Analisi sul consumo di vino e carne in base alle caratteristiche dei clienti

Metodi esplorativi per “Big Data”

II appello – 2 Febbraio 2022

Veronica Milazzo

INTRODUZIONE:

La seguente analisi è stata condotta per rispondere alla richiesta di un’azienda che vuole comprendere meglio i propri clienti e rendere più facile per loro modificare i prodotti in base alle esigenze, ai comportamenti e alle preoccupazioni specifici dei diversi tipi di clienti. In particolare, l’obiettivo finale è prevedere l’importo speso per due categorie di prodotti, vino e carne, in base ad informazioni generali sui clienti.

Il dataset fornito dall’azienda contiene 2240 osservazioni per 29 variabili, ma per l’analisi sono state selezionate 16 variabili che riguardano sia informazioni sui clienti che informazioni sull’importo speso per diverse categorie di prodotti. Le variabili sono le seguenti:

- ID: identificatore univoco di ogni cliente;
- Year_Birth: anno di nascita del cliente;
- Age: età del cliente;
- Education: livello di istruzione del cliente;
- Marital_Status: stato civile del cliente;
- Income: reddito familiare annuo del cliente;
- Kidhome: numero di bambini nel nucleo familiare del cliente;
- Teenhome: numero di adolescenti nel nucleo familiare del cliente;
- Dt_Customer: data di iscrizione del cliente all’azienda;
- Complain: variabile indicatrice delle lamentele del cliente. Assume valore 1 se il cliente si è lamentato negli ultimi due anni, 0 altrimenti;
- MntWines: importo speso per il vino negli ultimi due anni;
- MntFruits: importo speso per la frutta negli ultimi due anni;
- MntMeatProducts: importo speso per la carne negli ultimi due anni;
- MntFishProducts: importo speso per il pesce negli ultimi due anni;
- MntSweetProducts: importo speso per i dolci negli ultimi due anni;
- MntGoldProd: importo speso per l’oro negli ultimi due anni.

L’età del cliente è stata ricostruita partendo dall’anno di nascita del cliente. Le variabili “Education”, “Marital_Satus” e “Complain” sono qualitative, mentre tutte le altre sono quantitative. In particolare, la variabile “Education” comprende cinque livelli di istruzione che sono: Basic – 2n Cycle – Graduation – Master – PhD e la variabile “Marital_Status” comprende due livelli inammissibili (“Absurd” e “YOLO”), ai quali è stato assegnato il valore mancante, e altri sei livelli (“Alone”, “Divorced”, “Married”, “Single”, “Together” e “Widow”) che sono sati raggruppati in soli due livelli:

1. Alone: raggruppa le modalità Alone, Divorced, Single e Widow;
2. Couple: raggruppa le modalità Married e Together.

ANALISI ESPLORATIVA:

Prima di raggruppare le modalità della variabile “Marital_Status”, la verifica dei duplicati ha evidenziato la presenza di 201 clienti presenti più di una volta ma con diverso codice identificativo, i quali sono stati rimossi dal dataset riducendo il numero di osservazioni a 2039.

Inoltre, sono presenti dei valori anomali per le variabili “Income” e “Age”. Osservando la distribuzione del reddito familiare annuo del cliente (grafico 1) si può notare la presenza di un valore anomalo di €666.666 che è stato sostituito con il valore mancante.

Grafico 1

I valori inammissibili per l’età sono mostrati nella tabella 1.

ID	Year_Birth	Age
7829	1900	122
11004	1893	129
1150	1899	123

Tabella 1

Probabilmente in fase di immissione dei dati è stata riportato un anno di nascita diverso da quello comunicato dal cliente. In questo caso gli anni di nascita corretti potrebbero essere rispettivamente 2000, 1993 e 1999, ma comunque vengono considerati dati mancanti.

Oltre ai dati mancanti dovuti ad anomalie di alcuni valori o modalità di alcune variabili, ne sono presenti altri che verranno evidenziati nei grafici seguenti.

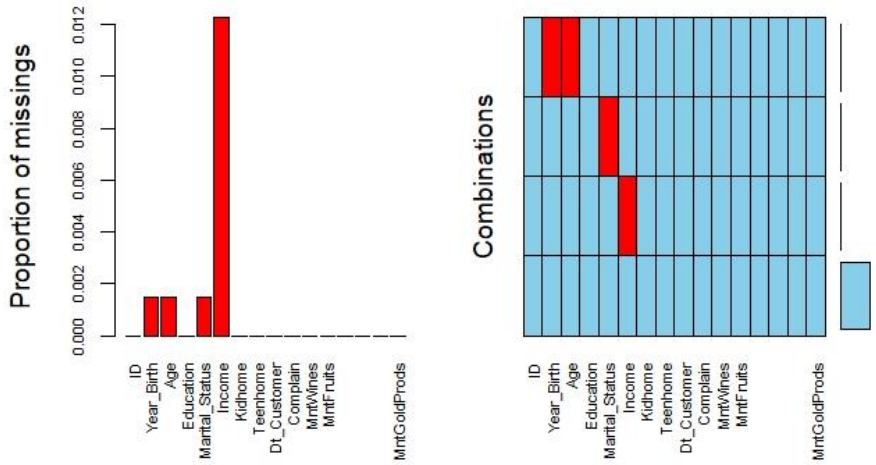

Grafico 2

Il grafico 2 mostra la proporzione di dati mancanti per ogni variabile. Solo la variabile “Income” presenta una proporzione più consistente rispetto alle altre variabili di circa 1,2 %.

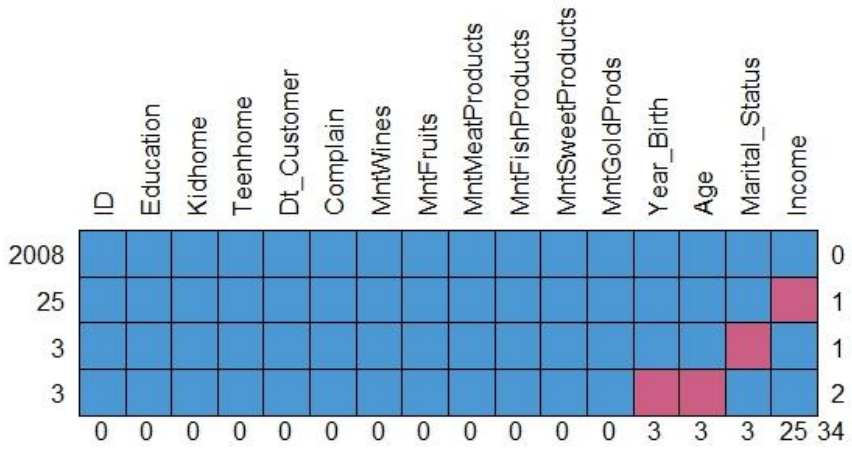

Grafico 3

Il pattern di dati mancanti, mostrato nel grafico 3, è arbitrario e inoltre:

- 2008 osservazioni sono complete;
- Per 25 osservazioni manca il valore del reddito familiare annuo del cliente;
- Per 3 osservazioni non è presente l’informazione sullo stato civile del cliente;
- Per 3 osservazioni non disponiamo dell’età del cliente, e di conseguenza dell’anno di nascita, per quanto visto nella tabella 1.

Comunque, la percentuale di dati mancanti per questo dataset è contenuta ed è circa l’1,5%.

Ricordando che l’obiettivo finale dell’analisi è prevedere l’importo speso per vino e carne, vediamo quali sono le relazioni tra le variabili di interesse e le altre variabili.

Nel grafico 4 possiamo vedere le correlazioni delle variabili numeriche, calcolate considerando solo le osservazioni complete.

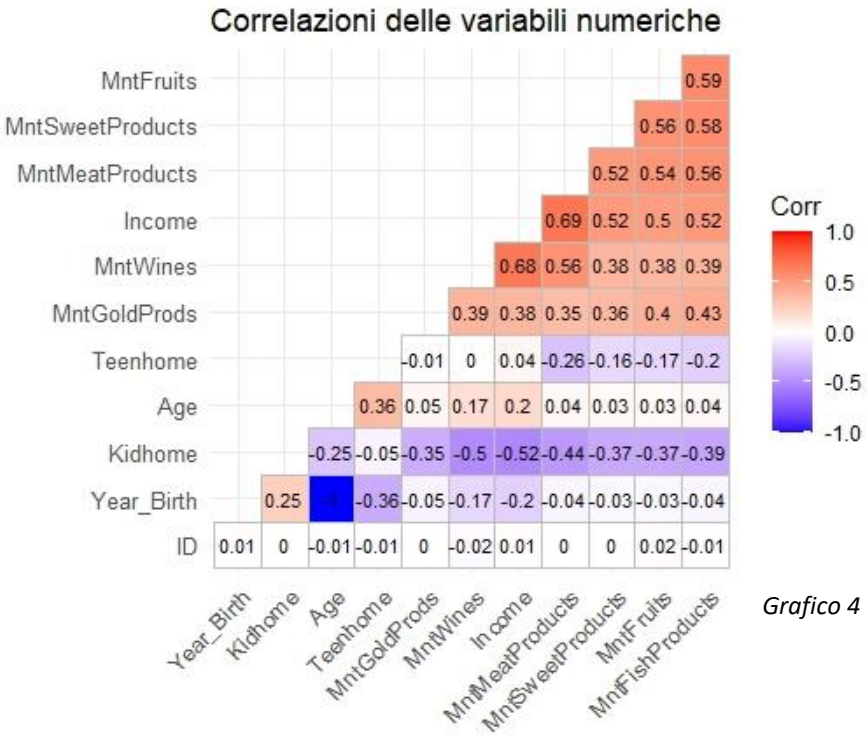

Grafico 4

Considerando le correlazioni tra l'importo speso per il vino e le altre variabili notiamo che l'importo speso per il vino:

- È correlato positivamente al reddito familiare annuo e all'importo speso per tutti gli altri prodotti, in particolare per la carne;
- È correlato debolmente anche all'età;
- È correlato negativamente al numero di bambini nel nucleo familiare;

Considerando le correlazioni tra l'importo speso per la carne e le altre variabili notiamo che l'importo speso per il carne:

- È correlato positivamente al reddito familiare annuo e all'importo speso per tutti gli altri prodotti, in particolare per quelli alimentari;
- È correlato negativamente al numero di bambini e adolescenti nel nucleo familiare.

Nel grafico 5 possiamo vedere la distribuzione dell'importo speso sia per il vino che per la carne condizionatamente alle variabili qualitative e l'andamento dell'importo speso sia per il vino che per la carne rispetto alla data di iscrizione del cliente all'azienda.

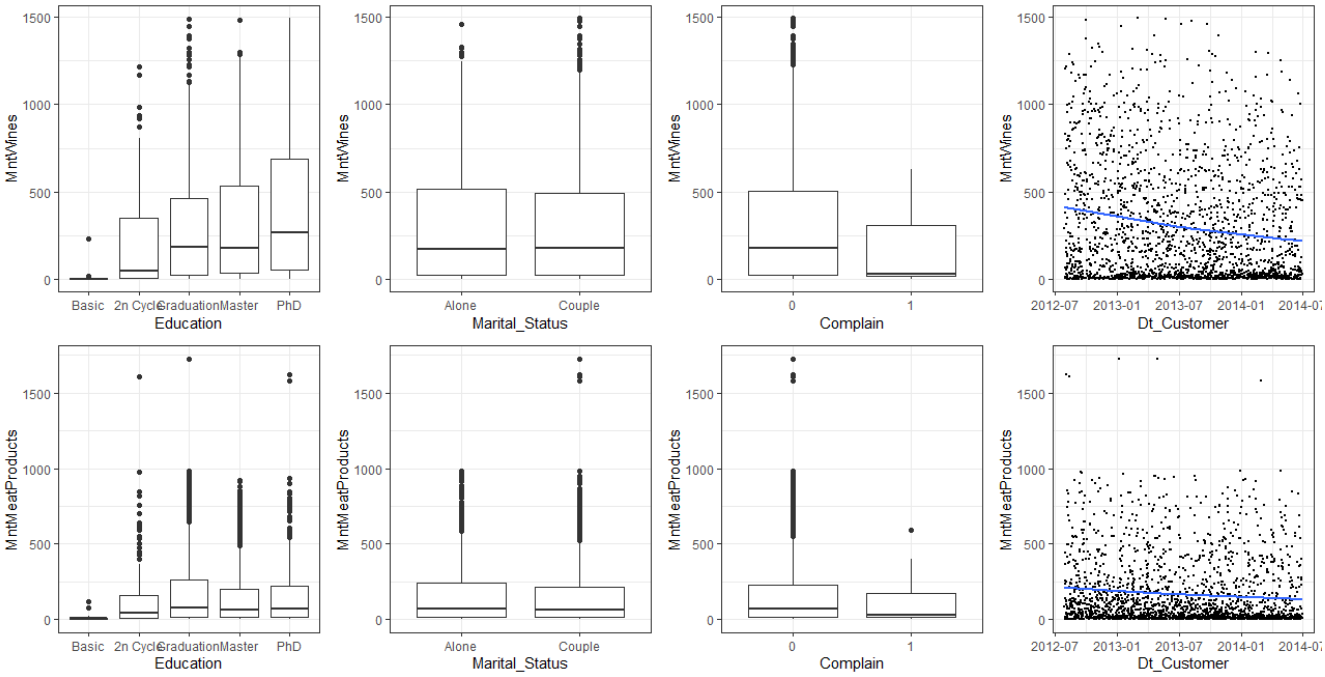

Grafico 5

Si evince che al crescere del livello di istruzione cresce l'importo speso per il vino; inoltre, chi si è lamentato spende di meno di chi non si è lamentato negli ultimi due anni e man mano che i clienti si fidelizzano all'azienda l'importo speso per il vino aumenta. Le stesse relazioni valgono anche per l'importo speso per la carne, ma sono molto più deboli.

PROCEDURA DI BOOSTING E ALBERO DI REGRESSIONE:

Per predire l'importo speso per il vino e per la carne partendo da alcune caratteristiche del cliente e dall'importo speso per alcune categorie di prodotti, è stata ottimizzata una procedura di Boosting utilizzando una frazione di training pari al 75%.

Inizialmente il dataset è stato partizionato in due sottoinsiemi, train e test, poiché si vuole utilizzare il test set per confrontare la procedura di Boosting con il singolo albero di regressione.

Per l'ottimizzazione, la scelta da fare riguarda tre parametri:

1. Il numero di alberi, in questo caso posto uguale a 5000;
2. La profondità, in questo caso scelta tra 1, 3 o 6;
3. Il parametro di shrinkage λ, in questo caso scelto tra 0.001, 0.005 e 0.01.

Importo speso per il vino:

Per quanto riguarda l'importo speso per il vino, la procedura ottima che porta a una maggiore riduzione di errore calcolato usando le osservazioni del test set è, come mostrato nel grafico 6, quella con profondità pari a 6 e λ pari a 0.001.

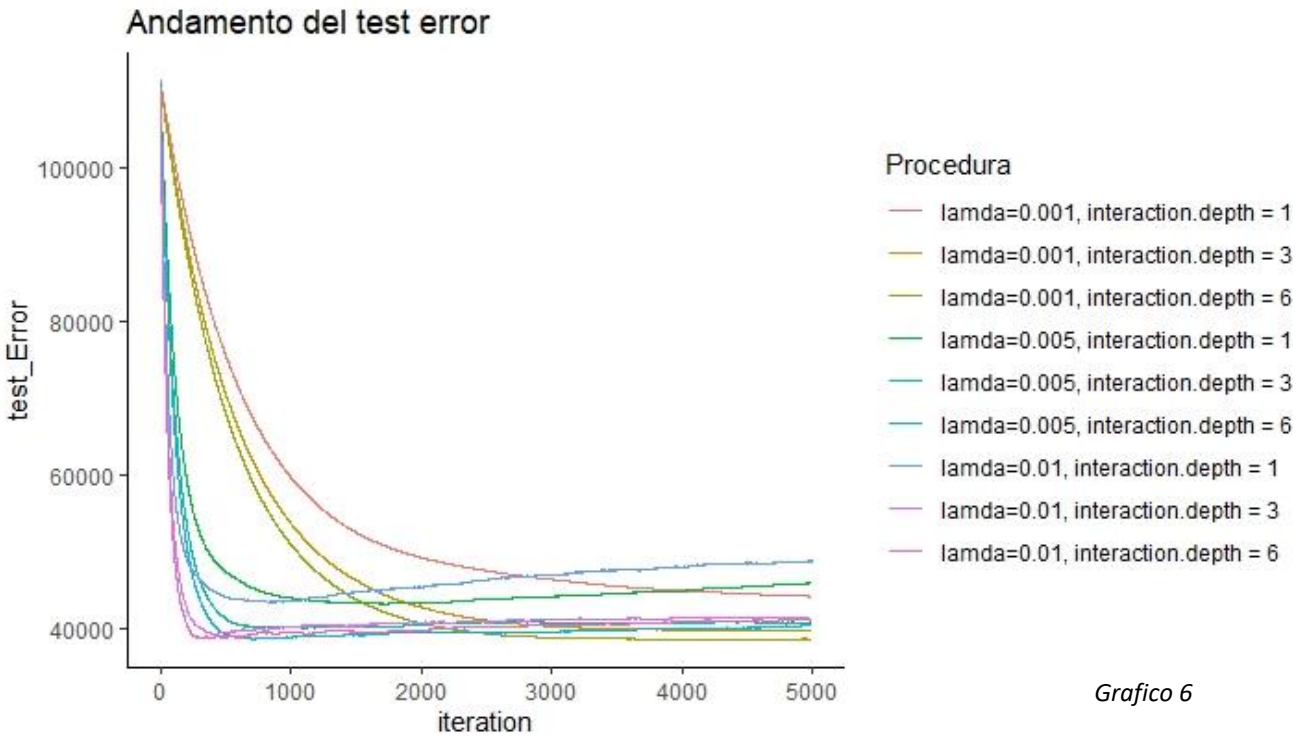

Grafico 6

Il numero di alberi ottimo è pari a 4264.

Le variabili più importanti sono presentate nella tabella 2 e nel grafico 7. Si nota che, come visto nell'analisi esplorativa, il reddito familiare annuo, l'importo speso per la carne, la data di iscrizione del cliente all'azienda ma anche il livello di istruzione influiscono molto sull'importo speso per il vino. Nonostante sembrava ci fosse una forte relazione tra la variabile "Complain" e l'importo speso per il vino, questa variabile non è molto importante a fini previsivi, ciò è dovuto al fatto che solo 1% dei clienti si è lamentato negli ultimi due anni.

Variabili	Influenza relativa
Income	51.19273403
MntMeatProducts	21.20559415
Dt_Customer	6.58859647
MntFruits	3.50902356
Education	3.47615671
MntGoldProds	3.42807243
MntFishProducts	3.23953366
MntSweetProducts	2.48266550
Kidhome	1.66595526
ID	1.64369907
Year_Birth	0.75929883
Age	0.64480362
Teenhome	0.10006494
Marital_Status	0.06380175
Complain	0.00000000

Tabella 2

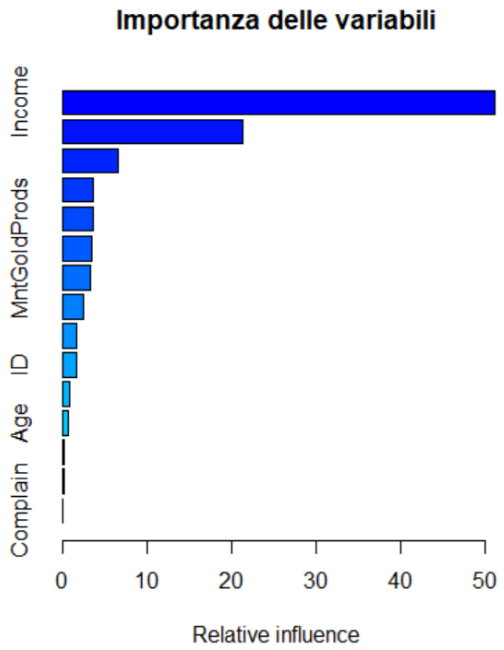

Grafico 7

L'albero di regressione ottimo, scelto ponendo il parametro di costo complessità uguale a 0.0124750 secondo il criterio una volta l'errore standard, è rappresentato nel grafico 8.

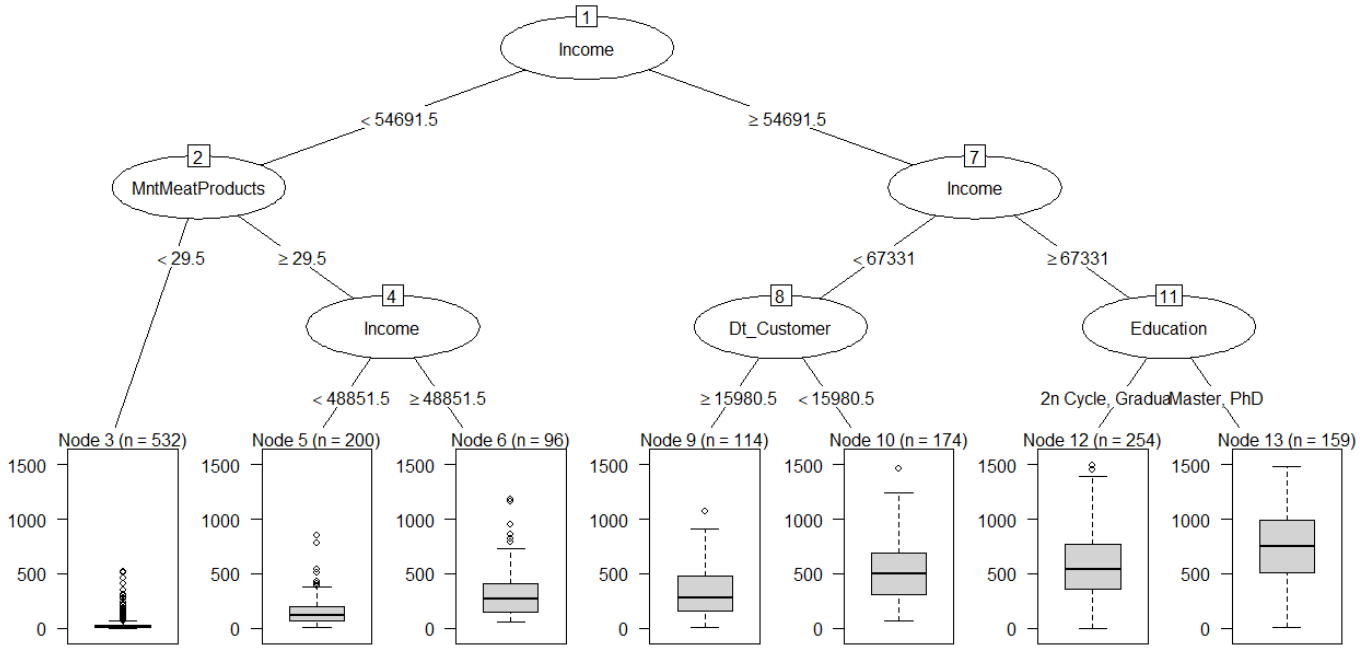

Grafico 8

Ancora una volta si nota che le variabili più importanti e utilizzate come split principale sono il reddito familiare annuo, l'importo speso per carne, la data di iscrizione del cliente all'azienda e il livello di istruzione. Queste variabili sono tra le più importanti della procedura di Boosting. Inoltre, la variabilità dei boxplot di alcune foglie è molto elevata, segno che questo albero non riesce a restituire buone previsioni.

Il singolo albero di regressione è instabile, per questo si preferisce usare la procedura di Boosting che ha una migliore capacità predittiva. Nel grafico 9 è possibile vedere il confronto tra la

performance sul test set della procedura di Boosting ottima e la performance sul test set dell'albero di regressione.

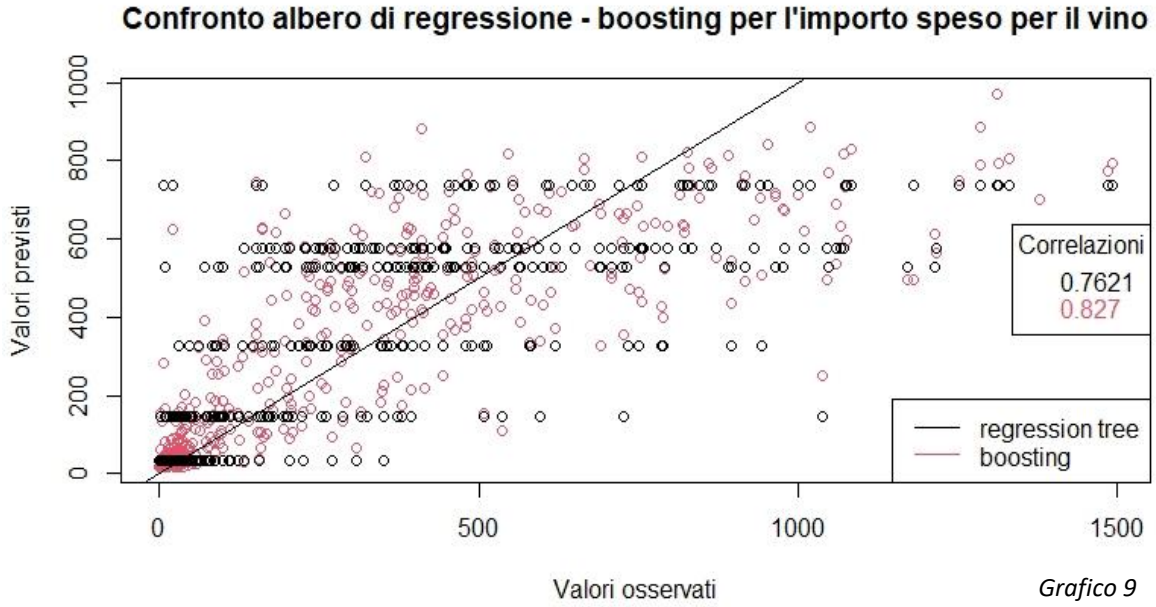

Grafico 9

La procedura di Boosting è la migliore fra le due perché, rispetto all'albero di regressione, la correlazione tra valori osservati e valori previsti è più alta (0.827) e riesce a prevedere meglio soprattutto i valori più bassi.

Importo speso per la carne:

Per quanto riguarda l'importo speso per la carne, la procedura ottima che porta a una maggiore riduzione di errore calcolato usando le osservazioni del test set è, come mostrato nel grafico 10, quella con profondità pari a 3 e λ pari a 0.001.

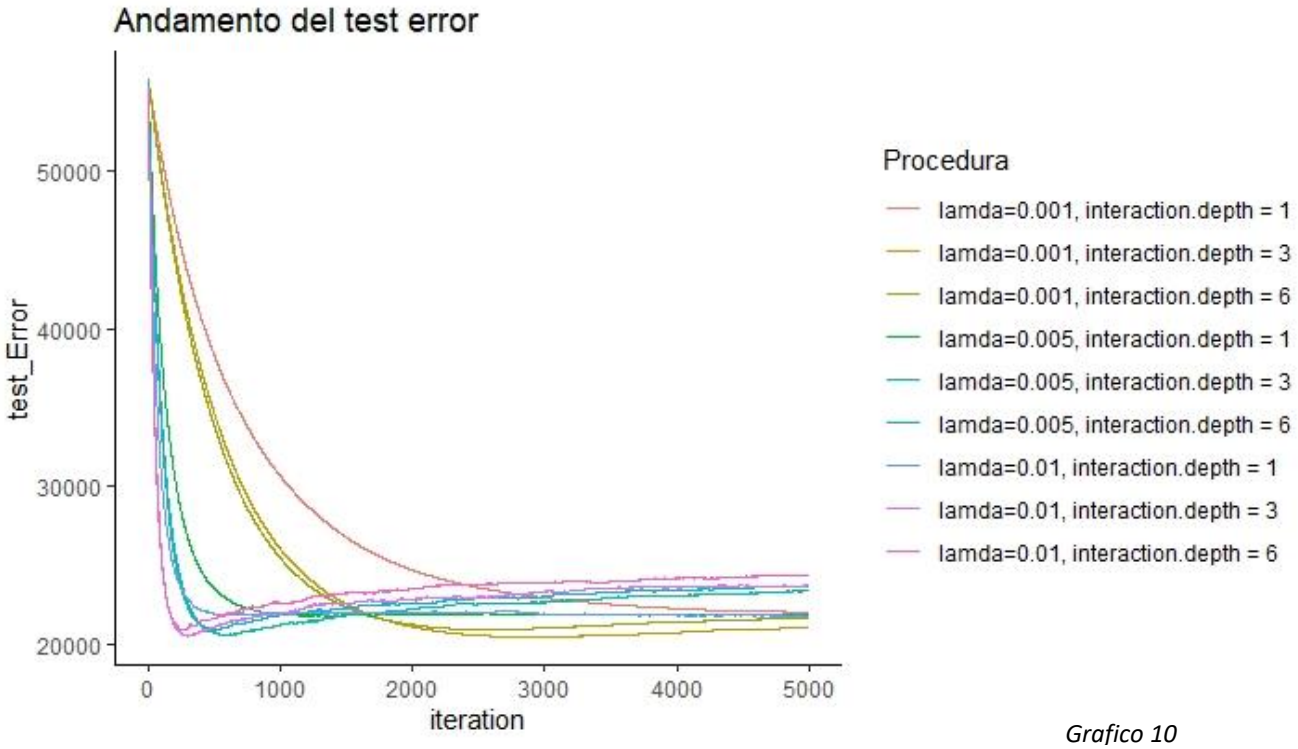

Grafico 10

Il numero di alberi ottimo è pari a 2888.

Le variabili più importanti sono presentate nella tabella 3 e nel grafico 11. Si nota che, come visto nell'analisi esplorativa, il reddito familiare annuo e l'importo speso per i prodotti alimentari influiscono molto sull'importo speso per la carne. Inoltre, il numero di adolescenti nel nucleo familiare è la seconda variabile più importante.

Variabili	Influenza relativa
Income	63.7763694
Teenhome	7.8728684
MntFruits	6.3572126
MntFishProducts	5.8377303
MntWines	5.2700155
MntSweetProducts	3.7285992
Dt_Customer	3.7152522
MntGoldProds	1.2320403
ID	0.8504738
Kidhome	0.3652078
Marital_Status	0.3236954
Year_Birth	0.2488017
Education	0.2118538
Age	0.2098797
Complain	0.0000000

Tabella 3

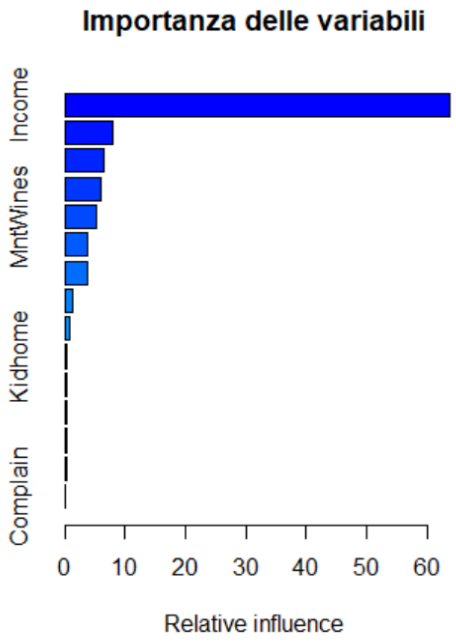

Grafico 11

L'albero di regressione ottimo, scelto ponendo il parametro di costo complessità uguale a 0.0182744 secondo il criterio una volta l'errore standard, è rappresentato nel grafico 12.

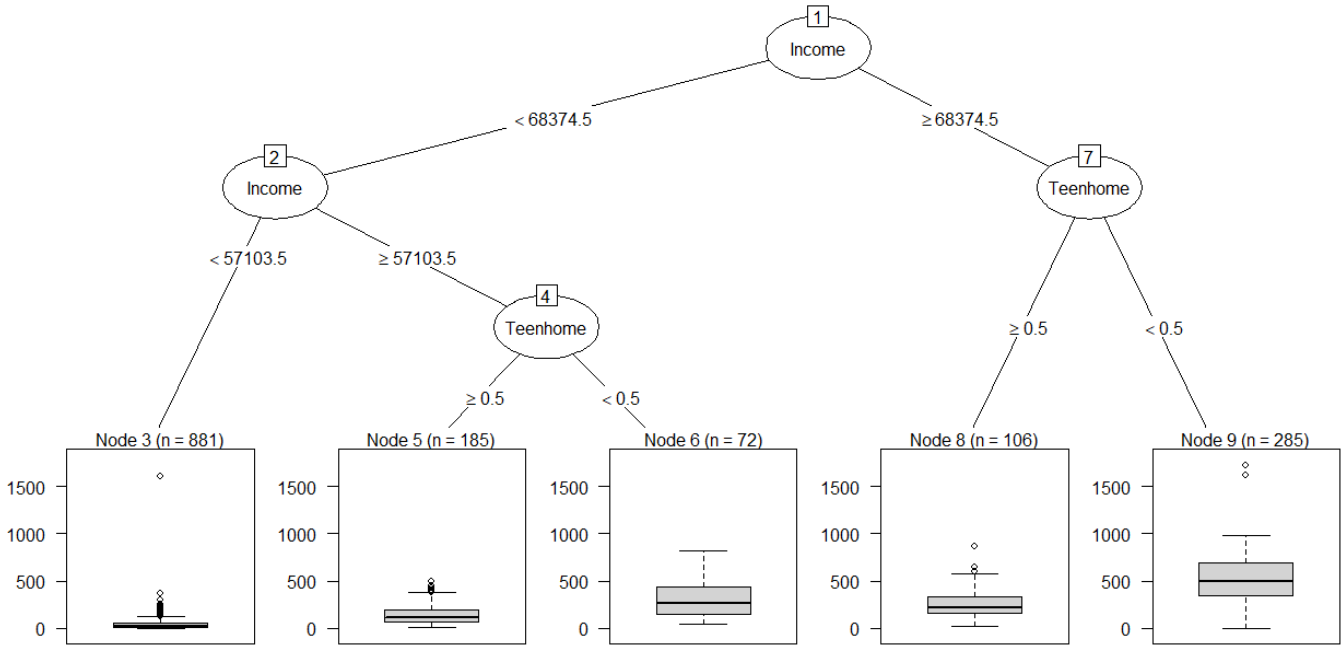

Grafico 12

È un albero più semplice rispetto al precedente. Ancora una volta si nota che le variabili più importanti e utilizzate come split principale sono il reddito familiare annuo e il numero di adolescenti nel nucleo familiare. Queste sono le prime due variabili più importanti della procedura di Boosting.

Nel grafico 13 è possibile vedere il confronto tra la performance sul test set della procedura di Boosting ottima e la performance sul test set dell'albero di regressione.

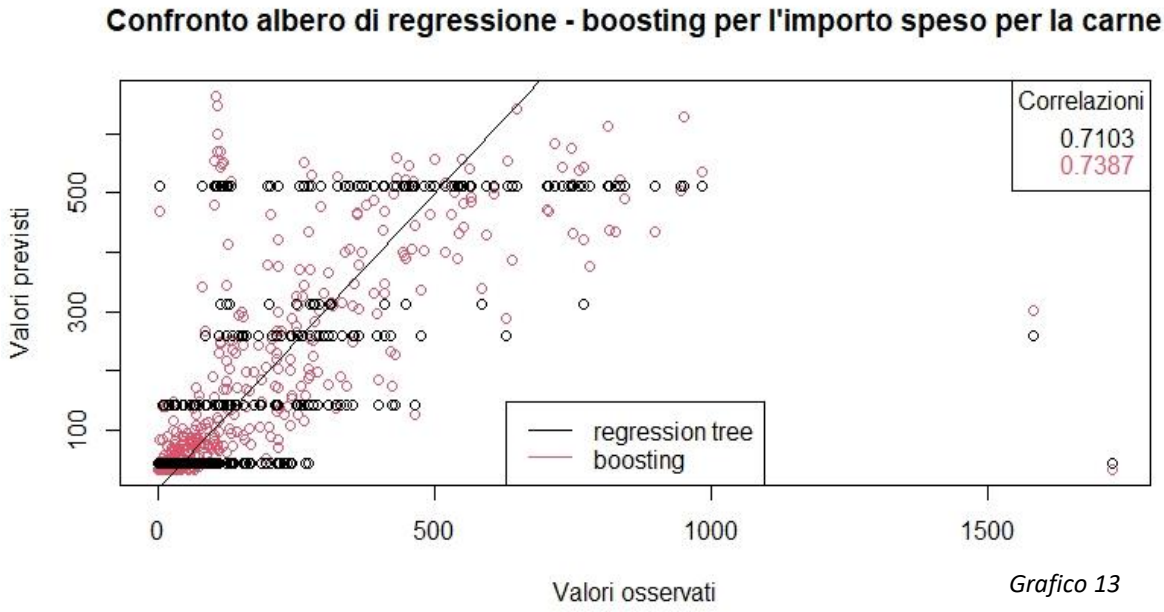

Grafico 13

La procedura di Boosting è leggermente migliore rispetto all'albero di regressione poiché la correlazione tra valori osservati e valori previsti è appena più alta (0.7387) e riesce a prevedere un po' meglio i valori molto bassi. In questo caso non c'è molta differenza nella performance predittiva delle due procedure.

Confronto dell'effetto del reddito familiare annuo con l'analisi esplorativa:

Per cercare di capire perché la procedura di Boosting, nonostante sia molto più stabile del singolo albero di regressione, non riesce a prevedere bene i valori alti ci concentriamo sulla variabile "Income", cioè la più importante per prevedere l'importo speso sia per il vino che per la carne. Nei grafici 14 e 15 è mostrato il confronto tra l'effetto marginale nella procedura di Boosting e l'andamento nei dati osservati.

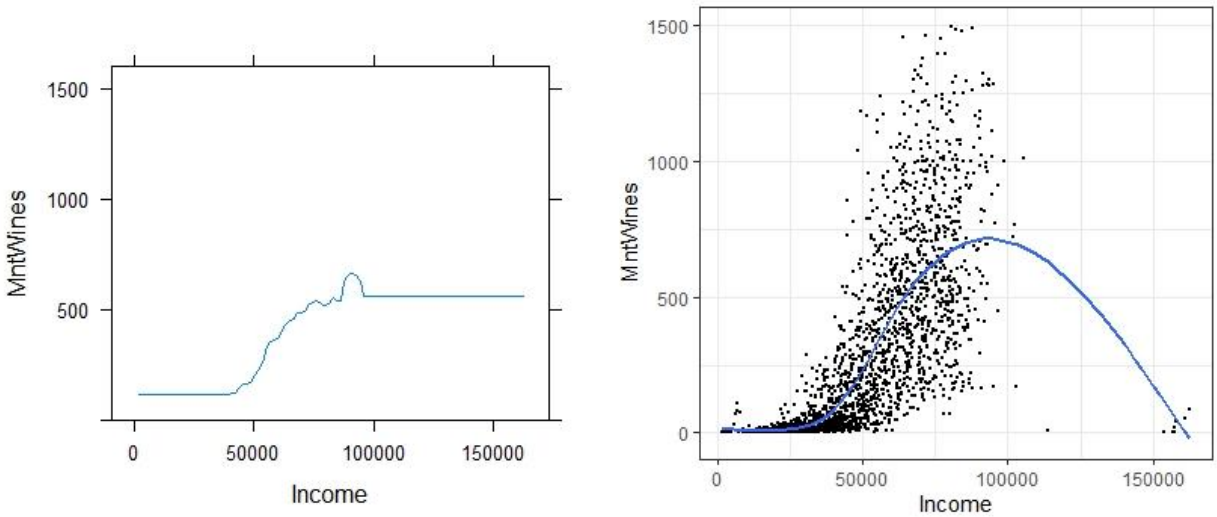

Grafico 14

Dal grafico 14 si nota che il reddito familiare annuo non permette di prevedere valori alti dell'importo speso per il vino perché il gruppo di clienti con reddito superiore a €150.000 spende molto poco per il vino. Ciò stabilizza l'effetto della variabile "Income" a una spesa intorno a €700.

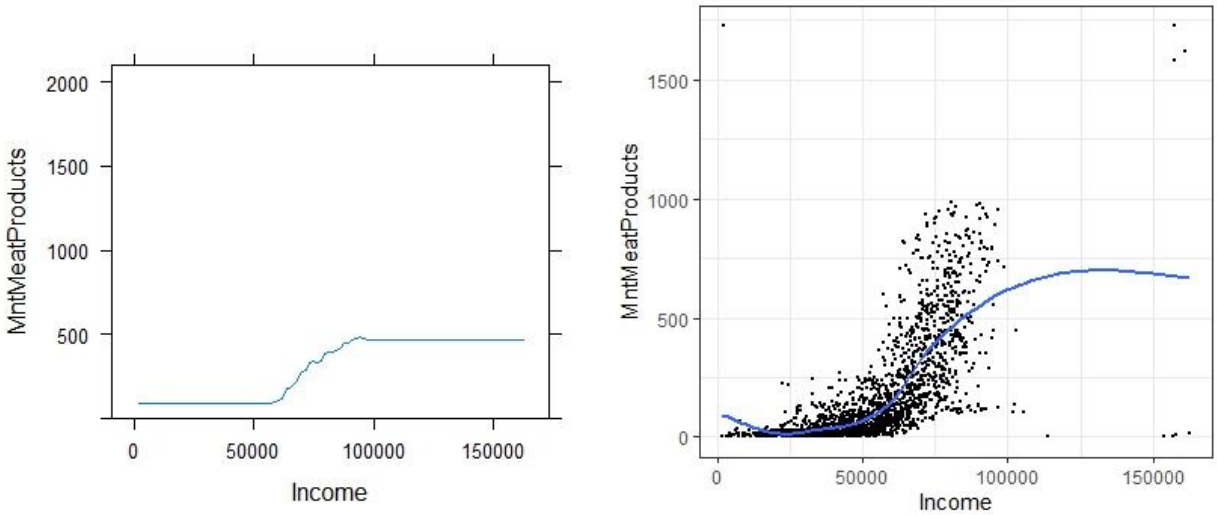

Grafico 15

Dal grafico 15 si nota che il reddito familiare annuo non permette di prevedere valori alti dell'importo speso per la carne perché il gruppo di clienti con reddito superiore a €150.000 si divide nettamente tra coloro che spendono più di €1500 e coloro che non spendono quasi nulla. Ciò stabilizza l'effetto della variabile "Income" a una spesa intorno a €500.

Se non ci fosse quel gruppo di clienti con reddito superiore a €150.000 l'effetto della variabile "Income" sarebbe molto più crescente e sarebbero previsti anche spese per il vino e per la carne superiori rispettivamente a €700 e a €500.

CONCLUSIONI:

Confrontando le tabelle 2 e 3, è possibile notare una certa analogia:

- Le variabili più importanti per entrambe le procedure sono il reddito familiare annuo e l'importo speso per tutte le altre categorie di prodotti;
- Le variabili meno importanti per entrambe le procedure sono l'età, e di conseguenza l'anno di nascita del cliente, lo stato civile e la presenza di lamentele negli ultimi due anni.

La differenza sostanziale è che il livello di istruzione è tra le variabili più importanti per prevedere l'importo speso per il vino e il numero di adolescenti nel nucleo familiare è tra le variabili più importanti per prevedere l'importo speso per la carne.