CORSO DI LAUREA MAGISTRTALE IN STATISTICA & DATA SCIENCE

DIPARTIMENTO DI SCIENZE ECONOMICO AZIENDALI E STATISTICHE

TITOLO INDAGINE:

ANALISI DEL COMPORTAMENTO E ESIGENZE DEI CLIENTI DI UN'AZIENDA, CON L'OBIETTIVO DI PREVEDERE L'IMPORTO SPESO PER DUE CATEGORIE DI PRODOTTO: VINO E CARNE

Marco Speciale

Anno accademico 2021/2022

Indice

1. Introduzione

I dati forniti a questo studio comprendono:

- informazioni generali dei clienti fornite dagli stessi in fase di registrazione con l'azienda

- spesa totale per ogni categoria di prodotto negli ultimi due anni

L'obiettivo finale dell'analisi è quello di prevedere l'importo speso per due differenti categorie di prodotto, vino e carne, sfruttando i dati dei clienti a disposizione. Oggetto di particolare attenzione gli elementi che maggiormente incidono sulle scelte di acquisto delle sopracitate categorie di prodotto, quali il reddito di ciascun cliente, la scelta di altre categorie di prodotto, lo stato civile e il livello di istruzione.

2. Data cleaning

Gli elementi di indagine sono pervenuti in formato csv.
Per analizzare i dati è stato utilizzato il software di programmazione "R".
Il dataset iniziale era composto da 2240 osservazioni e 29 variabili ma le variabili oggetto di questa analisi sono esclusivamente 15. (Vedi appendice)

Dopo avere visualizzato la tipologia di variabili presente si è immediatamente provveduto a verificare la presenza di eventuali ripetizioni nelle osservazioni.

Preliminarmente per potere procedere con la rimozione delle ripetizioni è stata esclusa la variabile ID in quanto non permetteva la visualizzazione di osservazioni ripetute dato che vi erano osservazioni identiche con ID diversi.
Sono state riscontrate 201 ripetizioni a fronte di 2240 osservazioni, successivamente eliminate sfruttando la funzione **unique** che ha appunto eliminato le stesse.
E' stata creata la variabile Enrollment la quale ha sostituito la variabile Dt_Customer, in quest'ultima nuova variabile viene ricavato il solo anno di registrazione dei clienti non tenendo conto né del mese né del giorno, poi è stata creata la variabile Age in cui è stata riportata l'età dei clienti ottenuta dall'anno di riferimento a questa indagine (2021) e la variabile Year_Birth; ulteriormente le modalità della variabile Marital_Status sono state opportunamente aggregate in una nuova variabile denominata Status.

Dunque Marital_Status è stata rimpiazzata con Status non considerando valide alcune delle modalità come: "YOLO" e "Absurd"; conseguentemente sono stati ordinati i livelli della variabile Education seguendo l'ordine gerarchico dei corsi di studio.
Inoltre, sono stati riscontrati dei valori che non potevano essere considerati validi ai fini dell'analisi per la variabile Age, dove tre clienti risultavano essere ultracentenari e per la variabile Income, dove un valore indicato era pari a "6666666". Quindi al fine di verificare che il dataset avesse valori validi per l'indagine, sono state applicate delle regole, con particolare riferimento alle variabili Age e Income in modo tale da localizzare tali errori e sostituirli con degli NA. Oltre gli errori sopracitati non ne sono stati riscontrati ulteriori.Il dataset finale oggetto di analisi è composto da 2039 osservazioni e 14 variabili (ID unica variabile rimossa).

3. Analisi esplorativa

Al fine di verificare la completezza delle osservazioni nel dataset è stata verificata l'eventuale presenza di dati mancanti. Grazie alle funzione **rsem.pattern** il cui output si presenta sotto forma di matrice indicatrice del dato mancante è stato possibile individuare tutti i pattern di dati mancanti.
Il pattern 1 è composto da 2008 righe su 2039, caratterizzate dal fatto che tutte le 14 variabili sono state osservate, poi troviamo i pattern 2 e 4 nei quali sono mancanti tre dati ciascuno per le variabili Age e Status ed infine nel pattern 3 sono mancanti 25 dati per la variabile Income.
Attraverso l'utilizzo della funzione **aggr** è stata individuata la percentuale di dati mancanti marginalmente per ognuna delle variabili (parte sinistra del grafico) mentre nella parte destra vi è la rappresentazione, non in scala, dei pattern.
L'output ha restituito le informazioni riportate di seguito:

- La percentuale di dati osservati è pari al 98.48%
- La percentuale di dati mancanti per la variabile Income è pari 1.22%
- La percentuale di dati mancanti per le variabili Age e Status è rispettivamente pari allo 0.14% (Figura 1 in appendice)

Dall'analisi sui pattern mancanti si evince come questi siano presenti esclusivamente nelle variabili esplicative e non sulle variabili di risposta, dunque, si è ritenuto non necessario effettuare un'imputazione per i dati mancanti. Usufruendo di rappresentazioni grafiche si sono confrontatela le distribuzioni delle risposte MntWines e MntMeatProducts quando l'esplicativa Income era osservata o mancante. (Figure 2 e 3 in appendice)

Al fine di verificare quali fattori abbiano inciso maggiormente sulle scelte d'acquisto nelle categorie di prodotto vino e carne, si è proceduto quantificando la correlazione tra le variabili risposta e le esplicative. Tramite l'ausilio di svariate rappresentazioni grafiche interattive si è dedotto che le variabili più correlate con le risposte Mnt Wines e MntMeatProducts sono:

- Income che ha una correlazione pari a 0.68 con MntWines e 0.69 con MntMeatProducts
- MntFruits, MntFishProducts, MntSweetProducts che hanno una correlazione di circa 0.40 con MntWines ma si dimostrano maggiormente correlate con MntMeatProducts, ognuna con valori diversi che si attestano attorno a 0.55
- MntGoldProducts che ha una correlazione pari a 0.39 con MntWines e 0.34 con MntMeatProducts

Si sottolinea come si sia evidenziata una correlazione pari a 0.56 tra le due variabili risposta MntWines e MntMeatProducts.

Dopo avere svolto una prima analisi esplorativa per i dati mancanti e avere valutato le correlazioni tra le variabili risposta e le esplicative, le indagini esplorative sono state svolte sulle distribuzioni delle risposte.

In merito quella di MntWines si è evidenziata un'asimmetria positiva che dimostra come vi sia un numero molto elevato di clienti con una spesa inferiore a 100 per questa categoria di prodotto. La distribuzione decresce all'aumentare del consumo di vino fino ad arrivare a sette clienti che superano la soglia di 1400. (Figura 7 in appendice)

Analoghe considerazioni sono state fatte per la distribuzione di MntMeatProducts sia per l'asimmetria positiva che per la quantità di clienti che si sono contenuti nella spesa, la distribuzione è anch'essa decrescente, infine si evidenziano cinque clienti che hanno speso più di 1500. (Figura 8 in appendice)

Visualizzando la distribuzione congiunta di MntWines e MntMeatProducts si manifesta che ad un basso consumo di vino corrisponde un basso consumo di carne, al crescere del consumo di queste due categorie di prodotto in alcuni casi si è verificato che ad un elevato consumo di vino non corrisponda un elevato consumo di carne e viceversa, anzi i clienti che hanno acquistato grandi quantità di vino o carne prediligono solo uno dei due prodotti a discapito dell'altro. Questa deduzione è confermata anche dalla correlazione tra le due variabili pari a 0.56 quindi il risultato è tutt'altro che inaspettato. (Figura 9 in appendice)

Confrontando le distribuzioni congiunte delle risposte con l'esplicativa Income si è evidenziato come nella zona iniziale delle distribuzioni all'aumentare del reddito, l'ammontare di spesa per vino e carne sia sempre contenuta entro determinati limiti per poi crescere vertiginosamente nel caso del vino, mentre per la carne si rileva una crescita moderata. Entrambe le distribuzioni presentano degli outlier, nel caso di MntWines vi sono clienti con elevati valori di reddito che però hanno acquistato esigue quantità di vino, invece per MntMeatProducts gli outlier sono più variegati, infatti vi sono clienti che come nel caso di MntWines hanno acquistato esigue quantità di carne pur avendo elevati valori di reddito; poi vi sono coloro i quali ad una copiosa spesa in carne corrispondono livelli alti di reddito per poi arrivare ad un unico cliente che ha si un valore di reddito basso ma che ha speso più di 1500 per la carne negli ultimi due anni. Il comportamento d'acquisto singolare di quest'ultimo cliente è stato posto sotto lente di ingrandimento, rilevando che comunque si trattasse di un cliente sposato e che il suo reddito fosse comunque superiore, se pur di poco (2447), alla spesa effettuata in carne (1725) (Figure 10-11 in appendice). Analizzando la distribuzione di Mnt Wines e MntMeatproducts condizionatamente ad Education e Status è evidente che chi abbia consumato più vino siano i divorziati e i vedovi con livelli di Education quali: PhD, 2nd Cycle, Master e Graduation. Più in generale coloro i quali abbiano conseguito PhD e Master sono consumatori abituali di vino con una spesa maggiore di 500. Per chi sceglie la carne il grafico ha evidenziato come i maggiori consumatori siano stati i single, spendendo almeno 250. (Figure 12-13 in appendice)

4. Boosting e albero di regressione

Per mezzo di questi due metodi è stato possibile raggiungere l'obiettivo finale dello studio; infatti, le procedure sopraelencate hanno permesso di effettuare delle previsioni per le variabili risposta (MntWines e MntMeatProducts).La procedura di boosting per MntWines è stata portata avanti implementando nove diversi tentativi in ciascuno dei quali si sono variati i valori dei parametri:

- Numero di alberi ottimo per la procedura
- λ, detto parametro shrinkage, il quale controlla la velocità di adattamento dei dati alla procedura
- Interaction depth, che controlla la profondità dell'albero

Dopo avere implementato tutte le procedure con l'ausilio della funzione **gbm** è stata selezionata quella che riportava il valore test error più piccolo con numero di alberi pari a 611, shrinkage pari a 0.05 e interaction depth pari a 3. Le variabili di maggiore rilievo per questa procedura sono Income e MntMeatProducts confermato anche da ulteriori rappresentazioni grafiche dalle quali si nota come entrambe abbiano un effetto marginale positivo su MntWines coerentemente con quanto visto in analisi esplorativa. (Figure 14-15-16-17-18-19 e Tabella 1 in appendice)

Per quanto concerne la creazione dell'albero di regressione, sempre per MntWines, è stato selezionato come training set il primo 75% delle osservazioni e come test set il restante 25% (è stato verificato preliminarmente che il dsataset non avesse nessun tipo di ordine crescente o decrescente delle variabili). Sfruttando la funzione **rpart** è stato generato l'albero poi successivamente potato con la funzione **prune** inserendo l'opportuno valore del parametro cp (parametro costo-complessità). L'albero potato ha un R^2 di 0.61 e una correlazione di 0.78, ancora le variabili più importanti si confermano essere Income e MntMeatProducts. Figure (20-21-22 e Tabella 2 in appendice)

In fase finale è stato fatto il confronto tra il singolo albero di regressione con la procedura di boosting ed è stato possibile individuare le previsioni (punti neri in orizzontale) del singolo albero indicate dalla media di tutte le osservazioni all'interno delle stesse foglie; non è stato però possibile prevedere valori alti nella variabile risposta, infatti, nell'asse delle ordinate y il limite superiore è di 800 mentre per le ascisse raggiunge 1500. Infine, è stato considerato come indicatore di performance la correlazione: quella della procedura di boosting è pari 0.81 che si dimostra essere superiore a quella del singolo albero, inoltre la correlazione fra le due previsioni è uguale a 0.90 dunque le procedure hanno previsioni simili ma il boosting è stato selezionato come metodo migliore per effettuare le previsioni su MntWines. (Figura 23 in appendice)

Per la procedura di boosting e dell'albero di regressione per MntMeatProdcucts si è seguito pedissequamente l'iter sopra riportato. Le variabili di maggiore rilievo per il boosting sono Income ed MntWines, ciò confermato anche da ulteriori rappresentazioni grafiche dalle quali si nota come entrambe abbiano un effetto marginale positivo su MntMeatProducts coerentemente con quanto visto in analisi esplorativa. (Figure 24-25-26-27-28-29 e Tabella 3 in appendice)

Per l'albero di regressione invece l'albero potato ha un R^2 pari a 0.60 e le variabili di maggiore rilievo sono Income e MntFishProducts e una correlazione di 0.77. (Figure 30-31-32 e Tabella 4 in appendice)

Nuovamente in fase finale, confrontando il singolo albero di regressione con la procedura di boosting è stato possibile affermare tutte le considerazioni precedentemente dedotte, in aggiunta c'è però da dire che la media dei valori nelle foglie in questo caso è influenzata da dei valori outlier. Anche in questo caso non è stato possibile prevedere valori alti nella variabile risposta dato che hanno scale di misura differenti, infatti, nell'asse delle ordinate il limite superiore della y è 500 mentre per le ascisse la x arriva a 1500. Infine, utilizzando sempre come indicatore di performance la correlazione: quella della procedura di boosting è pari 0.78 che si dimostra essere superiore di quella del singolo albero; inoltre la correlazione fra le previsioni delle due procedure è uguale a 0.94 dunque le procedure hanno previsioni simili ma il boosting è stato selezionato come metodo migliore per effettuare le previsioni su MntMeatProducts. (Figura 33)

5. Conclusioni

Dalla seguente analisi è stato possibile delineare alcuni degli aspetti fondamentali, tipici di un consumatore di vino e carne. In primo luogo si è visto come sia il reddito annuale di ciascun cliente a determinare, nella maggior parte dei casi, l'ammontare di spesa per le due categorie di prodotto, questo però allo stesso tempo non è determinante per la scelta fra vino e carne; infatti, l'analisi di confronto tra questi due prodotti ha permesso, nello specifico, di approfondire quali fossero i profili che maggiormente preferivano vino e carne. Tra i più frequenti consumatori di vino sono stati individuati clienti divorziati e vedovi con gradi di istruzione almeno pari alla laurea, mentre tra gli abituali consumatori di carne sono stati riscontrati clienti single ed ancora una volta i vedovi con gli stessi livelli di istruzione dei consumatori frequenti di vino. Si registra inoltre una bassa correlazione tra le due categorie di prodotto e l'acquisto di oro.

Infine, nonostante i metodi di previsione adottati abbiano restituito previsioni per determinati valori, tali metodologie non possono essere ritenute soddisfacenti ai fini di questa analisi, in quanto non hanno potuto permettere le previsioni di valori elevati nelle variabili risposta.

Appendice

Dataset finale:
Informazioni generali dei clienti

- ID: identificativo unico per cliente
- Year_Birth: anno di nascita del cliente
- Education: grado di istruszione del cliente
- Marital_Status: stato civile del cliente
- Income: reddito annuale del cliente
- Kidhome: numero di bambini in casa del cliente
- Tennhome: numero di adolescenti in casa del cliente
- Dt_Customer: data di registrazione del cliente alla azienda
- Complain: 1 se il cliente ha manifestato delle lamentele negli ultimi due anni, 0 il contrario

Spesa totale per ogni categoria di prodotto negli ultimi due anni

- MntWines: ammontare speso in vino negli ultimi due anni
- MntFruits: ammontare speso in frutta negli ultimi due anni
- MntMeatProducts: ammontare speso in carne negli ultimi due anni
- MntFishProducts: ammontare speso in pesce negli ultimi due anni
- MntSweetProducts: ammontare speso in dolci negli ultimi due anni
- MntGoldProds: ammontare speso in oro negli ultimi due anni

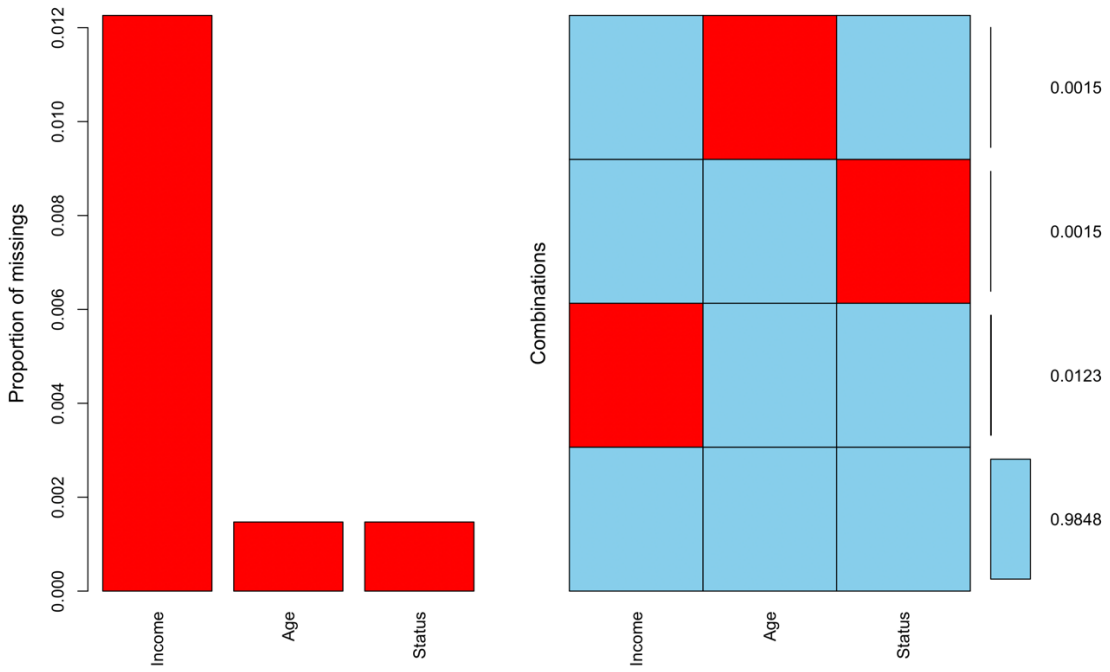

Figura 1

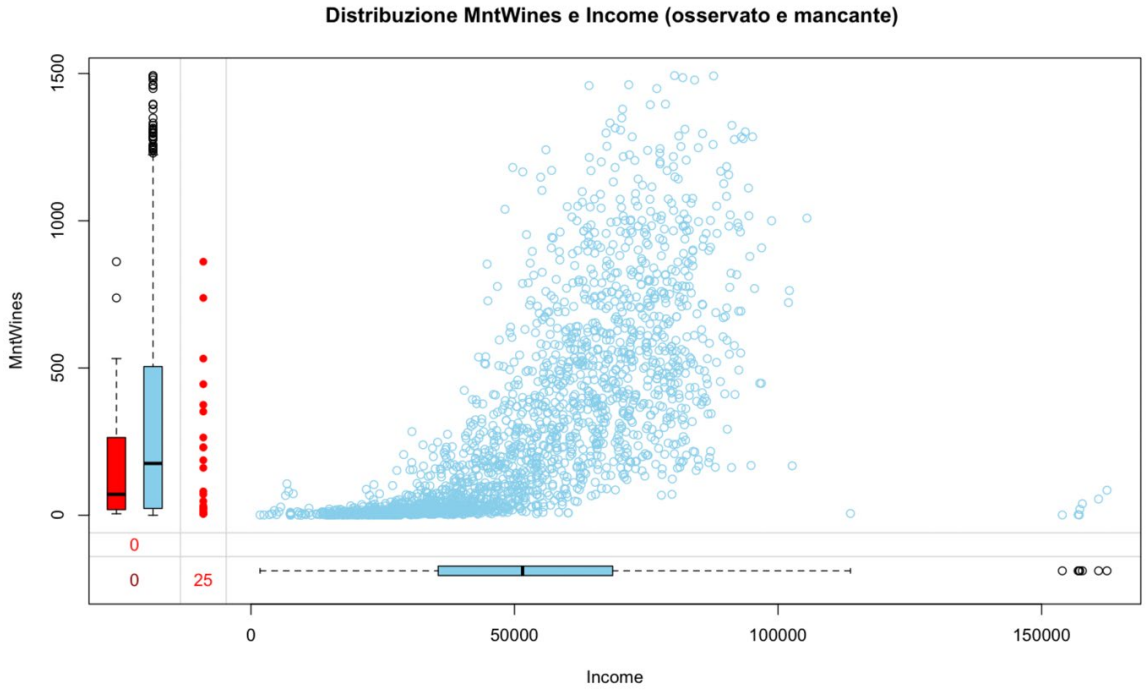

Figura 2

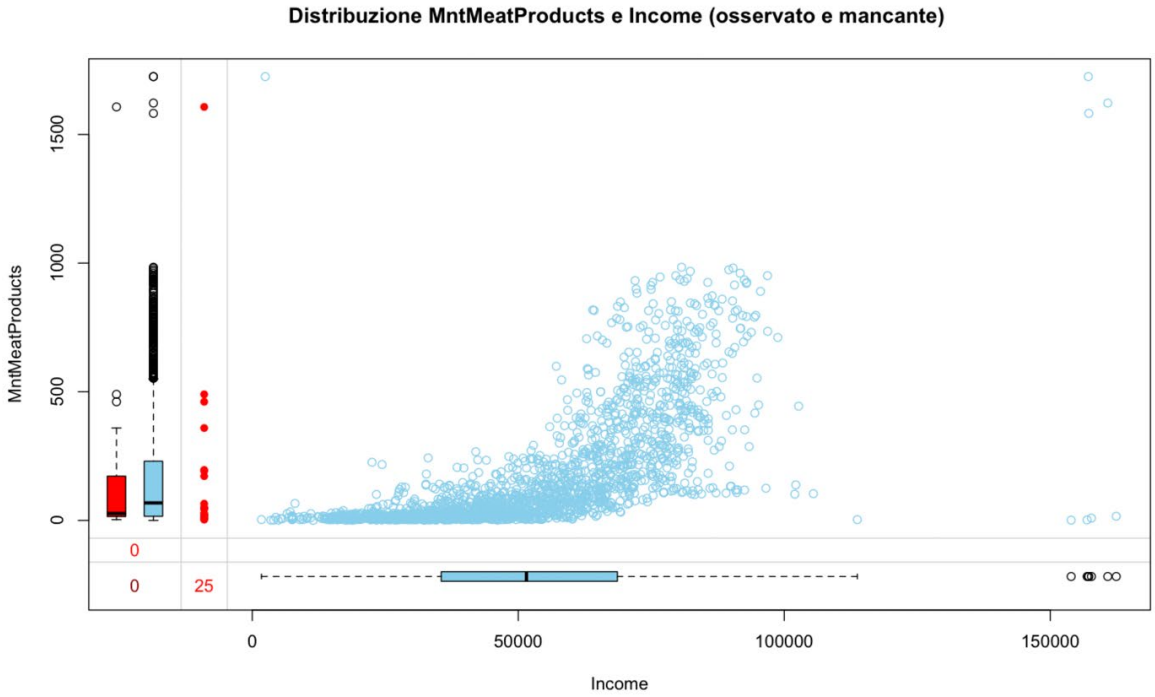

Figura 3

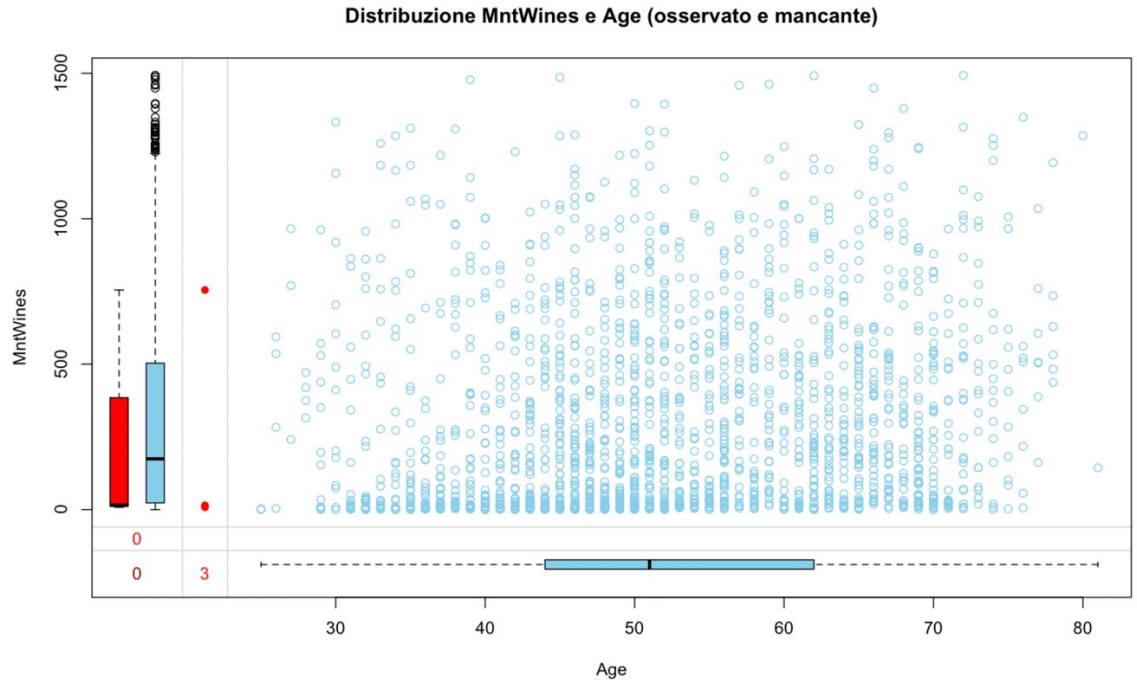

Figura 4

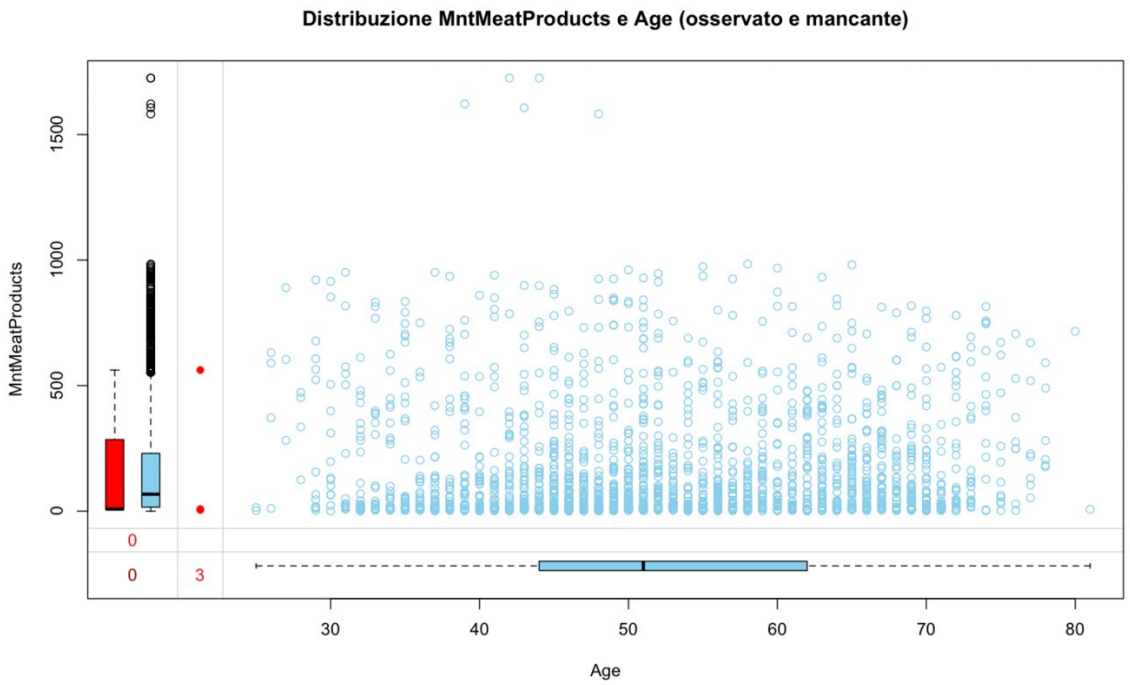

Figura 5

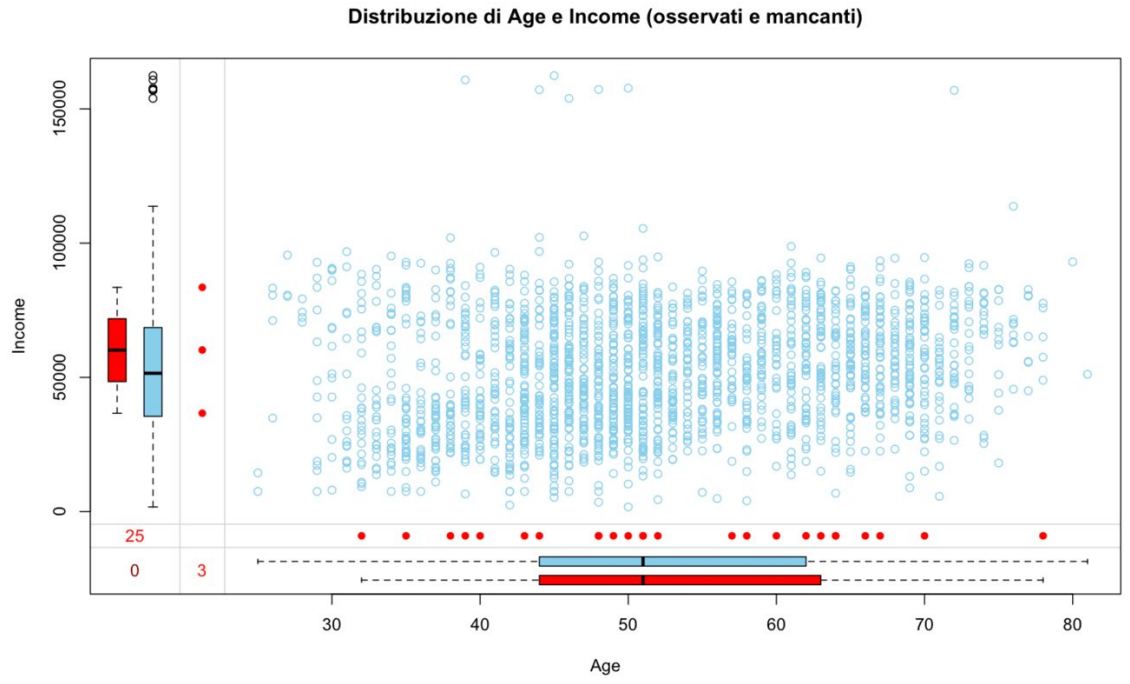

Figura 6

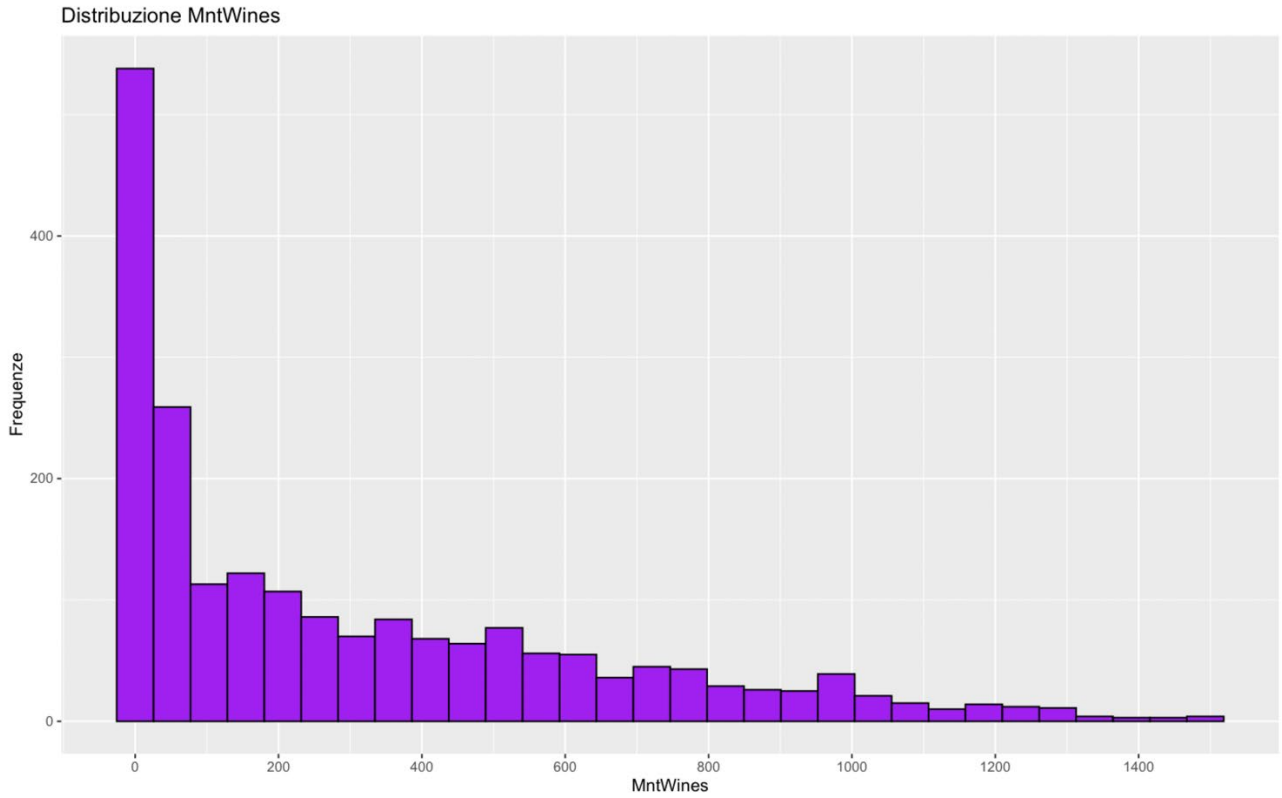

Figura 7

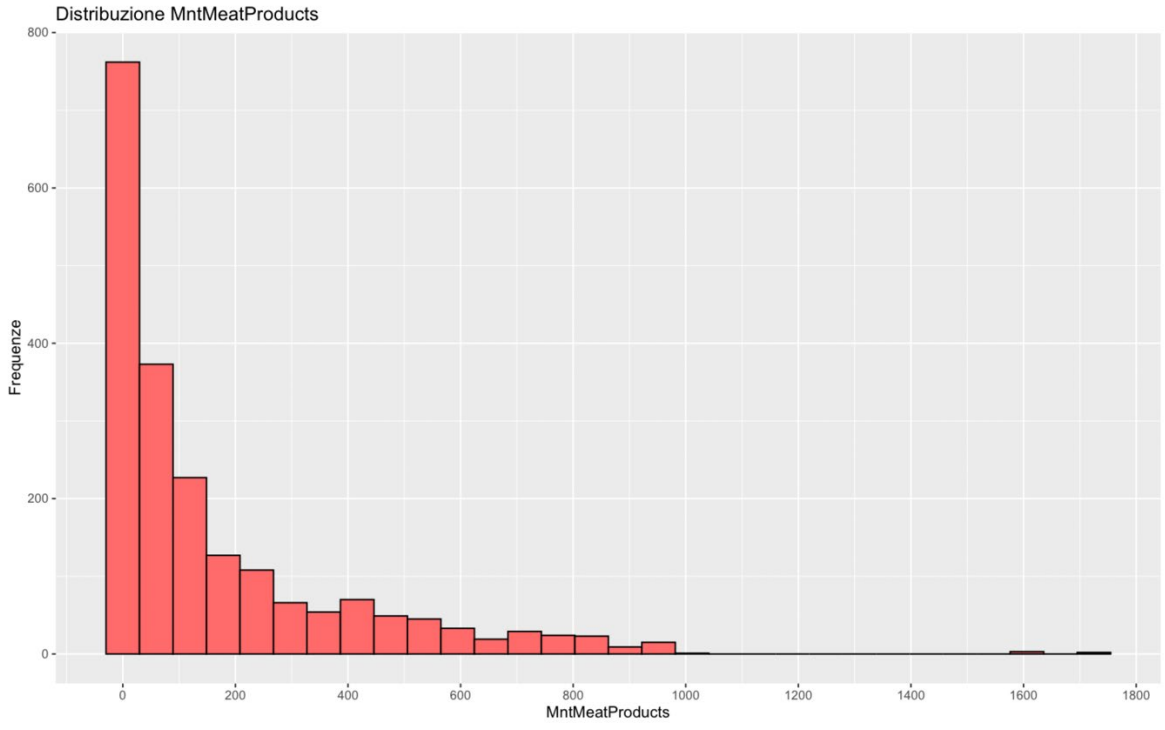

Figura 8

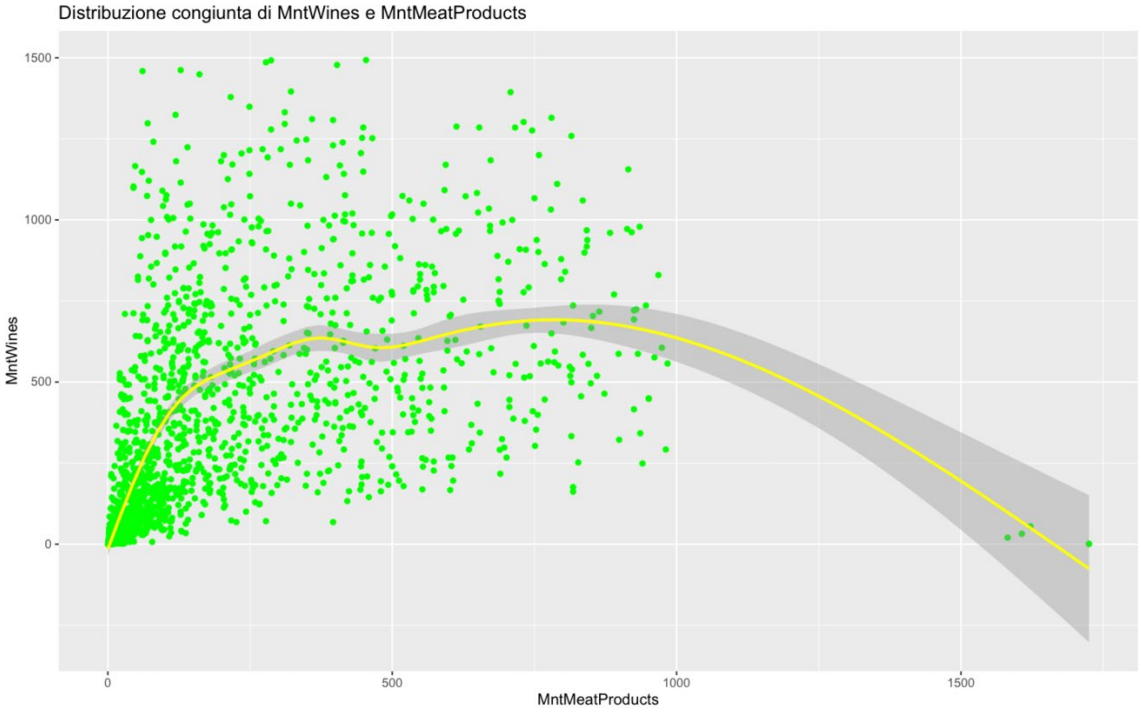
Distribuzione congiunta di MntWines e MntMeatProducts
MntWines
1500
1000
500
0
0
500
1000
1500
MntMeatProducts

Figura 9

Figura 10

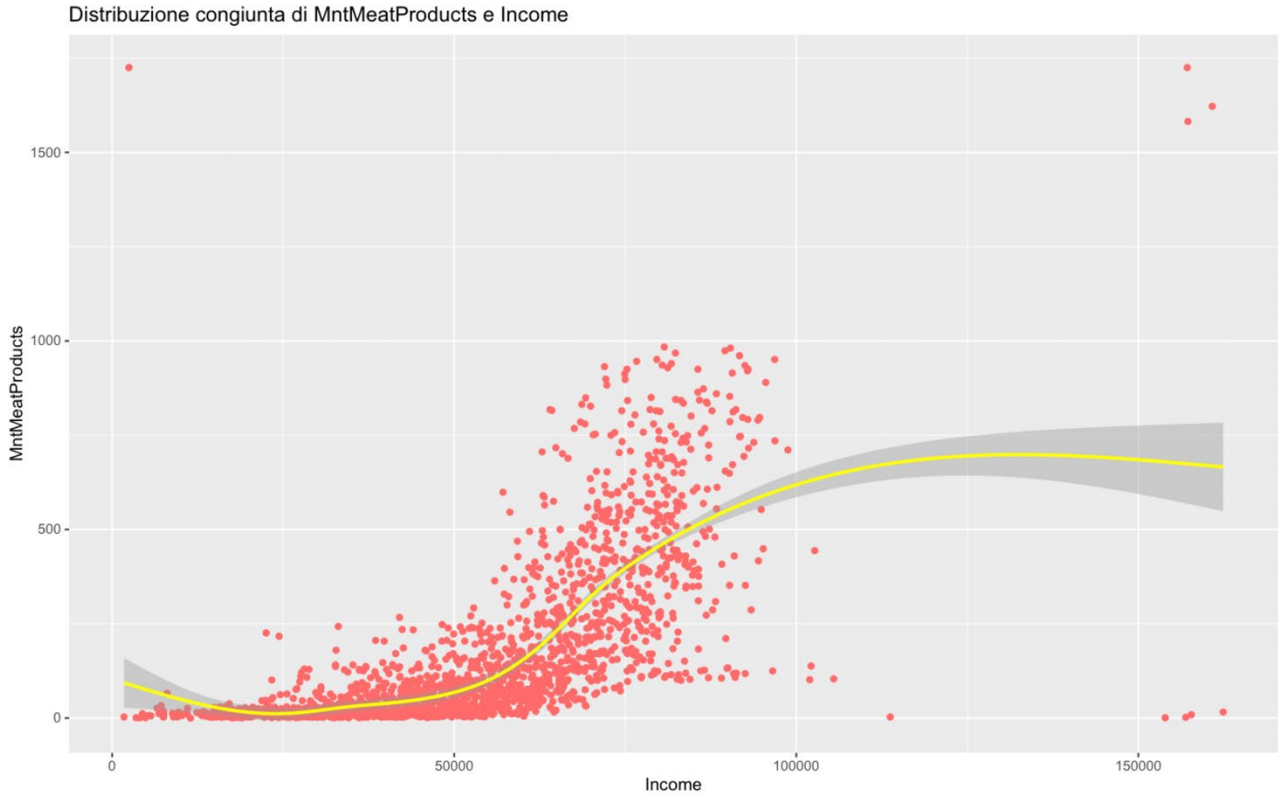

Figura 11

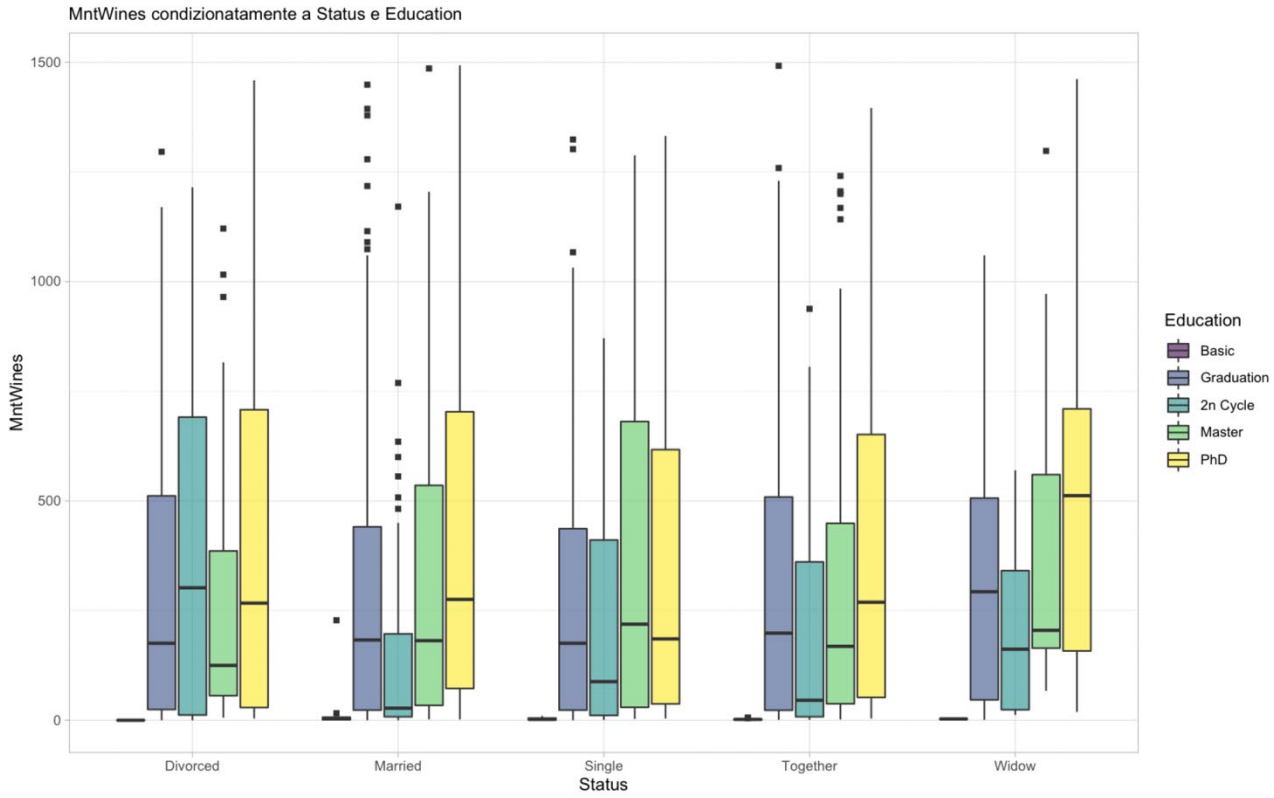

Figura 12

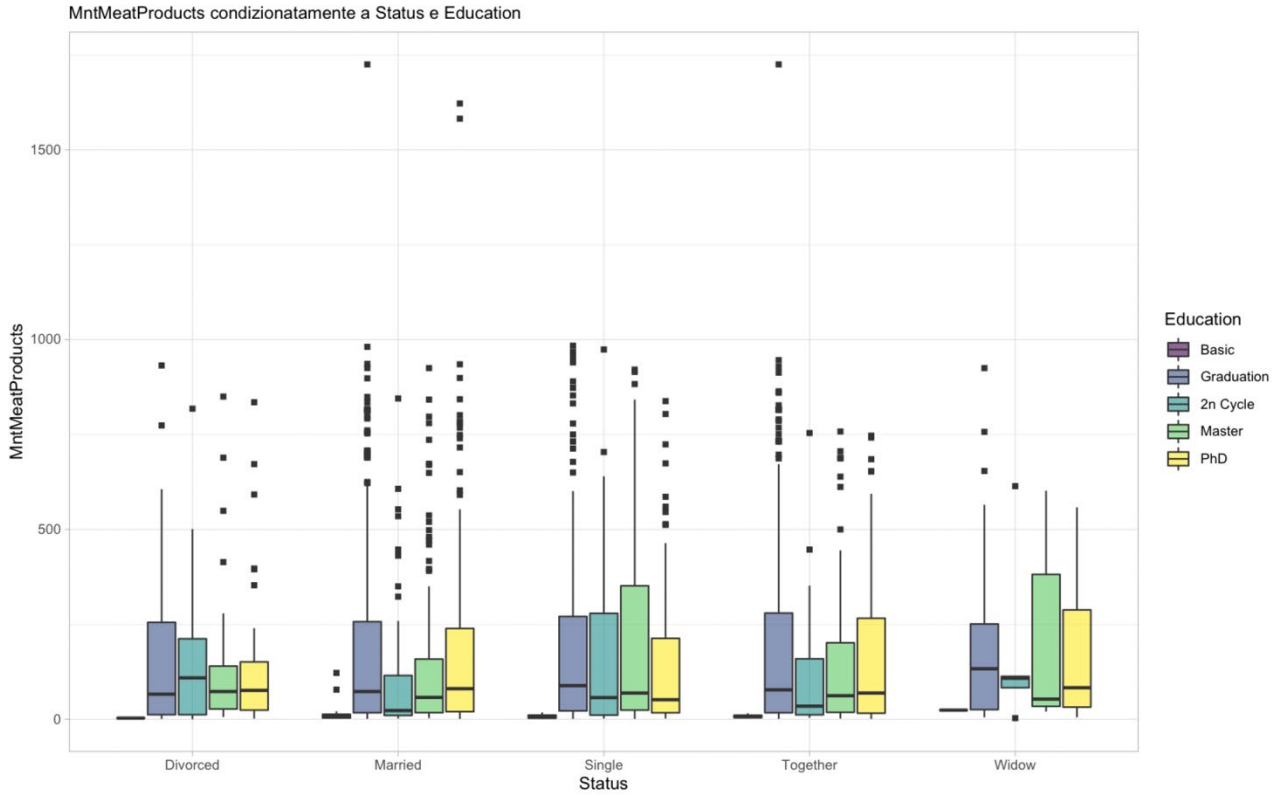

Figura 13

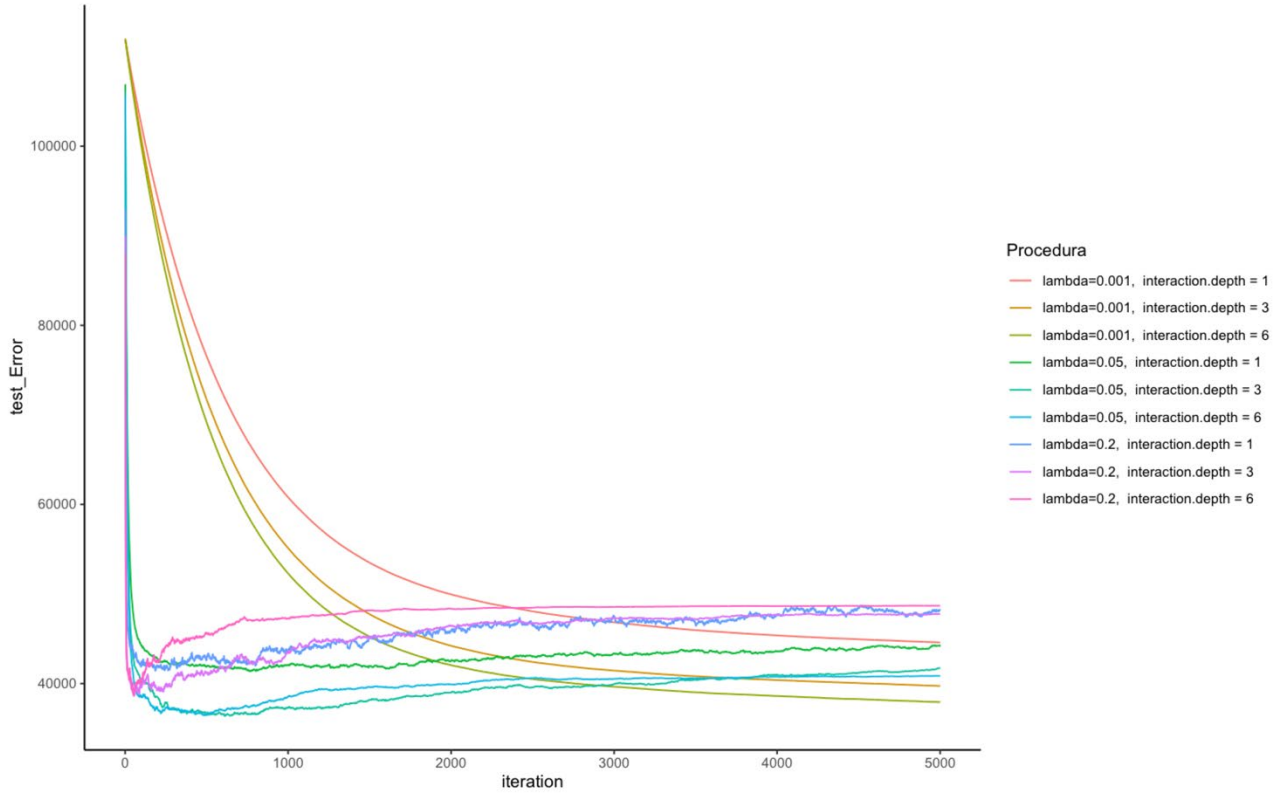

Figura 14 Test error per MntWines

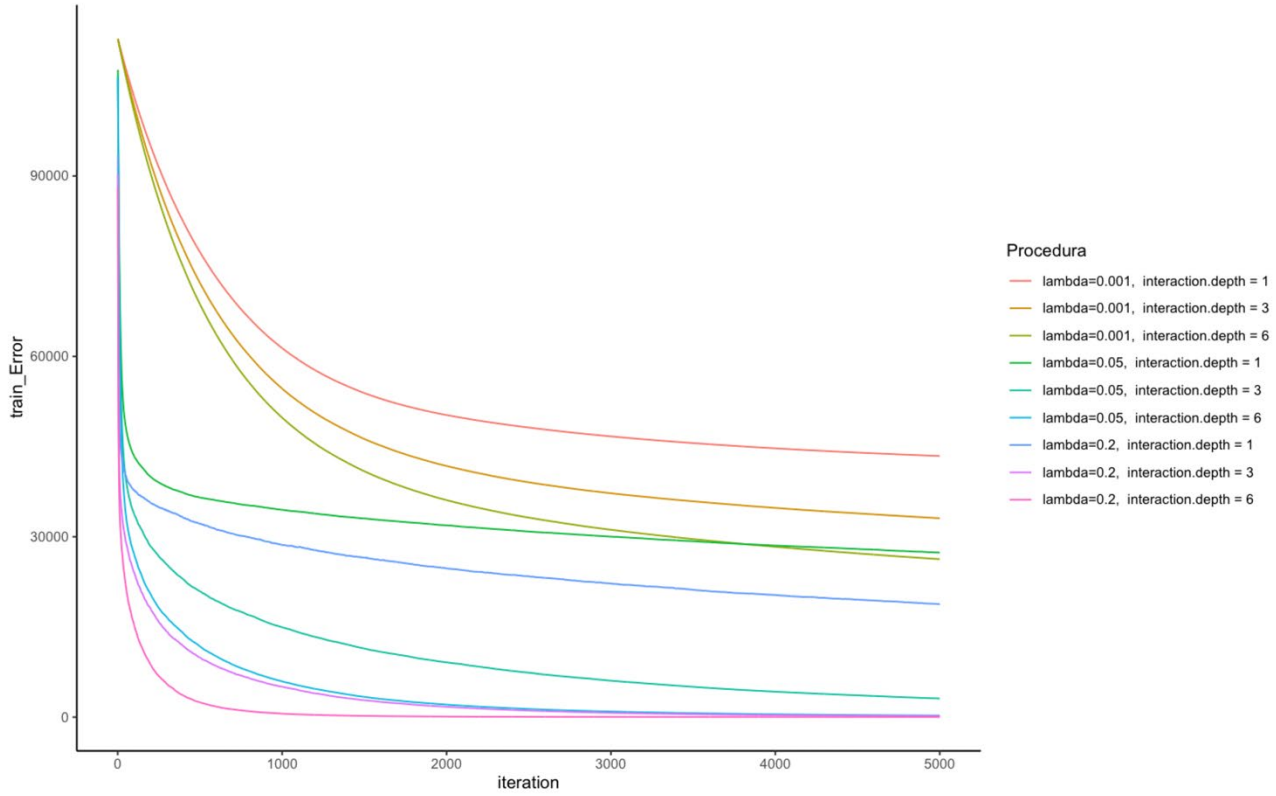

Figura 15 Train error per MntWines

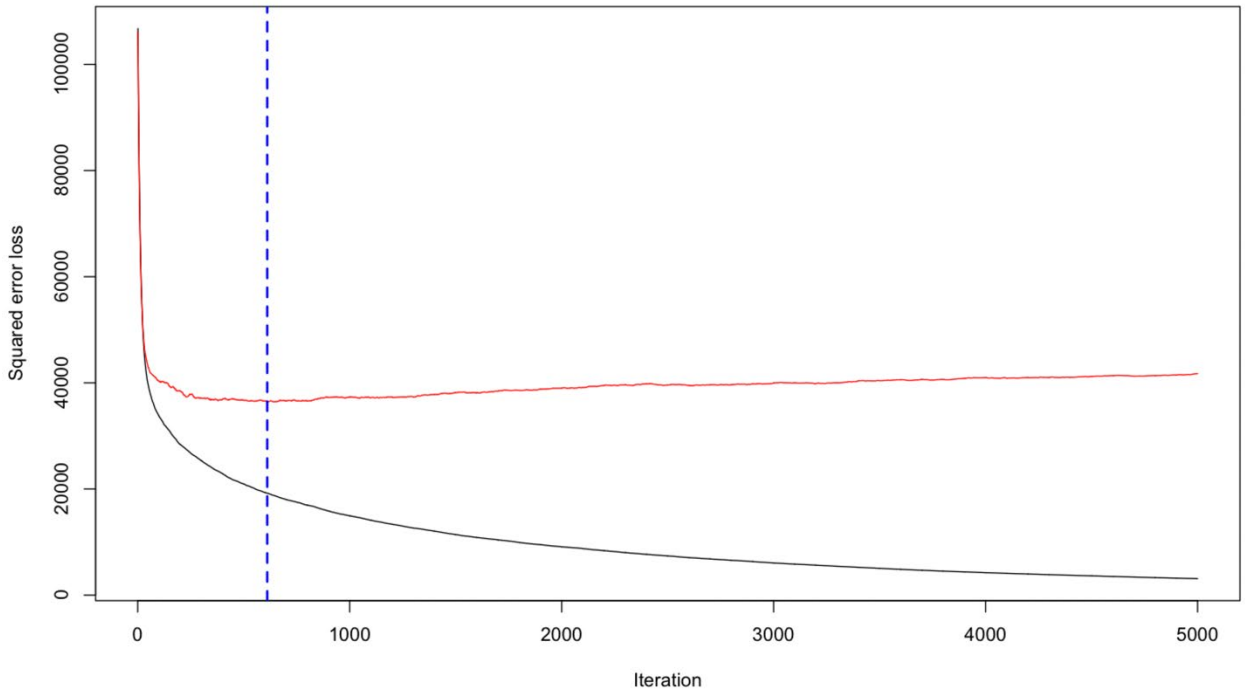

Figura 16 Numero ottimo di alberi per la procedura pari a 611

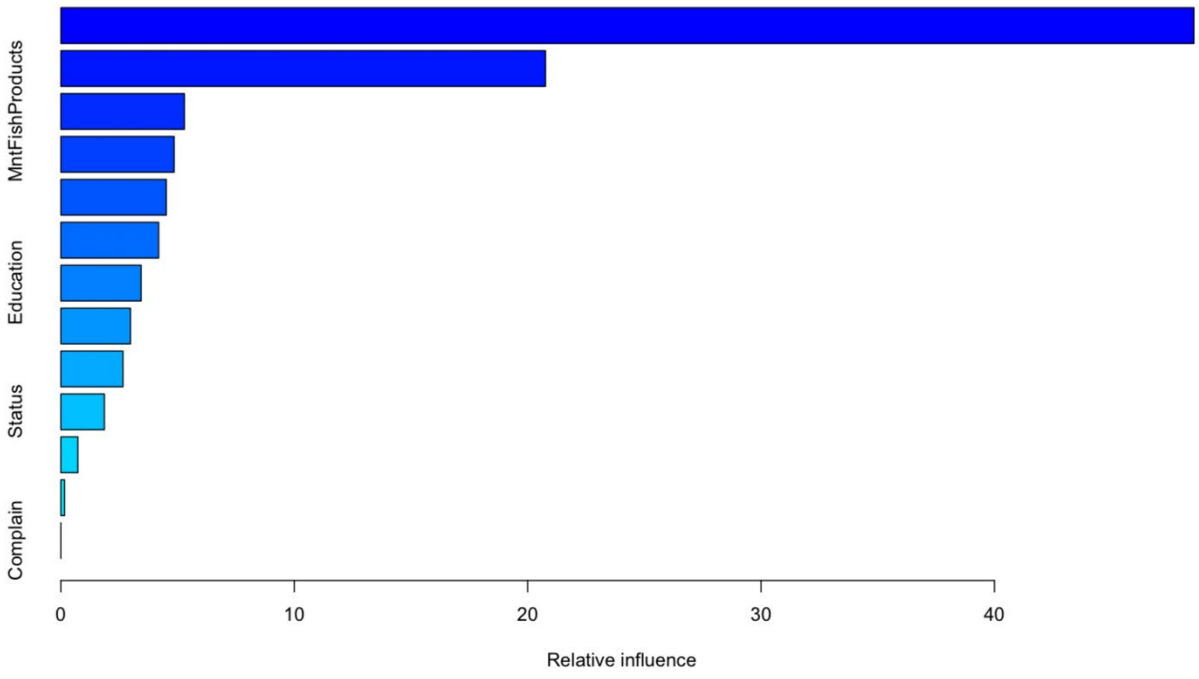

Figura 17 Importanza delle variabili del boosting per MntWines

	var	rel.inf
Income	Income	48.5498033
MntMeatProducts	MntMeatProducts	20.7642699
MntFishProducts	MntFishProducts	5.2940003
MntSweetProducts	MntSweetProducts	4.8528808
MntGoldProds	MntGoldProds	4.5149537
MntFruits	MntFruits	4.1897698
Education	Education	3.4379515
Enrollment	Enrollment	2.9816470
Age	Age	2.6615417
Status	Status	1.8639578
Kidhome	Kidhome	0.7295413
Teenhome	Teenhome	0.1596829
Complain	Complain	0.0000000

Tabella 1 Importanza variabili boosting per MntWines

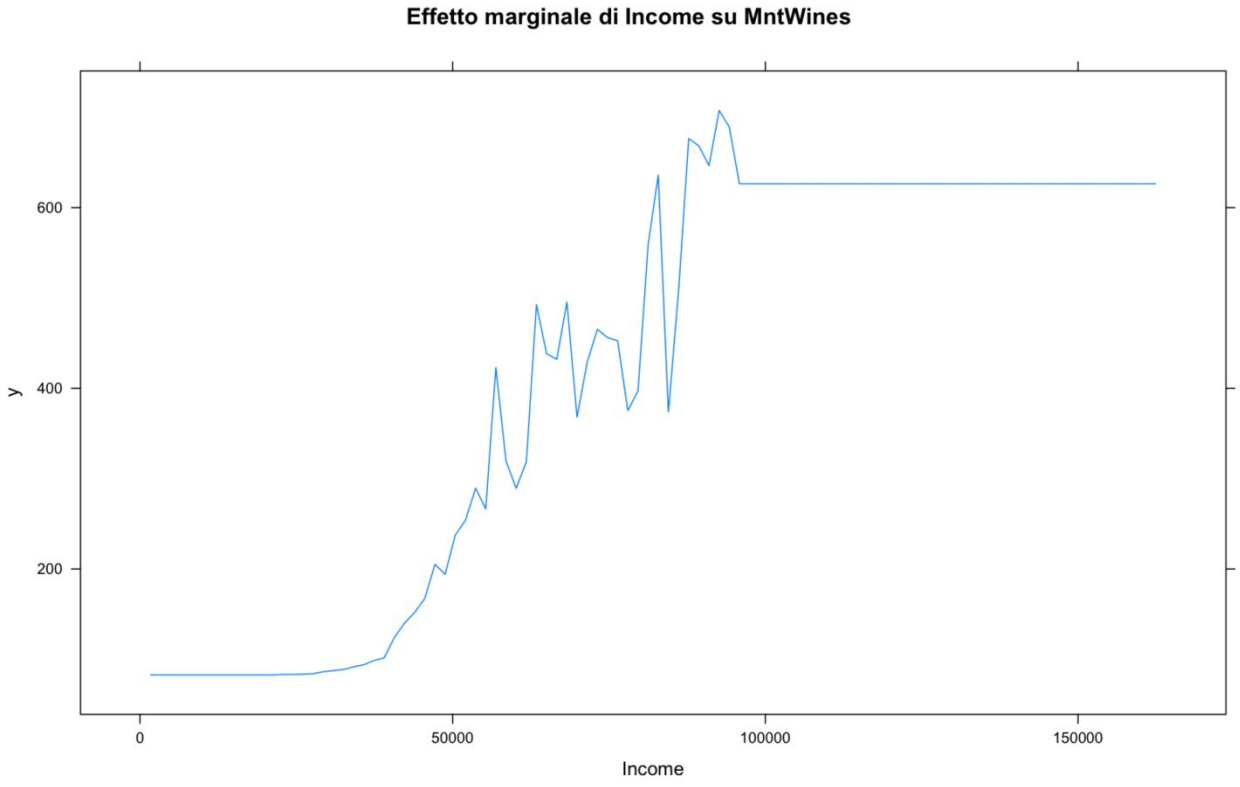

Figura 18

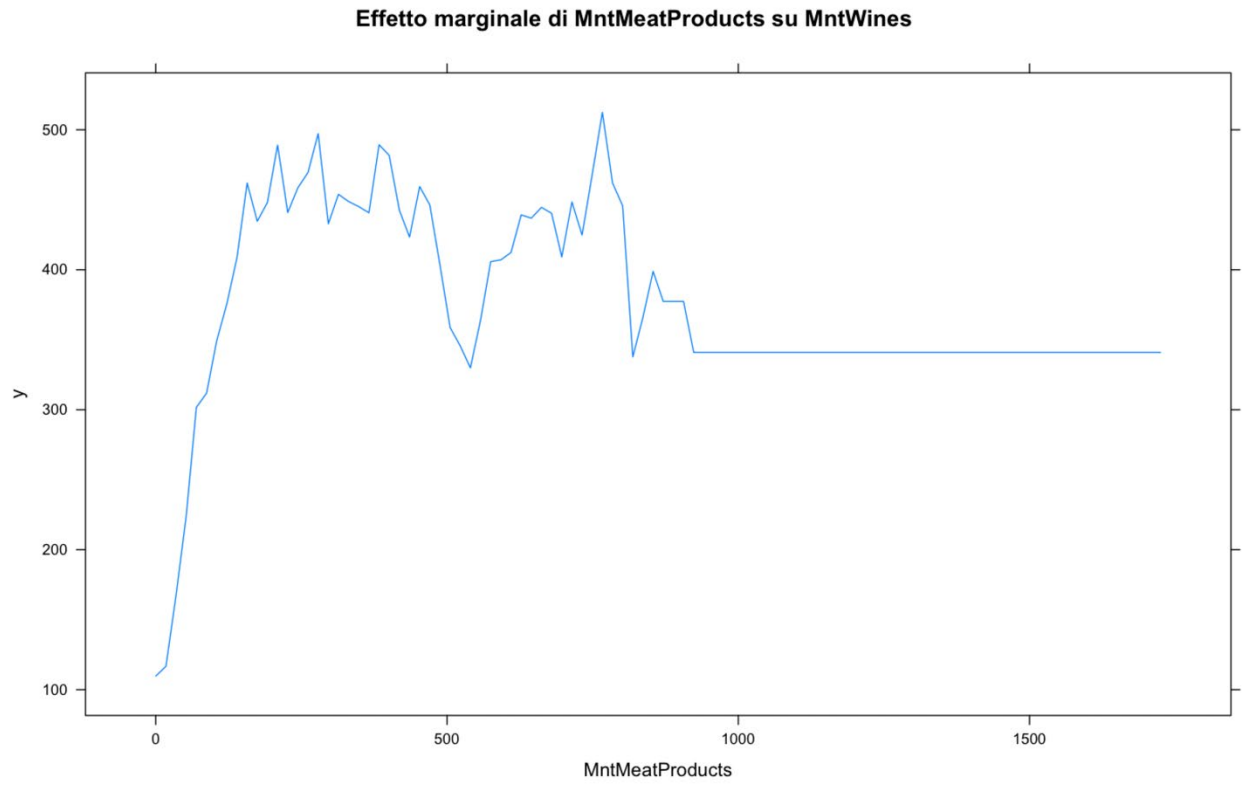

Figura 19

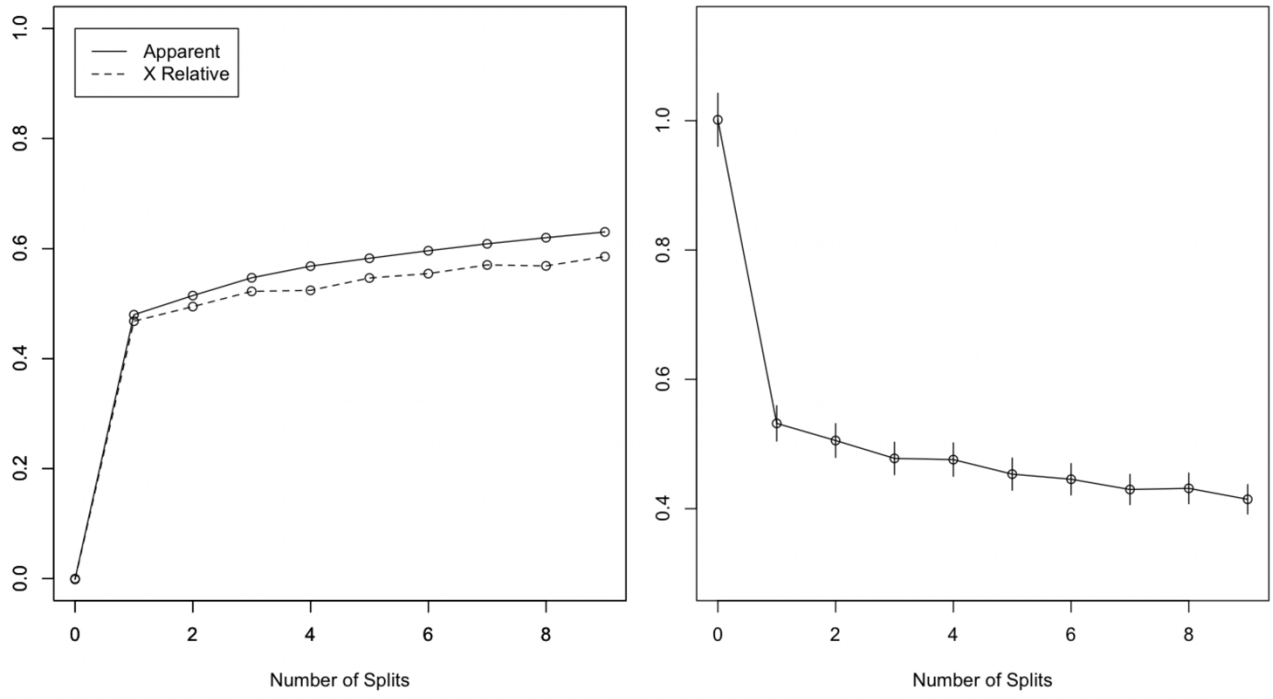

Figura 20 R^2 e cp dell'albero di regressione per MntWines

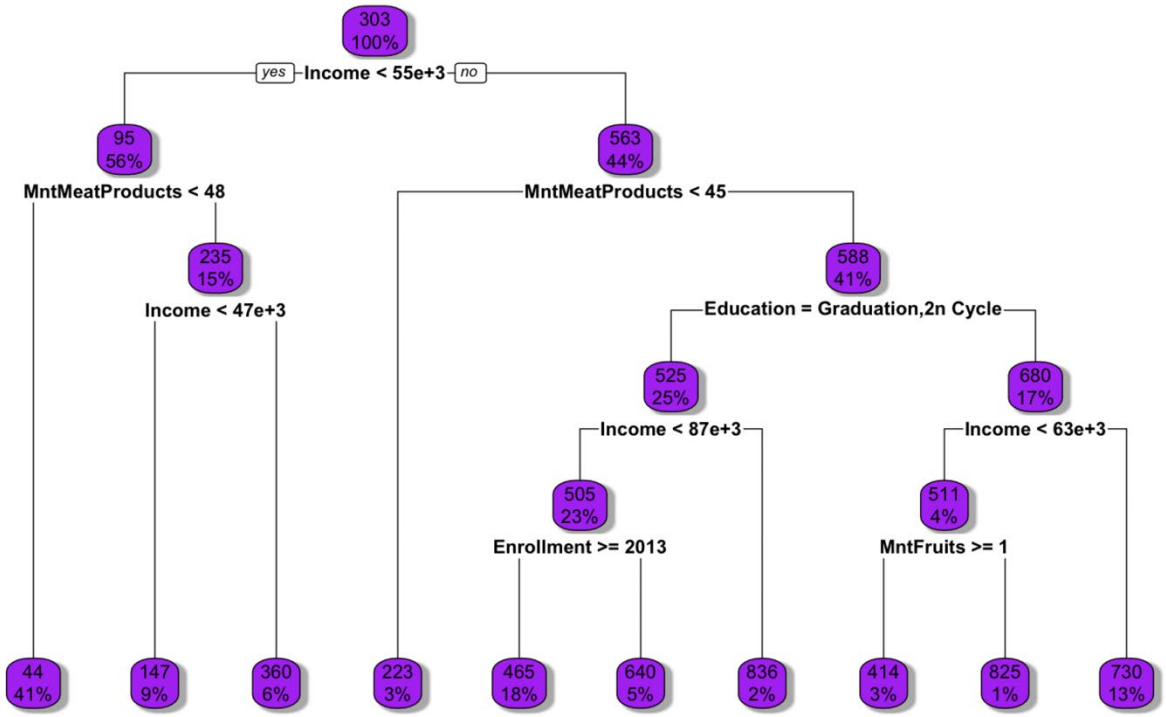

Figura 21 Albero di regressione per MntWines

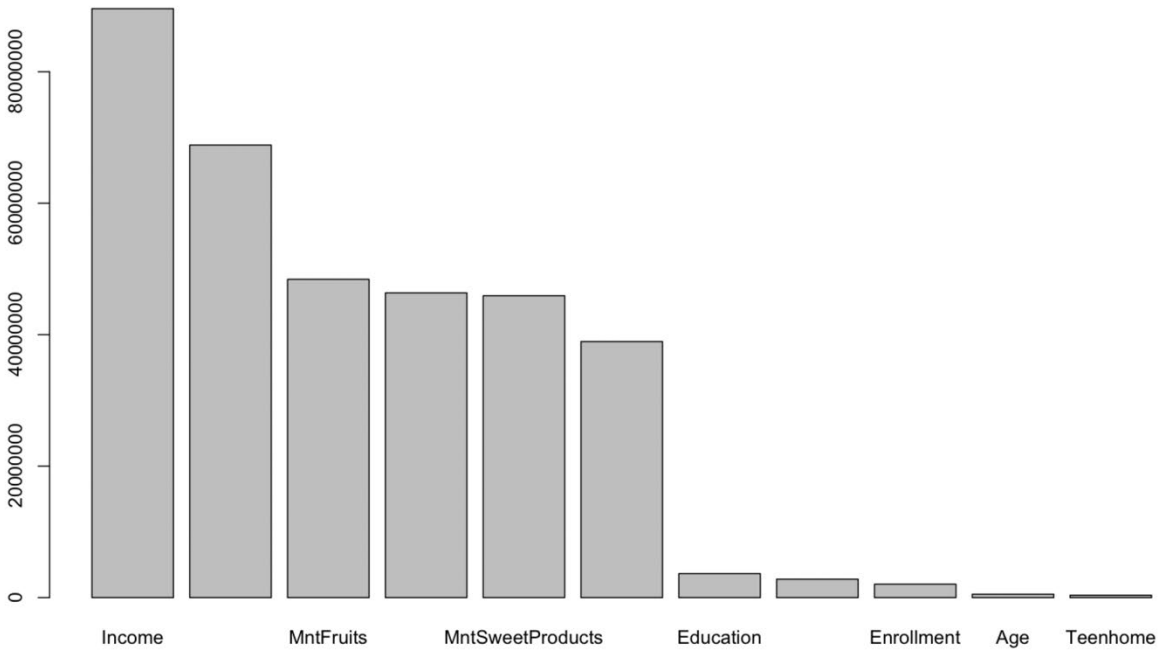

Figura 22 Importanza delle variabili dell'albero di regressione per MntWines

Income	MntMeatProducts	MntFruits	MntFishProducts
89598097.9	68849199.5	48425429.3	46362751.5
MntSweetProducts	Kidhome	Education	MntGoldProds
45931948.3	38959208.0	3653270.9	2808564.5
Enrollment	Age	Teenhome	
2044937.9	520013.9	364936.7	

Tabella 2

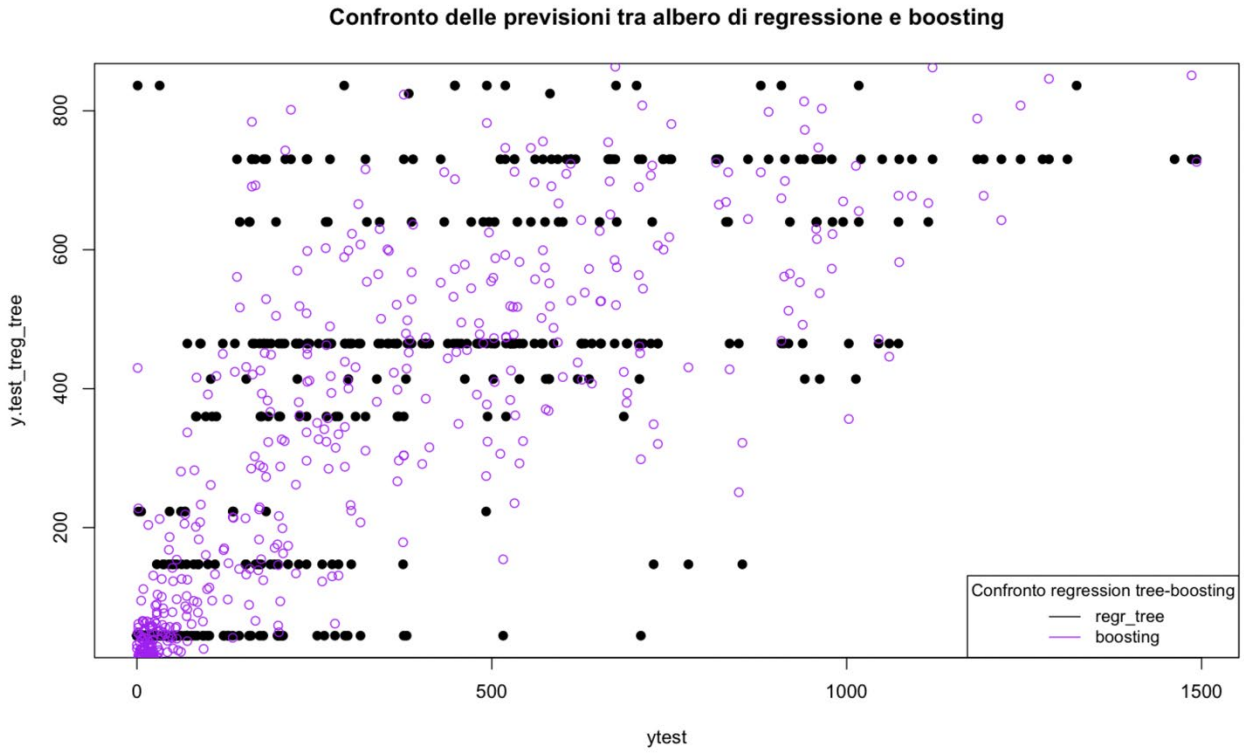

Figura 23

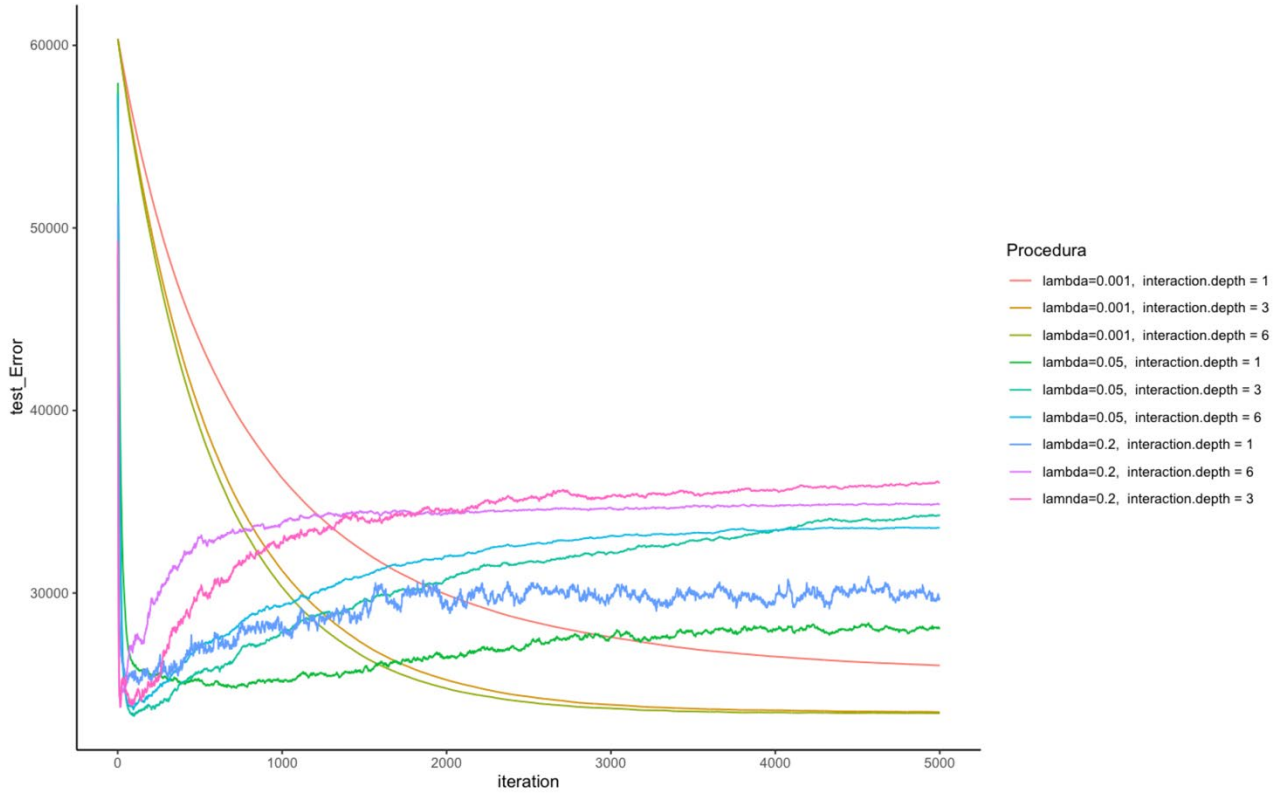

Figura 24 Test error per MntMeatProducts

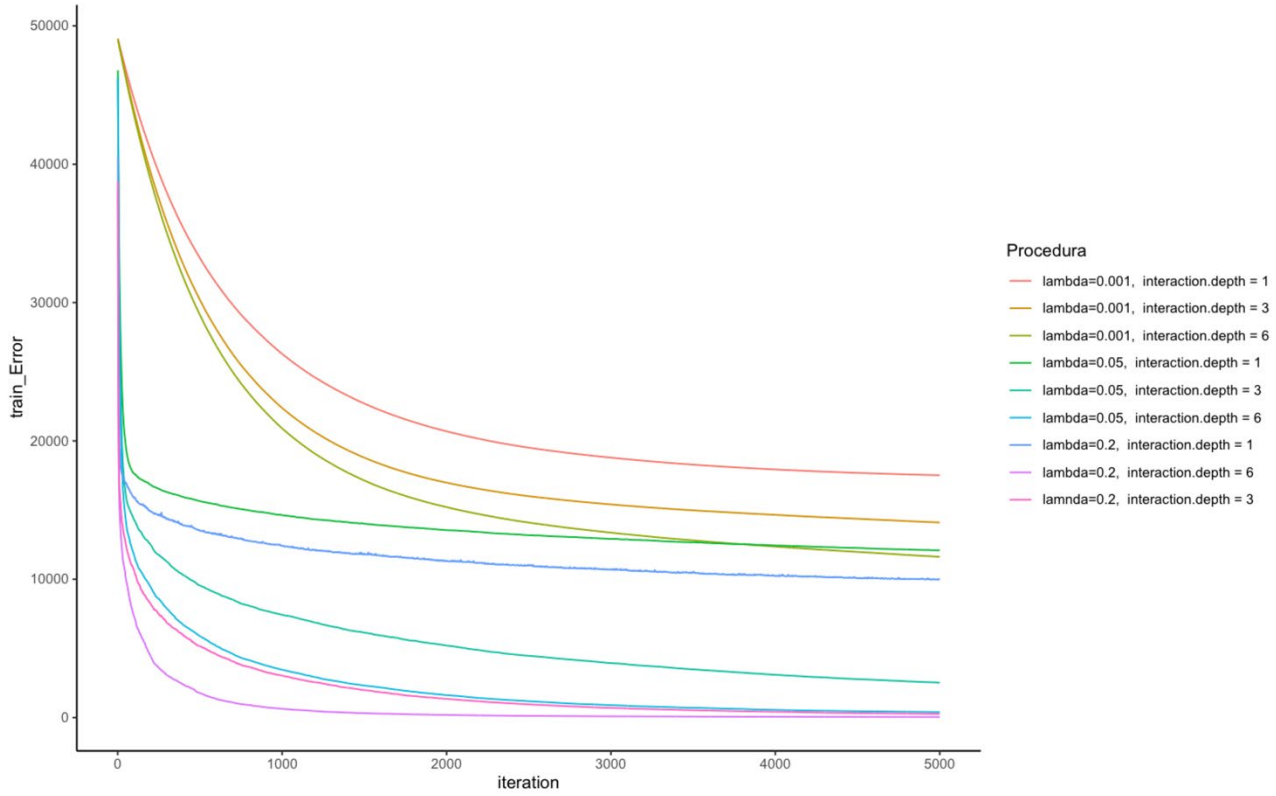

Figura 25 Train error per MntMeatProducts

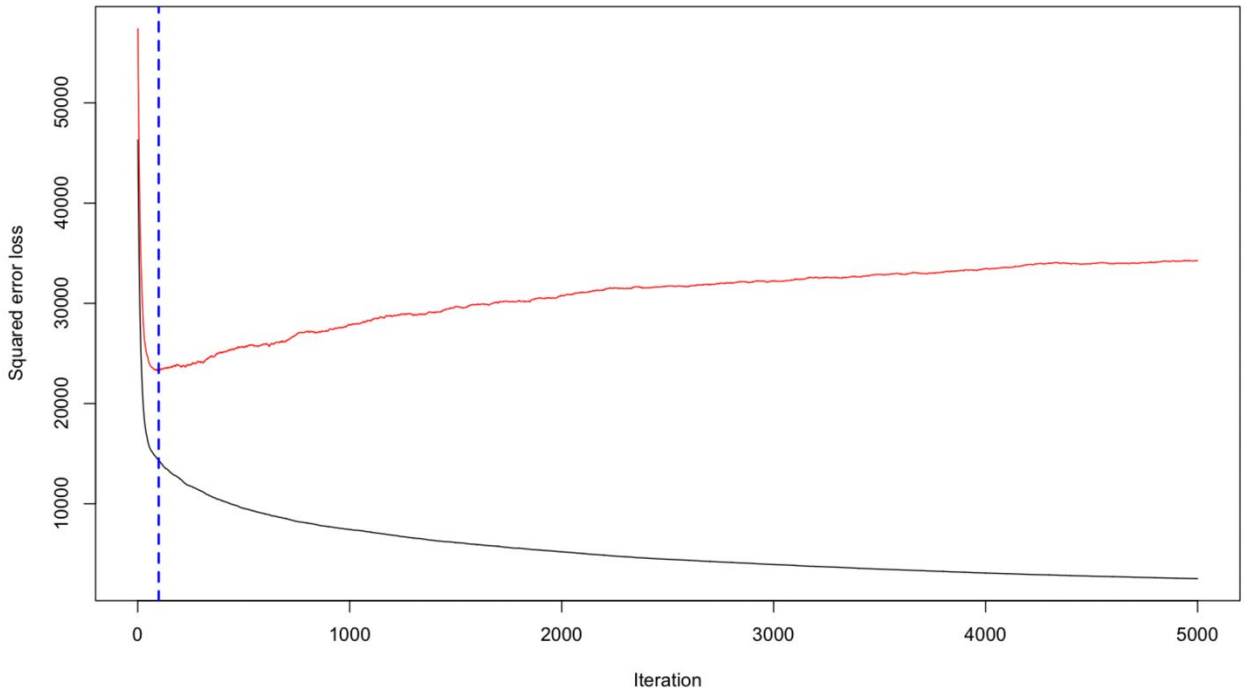

Figura 26 Numero ottimo di alberi la procedura pori a 99

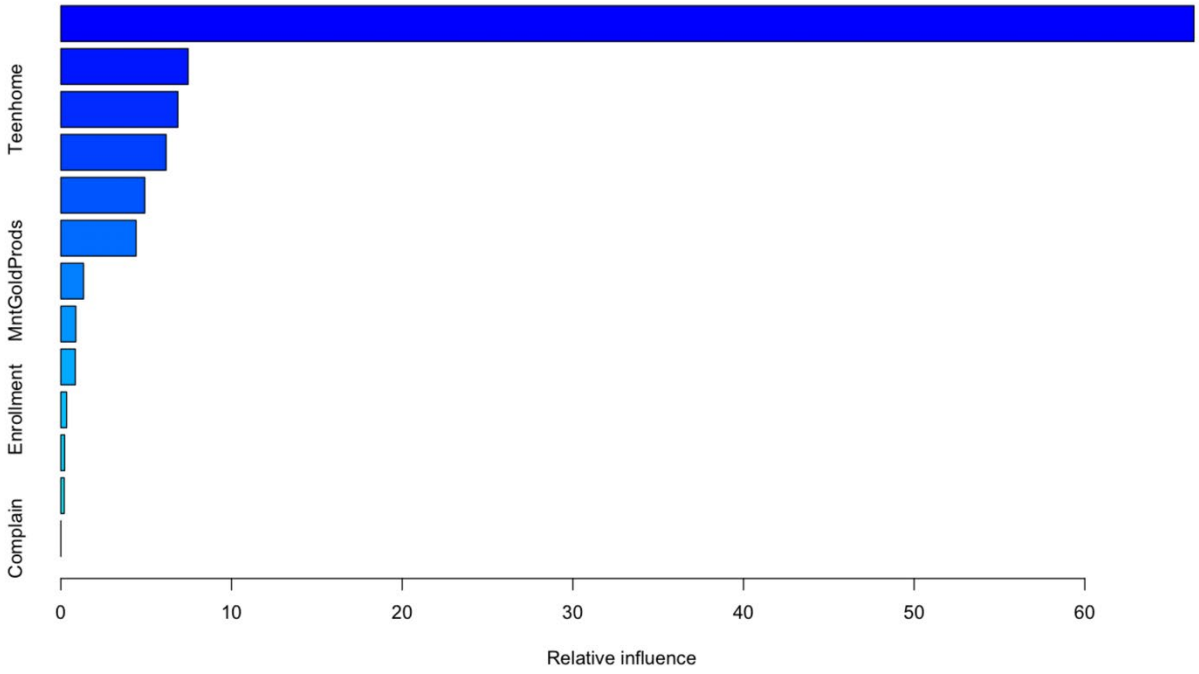

Figura 27 Importanza delle variabili del boosting per MntMeatProducts

	var	rel.inf
Income	Income	66.3968861
MntWines	MntWines	7.4568671
Teenhome	Teenhome	6.8571622
MntFishProducts	MntFishProducts	6.1656968
MntFruits	MntFruits	4.9155984
MntSweetProducts	MntSweetProducts	4.4085106
MntGoldProds	MntGoldProds	1.3290932
Status	Status	0.8731817
Age	Age	0.8474193
Enrollment	Enrollment	0.3353409
Education	Education	0.2189069
Kidhome	Kidhome	0.1953368
Complain	Complain	0.0000000

Tabella 3

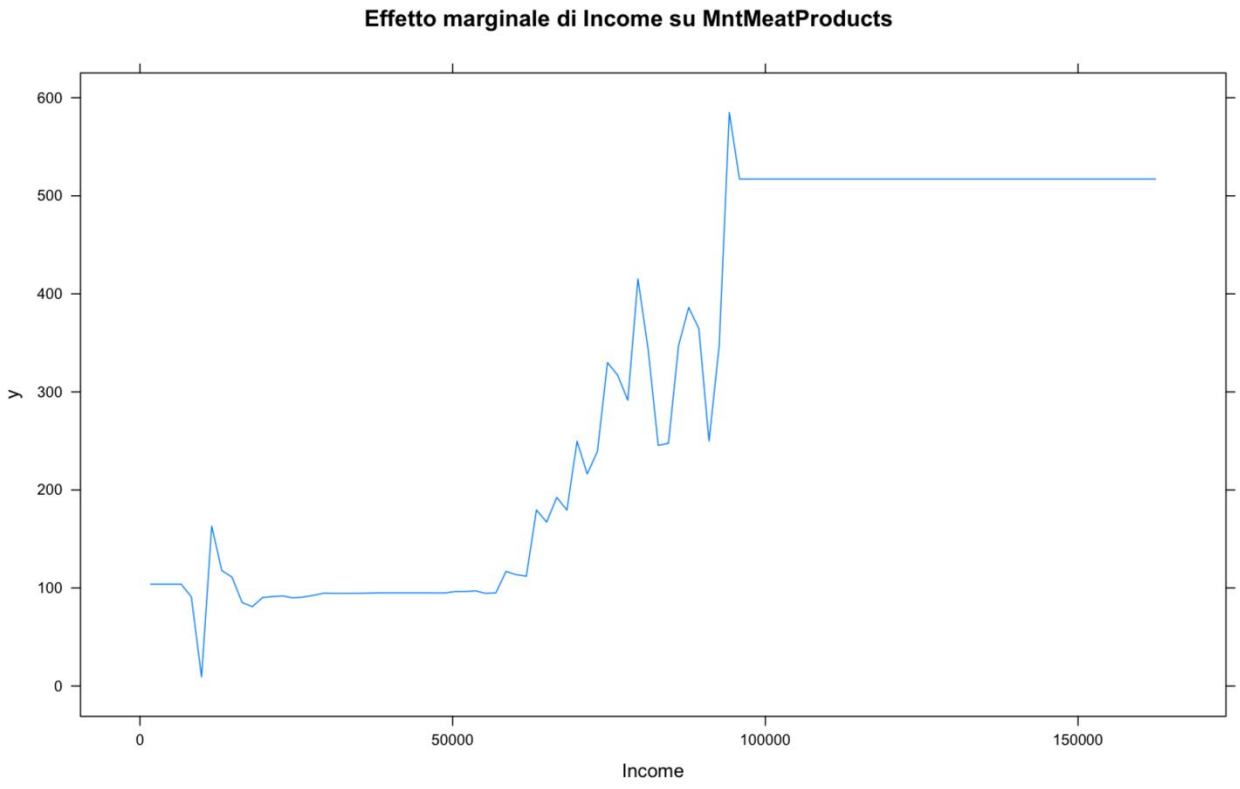

Figura 28

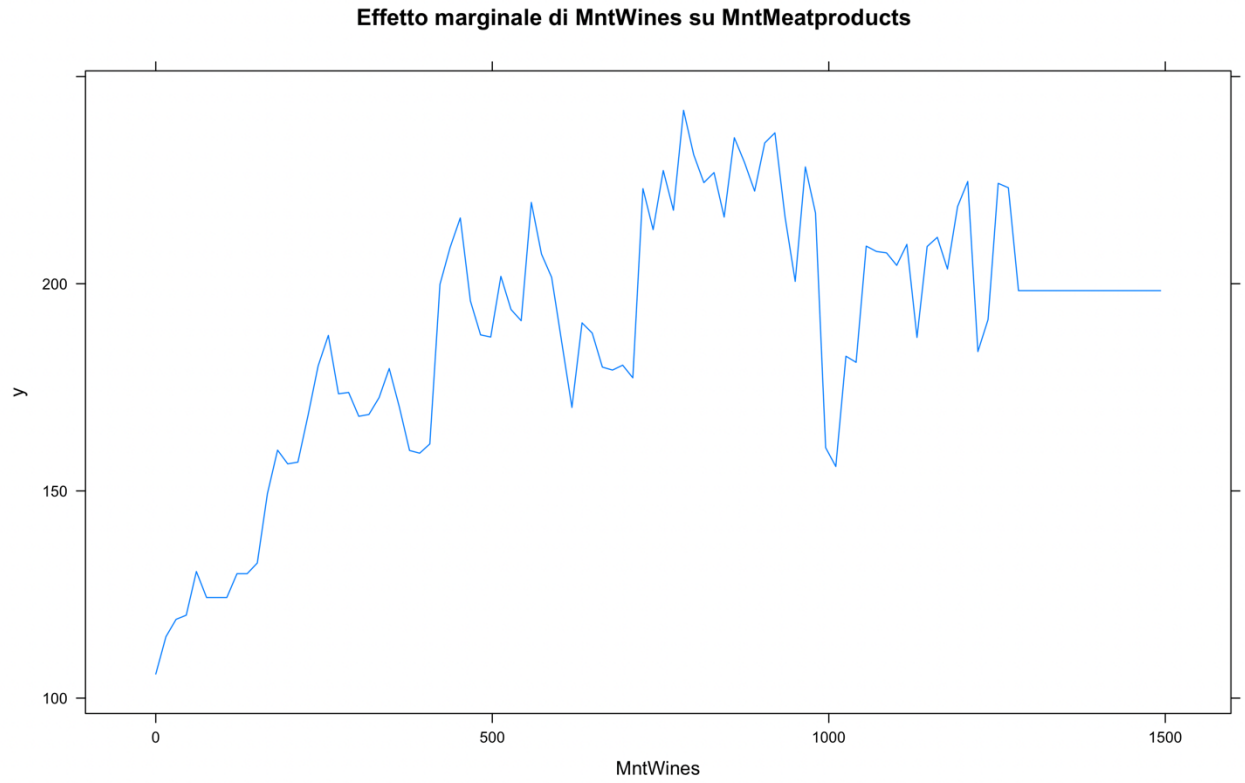

Figura 29 Effetto marginale di MntWines su MntMeatProducts

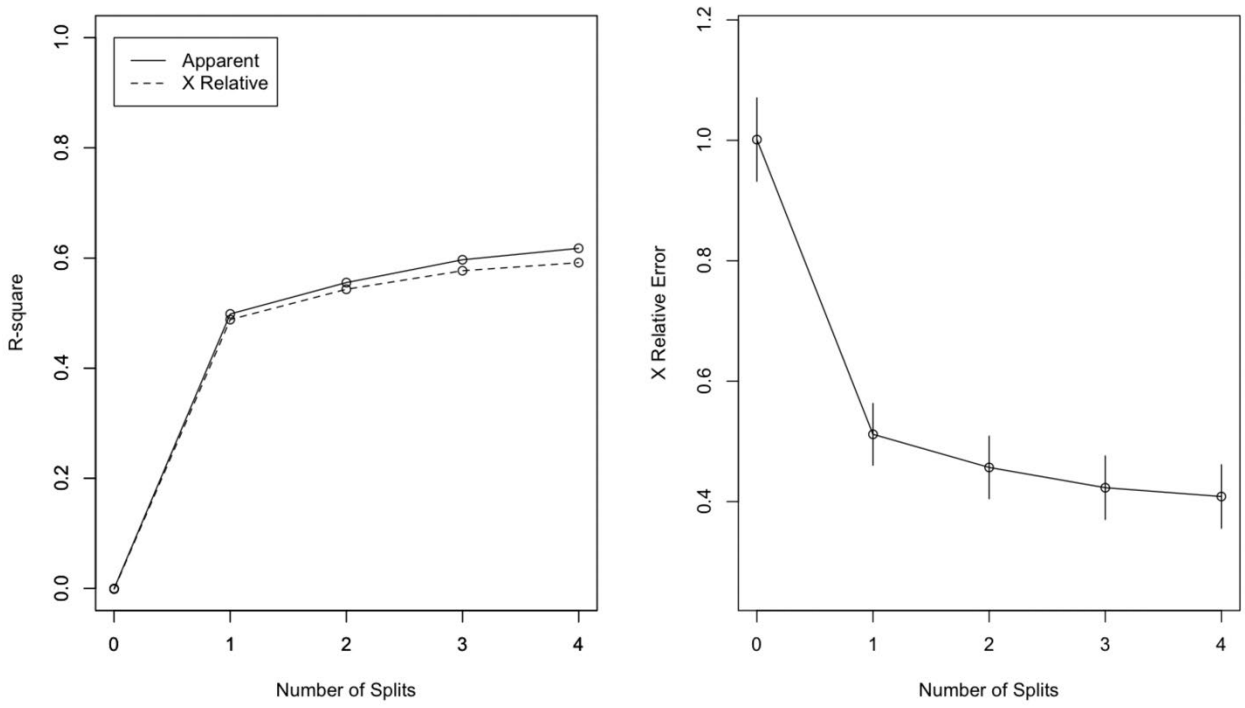

Figura 30 R^2 e cp dell'albero di regressione per MntMeatProducts

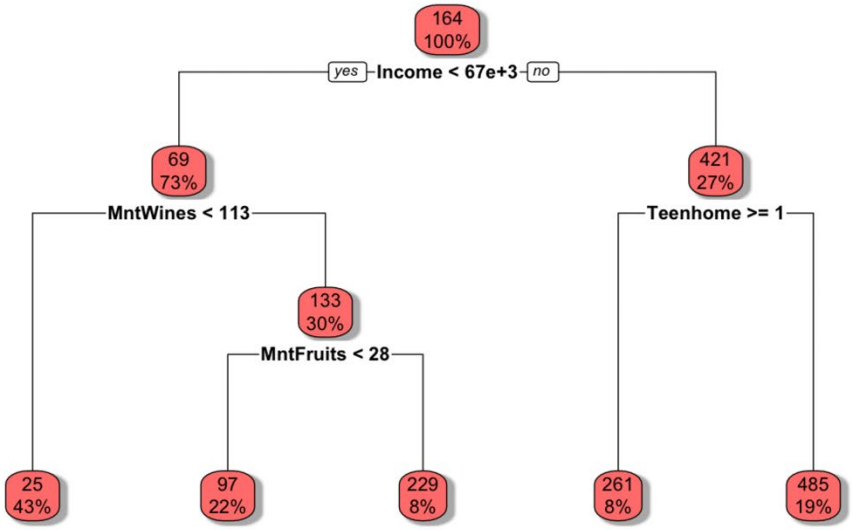

Figura 31 Albero di regressione per MntMeatProducts

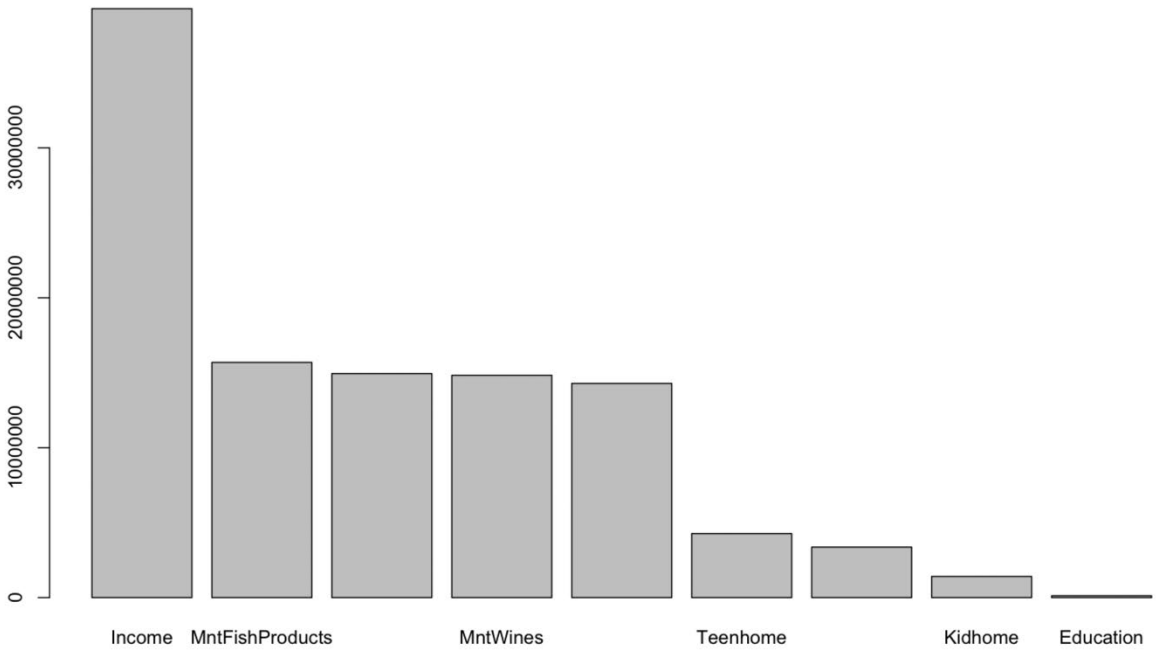

Figura 32 Importanza delle variabili dell'albero di regressione per MntMeatProducts

Income	MntFishProducts	MntFruits	MntWines
39285133.3	15689875.6	14946399.7	14831345.2
MntSweetProducts	Teenhome	MntGoldProds	Kidhome
14290285.8	4275109.0	3370895.8	1416605.0
Education			
127882.9			

Tabella 4

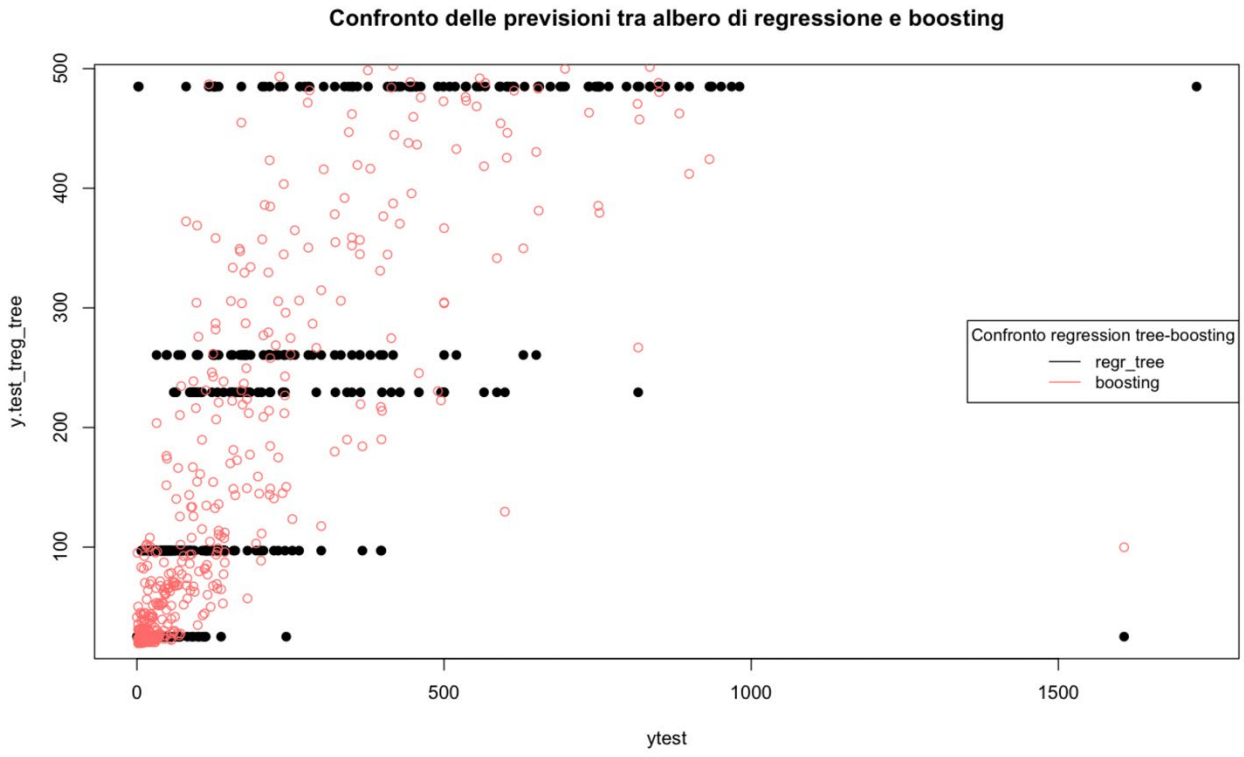

Figura 33