SCRIPT REGRESSION

Dataset su cui si lavora: **Top 20 products**
Variabile risposta: **Vendite in Volume**

Preparazione DataSet

Prepariamo il dataset scegliendo i tipi di sconti che vogliamo trattare mettendo tutti gli altri NULL.
Sconti selezionati (sono tutte variabili binarie 0/1):

- Sconto Solo Special Pack
- Sconto Solo Volantino
- Sconto Solo Display
- Sconto Solo Riduzione Prezzo
- Sconto Solo Loyalty

Aggiungiamo una colonna per i periodi di vacanze in modo da poter tenere in considerazione i diversi periodi della time series di cui analizziamo le vendite.
Vacanze selezionate:

- Pasqua ("2019-04-21", "2020-04-12", "2021-04-04", "2022-04-17", "2023-04-09", “2024-03-31”)
- Natale (dal “2019-12-20" al “2024-01-06”)

Aggiungiamo una colonna per il periodo estivo dal 21/06 al 23/09.
Sia le vacanze che l’estate sono variabili 0/1.
Aggiungiamo poi la suddivisione Leader/Follower che deriva dal cluster svolto in precedenza (in particolare prendiamo il risultato per cui Moretti 66 Cl, Ichnusa non filtrata 50 Cl e Heineken 66 Cl sono leader).
Vengono aggiunte due variabili relative al prezzo non scontato e al prezzo scontato, che otteniamo facendo le divisioni:
Vendite.in.Valore.Senza.promozione/Vendite.in.Volume.Senza.promozione
Vendite.in.Valore.Con.promozione/Vendite.in.Volume.Con.promozione
Un’ultima variabile che viene aggiunta è una variabile binaria che chiamiamo low.volumes, che ci serve per identificare gli outlier che avevamo detto in precedenza, cercando di mitigare il loro effetto. Low.volumes è 1 se le vendite in volume sono inferiori a 100, altrimenti è 0 (scelta fatta vedendo i dati settimanali).
Vedendo le distribuzioni delle vendite in volume e dei prezzi decidiamo di fare la regressione usando la loro trasformazione logaritmica (vedi immagine della distribuzione delle vendite).

Modello lineare

Iniziamo a costruire il modello lineare tenendo in considerazione tutte le variabili dette in precedenza e rimuoviamo quelle che risultano non significative. Alla fine il modello è fatto da:

- Price.NoDiscount.log
- Price.Discount.log
- Discount.Flyer
- Discount.Display
- Discount.PriceReduction
- is.Summer
- is.Leader
- Low.Volumes (aggiunta in seguito a un primo modello lineare per migliorarne il fit eliminando degli outlier bassi rispetto a una soglia di vendita posta arbitrariamente a 100)

Il modello ha un R2 adjusted del 65.05%. Riportiamo la diagnostica del modello (immagini) e vediamo che i residui sono poco normali (code pesanti) e anche l’omoschedasticità non è propriamente rispettata, in particolare i punti sulla sinistra sono proprio gli outlier che dicevamo inizialmente.

Linear mixed models

Plottando i residui del modello lineare rispetto alla categorizzazione prodotto e brand, vediamo che le varianze sono molto diverse (boxplot). Decidiamo quindi di implementare questi due modelli, mettendo prima solo l’intercetta random, poi il prezzo scontato e infine il prezzo non scontato: questi risultano essere i modelli migliori (accertato test anova). Riportiamo il PVRE dei due modelli (tabella).

Paragone dei modelli

Per scegliere quale modello è il migliore tra i tre proposti, oltre che fare un anova test e guardare l'AIC dei modelli, calcoliamo anche la pinball loss sul test set (questa parte ce la deve dire marco). Alla fine il modello migliore risulta essere il lmm per prodotto. Quindi facciamo la predizione usando questo modello, in particolare il nostro test set è costituito dall'ultimo 20% del periodo temporale e riportiamo l'MSE e il MAE, oltre che la figura per la predizione puntuale e gli intervalli di predizione per la Moretti 66 Cl (che è il top prodotto).

Le ultime 2 immagini sono: il dotplot del lmm per prodotto (dx) che ci fa vedere i 95% CI dei random effect e 95% CI dei beta dei tre modelli (sx).
Dal grafico di destra possiamo notare i comportamenti delle birre rispetto alla vendita scontata o meno e vale la pena concentrarsi sulla birra Corona e Dreher (estremi). Dalla corona che ha una stima dell'intercetta positiva alta ci aspettiamo che le vendite siano slegate dall'applicazione di eventuali sconti e che ci sia una alta domanda di questo prodotto. Un coefficiente positivo per il discounted price potrebbe volerci indicare, controintuitivamente, che un prezzo alto scontato ci dia comunque un buon numero di vendite (chatgpt propone anche un'interpretazione di credibilità del prodotto se mantiene prezzo alto). Per il prezzo non scontato invece se questo è più alto va ad abbassare le vendite tendenzialmente.
Da quest'ultimo plot possiamo dire che i prodotti leader (essendo il beta di is.leader positivo) vendono di più dei follower (coerente rispetto alle dinamiche di mercato); che le promozioni del tipo display, volantino e riduzione prezzo impattano in modo positivo sulle vendite e sono le promozioni più significative, mentre special pack e loyalty, non essendo significative per il modello, non sono cosi impattanti; che c'è un seasonal trend, per il fatto che is.summer è positivo, che significa cioè che le vendite in estate sono generalmente più alte; infine il prezzo scontato ha un beta negativo (giusto: più il prezzo è basso, più le vendite sono alte), quindi effettivamente il fatto di avere un prezzo scontato aiuta le vendite, mentre il beta positivo per il prezzo non scontato può essere giustificato dal fatto che questo prodotti vengono venduti comunque bene anche quando il prezzo non è scontato.