ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Alma Mater Studiorum - Università di Bologna

Dipartimento di Informatica - Scienza e Ingegneria

Corso di Laurea in Informatica per il Management

Tecniche di Machine Learning per l'analisi dell'impatto della pubblicità sulle vendite

Candidato:
Carlos Vasco Chironda

Relatrice:
Chiari.ma Prof.ssa Elena Piccolomini

Anno Academico 2023 - 2024

A mia madre...

Indice

INTRODUZIONE

In un contesto economico sempre più globale e competitivo, l'importanza della pubblicità come leva per il successo aziendale è cresciuta in modo esponenziale. Oggi, le aziende si trovano a fronteggiare la necessità di massimizzare i profitti e ottimizzare i costi, e la pubblicità si è affermata come uno degli strumenti principali per raggiungere questi obiettivi. Tuttavia, non tutte le campagne pubblicitarie sono uguali: la tipologia, il canale e la modalità di comunicazione possono avere impatti differenti sulle vendite

Questo elaborato si propone di analizzare in che modo le diverse forme di pubblicità influenzano le performance di vendita, con un'attenzione particolare alle tecniche di machine learning. Utilizzando modelli predittivi avanzati, si cercherà di identificare quali fattori pubblicitari abbiano un impatto significativo sulle vendite e come le aziende possano ottimizzare le loro strategie di marketing .

Nell'era digitale, i canali pubblicitari si sono moltiplicati e diversificati: televisione, radio, giornali, e soprattutto i social media hanno trasformato il modo in cui le aziende interagiscono con i consumatori. I social media, in particolare, hanno aperto nuove opportunità, ma anche sfide, per le aziende che devono capire come integrarli efficacemente nei loro piani strategici . L'uso strategico e ottimizzato della pubblicità non solo aiuta a raggiungere una più ampia base di clienti, ma può anche conferire un vantaggio competitivo significativo.

In questo contesto, il machine learning emerge come uno strumento prezioso per supportare le decisioni aziendali. Grazie alla capacità di analizzare grandi volumi di dati e di identificare pattern nascosti, l'intelligenza artificiale permette di prevedere l'impatto delle campagne pubblicitarie e di ottimizzare gli investimenti. In questo studio, esploreremo come i modelli di machine learning possano essere applicati per determinare quali canali pubblicitari siano più efficaci nel migliorare

le vendite, fornendo alle aziende strumenti concreti per migliorare le loro performance

L'obiettivo finale è fornire un quadro chiaro e pratico su come le aziende possano utilizzare questi modelli predittivi per prendere decisioni basate sui dati, massimizzando così l'efficacia delle loro strategie pubblicitarie.

Inoltre, sarà fornita una panoramica della letteratura rilevante sull'argomento per contestualizzare la ricerca e giustificare le scelte metodologiche adottate.
Il lavoro si sviluppa nei capitoli successivi. Il primo capitolo presenta il contesto in cui si inserisce la ricerca, con una descrizione degli obiettivi principali e delle motivazioni che hanno guidato lo studio. Il secondo capitolo descrive l'approccio metodologico adottato, con particolare attenzione alle fasi di pre-processing dei dati, che sono fondamentali per garantire l'affidabilità dei modelli di machine learning utilizzati. Nel terzo capitolo vengono illustrati i modelli di machine learning impiegati nell'analisi, con una discussione sulle loro caratteristiche principali e sull'applicazione pratica nella valutazione dell'impatto della pubblicità. Infine, il quarto capitolo presenta i risultati ottenuti, fornendo un'analisi dei dati e una riflessione sulle implicazioni pratiche di questi risultati per le aziende.

Capitolo 1

CONTESTO E OBIETTIVI

Nel panorama economico moderno, le aziende affrontano la sfida di competere in un mercato sempre più dinamico e globalizzato, dove le decisioni strategiche, come la pianificazione pubblicitaria, possono determinare il successo o il fallimento di un'organizzazione. La pubblicità rappresenta una leva fondamentale per raggiungere obiettivi di crescita, ottimizzare i costi e incrementare il volume delle vendite, ma non tutti i media pubblicitari contribuiscono allo stesso modo. Comprendere l'impatto specifico di ciascun canale (come TV, radio o giornali) è cruciale per massimizzare l'efficacia degli investimenti e ottenere un ritorno positivo.

Con la crescente disponibilità di dati, gli strumenti di machine learning hanno trasformato l'approccio all'analisi e alla previsione del comportamento delle vendite. Grazie alla loro capacità di analizzare grandi quantità di dati, identificare pattern nascosti e generare previsioni accurate, queste tecnologie permettono alle aziende di passare da decisioni intuitive a decisioni basate sui dati. Questo elaborato si inserisce in questo contesto e si propone di fornire un contributo concreto per rispondere alle seguenti domande chiave.

1.0.1 Research Questions (Domande di Ricerca)

1. **Quali media contribuiscono maggiormente alle vendite?**
   → Questa domanda mira a individuare l'efficacia relativa di ciascun canale pubblicitario (TV, radio e giornali) nel promuovere le vendite, distinguendo tra contributi significativi e trascurabili.

2. **Quanto potrebbero aumentare le vendite in relazione a un aumento dell'investimento pubblicitario?**
   → L'obiettivo è quantificare il ritorno incrementale sulle vendite derivante da un aumento degli investimenti pubblicitari in uno o più media.

3. **Con quale precisione possiamo stimare l'effetto di ciascun mezzo sulle vendite?**
   → La domanda si focalizza sulla capacità dei modelli di machine learning di stimare con accuratezza l'influenza di ciascun canale pubblicitario sulle vendite, anche tenendo conto delle possibili interazioni tra i media.

4. **Con quale precisione possiamo prevedere le vendite future basandoci sui dati storici degli investimenti pubblicitari?**
   → Questa domanda esplora la componente predittiva del modello, valutando la capacità del machine learning di fornire stime affidabili sulle vendite future.

5. **Quali strategie di allocazione ottimale degli investimenti pubblicitari possono essere derivate dai risultati?**
   → Una volta identificato l'impatto di ciascun canale, ci si interroga su come le aziende possano distribuire efficacemente il budget pubblicitario per massimizzare il ritorno sull'investimento.

[7]

1.0.2 Analisi dell'attuale contesto economico e di mercato

La pubblicità gioca un ruolo fondamentale nell'economia globale, fungendo da indicatore dell'andamento industriale e delle tendenze sociali. Previsioni che indicano un aumento fino a 754,4 miliardi di dollari nel 2024. Questo rappresenta un incremento del 5% rispetto al 2023, dimostrando un legame stretto tra la crescita del settore pubblicitario e l'andamento generale dell'economia globale. Non solo la spesa pubblicitaria sta superando la crescita del PIL globale, ma il settore si sta anche adattando a nuove dinamiche di consumo e comportamenti degli utenti, segnati dall'accelerazione della digitalizzazione e dall'adozione di nuove tecnologie.

Dominanza del digitale

Il digitale continua a dominare il panorama pubblicitario, influenzando profondamente le strategie delle aziende. Nel 2020, la pubblicità digitale ha registrato una crescita eccezionale del 13,2%, attestandosi a 290,1 miliardi di dollari. Le piattaforme di social media e i motori di ricerca, in particolare, hanno giocato un ruolo cruciale, contribuendo a più della metà della spesa pubblicitaria globale. Le proiezioni indicano che questo trend non solo si manterrà, ma crescerà ulteriormente: ad esempio, le entrate pubblicitarie di Alphabet (proprietaria di Google) sono previste in aumento del 10,5%, mentre Facebook si prevede cresca del 19%.

Le aziende stanno investendo sempre di più in campagne pubblicitarie su piattaforme digitali per raggiungere un pubblico sempre più connesso e mobile. Questa transizione è stata accelerata dalla pandemia di COVID-19, che ha modificato le abitudini di consumo e ha reso la pubblicità digitale un canale essenziale per le aziende che cercano di mantenere la loro visibilità e interazione con i clienti .

Settori in crescita

Particolare attenzione va data ai settori che beneficeranno di questa crescita. Le proiezioni indicano che il settore dei viaggi e trasporti crescerà dell'8,1%, un segno di ripresa post-pandemia, mentre il settore media e intrattenimento vedrà un aumento del 6,5%. Queste crescite sono in parte attribuibili agli eventi globali, come le Olimpiadi di Parigi e le elezioni presidenziali negli Stati Uniti, che tradizionalmente generano un aumento della spesa pubblicitaria.

Inoltre, la pubblicità televisiva, che ha storicamente dominato il mercato, sta subendo una trasformazione. Si prevede un incremento della spesa per la TV, specialmente per i canali connessi, con un aumento del 24,2%. Questo cambiamento riflette un'evoluzione nei gusti e nelle preferenze degli spettatori, che sempre più si orientano verso contenuti on-demand e piattaforme di streaming .

Media tradizionali vs. digitalizzazione

Nonostante la crescita del digitale, i media tradizionali non sono ancora scomparsi. Tuttavia, i loro tassi di crescita rimangono inferiori. La stampa, ad esempio, continua a contrarsi, con una previsione di

riduzione della spesa pubblicitaria. Questo fenomeno suggerisce una necessità per i media tradizionali di adattarsi a un ambiente in continua evoluzione, cercando di integrare strategie digitali per attrarre un pubblico più giovane e connesso.

Le aziende stanno investendo in esperienze pubblicitarie interattive, che incoraggiano l'engagement e la fidelizzazione del cliente, come testimoniano le campagne pubblicitarie innovative sui social media. Le campagne che sfruttano la realtà aumentata (AR) e la realtà virtuale (VR) sono sempre più comuni, offrendo agli utenti esperienze immersive che possono tradursi in un maggiore coinvolgimento e conversioni.

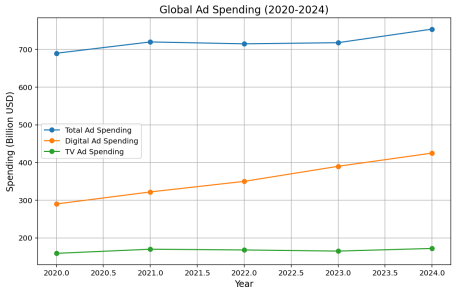

Figura 1.1: Global ad spending and its composition.

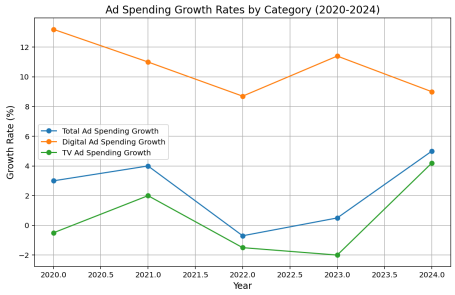

Figura 1.2: Growth rates in different categories.

In conclusione, il panorama della spesa pubblicitaria globale sta subendo un cambiamento radicale. Le aziende devono rimanere agili e pronte ad adattarsi a queste nuove dinamiche. La continua evoluzione delle tecnologie digitali, insieme alla necessità di rispondere rapidamente ai cambiamenti nelle preferenze dei consumatori, sarà cruciale per il successo nel panorama competitivo.

Con l'avvento di nuove tecnologie e piattaforme, i marketer hanno ora a disposizione strumenti più sofisticati per monitorare e analizzare l'efficacia delle loro campagne, consentendo una personalizzazione e una segmentazione del pubblico senza precedenti. Questo non solo migliora l'efficacia delle campagne pubblicitarie, ma offre anche un ritorno sugli investimenti più significativo.

L'attenzione per la sostenibilità e la responsabilità sociale sta diventando sempre più centrale nella strategia pubblicitaria delle aziende. In un mondo sempre più consapevole delle problematiche ambientali e

sociali, le campagne pubblicitarie che riflettono questi valori possono attrarre un pubblico più ampio e generare maggiore fiducia.[1] [3]

1.0.3 Panoramica del dataset

Il dataset utilizzato in questo progetto è stato raccolto dalla piattaforma Kaggle, un ampio repository di dataset gratuiti e una comunità attiva per analisti di dati, scienziati e sviluppatori. Il dataset analizzato riguarda le spese pubblicitarie su diversi canali media e il loro impatto sulle vendite. È pensato per aiutare analisti di marketing, data scientist e responsabili delle strategie aziendali a comprendere come le varie forme di spesa pubblicitaria influenzino i ricavi di vendita, agevolando decisioni più informate nelle campagne di marketing.

Caratteristiche principali:

1. **TV**: Investimenti in campagne pubblicitarie televisive (in migliaia di dollari).
2. **Radio**: Investimenti in campagne pubblicitarie radiofoniche (in migliaia di dollari).
3. **Newspaper (Giornali)**: Investimenti in campagne pubblicitarie sui giornali (in migliaia di dollari).
4. **Sales (Vendite)**: Ricavi generati dalle campagne di vendita (in migliaia di dollari). [6]

Figura 1.3: Immagine media

Capitolo 2

Approccio Metodologico e Pre-Processing dei Dati

2.0.1 Classificazione dei problemi

Iniziamo con un esempio semplice. Supponiamo di essere consulenti statistici assunti da un cliente per fornire consigli su come aumentare le vendite di un determinato prodotto. Il set di dati Advertising (Pubblicità) consiste nelle vendite di quel prodotto in 200 diversi mercati, insieme ai budget pubblicitari per il prodotto in ciascuno di questi mercati in tre diversi media: TV, radio e giornali. I dati sono mostrati nella Figura 4. Non è possibile per il nostro cliente aumentare direttamente le vendite del prodotto. D'altra parte, essi possono controllare la spesa pubblicitaria in ciascuno dei tre media. Quindi, se determiniamo che c'è un'associazione tra pubblicità e vendite, possiamo istruire il cliente su come regolare i budget pubblicitari, aumentando così indirettamente le vendite. In altre parole, il nostro obiettivo è sviluppare un modello accurato che possa essere utilizzato per prevedere le vendite in base ai budget per i tre media.

In questo contesto, i budget pubblicitari sono variabili di input, mentre le vendite sono la variabile di output. Le variabili di input vengono tipicamente indicate con il simbolo X, con un pedice per distinguerle. Ad esempio, x_1 potrebbe essere il budget per la TV, x_2 il budget per la radio e x_3 il budget per i giornali. Gli input sono noti con diversi nomi, come predittori, variabili indipendenti, caratteristiche o talvolta solo variabili. La variabile di output—in questo caso, le vendite—viene spesso chiamata risposta o variabile dipendente, ed è tipicamente indicata con il simbolo Y. [4]

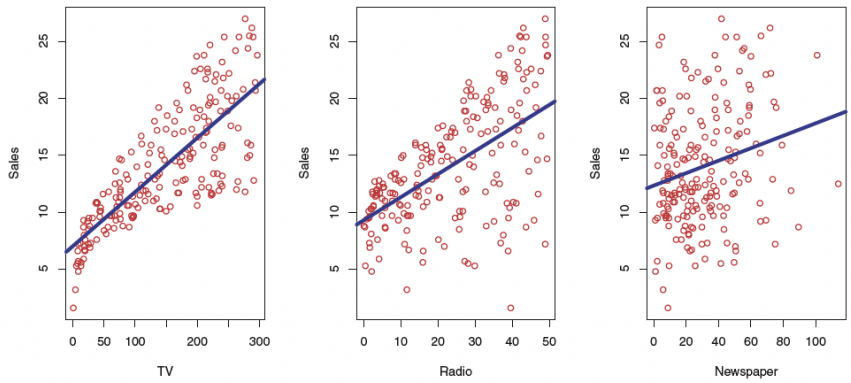

Figura 2.1: Il grafico mostra le vendite, in migliaia di unità, in funzione dei budget per TV, radio e giornali, in migliaia di dollari, per 200 mercati differenti. In ogni grafico mostriamo l'adattamento dei minimi quadrati semplici delle vendite a quella variabile. ciascuna linea blu rappresenta un modello semplice che può essere utilizzato per prevedere le vendite utilizzando rispettivamente TV, radio e giornali

2.0.2 Apprendimento Supervisionato vs non Supervisionato

La maggior parte dei problemi di apprendimento statistico rientra in una delle due categorie: **supervisionato** o non **supervisionato**. L'esempio che di cui discusso finora in questo capitolo appartiene al dominio dell'apprendimento supervisionato. Per ciascuna osservazione delle misurazioni dei predittori $x, \quad i = 1, \ldots, n$, c'è una corrispondente misurazione della risposta $y_1, y_2, \ldots, y_n$. Desideriamo adattare un modello che colleghi la risposta ai predittori, con l'obiettivo di prevedere accuratamente la risposta per future osservazioni (predizione) o comprendere meglio la relazione tra la risposta e i predittori (inferenza). Molti metodi classici di apprendimento statistico, come la regressione lineare e la regressione logistica , così come approcci più moderni come GAM, boosting e support vector machine, operano nel dominio dell'apprendimento supervisionato.

Al contrario, l'apprendimento non supervisionato descrive la situazione, un po' più sfidante, in cui per ogni osservazione $i, \quad i = 1, \ldots, n$, osserviamo un vettore di misurazioni xi, ma nessuna risposta associata yi. Non è possibile adattare un modello di regressione lineare, poiché non esiste una variabile risposta da prevedere. In questo contesto, in un certo senso stiamo lavorando alla cieca; la situazione è definita

non supervisionata perché ci manca una variabile risposta che possa supervisionare la nostra analisi. Che tipo di analisi statistica è possibile? Possiamo cercare di comprendere le relazioni tra le variabili o tra le osservazioni. Un metodo di apprendimento statistico che possiamo utilizzare in questo contesto è l'analisi dei cluster, o clustering. Lo scopo dell'analisi dei cluster è determinare, sulla base di $x, \quad i = 1, \ldots, n$, se le osservazioni appartengono a gruppi relativamente distinti. Per esempio, in uno studio di segmentazione del mercato, potremmo osservare diverse caratteristiche (variabili) per potenziali clienti, come il codice postale, il reddito familiare e le abitudini di acquisto. Potremmo credere che i clienti si suddividano in gruppi diversi, come grandi spendaccioni rispetto a piccoli spendaccioni. Se fossero disponibili informazioni sui modelli di spesa di ciascun cliente, allora sarebbe possibile un'analisi supervisionata. Tuttavia, queste informazioni non sono disponibili, cioè non sappiamo se ciascun potenziale cliente è un grande spendaccione o meno. In questo contesto, possiamo cercare di raggruppare i clienti sulla base delle variabili misurate, al fine di identificare gruppi distinti di potenziali clienti. Identificare tali gruppi può essere interessante perché potrebbe emergere che i gruppi differiscono rispetto a qualche proprietà di interesse, come le abitudini di spesa.

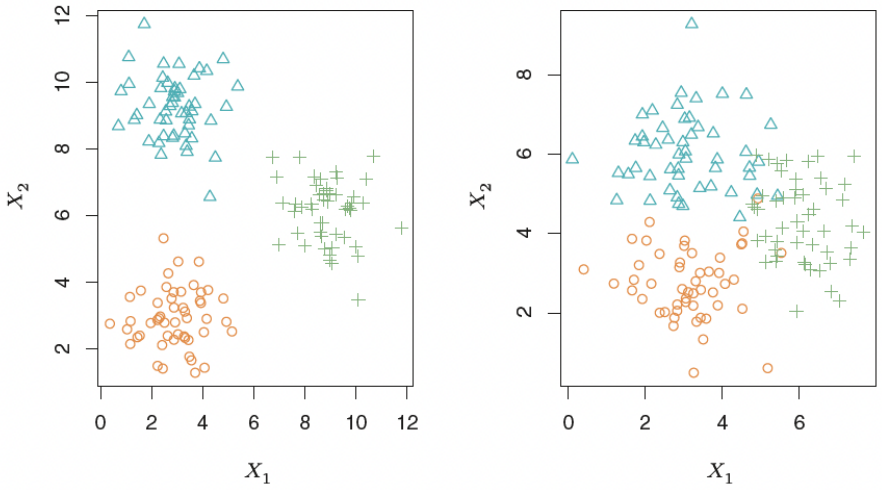

Figura 2.2: Queste due rappresentazioni illustrano come la separazione dei gruppi in un dataset di clustering influisca sulla difficoltà dell'analisi. In situazioni in cui i gruppi sono ben separati, il clustering è relativamente semplice, mentre in situazioni di sovrapposizione, l'analisi richiede metodi più sofisticati e precisi

La Figura 2.2 fornisce un'illustrazione semplice del problema del clustering. Vengono rappresentati 150 osservazioni con misurazioni su due variabili, X_1 e X_2. Ogni osservazione corrisponde a uno dei tre gruppi distinti. Per scopi illustrativi, si possono rappresentare i membri di ciascun gruppo usando colori e simboli diversi. Tuttavia, in pratica, le appartenenze ai gruppi sono sconosciute, e l'obiettivo è determinare il gruppo a cui ciascuna osservazione appartiene. Nel pannello a sinistra della Figura 5, questo è un compito relativamente facile perché i gruppi sono ben separati. Al contrario, il pannello a destra illustra un problema più difficile, in cui c'è una certa sovrapposizione tra i gruppi. Un metodo di clustering non potrebbe assegnare correttamente tutti i punti sovrapposti al loro gruppo (blu, verde o arancione). Negli esempi mostrati in Figura 5, ci sono solo due variabili, quindi si può semplicemente ispezionare visivamente i grafici a dispersione delle osservazioni per identificare i cluster. Tuttavia, in pratica, spesso ci troviamo di fronte a set di dati che contengono molte più di due variabili. In questo caso, non possiamo facilmente rappresentare graficamente le osservazioni. Per esempio, se ci sono P variabili nel nostro set di dati, allora possono essere realizzati $\frac{p(p-1)}{2}$ scatterplot distinti, e l'ispezione visiva non è un modo praticabile per identificare i cluster. Per questo motivo, i metodi di clustering automatizzati sono importanti.

Molti problemi rientrano naturalmente nei paradigmi di apprendimento supervisionato o non supervisionato. Tuttavia, a volte la questione se un'analisi debba essere considerata supervisionata o non supervisionata è meno chiara. Per esempio, supponiamo di avere un set di n osservazioni. Per m di queste osservazioni, dove m <n, abbiamo sia misurazioni dei predittori che una misurazione della risposta. Per le restanti $n - m$ osservazioni, abbiamo misurazioni dei predittori, ma nessuna misurazione della risposta. Tale scenario può verificarsi se i predittori possono essere misurati a basso costo, ma le risposte corrispondenti sono molto più costose da raccogliere. Definiamo questa situazione come un problema di apprendimento semi-supervisionato. In questo contesto, desideriamo utilizzare un metodo di apprendimento statistico che possa incorporare le m osservazioni per cui sono disponibili misurazioni della risposta, così come le $n - m$ osservazioni per cui non lo sono. Sebbene sia un argomento interessante, va oltre. [4]

[8]

2.0.3 Regressione vs Classificazione

Le variabili possono essere caratterizzate come quantitative o qualitative (note anche come categoriali). Le variabili quantitative assumono valori numerici. Esempi includono l'età, l'altezza o il reddito di una persona, il valore di una casa e il prezzo di un'azione. Al contrario, le variabili qualitative assumono valori in una delle K diverse classi o categorie. Esempi di variabili qualitative includono il genere di una persona (maschio o femmina), la marca del prodotto acquistato (marca A, B o C), se una persona è inadempiente su un debito (sì o no), o una diagnosi di cancro (Leucemia Mieloide Acuta, Leucemia Linfoblastica Acuta o Nessuna Leucemia). Tendiamo a riferirci a problemi con una risposta quantitativa come problemi di regressione, mentre quelli che coinvolgono una risposta qualitativa sono spesso chiamati problemi di classificazione. Tuttavia, la distinzione non è sempre così netta. La regressione lineare dei minimi quadrati viene utilizzata con una risposta quantitativa, mentre la regressione logistica è tipicamente usata con una risposta qualitativa (due classi, o binaria). Per questo motivo è spesso utilizzata come metodo di classificazione. Ma poiché stima le probabilità delle classi, può essere considerata anche un metodo di regressione. Alcuni metodi statistici, come i K-nearest neighbors e il boosting , possono essere utilizzati nel caso di risposte sia quantitative che qualitative. Tendiamo a selezionare i metodi di apprendimento statistico in base al fatto che la risposta sia quantitativa o qualitativa; cioè, potremmo usare la regressione lineare quando la risposta è quantitativa e la regressione logistica quando è qualitativa. Tuttavia, il fatto che i predittori siano qualitativi o quantitativi è generalmente considerato meno importante. La maggior parte dei metodi di apprendimento statistico possono essere applicati indipendentemente dal tipo di variabile predittore, a condizione che eventuali predittori qualitativi siano correttamente codificati prima che venga eseguita l'analisi.[4][9]

2.0.4 Pre-Processing dei dati

Il pre-processing dei dati è una fase fondamentale del processo di sviluppo di un modello di machine learning. In questa fase, vengono effettuate tutte le operazioni necessarie per pulire e trasformare i dati grezzi in un formato che possa essere facilmente utilizzato dagli algoritmi di machine learning. L'obiettivo principale è garantire che i dati siano privi di errori, incompleti o mal formattati, e che siano pronti per l'analisi successiva.

Pulizia dei dati:

```
# 1. Pulizia dei Dati (Data Cleaning)
# Controllo di eventuali valori nulli e rimozione delle righe con valori mancanti
print("Valori nulli prima del drop:")
print(df.isnull().sum())
df.dropna(inplace=True)
print("Valori nulli dopo il drop:")
print(df.isnull().sum())
"""
```

Figura 2.3: codice della pulizia del dataset

```
TV           0 | radio    0
newspaper    0 |sales     0
dtype: int64
Valori nulli dopo il drop:
TV           0  | radio    0
newspaper    0  |sales     0
```

Figura 2.4: output generato

Pulizia dei Dati - Gestione degli Outliers:

Durante la pulizia dei dati, è essenziale identificare e gestire gli outliers, cioè valori anomali che possono distorcere i risultati e ridurre l'accuratezza del modello. Per rilevare questi valori si utilizza l'Intervallo Interquartile (IQR): si calcolano i quantili Q1 (25° percentile) e Q3 (75° percentile), e l'IQR viene definito come la differenza tra Q3 e Q1. Gli outliers sono i valori che eccedono il range [Q1 - 1.5IQR, Q3 + 1.5IQR], e vengono rimossi per migliorare la qualità dei dati. [10]

```
# 3. Identificazione di outliers con l'uso di quantili
Q1 = df.quantile(0.25)
Q3 = df.quantile(0.75)
IQR = Q3 - Q1
#Rimozione  outliers(eventuali valori troppo distanti dal range interquartile)
df_clean = df[~((df < (Q1 - 1.5 * IQR)) | (df > (Q3 + 1.5 * IQR))).any(axis=1)]
print(f"Numero di righe dopo rimozione outliers:
+{df_clean.shape[0]} (rispetto a {df.shape[0]} iniziali)")
# 4. Standardizzazione delle feature
# Separiamo le feature e la variabile target
X = df_clean[['TV', 'radio', 'newspaper']]  # variabili indipendenti
y = df_clean['sales']  # variabile dipendente
# Split dei dati in training e test set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)
# Applicazione dello scaling (normalizzazione) per TV, radio e newspaper
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
# Visualizzazione delle prime righe del training set dopo la normalizzazione
print("Prime righe di X_train dopo scaling:")
print(X_train_scaled[:5])
```

Figura 2.5: input generato

```
Numero di righe dopo rimozione outliers:  +198 (rispetto a 200 iniziali)
Prime righe di X_train dopo scaling:
[[ 0.4267515   1.39195161 -1.35187881]
 [-1.62566033  1.7249443   2.15543509]
 [-0.10440616 -0.56604543 -0.94516438]
 [-0.87202216 -0.39954908 -0.81597273]
 [-1.43579476 -0.79914031 -0.01211361]]
```

Figura 2.6: output generato

Successivamente, si procede con la standardizzazione delle feature (ad esempio: TV, radio, newspaper), trasformandole in una scala con media 0 e deviazione standard 1. Questa fase garantisce che tutte le variabili abbiano lo stesso peso, fondamentale per algoritmi di machine learning che richiedono dati su scale comparabili per ottenere risultati accurati ed equi.

2.0.5 Analisi Esplorativa dei Dati (EDA)

Una volta che i dati sono stati puliti e trasformati, si esegue un'analisi esplorativa per comprendere meglio le relazioni tra le variabili e identificare pattern nascosti. In questa fase si utilizzano tecniche di visualizzazione (ad esempio, grafici a barre, scatter plot, heatmap) per verificare distribuzioni, correlazioni e tendenze. Queste analisi possono guidare eventuali modifiche aggiuntive nel pre-processing e aiutare nella scelta dell'algoritmo da utilizzare per il modello.

```
###################### Head ######################
      TV  radio  newspaper  sales
0  230.1   37.8       69.2   22.1
1   44.5   39.3       45.1   10.4
2   17.2   45.9       69.3    9.3
3  151.5   41.3       58.5   18.5
4  180.8   10.8       58.4   12.9
###################### Tail ######################
        TV  radio  newspaper  sales
195   38.2    3.7       13.8    7.6
196   94.2    4.9        8.1    9.7
197  177.0    9.3        6.4   12.8
198  283.6   42.0       66.2   25.5
199  232.1    8.6        8.7   13.4
###################### NA ######################
TV           0
radio        0
newspaper    0
sales        0
dtype: int64
###################### Description ######################
                TV       radio   newspaper       sales
count   200.000000  200.000000  200.000000  200.000000
mean    147.042500   23.264000   30.554000   14.022500
std      85.854236   14.846809   21.778621    5.217457
min       0.700000    0.000000    0.300000    1.600000
25%      74.375000    9.975000   12.750000   10.375000
50%     149.750000   22.900000   25.750000   12.900000
75%     218.825000   36.525000   45.100000   17.400000
max     296.400000   49.600000  114.000000   27.000000
###################### Quantiles ######################
              0.00    0.05    0.50     0.95     0.99   1.00
TV             0.7  13.195  149.75  280.735  292.907  296.4
radio          0.0   1.995   22.90   46.810   49.400   49.6
newspaper      0.3   3.600   25.75   71.825   89.515  114.0
sales          1.6   6.600   12.90   23.800   25.507   27.0
```

Figura 2.7: Missing Data: Non ci sono valori nulli (NA).Description: Le statistiche descrittive mostrano la media, la deviazione standard, il minimo, il massimo e i percentili per le variabili.

Matrice di correlazione

La matrice di correlazione rappresenta la relazione lineare tra le variabili presenti nel dataset. I valori della correlazione variano da -1 a 1. Un valore di correlazione vicino a 1 indica una forte correlazione positiva, ossia quando una variabile aumenta, anche l'altra tende ad aumentare. Al contrario, un valore vicino a -1 indica una forte correlazione negativa, dove all'aumento di una variabile corrisponde una diminuzione dell'altra. Un valore vicino a 0, invece, suggerisce che non vi è alcuna relazione lineare tra le variabili.

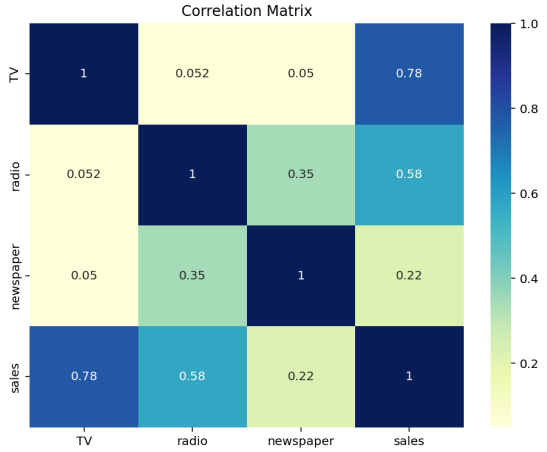

Figura 2.8: dal grafico si può vedere la forte relazione tra le variabile TV e sales.

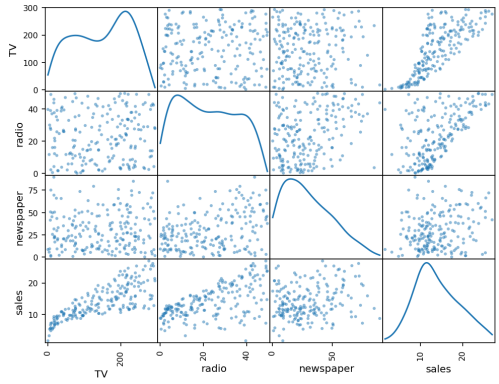

Figura 2.9: dal grafico si può vedere la relazione non molto forte tra le variabile newspaper e sales .

Epilogo dell'EDA

Nell'esplorazione del dataset, una visualizzazione della distribuzione degli investimenti è un passo cruciale per comprendere meglio la struttura dei dati e le caratteristiche di ogni variabile. Questa comprensione aiuta a identificare schemi o anomalie e a scegliere gli algoritmi di machine learning più appropriati. In particolare, la distribuzione delle variabili come gli investimenti in TV, Radio e Newspaper fornisce informazioni utili sulla concentrazione degli investimenti e sulle loro variazioni all'interno del dataset.

Nel grafico riportato di seguito, possiamo osservare la distribuzione

degli investimenti sui diversi canali pubblicitari. Questi istogrammi mostrano chiaramente come gli investimenti si distribuiscono su ciascun canale e rivelano eventuali picchi o aree di densità minore, supportando così una migliore analisi e modellazione.

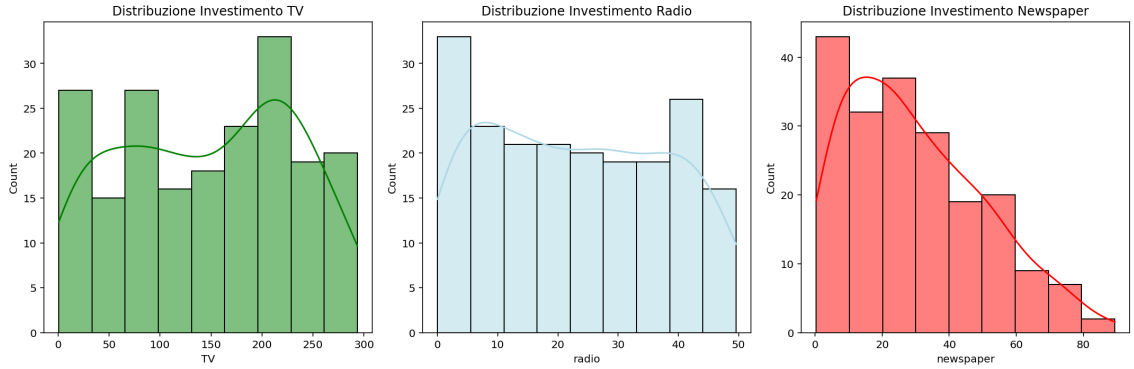

Figura 2.10: Gli istogrammi mostrano la distribuzione degli investimenti pubblicitari sui diversi canali, evidenziando quanti dati rientrano in ciascun range di valori, mentre le curve KDE sovrapposte mostrano la densità di probabilità degli investimenti. Questi grafici sono utili per capire se ci sono range di investimenti più comuni o aree di investimento meno frequentate nei diversi canali pubblicitari.

In conclusione, dalla matrice di correlazione e dal grafico possiamo osservare che esistono differenze significative tra gli investimenti nei diversi canali pubblicitari. La presenza di investimenti elevati nella televisione suggerisce un potenziale maggiore impatto sulle vendite, mentre l'investimento relativamente basso nei giornali potrebbe riflettere una strategia pubblicitaria meno efficace o una variazione nelle preferenze del pubblico. Questo potrebbe indicare che i giornali rappresentano un'opportunità potenzialmente inesplorata, oppure che il target di riferimento è cambiato nel tempo, preferendo canali come la TV o la radio.
Queste osservazioni offrono un utile punto di partenza per l'applicazione dei modelli di machine learning. Capire come vengono distribuiti gli investimenti aiuterà a prevedere meglio l'impatto futuro su vendite e risultati, e potrà guidare decisioni strategiche sull'allocazione delle risorse pubblicitarie.

Capitolo 3

MODELLI DI MACHINE LEARNING

Il machine learning è il settore dell'intelligenza artificiale che si occupa dello sviluppo di sistemi in grado di apprendere da dati, piuttosto che attraverso la programmazione esplicita. Il principale obiettivo del machine learning è rendere le macchine intelligenti, cioè capaci di imparare dalle esperienze e, in base a queste, prendere decisioni o modificare il proprio comportamento. Una delle più recenti definizioni di machine learning è quella di Tom M.

Mitchell, professore dell'Università di Carnegie Mellon: A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Traducendo: "Si dice che un programma impara da una certa esperienza E rispetto a una classe di compiti T ottenendo una performance P, se la sua perfomance nel realizzare i compiti T, misurata dalla performance P, migliora con l'esperienza E.

Quindi un programma è in grado di apprendere dall'esperienza se, nel realizzare un compito, le sue prestazioni migliorano durante il periodo di tempo in cui si ripete la stessa attività. La necessità di ricorrere al machine learning nasce dal fatto che prevedere a priori l'intero set di possibili comportamenti in base all'input è complesso da descrivere in un linguaggio di programmazione. Per questo motivo vengono usati algoritmi diversi che consentono agli elaboratori di evolvere il proprio comportamento basandosi sui dati per migliorare, descrivere e preve-

dere i risultati. In generale, un algoritmo di machine learning parte da un set di dati su cui si allena e in seguito, apportando le giuste modifiche, permette di ottenere un modello accurato. La continua presenza dei sistemi informativi nelle nostre vite, e la mole di dati che vengono creati in questo modo, rendono necessarie le tecnologie adatte a gestire lo storage e ad estrarre conoscenza dai dati stessi. Per questo motivo, il machine learning sta diventando sempre più un aspetto centrale all'interno di organizzazioni di sviluppo, al fine di poter sfruttare i dati ed acquisire informazioni su essi. Utilizzando il giusto modello, in breve tempo è possibile effettuare delle ana- lisi esplorative, per capire la realtà attuale degli investimenti dell'imprese, oppure delle analisi predittive, per prevedere e anticipare i cambiamenti.[5]

3.0.1 Introduzione agli algoritmi

Nel presente capitolo, vengono introdotti due dei modelli di machine learning utilizzati nell'analisi dei dati relativi all'impatto della pubblicità sulle vendite: la **Regressione lineare** e il **Random Forest**. Entrambi questi modelli sono ampiamente usati in ambito predittivo e, sebbene appartenenti a famiglie differenti di algoritmi, hanno caratteristiche che li rendono utili per affrontare il tipo di problema proposto in questo studio.

La regressione lineare è uno dei modelli statistici più semplici e più utilizzati per l'analisi delle relazioni lineari tra una variabile dipendente e una o più variabili indipendenti. L'idea di base è quella di trovare la retta (o iperpiano, nel caso di più variabili) che meglio si adatta ai dati, minimizzando la somma dei quadrati delle differenze tra i valori osservati e quelli predetti dal modello. Questo modello è particolarmente utile quando esiste una relazione lineare tra le variabili, come nel caso della pubblicità e delle vendite in questo studio, dove si assume che l'aumento della spesa pubblicitaria possa avere un impatto lineare sulle vendite. Nonostante la sua semplicità, la regressione lineare è uno strumento potente per ottenere un'interpretazione chiara dei dati e delle relazioni tra le variabili.

Il modello Random Forest, invece, è un algoritmo di tipo ensemble che utilizza una combinazione di alberi decisionali per fare previsioni.

L'algoritmo costruisce una "foresta" di alberi decisionali, ciascuno addestrato su un sottoinsieme casuale dei dati e delle variabili, e utilizza la media (per problemi di regressione) o la modalità (per problemi di classificazione) delle previsioni di ciascun albero per ottenere una previsione finale. Questo approccio riduce notevolmente il rischio di overfitting, che può essere un problema nei modelli basati su singoli alberi decisionali. La Random Forest è in grado di catturare relazioni complesse e non lineari nei dati e, grazie alla sua natura robusta, tende a dare risultati molto accurati. È particolarmente utile quando ci sono interazioni tra le variabili e quando il numero di feature (variabili) è elevato, come nel caso del nostro studio, dove sono coinvolte diverse variabili relative ai canali pubblicitari.

Entrambi i modelli sono stati scelti per questo studio in quanto rappresentano approcci complementari: la regressione lineare offre una comprensione diretta e semplice dei dati, mentre la Random Forest consente di esplorare relazioni più complesse e fornisce previsioni più precise.

3.0.2 Regressione Lineare

Definizione e Formula

questo capitolo tratta della regressione lineare, un approccio molto semplice per l'apprendimento supervisionato. In particolare, la regressione lineare è uno strumento utile per prevedere una risposta quantitativi la regressione lineare è ancora un metodo di apprendimento statistico utile e ampiamente utilizzato. Inoltre, funge da buon punto di partenza per approcci più recenti: Molti approcci statistici avanzati possono essere visti come generalizzazioni o estensioni della regressione lineare. Di conseguenza, l'importanza di avere una buona comprensione della regressione lineare prima di studiare metodi di apprendimento più complessi non può essere sottovalutata.

Prendendo in considerazione il nostro dataset . La figura 4 mostra le vendite (in migliaia di unità) di un determinato prodotto in funzione dei budget pubblicitari (in migliaia di dollari) per i media TV, radio e giornale. vogliamo analizzare come i diversi canali di pubblicità impattano sulla sulle vendite. Quali informazioni sarebbero utili

per fornire tale informazione? Ecco alcune domande importanti che potremmo cercare di affrontare:

1. **Quali media contribuiscono alle vendite?** Tutti e tre i media—TV, radio e giornale—contribuiscono alle vendite, oppure solo uno o due di essi? Per rispondere a questa domanda, dobbiamo trovare un modo per separare gli effetti individuali di ciascun mezzo quando si è speso denaro su tutti e tre i media.

2. **Con quale precisione possiamo stimare l'effetto di ciascun mezzo sulle vendite?** Per ogni dollaro speso in pubblicità su un determinato mezzo, di quanto aumenteranno le vendite? Con quale precisione possiamo prevedere questo aumento?

3. **Con quale precisione possiamo prevedere le vendite future?** Per qualsiasi livello di pubblicità su televisione, radio o giornali, qual è la nostra previsione di vendite, e quale è la precisione di questa previsione?

la regressione lineare rappresenta un strumento utile per rispodere queste domande.

Regressione Lineare semplice

La regressione lineare è la parte di statistica che studia la relazione tra due o più variabile in modo NON DETERMINISTICO per avere inferenza sul modello.

Si assume che ci sia approssimativamente una relazione lineare tra X e Y. Matematicamente, possiamo esprimere questa relazione lineare come:

$$y = \beta_0 + \beta_1 x + \epsilon$$

X potrebbe rappresentare la pubblicità in TV e Y le vendite ϵ è il termine di errore, che tiene conto di fattori non inclusi nel modello. Quindi possiamo fare la regressione delle vendite sulla TV adattando il modello:

$$sales = \beta_0 + \beta_1 TV$$

β_0 e β_1 sono due costanti sconosciute che rappresentano i termini di intercetta e coefficiente angolare nel modello lineare. Insieme, β_0 e β_1 sono noti come i coefficienti o parametri del modello

Estima dei coefficiente

in pratica i parametri β_0 e β_1 sono sconosciuti. Quindi, prima di poter utilizzare l'equazione per fare previsioni, dobbiamo utilizzare i dati per stimare i coefficienti. supponiamo di avere

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

dove n rappresenta il numero di osservazione, nel nostro caso 200. il nostro obiettivo è ottenere stime dei coefficienti β_0 e β_1 tali che il modello lineare si adatti bene ai dati disponibili, ovvero vogliamo che $y = \beta_0 + \beta_1 x$ per $i = 1, \ldots n$. In altre parole, vogliamo trovare un'intercetta β_0 e un coefficiente angolare β_1 tali che la linea risultante sia il più possibile vicina ai $n = 200$ punti dati. Esistono diversi modi per misurare questa vicinanza. Tuttavia, l'approccio più comune di gran lunga consiste nel minimizzare il criterio dei minimi quadrati. [4] [8] [10]

Regressione lineare multipla

la regressione lineare semplice è un approccio utile per prevedere una risposta basata su un singolo predittore. Tuttavia, in pratica, spesso abbiamo più di un predittore. Ad esempio, nel set di dati relativo alla pubblicità, abbiamo esaminato la relazione tra le vendite e la pubblicità televisiva, pero ci sono più variabile. Con regressione lineare multipla è possibile fare le stesse operazione avendo più di uno predittore. Il modello di regressione lineare multipla assume la forma:

$$y = \beta_0 + \beta_1 x + \beta_2 x \cdots + \beta_1 x + \epsilon$$

dove Xj rappresenta il $j - esimo$ predittore e β_j quantifica l'associazione tra quella variabile e la risposta. Interpretiamo $\beta_j$come l'effetto medio su Y di un aumento unitario in X_j, mantenendo fissi tutti gli altri predittori.
Nell'esempio pubblicitario, diventa:

$$y = \beta_0 + \beta_1 TV + \beta_2 radio + \beta_3 newspaper + \epsilon$$

Stima dei Coefficienti di Regressione

Come nel caso della regressione lineare semplice, i coefficienti di regressione $\beta_0, \beta_1 \ldots \beta_p$i sono sconosciuti e devono essere stimati. I

parametri vengono stimati utilizzando lo stesso approccio dei minimi quadrati che abbiamo visto nel contesto della regressione lineare semplice. [4]

Applicazione Pratica

```
#### Regressione Lineare ##########################################################
# 1. Creazione del modello di regressione lineare
model = LinearRegression()

# Addestramento del modello sui dati di training
model.fit(X_train_scaled, y_train)

# 2. Fare previsioni sui dati di test
y_pred = model.predict(X_test_scaled)
# 3. Valutazione del modello
# Mean Absolute Error (MAE)
mae = mean_absolute_error(y_test, y_pred)
# Mean Squared Error (MSE)
mse = mean_squared_error(y_test, y_pred)
# Root Mean Squared Error (RMSE)
rmse = np.sqrt(mse)
# Coefficiente di determinazione R^2
r2 = r2_score(y_test, y_pred)
# Stampa più ordinata dei risultati
print("===== Coefficienti di Regressione =====")
print(f"TV: {model.coef_[0]:.4f}")
print(f"Radio: {model.coef_[1]:.4f}")
print(f"Giornale: {model.coef_[2]:.4f}")

print("\n=== Metriche di Valutazione del Modello ===")
print(f"Mean Absolute Error (MAE): {mae:.2f}")
print(f"Mean Squared Error (MSE): {mse:.2f}")
print(f"Root Mean Squared Error (RMSE): {rmse:.2f}")
print(f"R² Score: {r2:.2f}")
```

Figura 3.1: Applicazione della regressione lineare, stima delle vendite in funzione degli investimenti in TV, Radio e Giornali

Le prestazioni del modello sono state valutate utilizzando le seguenti metriche, riassunte nella Tabella 4.4.

Tabella 3.1: Metriche di valutazione per il modello di Regressione Lineare

Metrica	**Valore**
Mean Absolute Error (MAE)	1.25
Mean Squared Error (MSE)	2.74
Root Mean Squared Error (RMSE)	1.66
R^2 Score	0.90

Interpretazione delle Metriche:

- Il valore di **MAE** di 1.25 suggerisce che, in media, le previsioni del modello differiscono di 1.25 unità rispetto ai valori reali.
- Il **MSE** (2.74) e il **RMSE** (1.66) indicano che l'errore quadratico medio e la sua radice sono contenuti, benché più alti rispetto ad altri modello avanzati.
- L'**R^2 Score** di 0.90 dimostra che il modello spiega il 90% della variabilità nelle vendite, un risultato positivo che conferma la buona correlazione tra predittori e target.

Coefficiente di Regressione

I coefficienti di regressione stimati dal modello indicano l'impatto di un aumento unitario negli investimenti pubblicitari (es. 1000$) su ciascun media. I risultati sono riportati nella Tabella 4.2.

Tabella 3.2: Coefficiente di regressione per ogni variabile

Variabile	**Coefficiente**
TV	3.88
Radio	2.74
Giornale	0.10

Interpretazione dei Coefficienti:

- La **TV** ha il coefficiente più alto (3.88), indicando che un aumento di 1000$ negli investimenti porta a un incremento medio di 3.88 unità nelle vendite. Questo conferma che è il media più influente.
- La **Radio** segue con un coefficiente di 2.74, mostrando un impatto positivo significativo, ma inferiore alla TV.
- La **pubblicità sui giornali**, con un coefficiente quasi nullo (0.10), contribuisce in modo trascurabile alle vendite.

Confronto tra Vendite Reali e Previste

Per valutare la qualità delle previsioni, è stato generato un grafico di dispersione che confronta le vendite reali con quelle previste dal modello (Figura 4.2).

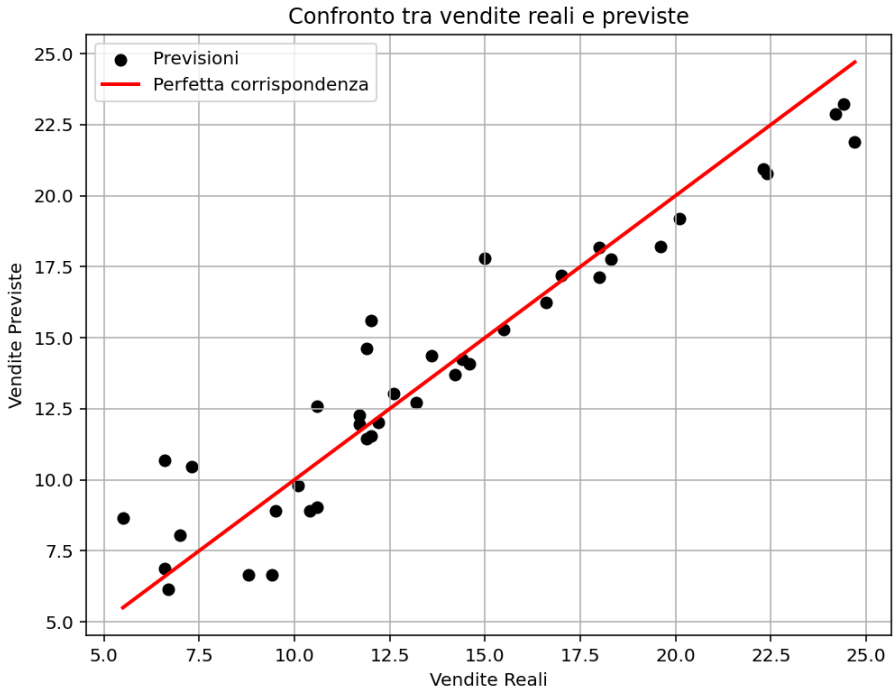

Figura 3.2: Confronto tra vendite reali e previste (Regressione Lineare). I punti rappresentano i dati reali e previsti; la linea rossa indica la perfetta corrispondenza.

Osservazioni:

- La distribuzione dei punti vicino alla linea di perfetta corrispondenza suggerisce una buona capacità del modello di fare previsioni accurate.
- Tuttavia, si osserva una maggiore dispersione dei punti , indicando una performance inferiore nel catturare alcune complessità dei dati.

possiamo concludere che Il modello di Regressione Lineare fornisce una buona stima dei risultati, spiegando il 90% della variabilità nei dati ($R = 0.90$). Tuttavia, i risultati mostrano che la capacità del modello di gestire errori grandi è inferiore rispetto a metodi più avanzati come Random Forest o Gradient Boosting . L'analisi dei coefficienti di regressione suggerisce che le risorse pubblicitarie dovrebbero essere

principalmente allocate su **TV** e **Radio**, mentre gli investimenti sui **giornali** dovrebbero essere rivalutati, poiché il loro impatto sulle vendite è minimo.

3.0.3 Random Forest Regressor

La **Random Forest** è uno degli algoritmi di machine learning più utilizzati per la sua capacità di ottenere previsioni altamente accurate. È un metodo supervisionato che può essere applicato sia a problemi di classificazione che di regressione. La Random Forest si basa sull'idea di combinare Bagging (Bootstrap Aggregating) con la costruzione di molteplici alberi decisionale per produrre una previsione finale che risulta robusta e generalizzabile.

Il concetto di Bagging

Il **bootstrap aggregating**, o **bagging**, è una tecnica che mira a ridurre la varianza di un metodo di apprendimento statistico. È particolarmente utile quando si lavora con alberi decisionali, che tendono ad avere un'alta varianza. L'idea alla base del Bagging è quella di costruire più modelli (ad esempio alberi decisionali) utilizzando diversi campioni casuali tratti dal dataset di addestramento. La previsione finale viene ottenuta facendo la media delle previsioni di tutti i modelli.

Formulando il concetto matematicamente, dato un insieme di dati $Z_0, Z_1, \ldots, Z_n$ con varianza σ^2, la varianza della media $\bar{Z}$ delle osservazioni è σ^2/B, dove B è il numero di modelli. Di conseguenza, un modo naturale per ridurre la varianza è costruire diversi modelli su insiemi di dati diversi e fare la media delle loro previsioni. In termini matematici, abbiamo:

$$\hat{f}_{\text{avg}}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x)$$

In pratica, non avendo accesso a molti set di addestramento, si utilizza il **bootstrap**, che consiste nell'estrarre ripetutamente campio-

ni con sostituzione dal set di dati originale per ottenere insiemi di addestramento diversi. Le previsioni finali vengono quindi mediate:

$$\hat{f}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_b^*(x)$$

Applicazione del Bagging negli alberi decisionali

Il **bagging** è particolarmente vantaggioso per gli alberi decisionali, in quanto aiuta a ridurre la loro alta varianza. Nel contesto della Random Forest, si costruiscono B alberi decisionali utilizzando B insiemi bootstrap. Poiché ogni albero è costruito senza potatura, ogni albero individuale avrà una bassa accuratezza (alto bias) ma una varianza elevata. Tuttavia, facendo la media delle previsioni degli alberi, la varianza complessiva viene ridotta, aumentando così l'accuratezza del modello complessivo. L'adozione di centinaia o migliaia di alberi in una Random Forest ha dimostrato di migliorare notevolmente le prestazioni predittive rispetto ai singoli alberi decisionali. [2] [4]

Random Forest: Il miglioramento rispetto al Bagging

La Random Forest migliora ulteriormente l'approccio del Bagging introducendo una piccola modifica che decorrela gli alberi. Oltre a costruire gli alberi utilizzando il campionamento bootstrap, la Random Forest introduce un'ulteriore casualità durante la selezione delle caratteristiche. Ogni volta che un albero prende una decisione su una divisione, anziché considerare tutte le caratteristiche disponibili, ne seleziona un sottoinsieme casuale m (dove tipicamente $m \approx \sqrt{p}$, con p che è il numero totale di predittori). Questo riduce la correlazione tra gli alberi e migliora la diversità del modello, rendendo le previsioni più robuste.

$$m = \sqrt{p} \quad \text{(dove } p \text{ è il numero di predittori)}$$

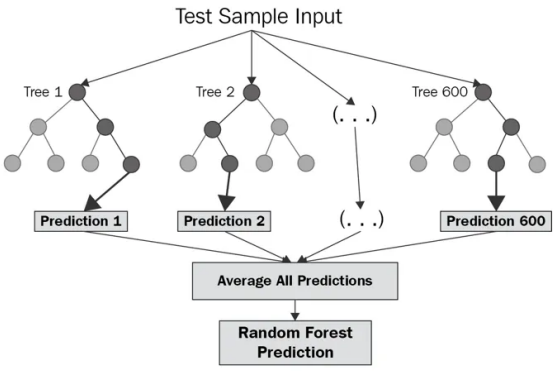

Figura 3.3: Ilustrazione del funzionamento del modello

Questa tecnica migliora la stabilità del modello e riduce il rischio di overfitting.[4]

Conclusione

In sintesi, la Random Forest è un potente strumento per il machine learning, particolarmente utile quando si cerca un equilibrio tra accuratezza predittiva e robustezza. La combinazione di Bagging e la selezione casuale delle caratteristiche durante la costruzione degli alberi consente alla Random Forest di ottenere risultati superiori rispetto ai singoli alberi decisionali. La sua capacità di ridurre la varianza e di evitare l'overfitting la rende una scelta ideale per problemi complessi, come quelli che si incontrano nella previsione delle vendite, dove la variabilità dei dati può essere elevata. Inoltre, la Random Forest offre un'importante capacità di interpretare l'importanza delle caratteristiche, permettendo di comprendere quali variabili hanno un impatto maggiore sul risultato predetto, rendendo il modello sia performante che interpretabile.

Random Forest Regressor Applicazione

l'adozione di questo modello permetterà di ottenere previsioni precise dell'impatto del investimento sui vari media (TV, radio, giornali) sulle vendite. contribuirà a rispondere la domanda di ricerca di questo elaborato.

Applicazione

```python

# Creazione del modello Random Forest
rf_model = RandomForestRegressor(n_estimators=100, random_state=42)

# Addestramento del modello
rf_model.fit(X_train_scaled, y_train)

# Previsioni sul test set
y_rf_pred = rf_model.predict(X_test_scaled)

# Valutazione del modello
rf_mae = mean_absolute_error(y_test, y_rf_pred)
rf_mse = mean_squared_error(y_test, y_rf_pred)
rf_rmse = np.sqrt(rf_mse)
rf_r2 = r2_score(y_test, y_rf_pred)
```

Figura 3.4: Codice Python per la Creazione del modello Random Forest.

```python
    # Stampa delle metriche di valutazione
print("===== Metriche di Valutazione del Modello Random Forest =====")
print(f"Mean Absolute Error (MAE): {rf_mae:.2f}")
print(f"Mean Squared Error (MSE): {rf_mse:.2f}")
print(f"Root Mean Squared Error (RMSE): {rf_rmse:.2f}")
print(f"R² Score: {rf_r2:.2f}")

# Grafico di confronto tra valori reali e previsi
plt.figure(figsize=(8, 6))
plt.scatter(y_test, y_rf_pred, color='green', label='Previsioni (Random Forest)')
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], color='red',
         linewidth=2, label='Perfetta corrispondenza')
plt.title('Confronto tra vendite reali e previste (Random Forest)')
plt.xlabel('Vendite Reali')
plt.ylabel('Vendite Previste')
plt.legend()
plt.grid(True)
plt.show()

# Visualizzazione dell'importanza delle feature
```

Le principali metriche di valutazione del modello Random Forest sono riassunte nella Tabella 3.3.

Tabella 3.3: Metriche di valutazione per il modello Random Forest

Metrica	**Valore**
Mean Absolute Error (MAE)	0.54
Mean Squared Error (MSE)	0.41
Root Mean Squared Error (RMSE)	0.64
R^2 Score	0.98

Interpretazione delle Metriche:

- Il valore di **MAE** di 0.54 indica che, in media, le previsioni del modello differiscono di meno di un'unità rispetto alle vendite reali.

- Il **MSE** (0.41) e il **RMSE** (0.64) confermano l'elevata accuratezza del modello, penalizzando in modo significativo gli errori maggiori.

- L'**R^2 Score** di 0.98 dimostra che il modello spiega il 98% della variabilità nei dati, garantendo un adattamento eccellente.

Importanza delle Variabili

Un'analisi dell'importanza delle variabili (*feature importance*) fornisce un'indicazione del contributo relativo di ciascun media pubblicitario nel prevedere le vendite (Figura 3.5).

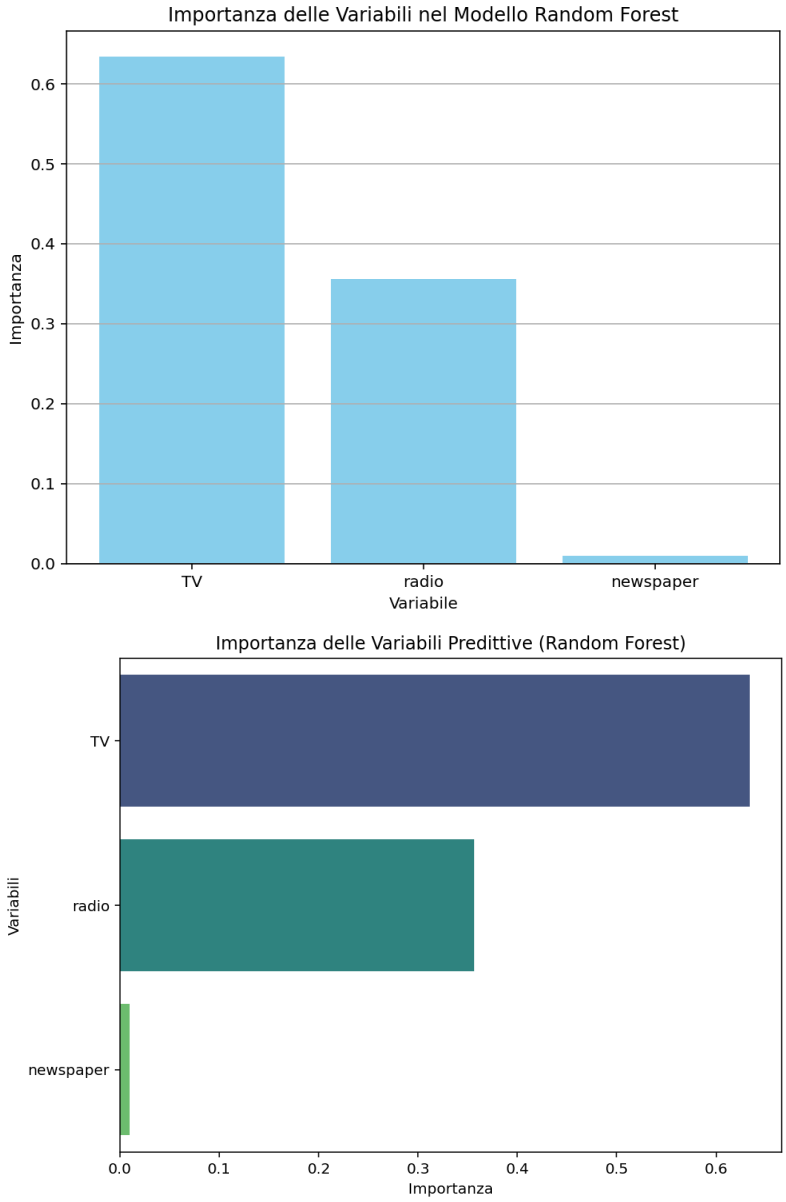

Figura 3.5: Importanza delle variabili nel modello Random Forest

I risultati evidenziano che:

- La pubblicità su **TV** contribuisce per il 63.4%, risultando il media più efficace nel migliorare le vendite.
- La **Radio** segue con un contributo del 35.6%, dimostrando un impatto significativo ma inferiore rispetto alla TV.
- I **giornali**, con un contributo trascurabile dello 0.98%, appaiono il media meno efficace per generare vendite.

Confronto tra Vendite Reali e Previste

Per valutare la qualità delle previsioni, è stato costruito un grafico di dispersione (Figura 3.6) che confronta le vendite reali con quelle previste dal modello Random Forest.

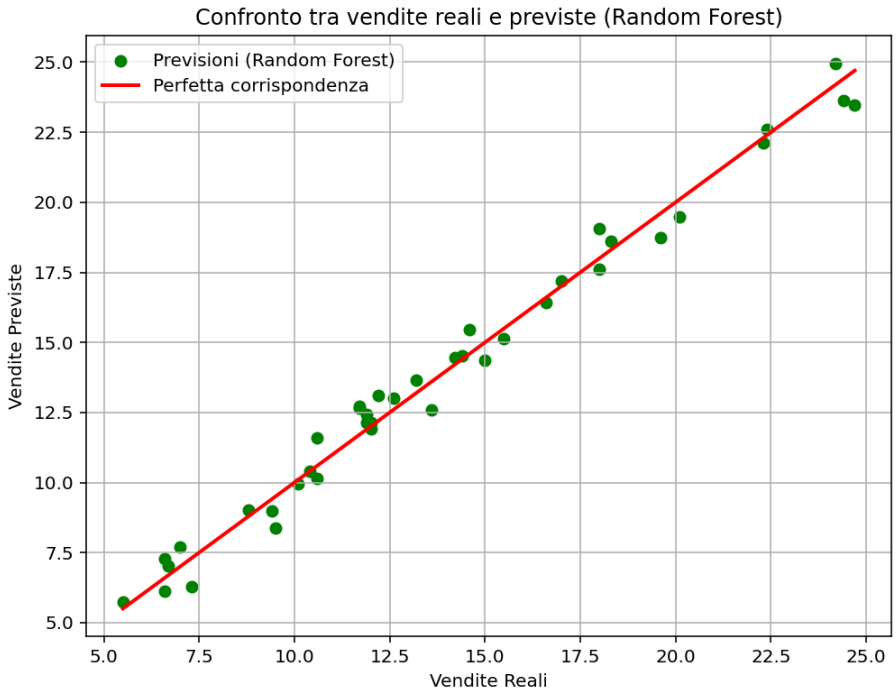

Figura 3.6: Confronto tra vendite reali e previste (Random Forest)

Osservazioni:

- I punti sono distribuiti vicino alla linea di perfetta corrispondenza (*linea rossa*), confermando l'elevata accuratezza del modello.
- La dispersione ridotta dei punti indica errori sistematici limitati e un modello ben calibrato.

In Conclusione possiamo dire che il modello Random Forest si è dimostrato superiore rispetto alla regressione lineare, con prestazioni nettamente migliori nelle metriche di valutazione e una maggiore capacità di modellare relazioni non lineari. Questi risultati suggeriscono che, per ottimizzare le vendite, gli investimenti pubblicitari dovrebbero essere concentrati principalmente su **TV** e **Radio**, riducendo al minimo la spesa sui **Giornali**.

Capitolo 4

I RISULTATI

Questo capitolo dell'elaborato si propone di fornire un riepilogo dei risultati ottenuti nell'analisi dei dati presenti nel dataset, applicando 2 modelli di machine learning. L'obiettivo principale della tesi è offrire una comprensione dell'impatto della pubblicità come leva strategica per migliorare le performance di vendita.
L'analisi presentata si basa su un dataset composto da 200 osservazioni e 4 variabili, ma la metodologia adottata è facilmente estendibile a modelli con un numero maggiore di osservazioni e variabili. Nello specifico, l'intento di questo capitolo è stimare l'effetto degli investimenti pubblicitari su diversi canali (TV, Radio, Giornali) in relazione alle vendite, fornendo un quadro chiaro e quantitativo dell'influenza di ciascun canale sul risultato finale in base all'accuratezza dei modelli utilizzati.

4.0.1 Analisi dei Risultati del Modello di Regressione Lineare

modello di **Regressione Lineare** è stato utilizzato per stimare l'effetto degli investimenti pubblicitari su *TV*, *radio* e *giornali* sulle vendite. Questo approccio presuppone una relazione lineare tra le variabili indipendenti e la variabile dipendente, risultando in un modello semplice e interpretabile. L'idea e quella di avere delle metriche e coefficiente che ci consentono di avere delle previsione per ogni singola variabile e di capire come queste previsione possono essere affidabile.

Metriche di Valutazione del Modello di Regressione

Le prestazioni del modello sono state valutate utilizzando le seguenti metriche, riassunte nella Tabella 4.4.

Tabella 4.1: Metriche di valutazione per il modello di Regressione Lineare

Metrica	**Valore**
Mean Absolute Error (MAE)	1.25
Mean Squared Error (MSE)	2.74
Root Mean Squared Error (RMSE)	1.66
R^2 Score	0.90

Interpretazione delle Metriche:

- **MAE (1.25)**: Indica che, in media, le previsioni del modello differiscono di 1.25 unità rispetto ai valori reali.
- **MSE (2.74)**: Penalizza maggiormente gli errori più grandi, evidenziando la capacità del modello di gestire le deviazioni.
- **RMSE (1.66)**: Fornisce una stima interpretabile dell'errore medio, con una deviazione media di 1.66 unità tra valori previsti e reali.
- **R^2 Score (0.90)**: Il modello spiega il 90% della variabilità nei dati, dimostrando un'ottima capacità di adattamento.

Coefficiente della Regressione

I coefficienti di regressione stimati dal modello rappresentano l'impatto di un incremento unitario negli investimenti pubblicitari su ciascun media. In altre parole, i coefficienti mostrano quante unità di vendita aggiuntive ci si può aspettare per ogni aumento di 1000 dollari spesi su un determinato canale pubblicitario. I risultati sono riassunti nella Tabella 4.2.

Tabella 4.2: Coefficiente di regressione per ciascun canale pubblicitario

Media	**Coefficiente**	**Interpretazione**
TV	3.88	Ogni 1000$ in TV generano in media 3.88 unità aggiuntive di vendita
Radio	2.74	Ogni 1000$ in Radio generano in media 2.74 unità aggiuntive di vendita
Giornali	0.10	Ogni 1000$ in Giornali generano solo 0.10 unità aggiuntive di vendita

Interpretazione dettagliata dei coefficienti:

- **TV (3.88):** Questo coefficiente indica che la pubblicità televisiva ha il maggiore impatto sulle vendite. Ad esempio, un'azienda che investe 5000$ in pubblicità televisiva si aspetterebbe un incremento medio di $3.88 \times 5 = 19.4$ unità di vendita.

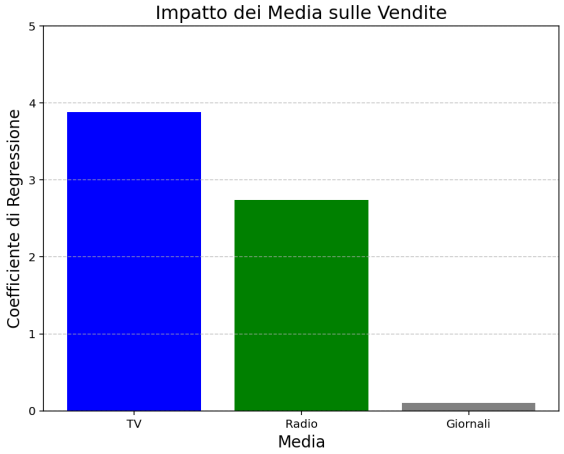

Figura 4.1: Impatto di ogni singola variabile

- **Radio (2.74):** La radio ha un impatto significativo, anche se inferiore alla TV. Ad esempio, un'azienda che investe 3000$ in pubblicità radiofonica può aspettarsi 2.74×3=8.22 unità aggiuntive di vendita.
- **Giornali (0.10):** L'impatto della pubblicità sui giornali è marginale. Un investimento di 10.000$ porterebbe a un incremento trascurabile di sole 0.10×10=1 unità di vendita.

Tabella 4.3: Esempi di incremento delle vendite per investimenti pubblicitari

Media	**Investimento (1000$)**	**Incremento di Vendite previsto**	**Interpretazione**
TV	2	3.88×2=7.76	L'impatto maggiore
Radio	3	2.74×3=8.22	Impatto moderato
Giornali	10	0.10×10=1	Impatto minimo

Il grafico e la tabella dimostrano che il canale televisivo offre il ritorno più significativo per dollaro speso, mentre i giornali mostrano un impatto trascurabile. Questi risultati suggeriscono che le aziende dovrebbero concentrare i loro budget pubblicitari principalmente su TV e radio per massimizzare il ritorno sugli investimenti.

Confronto tra il valori reali e quelli previsti

Per valutare la qualità delle previsioni del modello di regressione lineare, è stato generato un grafico di dispersione che mette a confronto i valori osservati (vendite reali) con quelli previsti dal modello. Questo tipo di analisi è fondamentale per comprendere l'accuratezza e le eventuali limitazioni del modello. La Figura 4.2 mostra i risultati:

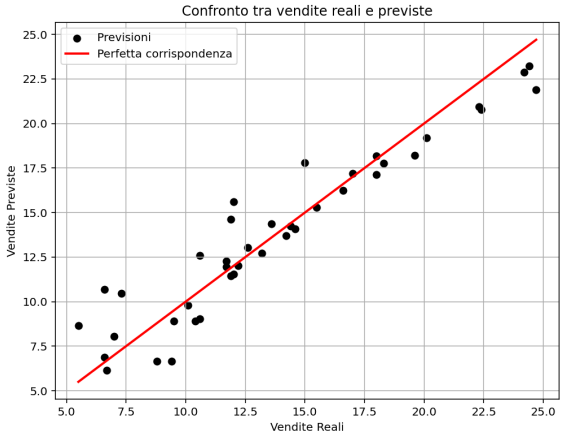

Figura 4.2: Confronto tra vendite reali e previste (Regressione Lineare). Ogni punto rappresenta una coppia di valori reale-previsto; la linea rossa indica la perfetta corrispondenza tra previsione e realtà.

- **Distribuzione dei punti:** La maggior parte dei punti si distribuisce vicino alla linea rossa (corrispondenza ideale), indicando che il modello fornisce previsioni ragionevolmente accurate per la maggior parte dei dati.

- **Dispersione:** In alcune aree, si nota una dispersione più ampia, dove i punti si allontanano significativamente dalla linea di perfetta corrispondenza. Questo suggerisce che il modello potrebbe non catturare adeguatamente alcune complessità dei dati.

- **Interpretazione pratica:** Ad esempio, un punto che si trova molto distante dalla linea potrebbe indicare che il modello sottostima o sovrastima le vendite per specifiche combinazioni di spesa pubblicitaria (ad esempio, un investimento elevato in giornali con basso impatto effettivo sulle vendite).

Esempio Numerico

Consideriamo un caso specifico per interpretare i risultati:

Vendite Reali: 20 unità

Vendite Previste: 18 unità

In questo caso, l'errore assoluto sarebbe pari a $|20 - 18| = 2$ unità. Tale errore, pur essendo relativamente piccolo, potrebbe diventare più significativo per valori più elevati, ad esempio con vendite reali di 100 e previsioni di 80, dove l'errore assoluto sarebbe di 20 unità.

Un grafico di dispersione aiuta a identificare sistematicamente tali discrepanze e a valutare se esse si verificano in aree specifiche (ad esempio, per alte o basse vendite).

Conclusioni sul Modello di Regressione Lineare
Il modello di regressione lineare ha dimostrato di essere in grado di spiegare il 90% della variabilità nei dati (R2=0.90), una misura che riflette una buona capacità di adattamento. Tuttavia, emergono alcune limitazioni:

Limitazioni con dati complessi: La maggiore dispersione rispetto a modelli più avanzati (ad esempio, Random Forest) evidenzia difficoltà nel gestire relazioni non lineari o interazioni tra le variabili.

Allocazione delle risorse pubblicitarie: Come suggerito dai coefficienti di regressione, il budget dovrebbe essere orientato principalmente su **TV** (coefficiente 3.88) e **Radio** (coefficiente 2.74), mentre gli investimenti in **giornali** dovrebbero essere rivalutati (coefficiente 0.10), poiché il loro contributo alle vendite è trascurabile.

4.0.2 Analisi dei risultati del modello random forest

Questo è un metodo che combina più alberi decisionali, garantendo una maggiore robustezza e precisione rispetto ai modelli lineari. Random Forest è particolarmente utile per catturare relazioni non lineari e interazioni tra variabili, offrendo prestazioni migliori rispetto alla regressione lineare in contesti complessi.
L'analisi del modello Random Forest Regressor ha permesso di valutare in maniera più approfondita l'impatto degli investimenti pubblicitari su diversi media (TV, Radio, Giornali) rispetto al modello di regressione lineare. Random Forest, essendo un modello basato su ensemble, supera alcune limitazioni della regressione lineare, come l'assunzione di una relazione lineare tra variabili, e consente di gestire relazioni complesse, interazioni tra le variabili e dati rumorosi. Di seguito, è riportata un'analisi dettagliata delle sue prestazioni.

Metriche di Valutazione del Modello Random Forest

Le prestazione del modello sono state valutate utilizzando le seguenti metriche,

Tabella 4.4: Metriche di valutazione per il modello di Regressione Lineare

Metrica	**Valore**
Mean Absolute Error (MAE)	0.54
Mean Squared Error (MSE)	0.41
Root Mean Squared Error (RMSE)	0.61
R^2 Score	0.98

Interpretazione delle Metriche:
I valori delle metriche del modello random forest sono molto positivi in confronto hai valori del modelle di regressione lineare, consento di avere più fiducia sui risultati e l'importanza di ogni variabile nel nostro dataset. possiamo interpretare queste metriche nel seguente modo:

- MAE (0.54): Indica che, in media, le previsioni del modello differiscono di 0.54 unità rispetto ai valori reali, evidenziando un errore medio molto basso rispetto alla regressione lineare (1.25).
- MSE (0.41): Penalizza maggiormente gli errori grandi, suggerendo che il modello è efficace nel ridurre le deviazioni.
- RMSE (0.64): Stima interpretabile dell'errore medio: in media, le previsioni si discostano di 0.64 unità dai valori reali.
- R^2 Score (0.98): Indica che il modello spiega il 98% della variabilità nei dati, una misura eccellente di adattamento e predizione.

Importanza delle Variabili

Random Forest fornisce una misura dell'importanza delle variabili nel determinare le previsioni. I risultati sono riassunti nella seguente tabella 4.5.

Tabella 4.5: Coefficiente di regressione per ciascun canale pubblicitario

Media	**Importanza (%)**	**Interpretazione**
TV	63.40	La TV rappresenta la variabile più influente nel determinare le vendite.
Radio	35.61	La radio ha un peso significativo, ma inferiore rispetto alla TV.
Giornali	0.98	L'impatto della pubblicità sui giornali è trascurabile.

Interpretazione

- L'importanza delle variabili conferma che TV e Radio sono i canali pubblicitari più efficaci. La TV domina sia nella regressione lineare (coefficiente: 3.88) che nel modello Random Forest (63.4% di importanza relativa).
- L'impatto dei Giornali rimane marginale (0.10 nella regressione lineare e 0.98% in Random Forest), suggerendo che questo canale pubblicitario potrebbe essere de-prioritizzato.

Considerando queste metriche che ci forniscono una buona interpretazione sull'importanza delle variabile, possiamo immaginare che se un'azienda decide di ridurre 50% dell'investimento sulla radio e investire lo stesso budget nella TV, l'effetto positivo sulle vendite potrebbe essere amplificato grazie all'importanza relativa della TV.

Confronto tra Vendite Reali e Previste

Il confronto tra le vendite reali e quelle previste dal modello Random Forest è illustrato nella Figura 4.3. Questo grafico di dispersione mostra la relazione tra i valori reali e quelli stimati, con una linea rossa che rappresenta la corrispondenza perfetta

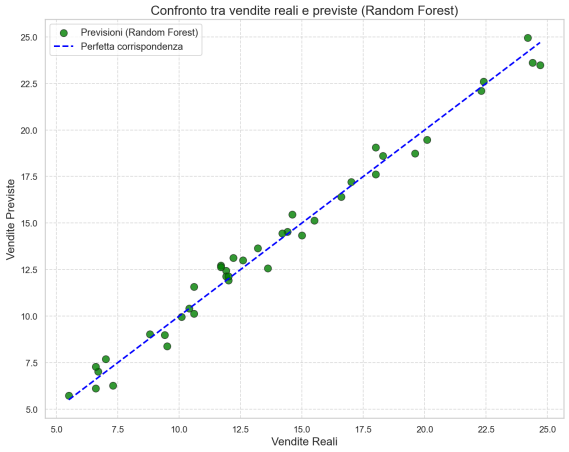

Figura 4.3: Grafico di dispersione

Osservazioni dal grafico di dispersione

- Distribuzione vicina alla linea ideale: La maggior parte dei punti è molto vicina alla linea di corrispondenza perfetta, indicando che il modello è in grado di fare previsioni molto accurate.

- Minima dispersione: La dispersione è significativamente inferiore rispetto alla regressione lineare, specialmente nei valori estremi (vendite alte o basse). Questo evidenzia la capacità del modello di catturare anche le complessità dei dati.
- Errori residuali ridotti: I pochi punti più lontani dalla linea suggeriscono errori molto bassi, che sono comunque significativamente minori rispetto a quelli della regressione lineare.

4.0.3 Confronto tra i modelli

Confronto delle Performance

Tabella 4.6: Confronto delle metriche di valutazione tra Regressione Lineare e Random Forest

Metrica	**Regressione Lineare**	**Random Forest**
Mean Absolute Error (MAE)	1.25	0.54
Mean Squared Error (MSE)	2.74	0.41
Root Mean Squared Error (RMSE)	1.66	0.64
R^2 Score	0.90	0.98

Interpretazione del Confronto:

- MAE e MSE: La Random Forest ha valori significativamente inferiori di MAE e MSE rispetto alla Regressione Lineare, indicando una maggiore accuratezza nelle previsioni e una minore penalizzazione degli errori.
- RMSE: Anche se il valore di RMSE della Random Forest è leggermente più alto rispetto alla MAE, rimane inferiore a quello della Regressione Lineare, confermando una migliore performance complessiva.
- R^2 Score: La Random Forest spiega il 98% della variabilità nelle vendite, rispetto al 90% della Regressione Lineare, dimostrando una capacità superiore di adattamento ai dati.

4.0.4 Interpretazione dell'Importanza delle Variabili

Confrontando i coefficienti di regressione e le feature importance della Random Forest, emergono alcune similitudini e differenze:

- TV: Rimane la variabile più influente in entrambi i modelli, confermando la sua importanza strategica nelle campagne pubblicitarie.
- Radio: Anche se meno influente della TV, continua a mostrare un impatto significativo sulle vendite.
- Giornali: In entrambi i modelli, l'impatto è minimo, suggerendo una necessità di riconsiderare la spesa pubblicitaria in questo canale.

4.0.5 Conclusioni del Confronto

Il confronto tra Regressione Lineare e Random Forest evidenzia come i modelli non lineari possano offrire migliori performance predittive, soprattutto quando si tratta di catturare relazioni complesse tra variabili. La Random Forest non solo migliora le metriche di errore, ma fornisce anche un'analisi più dettagliata dell'importanza delle variabili, essenziale per le decisioni strategiche di marketing.

Risposta alle Domande di Ricerca

1. Quali media contribuiscono maggiormente alle vendite?
Entrambi i modelli identificano la TV e la Radio come i principali contributori alle vendite, mentre i Giornali hanno un impatto trascurabile. La Random Forest, con l'importanza delle feature, conferma che la TV è il mezzo pubblicitario più efficace, seguito dalla radio.

2. Quanto potrebbero aumentare le vendite in relazione a un aumento dell'investimento pubblicitario?
Basandosi sui coefficienti della Regressione Lineare:

- Un aumento di 1000$ in TV porta a un incremento medio di 3.88 unità di vendita.
- Un aumento di 1000$ in Radio porta a un incremento medio di 2.74 unità di vendita.
- Un aumento di 1000$ in Giornali porta a un incremento medio di 0.10 unità di vendita.

La Random Forest conferma che gli investimenti in TV e Radio sono più efficaci nel generare vendite.

3. Con quale precisione possiamo stimare l'effetto di ciascun mezzo sulle vendite?

Le metriche di valutazione mostrano che entrambi i modelli sono precisi, con la Random Forest che offre una precisione superiore (R^2 = 0.98 Vs R^2 = 0.90). Questo significa che la Random Forest è più affidabile nello stimare l'impatto di ciascun mezzo sulle vendite.

4. Con quale precisione possiamo prevedere le vendite future basandoci sui dati storici degli investimenti pubblicitari?

La Random Forest, con un R^2 = 0.98 dimostra una capacità eccezionale nel prevedere le vendite future basate sugli investimenti storici. Il grafico di dispersione (Figura 3.6) evidenzia la precisione delle previsioni, mostrando che le vendite previste sono molto vicine ai valori reali.

CONCLUSIONE

Il progetto svolto ha avuto come obiettivo principale l'analisi dell'impatto di diversi canali pubblicitari sulle performance di vendita, utilizzando due modelli di machine learning per prevedere l'andamento delle vendite in base agli investimenti pubblicitari. Attraverso un approccio metodico e una combinazione di tecniche statistiche avanzate, sono stati ottenuti risultati significativi che offrono una panoramica del ruolo cruciale della pubblicità nel guidare le decisioni aziendali.

Il primo modello implementato è stato una regressione lineare, scelto per la sua semplicità e la trasparenza interpretativa. Questo modello, grazie alla sua capacità di identificare relazioni lineari tra variabili indipendenti (spesa pubblicitaria per TV, Radio e Giornali) e la variabile dipendente (vendite), ha fornito buoni risultati, con un coefficiente di determinazione R^2=0.90. Sebbene soddisfacente, la regressione lineare presenta delle limitazioni nell'adattarsi a relazioni non lineari e a interazioni complesse tra le variabili.

Per migliorare l'accuratezza delle previsioni, è stato implementato il modello Random Forest Regressor, una tecnica di ensemble che combina più decision tree per produrre previsioni più robuste e precise. Questo modello si è dimostrato significativamente più performante, con un R^2 pari a 0.98 e metriche di errore drasticamente ridotte (MAE: 0.54, RMSE: 0.64). Inoltre, il Random Forest ha permesso di analizzare l'importanza relativa delle variabili, evidenziando che la pubblicità televisiva ha un peso predominante, seguita dalla radio, mentre i giornali mostrano un'influenza marginale.

I risultati ottenuti dai modelli indicano chiaramente che la pubblicità televisiva ha l'impatto più significativo sulle vendite, rappresentando

oltre il 60% dell'importanza relativa nelle previsioni del modello Random Forest. La radio segue con circa il 36%, mentre i giornali si attestano a meno dell'1%. Questa gerarchia di impatto è coerente con le tendenze osservate nei media tradizionali, dove la TV continua a essere uno strumento fondamentale per raggiungere un vasto pubblico, nonostante la crescente digitalizzazione.

I risultati sottolineano anche le differenze tra i due modelli utilizzati. Mentre la regressione lineare fornisce una visione basilare e interpretabile, il Random Forest si dimostra più adatto a catturare la complessità dei dati e a produrre previsioni di maggiore accuratezza. Tuttavia, il trade-off tra interpretabilità e precisione rimane un aspetto cruciale nella scelta del modello da utilizzare.
L'analisi presentata in questo progetto offre una base solida per comprendere l'impatto della pubblicità sulle vendite, ma rappresenta solo un punto di partenza. Ci sono diversi modi in cui questo studio potrebbe essere esteso e approfondito integrazione di nuove variabili (es: Dati demografici, Fattori temporali e Indicatori macroeconomici ecc).

Questo elaborato dimostra come i modelli di machine learning possano essere utilizzati efficacemente per analizzare e prevedere l'impatto degli investimenti pubblicitari sulle vendite. La combinazione di metodi tradizionali, come la regressione lineare, e tecniche avanzate, come il Random Forest, evidenzia le potenzialità del machine learning nel fornire insight utili per le decisioni aziendali.

In un mondo sempre più dominato dal digitale, l'utilizzo di algoritmi di machine learning rappresenta non solo un'opportunità, ma una necessità per le aziende che vogliono ottimizzare le loro strategie pubblicitarie. Questo progetto ha cercato di offrire una visione introduttiva delle possibilità offerte da questi strumenti, lasciando spazio a future ricerche che potrebbero approfondire ulteriormente le dinamiche complesse tra pubblicità e vendite .

Bibliografia

[1] URL ttps://www.adcgroup.it/adv-express/big-data/scenari/warc-nel-2024-la-spesa-pubblicitaria-globale-crescera-del html.

[2] Matthew Wiener Andy Liaw. Classification and regression by randomforest. 1/2, 2002. doi: https://datajobs.com/data-science-repo/Random-Forest-%5BLiaw-and-Weiner%5D.pdf.

[3] Massimiliano Carrà. Titolo del sito web, 2023. URL https://forbes.it/2024/08/05/spesa-pubblicitaria-mondiale-nel-2024-le-cifre-di-dentsu/. Ultimo accesso: 29 Novembre 2024.

[4] Trevor Hastie Robert Tibshirani Gareth James, Daniela Witten. *An Introduction to statical learning*. Springer New York Heidelberg Dordrecht London, 2013.

[5] Kirsch D. Hurwitz J. *Machine Learning*. John Wiley Sons, Inc, 2018.

[6] kaggle. fonte dataset, 2024. URL https://www.kaggle.com/datasets/mehmetisik/advertisingcsv/data. kaggle.

[7] Chenglin Ye James Paul Lehana Thabane, Tara Thomas. Posing the research question: not so simple. 2008. doi: https://link.springer.com/content/pdf/10.1007/s12630-008-9007-4.pdf.

[8] Eleana Loli Piccolomini. Regressione lineare. Appunti presi a lezione o materiale su piattaforma virtualeUnibo.it, 2023. URL https://virtuale.unibo.it/pluginfile.php/1604154/mod_resource/content/0/regressione_lineare.pdf. Università di bologna, Statistica Numerica.

[9] Eleana Loli Piccolomini. Regressione e classificazione, 2023. URL `https://virtuale.unibo.it/pluginfile.php/1608791/mod_resource/content/0/slides_ML_I.pdf`. Università di bologna, Statistica Numerica.

[10] Josè Unpingco. *Python for probability, statistcs, and machine learning*. Springer, San Diego, CA USA, 2022.

[4] [10] [5] [8] [9] [6] [1] [2] [7]

RINGRAZIAMENTI

Desidero innanzitutto esprimere la mia sincera gratitudine alla Professoressa Elena Piccolomini per avermi seguito e reso possibile questo elaborato.

Ci tengo ad esprime la mia profonda gratitudine ai miei genitori, Ana, Vasco e Carolina, che con la loro dedizione e amore incondizionato mi hanno sostenuto in ogni momento della mia vita. Hanno reso possibile il mio percorso educativo, instillandomi valori fondamentali che mi hanno guidato fino a questo importante traguardo.

Un pensiero affettuoso ai miei fratelli, in particolare Zacarias, Vasquinho e Fernando, per il loro continuo supporto e per avermi dato forza nei momenti più difficili. Soppratutto per per avere creduto in me. La loro presenza costante ha reso questo lungo viaggio più leggero e pieno di significato.

Desidero ringraziare anche la mia numerosa famiglia, con un pensiero speciale ad Amani, Francesca Diana e Arcenio , per il loro sostegno, che hanno contribuito a rendere possibile questo traguardo.

Un sentito grazie va anche tutti miei amici, che sono stati parte integrante di questo percorso universitario. Un ringraziamento particolare a Dorian e Khalil per aver condiviso con me gioie e difficoltà di questi anni di studio, rendendo l'esperienza universitaria ancora più memorabile.

Un pensiero di riconoscenza va a tutti i miei professori e all'Università di Bologna, che con la loro competenza e professionalità mi hanno fornito gli strumenti e le conoscenze necessarie per la mia formazione accademica. Voglio anche ringraziare l'Italia, un Paese

che mi ha accolto calorosamente e mi ha fatto sentire a casa, offrendomi un ambiente dove crescere, imparare e costruire il mio futuro.

Infine, un ringraziamento profondo e sincero a tutti coloro che, in modi grandi o piccoli, hanno contribuito a questo importante capitolo della mia vita.

khanimambo...