito ad eguagliare, con un vantaggio di 5,57 punti percentuali sul secondo leader distributivo. Questa capillare presenza sul territorio è un elemento chiave nella strategia di penetrazione di mercato dell'azienda, che si traduce in una maggiore visibilità e accessibilità per i consumatori. La stessa strategia viene applicata alla confezione da 1000g, che sfiora il 50% di distribuzione ponderata, a dimostrazione di un approccio diversificato ma altrettanto efficace.

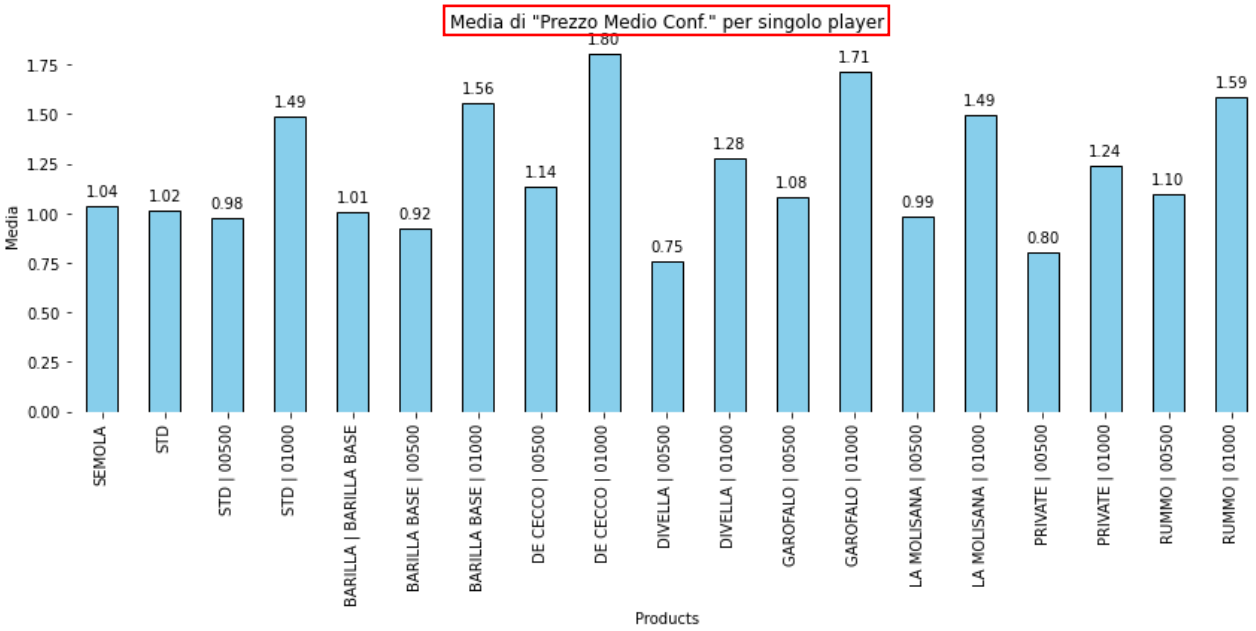
In termini di assortimento degli store, *Barilla* conferma il suo ruolo di leader: è infatti seconda solo alle *Private Label* per numero medio di referenze per punto vendita. Questo dato riflette la volontà dell'azienda di offrire un'ampia gamma di prodotti, rispondendo alle diverse esigenze dei consumatori e migliorando la capacità di attrarre target diversificati. Un assortimento più ampio non solo rafforza la posizione dell'azienda nel punto vendita, ma aumenta le opportunità di cross-selling e fidelizzazione del cliente.

In aggiunta, questi KPI indicano un'ottima gestione delle relazioni con la distribuzione, segno di una negoziazione efficace con la GDO (Grande Distribuzione Organizzata) e di una strategia ben strutturata per massimizzare la presenza sullo scaffale. La combinazione di una vasta copertura distributiva e di un portafoglio prodotti ricco posiziona *Barilla* in modo estremamente competitivo,

rendendola non solo leader di mercato, ma anche un esempio di best practice nell'ottimizzazione dei KPI distributivi e assortimentali.

Per quanto riguarda il posizionamento di prezzo, è utile distinguere i produttori di pasta in due cluster principali: da un lato i "premium player", come *Garofalo*, *La Molisana*, *Rummo* e *De Cecco*, che producono pasta trafilata al bronzo; dall'altro, i "player mainstream", come *Barilla*, *Divella* e le *Private Label*, che offrono pasta di semola senza trafilatura al bronzo.

Figura 10: Prezzo



Analizzando la *figura 10*, emergono differenze di prezzo significative tra i produttori. I player mainstream, come *Barilla*, *Divella* e le *Private Label*, tendono a presentare un prezzo medio per confezione inferiore alla media del mercato, riflettendo un posizionamento più accessibile e orientato al grande pubblico. Al contrario, i produttori premium, come *Garofalo*, *La Molisana*, *Rummo* e *De Cecco*, si collocano su fasce di prezzo più elevate, giustificando il costo con la qualità superiore, come la trafilatura al bronzo e un'attenzione particolare ai processi produttivi. Questa segmentazione evidenzia la differente strategia di pricing adottata dai due cluster, mirata a soddisfare esigenze di consumo diverse: da un lato l'accessibilità, dall'altro la qualità premium.

Per quanto riguarda la leva promozionale (*vedi figura 11*), emerge una chiara distinzione tra i player mainstream e i player premium. Questi due cluster, che già presentano differenze di prezzo, gestiscono la loro promozionalità in modo differente. I player mainstream, come *Barilla* e *Divella*, tendono a esercitare una pressione promozionale meno intensa rispetto ai premium player, sia per le confezioni da 500g che per quelle da 1000g. Al contrario, i premium player come *Garofalo*, *Rummo* e *La Molisana* adottano politiche promozionali più aggressive, con sconti medi più elevati, specialmente per le confezioni da 500g. Tali politiche promozionali aggressive dei premium player sono guidate da una serie di ragioni strategiche. In primis, questi sconti aiutano a giustificare i prezzi più elevati dei loro prodotti, rendendoli più accessibili ai consumatori e riducendo il rischio percepito di acquisto. In secundis, offrire promozioni incoraggia la prova del prodotto e può portare alla fidelizzazione dei clienti, che, una volta soddisfatti, potrebbero diventare acquirenti abituali.

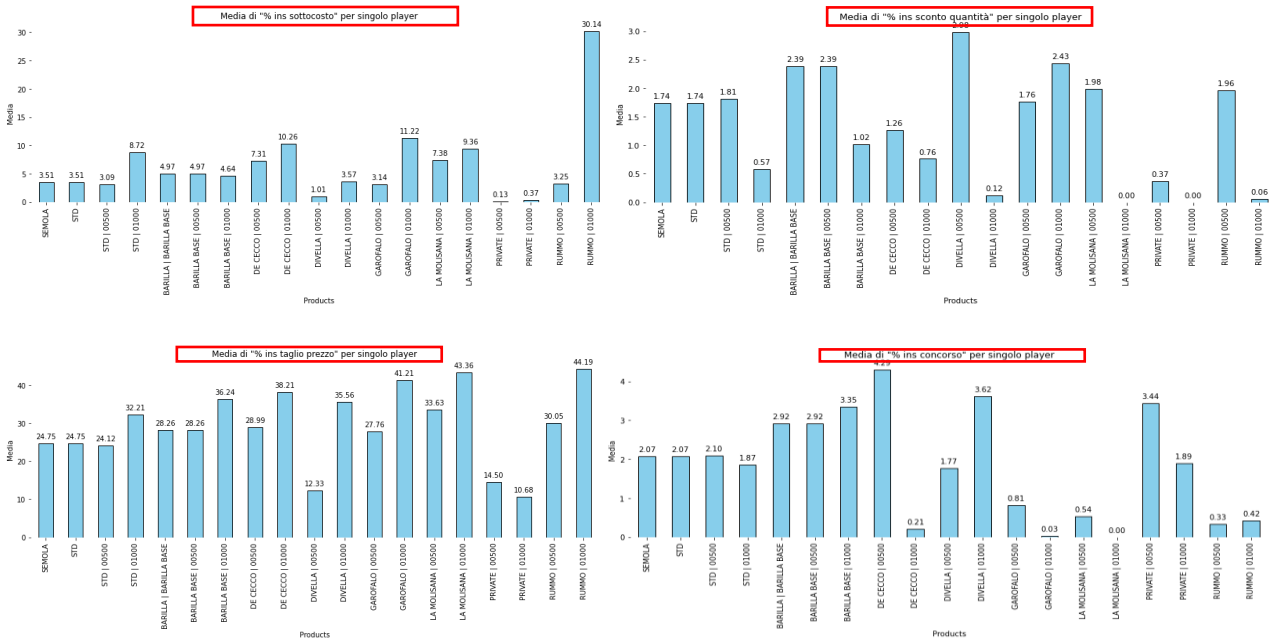
Una notevole eccezione si osserva nel segmento della confezione da 1000g, in cui Barilla riesce a distinguersi, mantenendo un buon presidio su questo KPI nonostante faccia parte del cluster mainstream.

Infine, riguardo alla visibilità promozionale nei volantini, Barilla e De Cecco si contendono il primato, essendo i due produttori con il maggior numero di apparizioni nei volantini dei retailer. Questo dimostra la capacità di entrambi i brand di adottare strategie promozionali efficaci, che consentono loro di ottenere un'ampia esposizione e di raggiungere un pubblico più vasto attraverso questo canale di comunicazione commerciale. Questo dato relativo al numero di volantini può essere interpretato come un indicatore di penetrazione del mercato e di efficacia nella comunicazione promozionale. Un elevato numero di apparizioni nei volantini suggerisce una strategia promozionale efficace e un forte impegno nella visibilità del prodotto presso i retailer. Per i produttori come Barilla e De Cecco, questo non solo indica una buona capacità di negoziazione con i retailer, ma anche un’efficace strategia di marketing che riesce a catturare l'attenzione dei consumatori attraverso promozioni mirate e visibilità costante.

Inoltre, entrando nello specifico nei volantini, leggiamo diverse strategie di atterraggio a volantino da parte dei vari produttori (vedi *figura 12*).

Figura 11: Trend promozionali

Figura 12: Trend a volantino



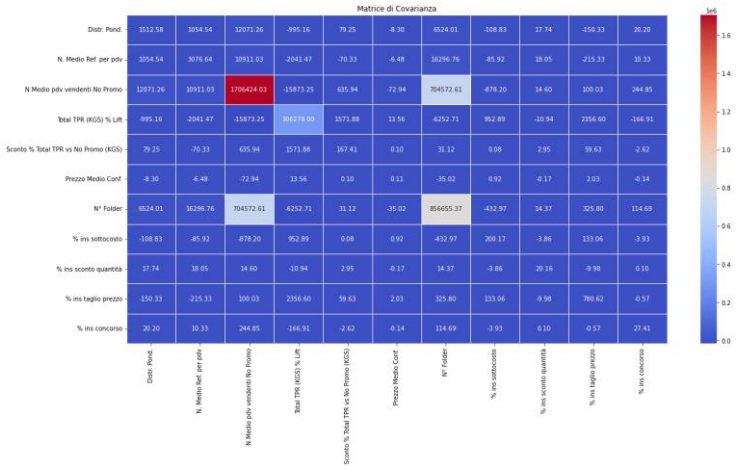
4.2 RELAZIONE TRA I DATI

Conoscere le relazioni tra i dati è fondamentale per comprendere come le variabili influenzano la variabile target, migliorando l'accuratezza del modello predittivo. Identificando correlazioni e interazioni tra le variabili si evitano ridondanze e si costruiscono modelli più interpretabili e affidabili. Ad esempio, includere troppe variabili (come nel database iniziale) può portare a un fenomeno noto come "overfitting", dove il modello diventa troppo complesso e si adatta troppo bene ai dati di addestramento, perdendo capacità di generalizzazione su nuovi dati. Allo stesso modo, attraverso l’analisi stati possiamo individuare variabili che sono collegate tra loro in modo eccessivo (multicollinearità), il che può confondere il modello e ridurre la sua interpretabilità e affidabilità.

Considerando tutte le variabili utilizzate per lo studio, sono state analizzate le covarianze e le correlazioni tra esse.

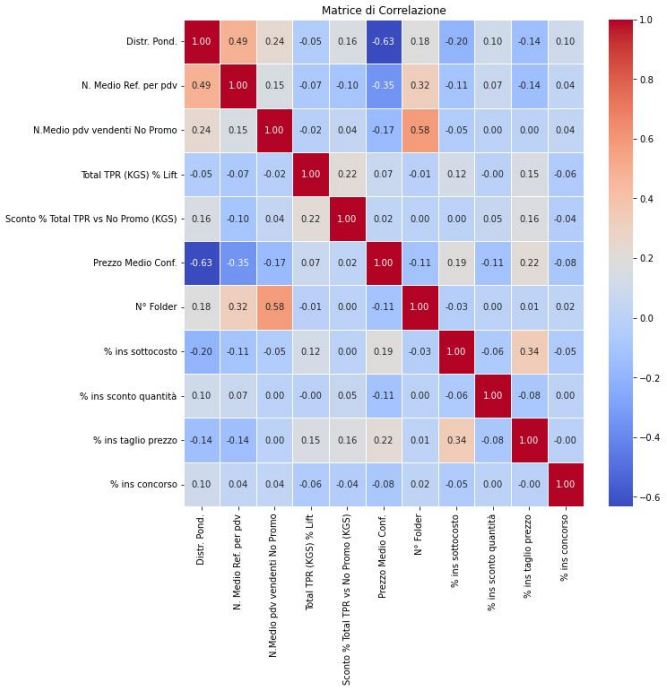
I risultati della covarianza per le variabili di interesse vengono descritti e riassunti dalla tabella di seguito.

Figura 13: Matrice di Covarianza



La matrice sopra riportata, mostra la presenta di due valori elevati presenti nella diagonale della tabella. Questi dati spiegano come ci sia un'alta varianza per le variabili "N. medio pdv vendenti No promo" e "N° folder", il che dimostra che le feature hanno una grande variabilità, o, in altre parole, presentano una grande dispersione dei valori della variabile intorno alla media. In generale, la matrice di covarianza ci consente di affermare che le feature considerate non presentano degli alti valori di covarianza. Quindi, non c'è una forte relazione lineare tra le variabili. Questa prima diagnosi dei dati è confermata anche dalla matrice di correlazione (*figura 14*).

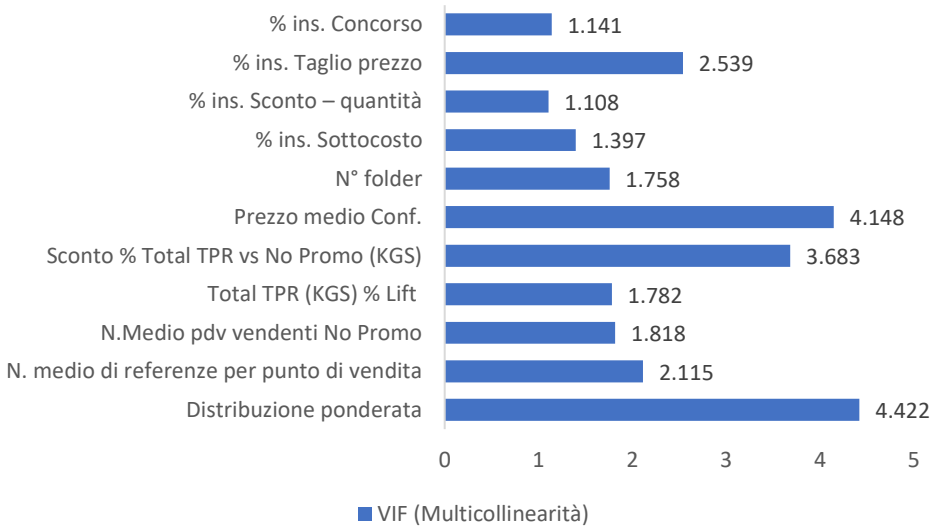
Figura 13: Matrice di Correlazione



Come riassunto dalla *figura 13*, non esiste una forte relazione, positiva o negativa, tra le variabili. Nonostante ciò, è evidente che, in alcuni casi, esiste una relazione moderata e positiva tra due variabili. In questo caso, le correlazioni (vicine a 0.5) indicano che esiste una certa associazione tra le due variabili, ma non è così forte da indicare una dipendenza lineare molto intensa. Questo può significare che le variabili sono in parte correlate, ma esistono anche altre influenze o variabilità non spiegate dalla relazione lineare.

Per ottenere una comprensione più completa, l'analisi è stata approfondita attraverso la verifica della multicollinearità.

Figura 14: Multicollinearità



Questa verifica ha confermato che non sussiste un problema di multicollinearità, pertanto possiamo avanzare con maggiore fiducia nella costruzione del modello. L'assenza di multicollinearità significativa implica che le variabili selezionate sono sufficientemente indipendenti, riducendo il rischio di stime distorte dei coefficienti di regressione e migliorando la stabilità e l'interpretabilità del modello.

4.3 PERFORMANCE DEI MODELLI

I risultati ottenuti dai modelli predittivi alimentati con il dataset finale, composto da 11 feature ed una variabile target, sono stati riassunti di seguito nella *tabella I*.

Tabella I: Comparazione delle performance dei modelli di machine learning

Algoritmi	Mean squared error	Mean absolute error	R2 score	Root Mean Squared Error (RMSE)
Decision Tree Regressor	14,590	0,874	0,904	3,819
Random Forest Regression	5,319	0,580	0,965	2,306
Extra Trees Regression	3,403	0,428	0,978	1,845
Linear Regression	17,230	1,690	0,887	4,151
Gradient Boosting	4,473	0,607	0,970	2,115
Support Vector Machine (SVM)	87,373	1,819	0,431	9,347
K-Nearest Neighbors (K-NN)	5,992	0,583	0,960	2,447
Stepwise Regression	17,578	1,649	0,885	4,192

Dai risultati presentati nella Tabella I, emerge con chiarezza che l'algoritmo Extra Tree Regression si distingue come il più performante per la previsione dei volumi di vendita di un produttore di pasta di semola, quando la previsione è alimentata da variabili direttamente collegate alle leve operative del trade marketing. L'Extra Tree Regression, con un valore di R^2 pari a circa 0.978, dimostra di spiegare il 97.8% della variabilità totale dei dati. Questo risultato è estremamente positivo e suggerisce una capacità predittiva eccellente del modello, in quanto è indice di una forte capacità di adattamento ai dati e di una significativa capacità del modello di catturare quasi tutta la variabilità presente nei dati. Inoltre, l'errore medio quadratico (RMSE), prossimo al valore 1, indica che il modello presenta un errore medio molto basso, confermando così la precisione nelle previsioni e la qualità nella rappresentazione delle relazioni tra le variabili indipendenti e la variabile dipendente.

Tuttavia, è importante osservare che molti degli algoritmi selezionati durante l'analisi previsionale sono risultati altamente performanti e in grado di fornire previsioni di alta precisione. Infatti, la

maggior parte dei modelli analizzati riesce a spiegare una percentuale molto elevata della variabilità dei dati, con l'unica eccezione del Support Vector Machine (SVM), che è l’unico modello a presentare un R^2 inferiore all'85%.

Questa chiave di lettura ci fornisce un insight fondamentale, in quanto evidenziano in maniera significativa la qualità dei dati e la pertinenza delle feature. L’ottimo risultato è quindi frutto di un’ottimo lavoro fatto durante la fase di feature selection, in cui le tecniche di selezione applicate sono state efficaci a catturare i fattori determinanti delle vendite. In altre parole, le variabili operative del trade marketing sono state adeguatamente identificate e integrate nel modello, permettendo ai modelli di regressione di spiegare con precisione la variabilità dei volumi di vendita.

Tra i modelli di regressione analizzati, i modelli basati su metodi di ensemble come Extra Trees Regression e Random Forest Regression si distinguono per la loro performance superiore. Questi modelli, con R^2 che raggiungono rispettivamente 0,978 e 0,965, hanno dimostrato una notevole capacità di adattamento ai dati. Questi risultati suggeriscono che i modelli di ensemble sono particolarmente adatti a catturare e modellare le relazioni complesse presenti nei dati di vendita, offrendo previsioni più accurate e affidabili. Al contrario, il Support Vector Machine (SVM) ha mostrato performance significativamente inferiori, con un R^2 di 0,431 e un RMSE di 9,347, indicando di non essere in grado di gestire efficacemente la variabilità dei dati, probabilmente a causa di caratteristiche non lineari o di rumori nel dataset che non sono stati adeguatamente gestiti dal modello. Perciò, il modello SVM non è la scelta ottimale per la previsione dei volumi di vendita. Invece, i modelli tradizionali come la Linear Regression e la Stepwise Regression offrono comunque previsioni valide ma la loro capacità di spiegare la variabilità dei dati è inferiore rispetto ai modelli più avanzati.

4.4 IMPLICAZIONI MANAGERIALI

In termini commerciali e manageriali, questi risultati hanno implicazioni significative. Primo su tutto perché ci spiegano quanto è forte l’impatto delle attività di trade marketing sulle vendite finali. Mentre, in secondo luogo, l'adozione di modelli avanzati permette una comprensione più profonda delle dinamiche di vendita e una previsione più accurata delle performance future.

Questo può tradursi in una pianificazione più efficace delle strategie di trade marketing, ottimizzazione delle campagne promozionali e una gestione più precisa delle scorte. Le aziende che

riescono a sfruttare al meglio le tecniche di modellazione avanzata possono ottenere un vantaggio competitivo significativo, migliorando la loro capacità di anticipare le tendenze del mercato e di adattarsi alle variazioni della domanda con maggiore agilità.

Grazie ai risultati ottenuti da questo studio, i professionisti del trade marketing e l'intera unità commerciale delle aziende che operano nel mercato della pasta di semola, e, in generale, nel settore dei beni di largo consumo, dispongono ora di un framework avanzato per implementare strategie altamente sofisticate e personalizzate.

Questo studio si configura come una risorsa strategica per la definizione di modelli commerciali data-driven, supportando le aziende con tecniche analitiche avanzate. Ciò avviene attraverso una serie di tecniche di machine learning (su tutte, l'Extra Tree Regression) in grado di fornire previsioni altamente accurate sull'andamento delle vendite.

L'integrazione di tali tecnologie rappresenta un'evoluzione significativa verso una gestione commerciale basata sui dati, orientata all'ottimizzazione delle performance e alla riduzione dell'incertezza decisionale. Con l'adozione delle informazioni e degli insights derivanti da questa analisi, i trade marketeer saranno a conoscenze delle tecniche di machine learning in grado di rendere più efficaci i processi decisionali legati alle strategie promozionali, alla gestione distributiva e alle politiche di pricing. Infatti le previsioni generate consentiranno di misurare e prevedere l'effetto delle campagne promozionali sulle vendite, permettendo una pianificazione mirata e l'allocazione efficiente del budget. Questo comporta un miglioramento del ROI e una maggiore precisione nel targeting delle promozioni. Ancora, le indicazioni fornite dai modelli predittivi potranno supportare una migliore pianificazione delle scorte, ottimizzando la distribuzione, aumentando l'efficienza operativa e la soddisfazione del cliente. Il tutto, potrà inoltre essere filtrato per area geografica (es. solo per il Lazio, o per la Campania, Lombardia ecc…) e per periodo di riferimento (es. totale anno o per trimestre).

Inoltre, rappresenta anche un supporto fondamentale alle attività di stima di budget. Infatti, come evidenziato nel *capitolo 1.2*, la possibilità di prevedere i volumi di vendita al *sell-out* consente di stimare in modo più accurato i volumi al *sell-in*. Questo approccio permette ai responsabili del budget di sviluppare previsioni più solide e razionali, riducendo l'elevato grado di incertezza tipico di queste attività.

Possiamo concludere quindi che questa trasformazione, resa possibile grazie all’uso intelligente dei dati e all’applicazione di tecniche di apprendimento automatico, non solo innalza la qualità dei processi appena spiegati, ma offre anche un approccio più flessibile e resiliente per affrontare le sfide future del mercato.

4.5 LIMITI E RICERCHE FUTURE

Nonostante i numerosi risultati positivi emersi dallo studio, è necessario riconoscere l'esistenza di alcune limitazioni intrinseche che possono influenzare l'interpretazione dei dati e delle conclusioni tratte. In particolare, l'andamento delle vendite di beni di largo consumo, come la pasta di semola, non è determinato unicamente dalle iniziative e dalle strategie messe in atto dalle aziende operanti nel mercato, ma risente anche di una serie di fattori esterni, spesso imprevedibili, che sfuggono al controllo diretto delle imprese.

Un esempio emblematico di questi fattori esterni è rappresentato dall'impatto della pandemia da COVID-19, che ha sconvolto le dinamiche di mercato a livello globale, provocando profonde perturbazioni nelle catene di approvvigionamento, mutamenti significativi nei comportamenti di consumo e un contesto di incertezza economica generalizzata. A ciò si aggiunge la recente instabilità geopolitica, come i conflitti armati in diverse regioni del mondo, che hanno contribuito a un'impennata dei costi delle materie prime, dell'energia e dei trasporti, con conseguenti riflessi sui prezzi finali al consumatore. L'aumento dei costi operativi derivante da tali dinamiche ha inciso negativamente sulla capacità di previsione delle vendite, distorcendo le proiezioni formulate nel breve e medio termine.

Oltre a questi macro-fattori di natura globale, vi sono specificità locali che giocano un ruolo altrettanto rilevante nel determinare l'andamento delle vendite di beni di largo consumo, in particolare nel contesto italiano. In questa analisi, attraverso la scomposizione dei dati per regione geografica, si propone di permettere al modello di machine learning di riuscire a trovare pattern nascosti all’interno dei dati, che potrebbero rivelare informazioni fondamentali sulle peculiarità locali di acquisto e di consumo. Tuttavia, sebbene questa suddivisione per regioni fornisca una visione più dettagliata, la vista per singola regione deve essere interpretata come una sintesi, e non come una rappresentazione esaustiva delle complessità locali. La segmentazione territoriale adottata nel modello di machine learning è utile per cogliere variazioni su scala regionale, ma non può

catturare appieno le micro-dinamiche commerciali che avvengono all'interno di specifiche aree metropolitane o rurali.

Le strategie commerciali e di trade marketing di un'azienda sono differenziate in base alle specificità delle diverse aree di mercato in cui essa opera

L’approccio adottato nella scomposizione per regione rappresenta un tentativo valido di adattare il modello di machine learning alle specificità del mercato italiano. Per massimizzare l'efficacia predittiva del modello, in futuro, sarebbe opportuno svolgere un’integrazione con dati di natura qualitativa e con una conoscenza approfondita del contesto locale, tenendo conto delle particolarità culturali, economiche e politiche che influenzano le scelte di acquisto dei consumatori in ciascuna regione. Tuttavia questo dettaglio richiederebbe una particolare attenzione, in quanto è essenziale mantenere un equilibrio tra l'affinamento dei parametri locali e la consapevolezza delle limitazioni insite in ogni processo di modellizzazione. Il rischio infatti è quello di complicare troppo il dataset con variabili che possono distorcere ulteriormente il quadro e complicare la capacità del modello di prevedere accuratamente i trend futuri.

Per quanto riguarda le potenziali opportunità di sviluppo futuro, lo studio apre la strada a una serie di implementazioni che potrebbero perseguire lo stesso obiettivo, ma con protagonisti e approcci differenti. In particolare, mentre l'analisi attuale si è concentrata su una categoria merceologica specifica, come la pasta di semola, sarebbe interessante orientare le future ricerche verso un livello di dettaglio maggiore, focalizzandosi su una singola referenza all'interno della categoria. Un'analisi più specifica su un singolo prodotto consentirebbe di ottenere una visione più granulare delle dinamiche di mercato, catturando informazioni che potrebbero essere oscurate in un'analisi a livello di categoria. Ciò potrebbe rivelarsi particolarmente utile per comprendere meglio come ottimizzare le strategie commerciali e le potenzialità di vendita di quella referenza.

Inoltre, un’altra direzione di sviluppo potrebbe essere condurre l’analisi su un singolo cliente nazionale di grande rilevanza per il mercato (es. Conad, Esselunga, Coop). Questo approccio personalizzato potrebbe offrire un quadro più dettagliato delle opportunità di crescita combinate combinate tra produttore e distributore, permettendo di affinare ulteriormente le strategie commerciali e di trade marketing.

5 CONCLUSIONI

Questo studio ha mirato ad esplorare e ottimizzare le previsioni di vendita nel mercato della pasta di semola attraverso l'applicazione di algoritmi di machine learning. L'analisi è stata condotta utilizzando un dataset comprensivo di 51864 valori numerici, relativi a KPI strettamente legati alle 3 principali leve del trade marketing (prezzo, promozioni e distrubizione & assortimento).

Tra gli algoritmi testati, l'Extra Trees Regression si è dimostrato il più performante, con un R^2 di 0,978, indicando che il modello spiega il 97,8% della variabilità dei dati. Questo risultato sottolinea una capacità predittiva eccellente, con un errore medio quadratico (RMSE) molto basso di 1,845. L'Extra Trees Regression ha superato altri modelli, come il Random Forest Regression e il Gradient Boosting, che, sebbene anch'essi altamente performanti, non hanno raggiunto la stessa precisione predittiva. Al contrario, il modello di Support Vector Machine (SVM) ha mostrato una performance significativamente inferiore, con un R^2 di 0,431 suggerendo limitazioni nella gestione delle variabilità dei dati.

Le risposte alle domande di ricerca presentate nel capitolo della metodologia hanno confermato che l'implementazione di modelli avanzati di machine learning offre una previsione più precisa e dettagliata, evidenziando come le variabili di trade marketing influenzano significativamente le performance di vendita.

I risultati di questo studio apportano un significativo contributo alla letteratura esistente sull'analisi predittiva delle vendite nel settore FMCG. L'utilizzo di tecniche di machine learning avanzate, arricchisce il corpo teorico riguardante l'accuratezza delle previsioni di vendita e l'efficacia delle leve di trade marketing. Questo studio conferma l'importanza di integrare metodologie quantitative per una comprensione più profonda dei fenomeni commerciali e sottolinea il valore delle tecniche di ensemble nella modellazione delle vendite.

Dal punto di vista pratico, i risultati ottenuti offrono alle aziende del settore dei beni di largo consumo, in particolare nel mercato della pasta di semola, strumenti analitici avanzati per ottimizzare le strategie commerciali. Le implicazioni pratiche includono una pianificazione più mirata delle campagne promozionali, una gestione più precisa delle scorte e una personalizzazione delle offerte basata su previsioni di vendita altamente accurate. L'adozione di modelli come l'Extra Trees Regression consente una maggiore precisione nella previsione delle vendite e nella valutazione dell'impatto delle attività di trade marketing, migliorando così il ritorno sugli investimenti e la competitività aziendale.

Sulla base dei risultati ottenuti, quindi, si raccomanda di adottare modelli avanzati di machine learning per la stima delle vendite e la pianificazione strategica. È essenziale integrare tecniche di ensemble come l’Extra Trees Regression, che ha dimostrato una notevole accuratezza e capacità di adattamento ai dati.

Le aziende dovrebbero considerare l'implementazione di un framework analitico che consenta di eseguire previsioni granulari basate su dati segmentati per area geografica e periodo di riferimento. Questa segmentazione permette una pianificazione più precisa e una migliore allocazione delle risorse. Infine, è importante investire nella formazione del personale per l’uso efficace delle tecniche di apprendimento automatico, migliorando la capacità decisionale e la reattività alle dinamiche di mercato.

Le scoperte di questo studio evidenziano l’importanza di un approccio data-driven nella gestione commerciale. La trasformazione verso una gestione basata su dati analitici rappresenta un’evoluzione cruciale per affrontare le sfide future del mercato, offrendo alle aziende strumenti avanzati per migliorare le loro performance e la loro competitività.

BIBLIOGRAFIA

- Marshall Fisher, Ananth Raman, & Anna McClelland (2019). How Zara Uses Operations Research to Reengineer Its Global Distribution Process. Harvard Business Review
- Christopher, M. (2016). Logistics & Supply Chain Management (5th ed.). Pearson.
- W.K. Wong n , Z.X. Guo (2010). A hybrid intelligent model for medium-term sales forecasting in fashion retail supply chains using extreme learning machine and harmony search algorithm. International Journal of Production Economics
- Kotler, P., & Keller, K. L. (2020). Marketing Management (15th ed.). Pearson
- Alexander René Schmutzler (2016). Organizing Trade Marketing, Category Management, and Shopper Marketing. Otto Beisheim School of Management
- Dewsnap, Belinda and David Jobber (2009), "An Exploratory Study of Sales-Marketing Integrative Devices", European Journal of Marketing, 43 (7/8), 985–1007
- Cespedes, Frank V. (1993), "Coordinating Sales and Marketing in Consumer Goods Firm", Journal of Consumer Marketing, 10 (2), 37–55.
- Cespedes, Frank V. (1995), Concurrent Marketing – Integrating Product, Sales, and Service, Boston: Harvard Business School Press.
- Davies, Gary (1993), Trade Marketing Strategy, London: P. Chapman Pub.
- Davies, Gary (1994), "Maintaining Relationships with Retailers", Journal of Strategic Marketing, 2 (3), 189–210.
- NIQ (nielseniq.com) (2024). Il Barometro dei Consumi di NIQ: nel 2023 le famiglie italiane hanno speso 187 miliardi di euro in beni di consumo
- Dewsnap, Belinda and David Jobber (2003), "Re-thinking Marketing Structures in the Fast Moving Consumer Goods Sector: An Exploratory Study of UK Firms", Bradford University School of Management Working Paper Series (Working Paper No 03/06).
- Randall, Geoffrey (1994), Trade Marketing Strategies – The Partnership between Manufacturers, Brands, and Retailers, 2nd ed., Oxford: Butterworth-Heinemann.
- Piercy, Nigel F. (1985), Marketing Organisation – An Analysis of Information Processing, Power, and Politics, London: George Allen & Unwin.
- Thain, Greg and John Bradley (2012), Store Wars – The Worldwide Battle for Mindspace and Shelfspace, Online and In-Store, 2nd ed., Chichester: Wiley.
- Zentes, Joachim (1989), "Trade Marketing: Eine neue Dimension in den Hersteller-Händler Beziehungen", Marketing: Zeitschrift für Forschung und Praxis, 4 (IV. Quartal), 224–229.
- Kotler, P., & Keller, K. L. (2016). Marketing Management. Pearson.

- Fornari, E., Fornari, D., Grandi, S., & Menegatti, M. (2013). The Influence of Retailing-Mix Levers on Private Label Market Share: The Case of the Italian FMCG Market. Journal of Retailing and Consumer Services, 20(8), 617-624.
- Shankar, V., & Bolton, R. N. (2004). An empirical analysis of determinants of retailer pricing strategy. *Journal of Marketing*, 68(3), 17-33.
- Raschka, S. and Mirjalili, V., 2017 - Python Machine Learning, 2nd Ed.- Packt Publishing, Birmingham
- Mitchell, T. M., 1997 - Machine Learning - McGraw Hill
- Domingos, Pedro. "A Few Useful Things to Know About Machine Learning." *Communications of the ACM*, vol. 55, no. 10, 2012, pp. 78-87
- Bajari, P., Nekipelov, D., Ryan, S. P., & Yang, M. (2015). Machine learning methods for demand prediction and sales forecasting. American Economic Review: Papers & Proceedings 2015, 105(5): 481-485
- Tenure Track Assistant Professor @SPOCRI UNIMC
- Emerson, M., Napoleão, V. G. (2023). Sales forecasting using machine learning algorithms. Revista Gestão e Secretariado (GeSec), São Paulo, SP, v. 14, n. 7, 2023, p. 11294-11308.
- Tarallo, E., Akabane, G. K., Shimabukuro, C. I., Mello, J., & Amancio, D. (2019). Machine learning in predicting demand for fast-moving consumer goods: An exploratory research. IFAC-PapersOnLine, 52(13), 737-742.
- J. Huber and H. Stuckenschmidt (2020). Daily retail demand forecasting using machine learning with emphasis on calendric special days. International Journal of Forecasting, 36, 1420–1438
- Raizada, S., & Saini, J. R. (2021). Comparative analysis of supervised machine learning techniques for sales forecasting. International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 12, No. 11
- Fornari, D. (2009). Trade marketing: Relazioni di filiera e strategie commerciali. Egea.
- Vishal Das, Tianyi Mao, Zhicheng Geng, Carmen Flores, Diego Pelloso, Fang Wang (2024). Enhancing sell-in and sell-out forecasting using ensemble machine learning method.
- Chopra, S., & Meindl, P. (2013). Supply chain management: Strategy, planning, and operation (5th ed.). Pearson Education, Inc.
- Ke, J., Zheng, H., Yang, H., & Chen, X. (2017). Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach. Transportation Research Part C: Emerging Technologies, 85, 591-608
- Asma ul Husna, Hassanzadeh Amin, S., & Ghasempoor, A. (2023). Demand forecasting using machine learning and deep learning approaches in the retail industry: A comparative study. SN Computer Science, 4(5), Article 392. SpringerLink

- Makridakis, S., & Hibon, M. (2000). The M3-Competition: results, conclusions and implications. International Journal of Forecasting, 16(4), 451-476
- Box, G. E., & Jenkins, G. M. (1970). Time Series Analysis: Forecasting and Control. Holden-Day.
- Fildes, R., & Goodwin, P. (2007). Good and bad judgment in forecasting: Lessons from four companies. Foresight: The International Journal of Applied Forecasting, 8(5), 5-10
- Chatfield, C. (2000). Time-series forecasting. CRC Press
- Kolkman, D. A., & Van Witteloostuijn, A. (2019). The non-parametric nature of various machine learning techniques: Flexibility in learning and adapting. Journal of Business Research, 104, 350-360
- Leo, M., Sharma, M., & Madan, I. (2019). Discovery of subtle insights in data through machine learning algorithms. International Journal of Scientific & Technology Research, 8(10), 904-907
- Ryll, L., & Seidens, M. (2019). Interactive process and structural adaptation in machine learning systems. Journal of Artificial Intelligence Research, 65, 725-742
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.
- Lantz, B. (2019). Machine learning with R: Expert techniques for predictive modeling (3rd ed.). Packt Publishing.
- Mohammed, S., Khan, M. B., & Bashier, E. B. M. (2016). Machine learning: Algorithms and applications. Springer
- Crane-Droesch, A. (2017). Machine learning applications in various domains including financial modeling and fraud detection. Annual Review of Financial Economics, 9, 101-121
- Enkono, M., & Suresh, K. P. (2020). Health and safety analysis using machine learning techniques. Journal of Healthcare Engineering
- Gholizadeh, M., Asadi, S., & Salehi, M. (2018). Application of machine learning in fraud detection. Journal of Financial Crime, 25(3), 641-655
- Sekban, S. (2019). Predictive modeling of sales performance using machine learning techniques. International Journal of Sales & Marketing Management Research and Development, 9(4), 25-34
- Tsoumakas, G. (2019). Forecasting consumer behavior and sales with machine learning. Journal of Consumer Behaviour, 18(5), 411-424
- Venishetty, V. K. (2019). Market demand prediction using machine learning algorithms. Journal of Market Research, 56(6), 901-920
- Raizada, S., & Saini, J. R. (2021). Comparative analysis of supervised machine learning techniques for sales forecasting. International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 12, No. 11
- Sujatha, G., & Sangeetha, R. (2021). A novel approach to identify the promotion of sales in supermarkets using machine learning algorithms. International Journal of Business Analytics & Intelligence, 9(1 & 2), 41-46.

- Rohaan, D., Topan, E., & Groothuis-Oudshoorn, C. G. M. (2021). Using supervised machine learning for B2B sales forecasting: A case study of spare parts sales forecasting at an after-sales service provider. Expert Systems with Applications, 185, 115565
- Chen, Z.-Y., Fan, Z.-P., & Sun, M. (2022). Machine learning methods for data-driven demand estimation and assortment planning considering cross selling and substitutions. Journal on Computing
- Martins, E., & Galegale, N. V. (2023). Sales forecasting using machine learning algorithms. GeSec
- Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. Belmont, CA: Wadsworth Publishing Company.
- Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324
- Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25(2), 197-227. https://doi.org/10.1007/s11749-016-0481-7
- Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning: Data mining, inference, and prediction. New York: Springer.
- Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451
- Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.
- Geurts, P., Demaret, L., & Goecks, J. (2006). Extremely randomized trees. Machine Learning, 63(1), 3-42. https://doi.org/10.1007/s10994-006-6226-1
- Raizada, S., & Saini, J. R. (2021). A survey of ensemble machine learning: Techniques, applications, and trends. Artificial Intelligence Review, 54(6), 4643-4687. https://doi.org/10.1007/s10462-020-09878-5
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297. https://doi.org/10.1007/BF00994018
- Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge: Cambridge University Press.
- Zhang, H. (2016). The optimal choice of k in k-nearest neighbor classifier. Journal of Machine Learning Research, 17(1), 1141-1150.
- Bhatia, N., & Vandana. (2010). K-nearest neighbor classifier: Review and applications. International Journal of Computer Applications, 1(9), 5-10. https://doi.org/10.5120/2222-2908
- Smith, M. (2018). Applied regression analysis: A second course in business statistics. New York: Wiley.
- Bajari, P., Nekipelov, D., Ryan, S. P., & Yang, M. (2015). Machine learning methods for demand estimation. American Economic Review, 105(5), 171-175. https://doi.org/10.1257/aer.p20151020

- Ailawadi, K. L., Neslin, S. A., & Gedenk, K. (2014). Retailing in the Age of Disruption. Cambridge University Press.
- Cadavid, J., Lamouri, S., & Grabot, B. (2019). The impact of promotional campaigns on sales forecasting using machine learning. Journal of Business Research, 98, 319-329
- Cespedes, F. V. (1993). Managing the Trade Marketing Function. Harvard Business School Press.
- Corstjens, M., & Corstjens, J. (1999). Store Wars: The Battle for Mindspace and Shelfspace. Wiley.
- Davies, B. (1993). The role of trade marketing in the changing landscape of retailing. Journal of Marketing, 57(3), 69-79
- Davies, B. (1994). Trade marketing: A review and agenda for future research. Marketing Intelligence & Planning, 12(4), 28-35
- Dewsnap, B., & Jobber, D. (2003). The role of trade marketing in the UK grocery market. European Journal of Marketing, 37(11/12), 1789-1809
- Dewsnap, B., & Jobber, D. (2004a). The use of trade marketing by manufacturers in the UK grocery market. Journal of Business Research, 57(4), 1-12
- Dewsnap, B., & Jobber, D. (2004b). The evolution of trade marketing. Marketing Review, 4(3), 291-312
- Dewsnap, B., & Jobber, D. (2009). Trade marketing and its role in the development of trade marketing strategies. European Journal of Marketing, 43(11/12), 1564-1584
- Kotler, P., & Armstrong, G. (2018). Principles of Marketing (17th ed.). Pearson.
- John R. Anderson, E. K., & Roberts, E. K. (2017). The impact of trade marketing on sales performance: A study of the UK grocery sector. Journal of Retailing and Consumer Services, 38, 64-72
- P. J. Johnson, & H. R. Thompson. (2019). Key performance indicators and trade marketing effectiveness. Journal of Marketing Analytics, 7(1), 45-60
- Sethuraman, R., Gázquez-Abad, J. C., & Martínez-López, F. J. (2022). Evaluating the effectiveness of trade marketing: An empirical analysis. European Journal of Marketing, 56(3), 672-694
- Thain, S., & Bradley, S. D. (2012). Trade Marketing: A Comprehensive Guide. Routledge.
- Zentes, J. (1989). The impact of trade marketing on retail performance. Journal of Retailing and Consumer Services, 10(4), 281-289
- McKinney, W. (2012). Python for Data Analysis: data wrangling with pandas, numpy, and ipython. O'Reilly
- Khan, S. I., & Hoque, A. S. M. L. (2020). SICE: An improved missing data imputation technique. Journal of Big Data, 7(1), 1-25
- Joel, L.O., Doorsamyha, W., Paul, B.S. On the Performance of Imputation Techniques for Missing Values on Healthcare Datasets

- Huber P. J. (1964). Robust Estimation of a Location Parameter. Breakthroughs in Statistics – Springer, 479–491
- Raizada, S., & Saini, J. R. (2021). Comparative analysis of supervised machine learning techniques for sales forecasting. International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 12, No. 11
- Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
- G. Sujatha, R. Sangeetha (2021). A Novel Approach to Identify the Promotion of Sales in Supermarkets using Machine Learning Algorithms. International Journal of Business Analytics & Intelligence 9 (1 & 2) 2021, 41-46
- Bajari, P., Nekipelov, D., & Waehrer, K. (2015). "Machine Learning Methods for Demand Estimation and Pricing Optimization." Handbook of Econometrics, 7A, 405-477