ANALYSIS OF THE BEHAVIOUR AND NEEDS OF A COMPANY'S CUSTOMERS, WITH THE AIM OF PREDICTING THE AMOUNT SPENT ON TWO PRODUCT CATEGORIES:

WINE AND MEAT

Marco Speciale

Index

1. Introduction

Data provided to this study includes:

- general customer information provided by customers when registering with the company

- total expenditure for each product category in the last two years

The final goal of the analysis is to predict the amount spent on two different product categories, wine and meat, using the available customers data. Particular attention will be paid to the elements that most influence the purchase choices of the above-mentioned product categories, such as the Income of each customer, the choice of other product categories, Marital Status and Education levels.

2. Data cleaning

The analysis' elements were received in csv format.
The programming software 'R' was used to analyse the data.
The starting dataset consisted of 2240 observations and 29 variables, but only 15 variables are included in this analysis.

After displaying the type of variables present, the observations were immediately checked for repetition.
In order to proceed with the removal of repetitions, the variable ID was first excluded because did not allow the display of repeated observations as there were identical observations with different IDs.
201 repetitions were found against 2240 observations, which were subsequently eliminated using the **unique** function.
The variable Enrollment was created to replace the variable Dt_Customer, in this new variable only the year of registration of the customers is obtained, taking into account neither the month or the day, then the variable Age was created in which the age of the customers was reported, obtained from the year reference to this analysis (2021) minus the variable Year_Birth; further the modalities of the variable Marital_Status were appropriately aggregated in a new variable named Status.

So, Marital_Status was replaced by Status and some of the modes such as: "YOLO" and "Absurd" were not considered valid; consequently, the Education levels were ordered according to the hierarchical order of study courses.
In addition, values that could not be considered valid for the analysis were found for the variable Age, where three clients were over 100 years old, and for the variable Income, where a value indicated was "6666666".
Therefore, in order to verify that the dataset had valid values for the analysis, rules were applied,with particular reference to the variables Age and Income in such a way as to locate these errors and replace them with NA. Apart from the above-mentioned errors, no other errors were found. The final dataset supplied to this analysis is composed of 2039 observations and 14 variables (only variable ID was removed).

3. Exploratory analysis

In order to verify the completeness of the observations in the dataset, it was checked whether there were any missing data. Thanks to the **rsem.pattern** function, the output of which is presented in the form of a missing datamatrix, it was possible to identify all missing data patterns.
Pattern 1 consists of 2008 rows out of 2039, characterised by the fact that all 14 variables were observed, then patterns 2 and 4 in which three data are missing each for the variables Age and Status and finallyin pattern 3, 25 data are missing for the variable Income.
Using the **aggr** function, the percentage of marginally missing data for each of the variables was identified (left-hand side of the graph), while the right-hand side shows the patterns, not to scale.

The output returned the following information:

- The percentage of observed data is 98.48%.
- The percentage of missing data for the variable Income is 1.22%.
- The percentage of missing data for the variables Age and Status is 0.14% respectively. (Figure 1 in the appendix)

From the analysis of the missing patterns, it can be seen that these are only present in the explanatory variables and not on the response variables, therefore, it was not considered necessary to perform an imputation for the missing data. Using graphical representations, the distributions of the MntWines and MntMeatProducts responses were compared when the Income explanatory variable was observed ormissing. As far as MntWines is concerned, it was possible to state that the distribution reduces considerablywhen Income is missing; in fact, in this circumstance the median value is lower and only two outliers are evident. (Figure 2 in the appendix)

The same applies to MntMeatProducts, the distribution is visibly more compact, when Income is missing, as this graph also shows that the median value is lower and the outliers of the distribution are three. (Figure 3 in the appendix)

No relationship was found between the response variables and the explanatory variable Age, while betweenAge and Income it was noted that the distribution of Income narrows considerably when Age data are missing. (Figures 4-5-6 in the appendix)

In order to verify which factors had the greatest impact on purchase choices in the wine and meat product categories, I proceeded to quantify the correlation between the response and explanatory variables. Using various interactive graphical representations it was deduced that the variables most correlated with the Mnt Wines and MntMeatProducts responses were:

- Income which has a correlation of 0.68 with MntWines and 0.69 with MntMeatProducts
- MntFruits,MntFishProducts,MntSweetProducts which have a correlation of about 0.40 with MntWines but prove to be more correlated with MntMeatProducts, each with different values around 0.55
- MntGoldProducts which has a correlation of 0.39 with MntWines and 0.34 with MntMeatProducts

It should be noted that there was a correlation of 0.56 between the two response variables MntWines and MntMeatProducts; Age, demonstrating the above, has a low correlation of 0.17 with both responses. (Table1 in the appendix)

After carrying out an initial exploratory analysis for missing data and assessing the correlations between the response and explanatory variables, exploratory investigations were carried out on the distributions of the responses.

In the case of MntWines, there is a positive asymmetry showing that there is a very high number of customers spending less than 100 on this product category. The distribution decreases as wine consumption increases until seven customers exceed the 1400 threshold. (Figure 7 in Appendix)

Similar considerations were made for the distribution of MntMeatProducts for both the positive asymmetry and the amount of customers who contained themselves in spending, the distribution is also decreasing, in the tail graph shows five customers who spent more than 1500. (Figure 8 in the appendix)

Viewing the joint distribution of MntWines and MntMeatProducts shows that a low consumption of wine corresponds to a low consumption of meat. As the consumption of these two product categories increases, in some cases a high consumption of wine does not correspond to a high consumption of meat, and vice versa. It is also confirmed by the correlation between the two variables of 0.56, so the result is far from unexpected. (Figure 9 in the appendix)

Comparing the joint distributions of the responses with the Income explanatory factor showed that in the initial area of the distributions as income increases, the amount spent on wine and meat is always contained within certain limits and then increases dramatically in the case of wine, while for meat there is moderate growth. Both distributions have outliers, in the case of MntWines there are customers with high incomes who have purchased small quantities of wine, while for MntMeatProducts the outliers are more varied, in fact there are customers who, as in the case of MntWines, have purchased small quantities of meat despite having high incomes; then there are those who have a large expenditure on meat corresponding to high levels of income and then there is only one customer who has a low income but has spent more than 1500 on meat in the last two years. The singular purchasing behaviour of this last customer was put under the magnifying glass and it was found that he was a married customer and that his Income was slightly higher (2447) than his meat expenditure (1725). Analysing the distribution of Mnt Wines and MntMeatproducts conditional on Education and Status it is evident in both cases that the lowest level of Education "Basic" is definitely compressed for very low values of expenditure and that many consumers have purchased few quantities of both products; it was rather interesting to see that those who consumed more wine were divorcees and widowers with levels of Education such as: PhD, 2nd Cycle, Master and Graduation. More generally, those with PhD and Master's degrees are regular wine drinkers with an expenditure of more than
500. For those who chose meat, the graph showed that the biggest consumers were single people and again widowed people with the same levels of education as for wine, spending at least 250. (Figures 12-13 in Appendix)

4. Boosting and regression tree

By using these two methods, it was possible to achieve the final objective of the study; in fact, the procedures listed above allowed predictions to be made for the response variables (MntWines and MntMeatProducts). The boosting procedure for MntWines was carried out by implementing nine different attempts in each of which the parameter values were varied:

- Number of trees optimal for the procedure
- λ, shrinkage parameter controls the speed of adaptation of the data to the procedure.
- Interaction depth, which controls the depth of the shaft

After implementing all the procedures with the help of the **gbm** function, the one with the smallest test error value was selected, with a number of trees of 611, shrinkage of 0.05 and interaction depth of 3. The most important variables for this procedure are Income and MntMeatProducts, which is also confirmed by further graphical representations from which it can be seen that both have a marginal positive effect on MntWines,in line with what was seen in the exploratory analysis. (Figures 14-15-16-17-18-19 and Table 2 in the appendix).

As regards the creation of the regression tree, again for MntWines, the first 75% of the observations were selected as the training set and the remaining 25% as the test set (it was first verified that the dsataset did not have any type of increasing or decreasing order of the variables). Using the **rpart** function, the tree was generated and then pruned with the **prune** function, inserting the appropriate value of the cp parameter (cost-complexity parameter). The pruned tree has an R^2 of 0.61 and a correlation of 0.78, still the most important variables are confirmed to be Income and MntMeatProducts. Figures (20--2221 and Table 3 in theappendix)

In the final phase, the comparison between the single regression tree and the boosting procedure was madeand it was possible to identify the predictions (black dots in the horizontal) of the single tree indicated by the average of all the observations within the same leaves; however, it was not possible to predict high values in the response variable, this is demonstrated by the different measurement scales, in fact, in the y-axis the upper limit is 800 while the x-axis reaches 1500. Finally, the correlation was considered as a performance indicator: that of the boosting procedure is equal to 0. 81 which proves to be higher than that of the single tree, moreover the correlation between the two predictions is equal to 0.90 so the procedures have similar predictions, but boosting was selected as the best method to make predictions on MntWines. (Figure 23 in appendix)

For the boosting procedure and the regression tree for MntMeatProdcucts the above procedure was followed. The most important variables for boosting are Income and MntWines, which is also confirmed by further graphical representations from which it can be seen that both have a marginal positive effect on MntMeatProducts consistently with what was seen in the exploratory analysis. (Figures 2-2-2-28-294567 and Table 4 in Appendix)

For the regression tree, on the other hand, the pruned tree has an R^2 of 0.60 and the most important variables are Income and MntFishProducts and a correlation of 0.77. (Figures 30--3231 and Table 5 in the Appendix).

Again, in the final phase, comparing the single regression tree with the boosting procedure, it was possible toaffirm all the previously deduced considerations, however, in addition it must be said that the average of thevalues in the leaves in this case is influenced by outliers. Also, in this case it was not possible to predict high values in the response variable since they have different measurement scales, in fact, on the y-axis the upperlimit of the y is 500 while on the x-axis the x reaches 1500. Finally, always using the correlation as a performance indicator: that of the boosting procedure is equal to 0. 78which proves to be higher than that of the single tree; moreover, the correlation between the predictions of the two procedures is equal to 0.94so the procedures have similar predictions, but boosting was selected as the best method to make predictionson MntMeatProducts. (Figure33)

5. Conclusions

From the following analysis, it was possible to outline some of the fundamental aspects of a wine and meat consumer. First of all, it is the annual Income of each customer that determines, in most cases, the amount of expenditure for the two product categories, but this is not decisive for the choice between wine and meat; in fact, the analysis of the comparison between these two products made it possible, in particular, to investigate the profiles that most preferred wine and meat. The most frequent wine drinkers included divorced and widowed customers with at least a bachelor's degree, while regular meat eaters included singlecustomers and again widowed people with the same level of education as frequent wine drinkers. The results show that the presence of children or teenagers is negatively correlated, or not correlated at all, with the purchase of the two product categories, while other product categories such as fruit, fish and sweets are also correlated with meat and wine (more strongly with the former than with the latter). There is also a low correlation between the two product categories and the purchase of gold.

Finally, although the forecasting methods adopted returned predictions for certain values, these methods cannot be considered satisfactory for the purposes of this analysis, as they could not allow for predictions of high values in the response variables.

Appendix

Dataset :

Customers general informations

- ID: Customer’s unique identifier
- Year_Birth: Customer’s birth year
- Education: Customer’s education level
- Marital_Status: Customer’s marital status
- Income: Customer’s yearly household income
- Kidhome: Number of children in customer’s household
- Tennhome: Number of teenagers in customer’s household
- Dt_Customer: Date of customer’s enrollment with the company
- Complain: 1 if a customer complained in the last 2 years, 0 otherwise

Amount spent for each category in the last 2 years

- MntWines: Amount spent on wine in the last 2 years
- MntFruits: Amount spent on fruits in the last 2 years
- MntMeatProducts: Amount spent on meat in the last 2 years
- MntFishProducts: Amount spent on fish in the last 2 years
- MntSweetProducts: Amount spent on sweets in the last 2 years
- MntGoldProds: Amount spent on gold in the last 2 years

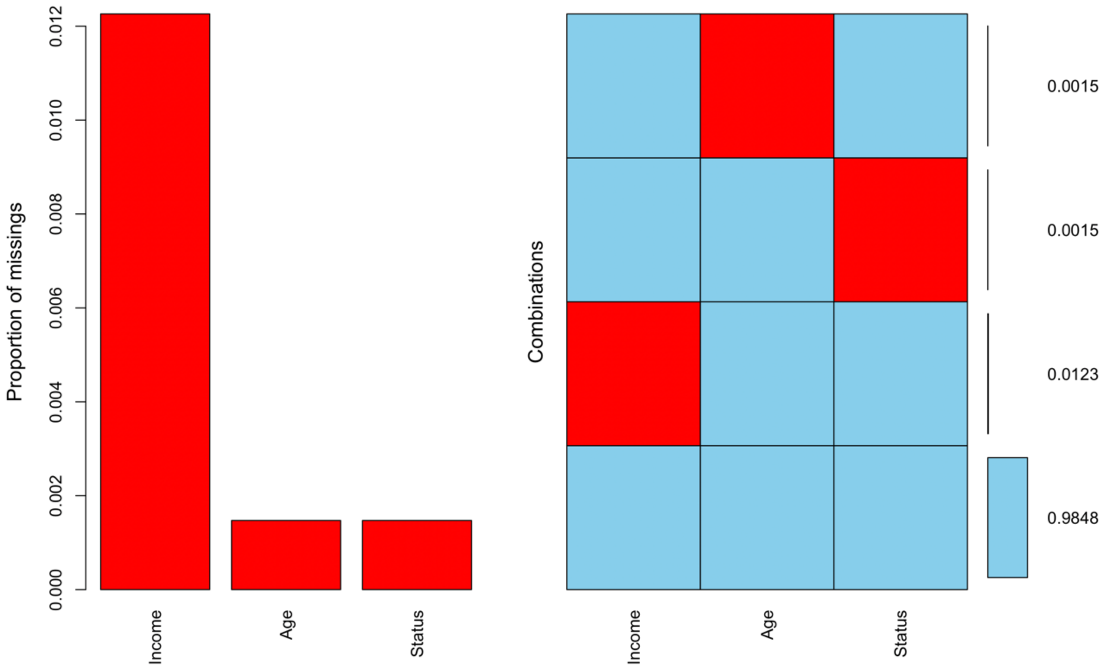


Figura 1

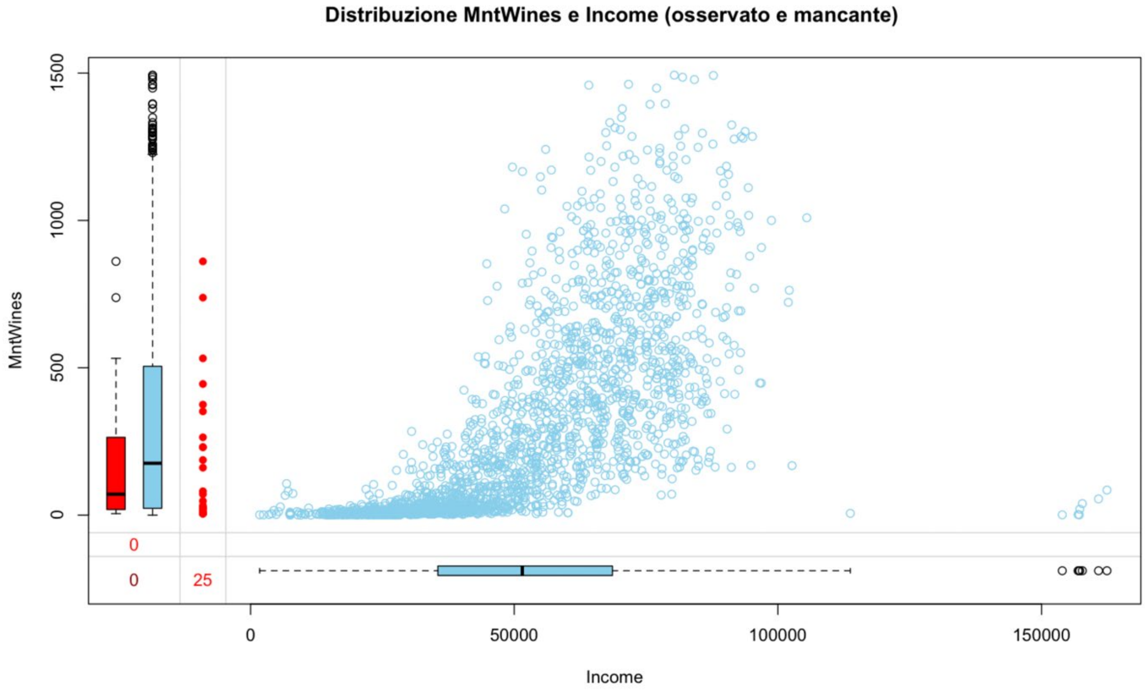


Figura 2

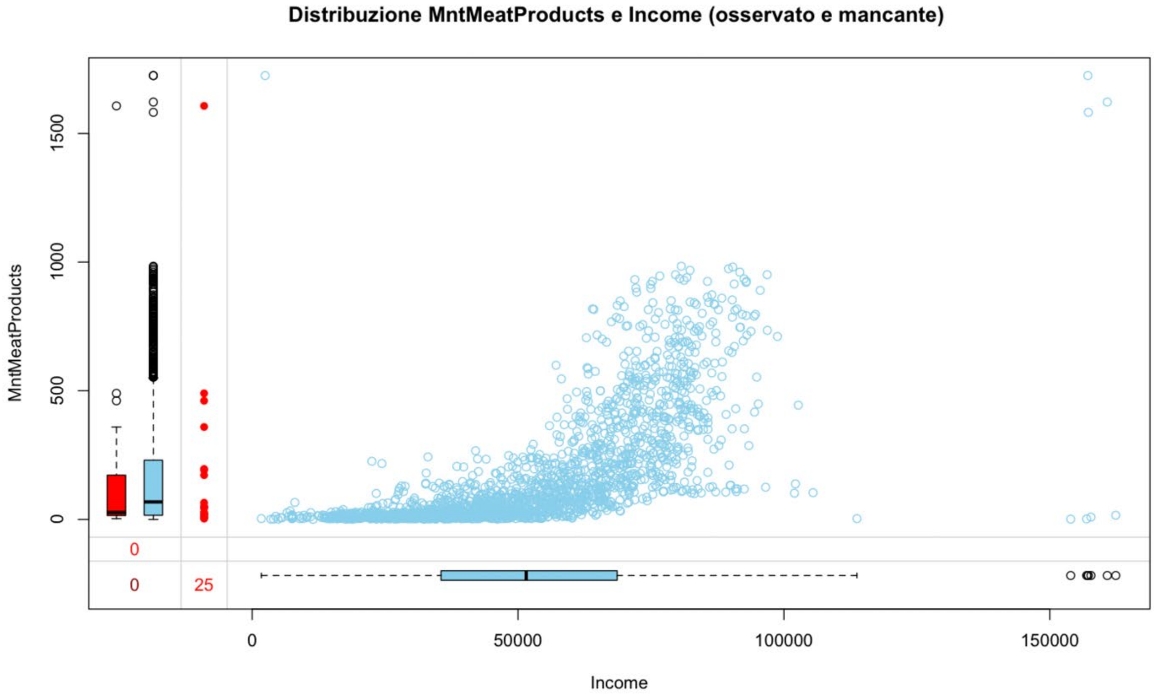


Figura 3

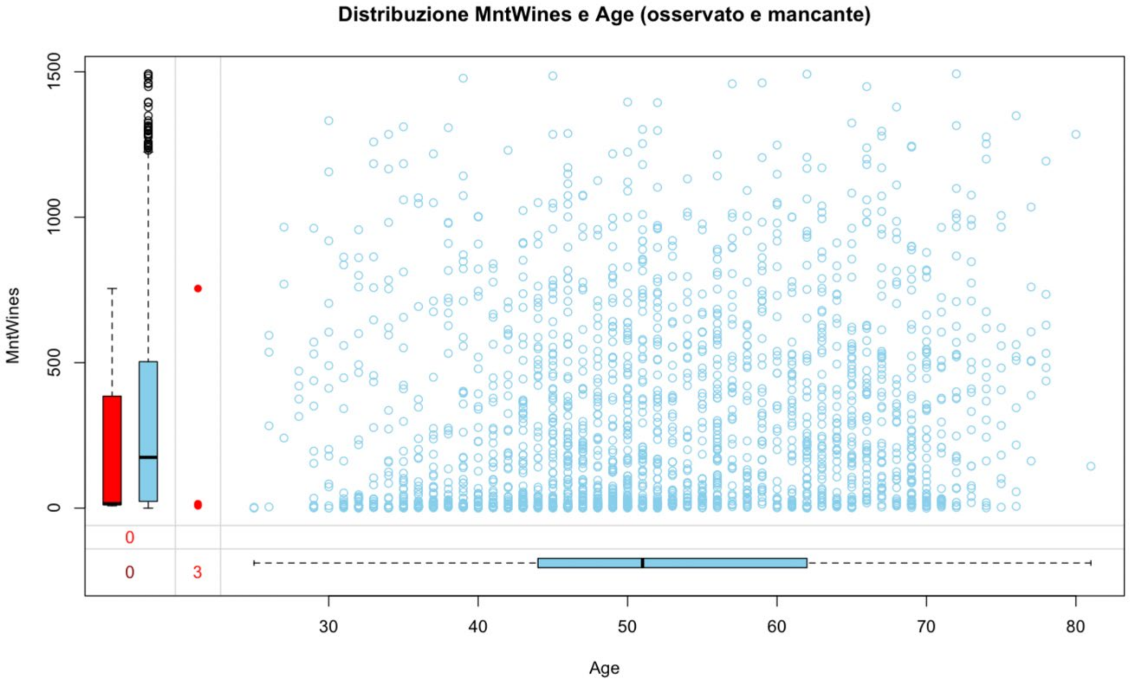


Figura 4

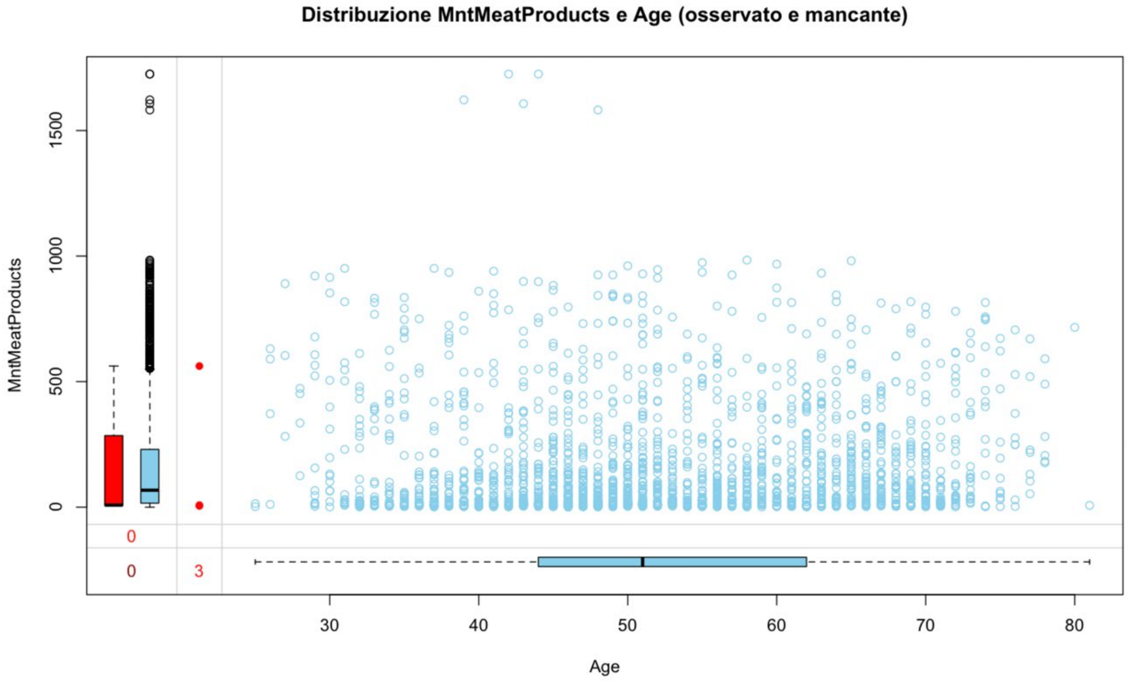


Figura 5

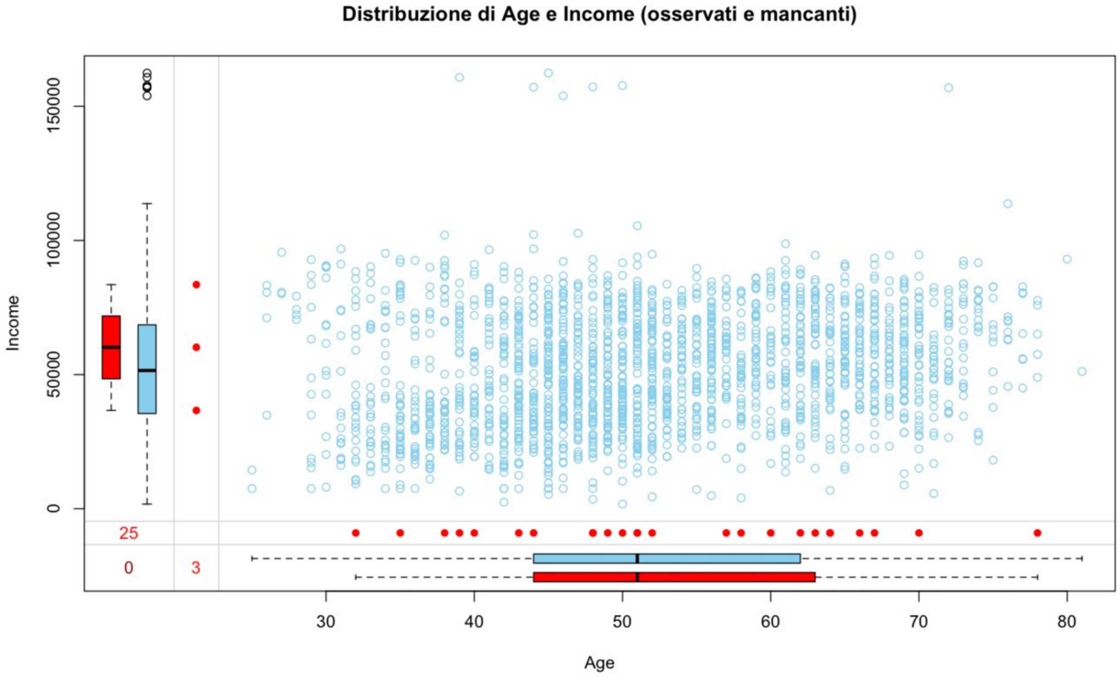


Figura 6

MntWines	MntMeatProducts
Age 0.17	Age 0
Income 0.68	Income 0.69
Kidhome -0.5	Kidhome -0.44
Teenhome 0	Teenhome -0.26
Education 0.17	Education 0
MntFruits 0.39	MntFruits 0.54
MntMeatProducts 0.56	MntWines 0.56
MntFishProducts 0.40	MntFishProducts 0.56
MntSweetProducts 0.38	MntSweetProducts 0.51
MntGoldProds 0.39	MntGoldProds 0.34

Tabella 1

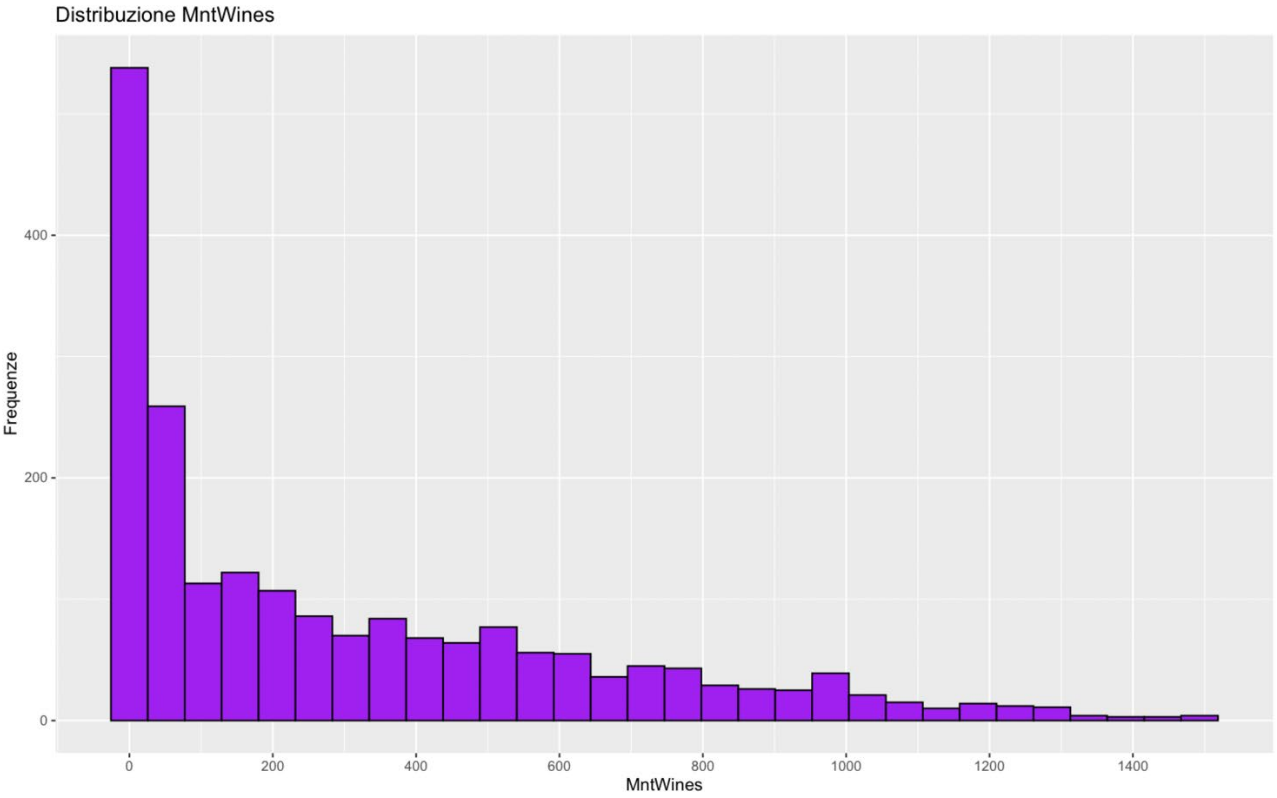


Figura 7

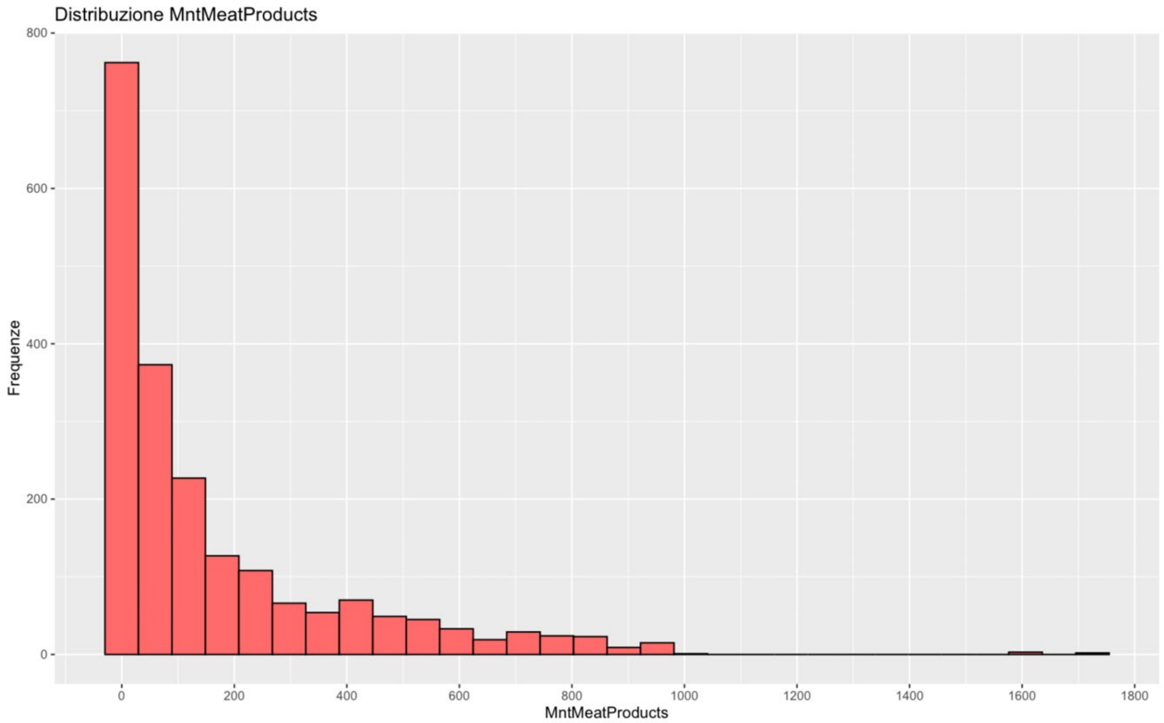


Figura 8

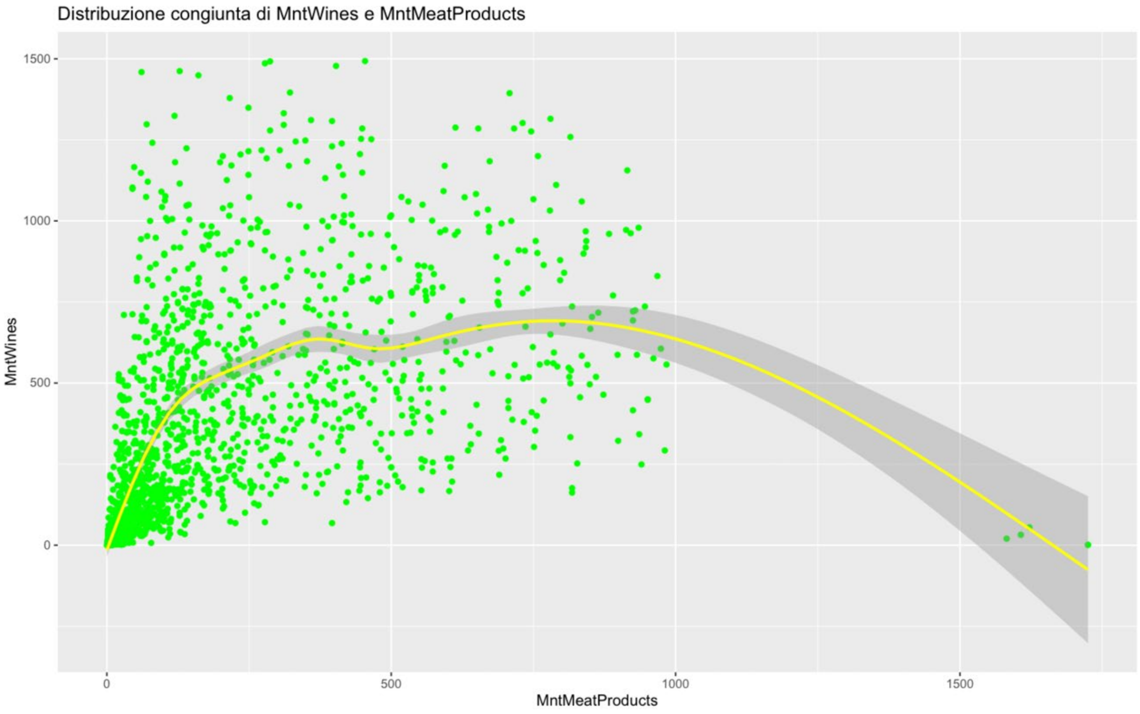


Figura 9

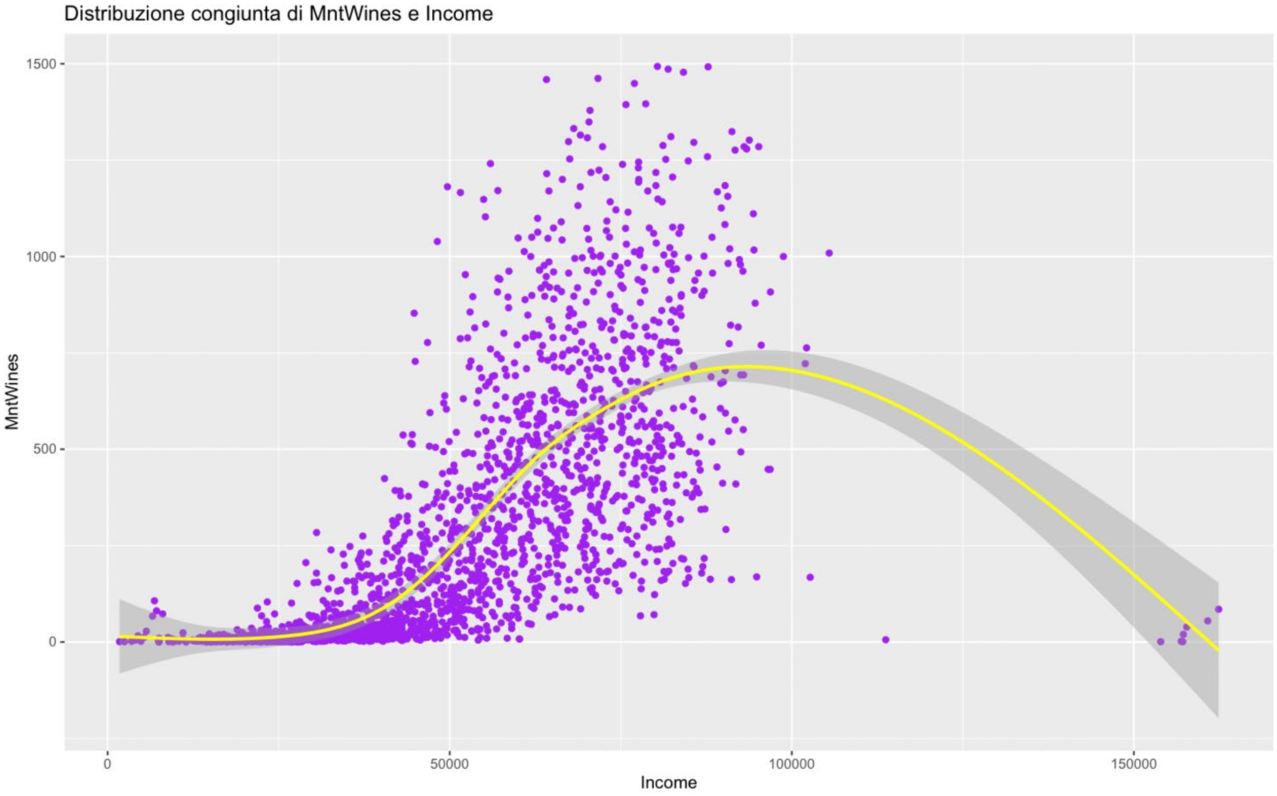


Figura 10

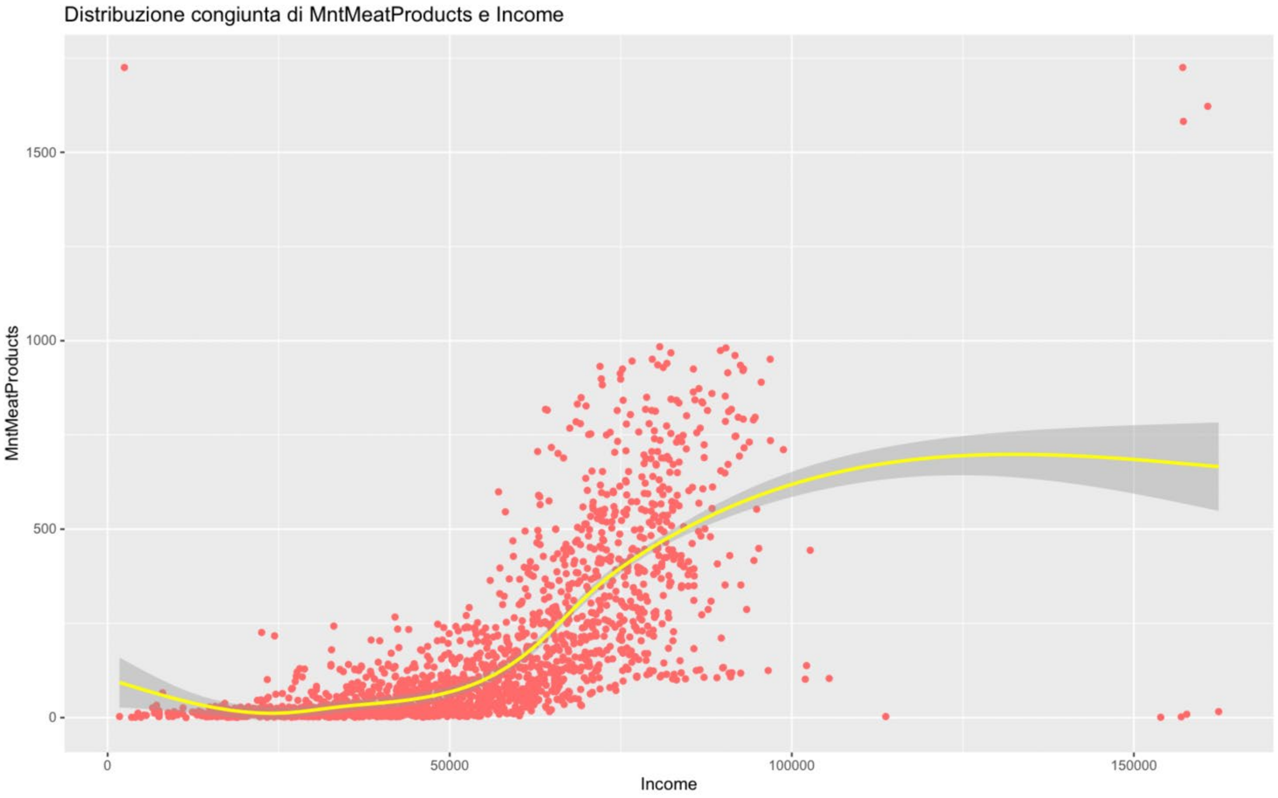


Figura 11

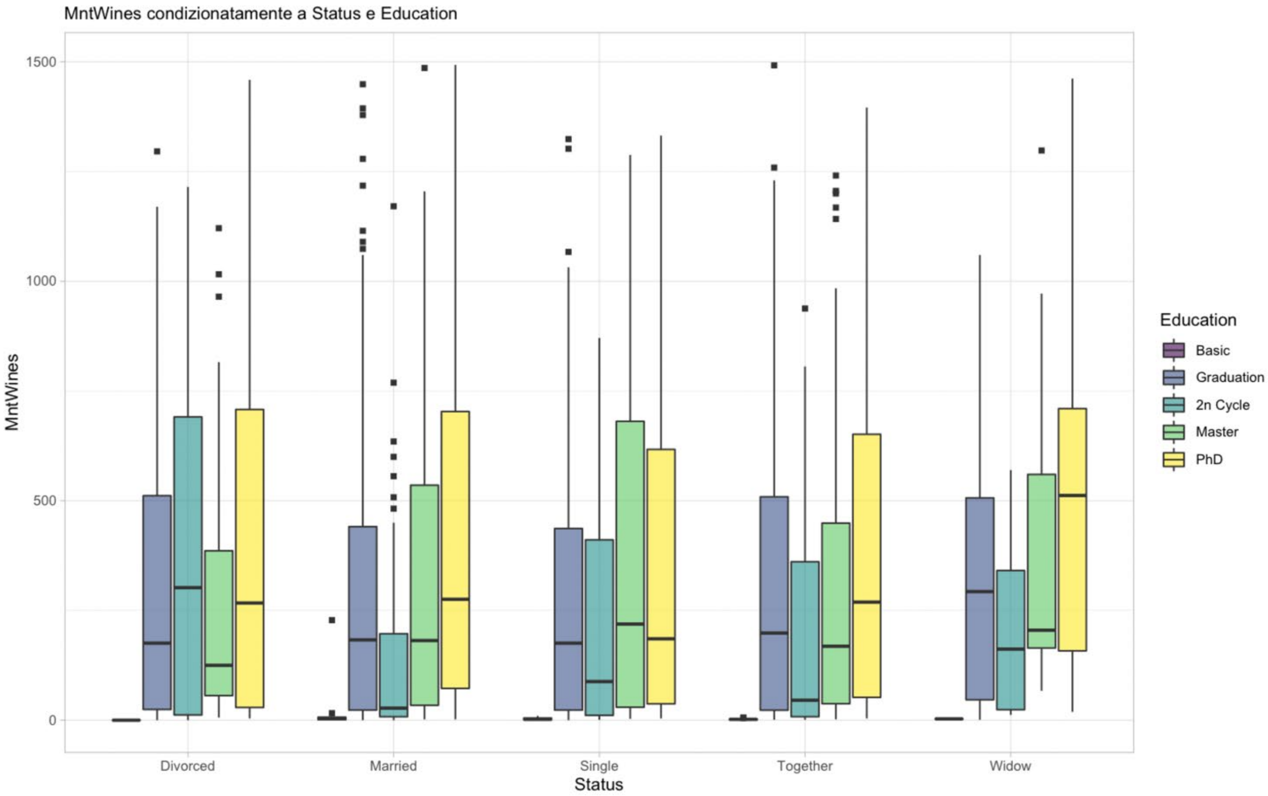


Figura 12

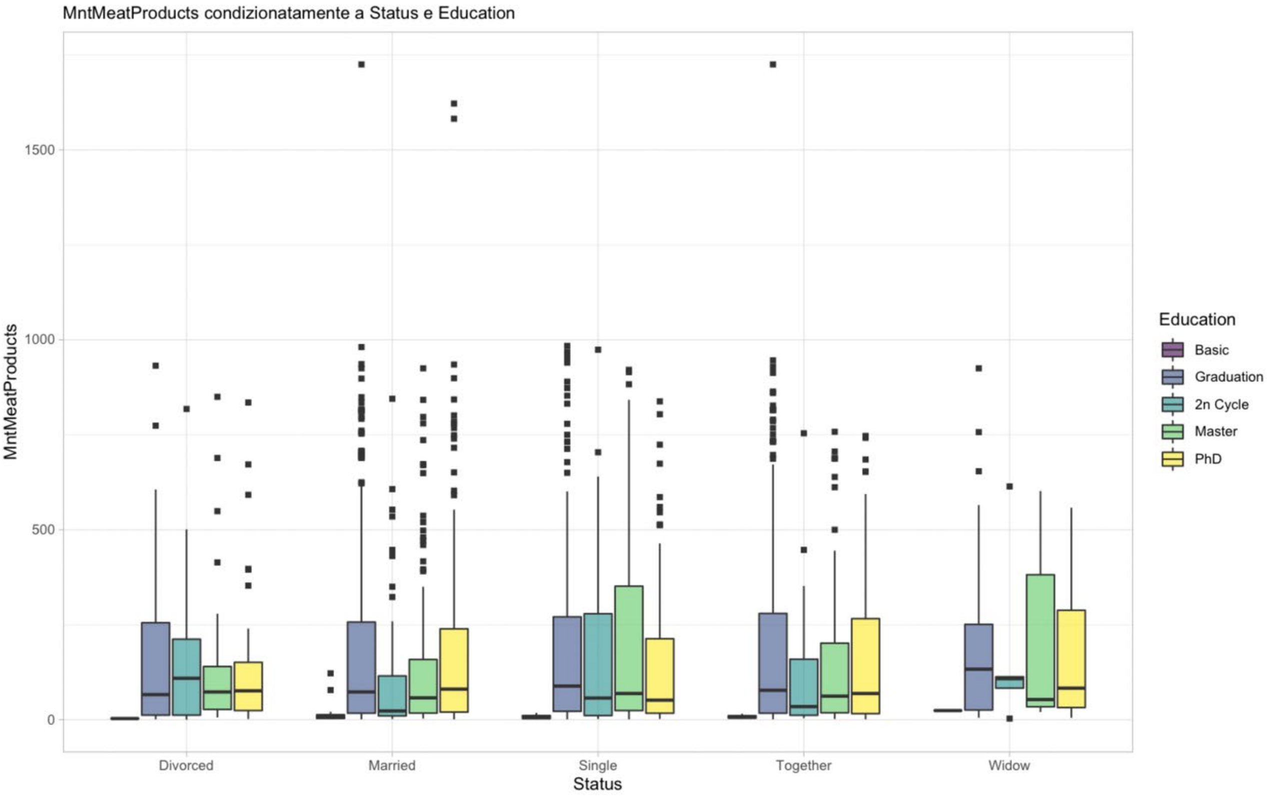


Figura 13

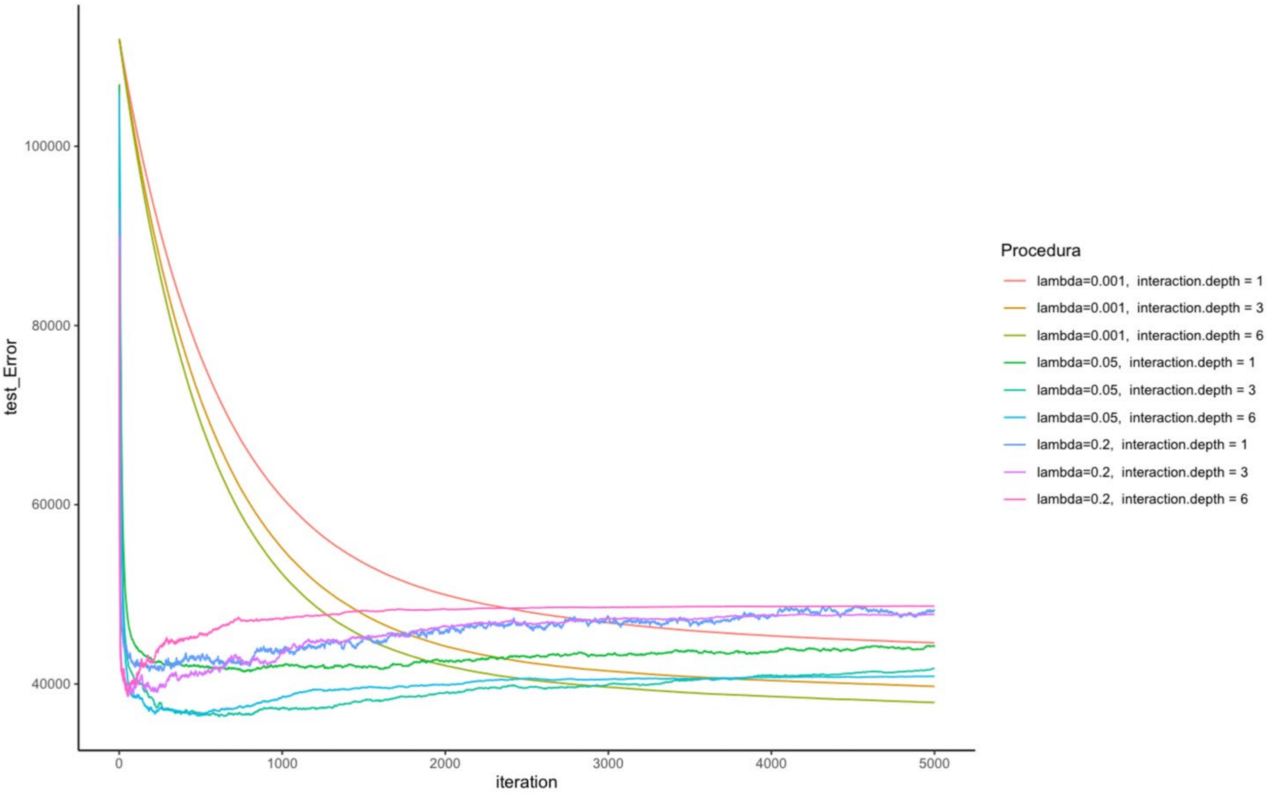


Figura 14 Test error per MntWines

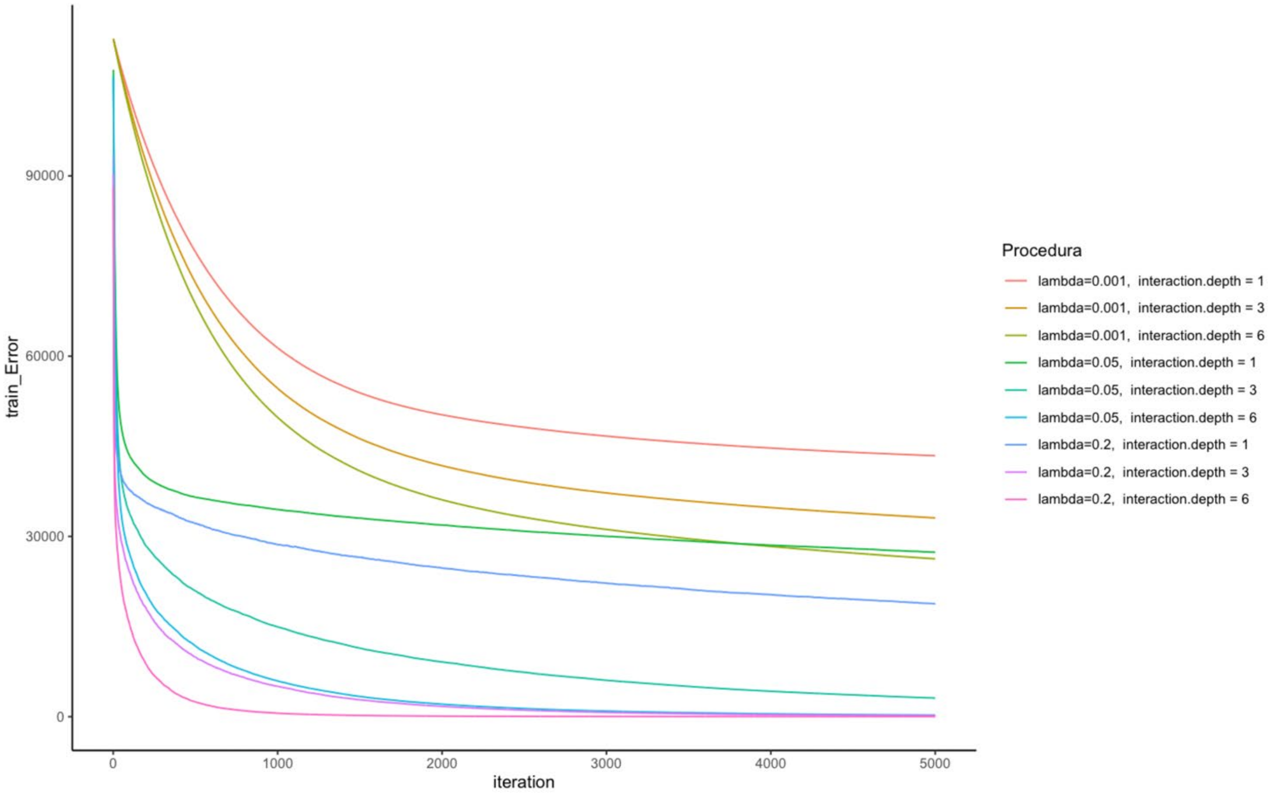


Figura 15 Train error per MntWines

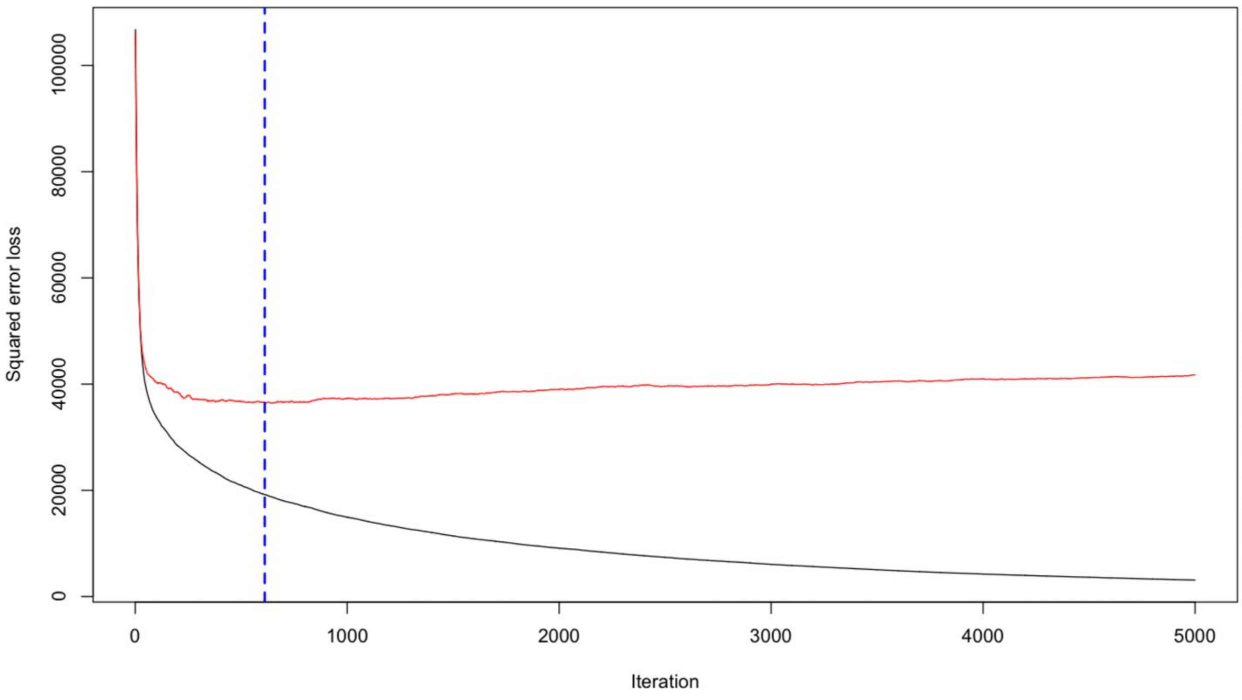


Figura 16 Numero ottimo di alberi per la procedura pari a 611

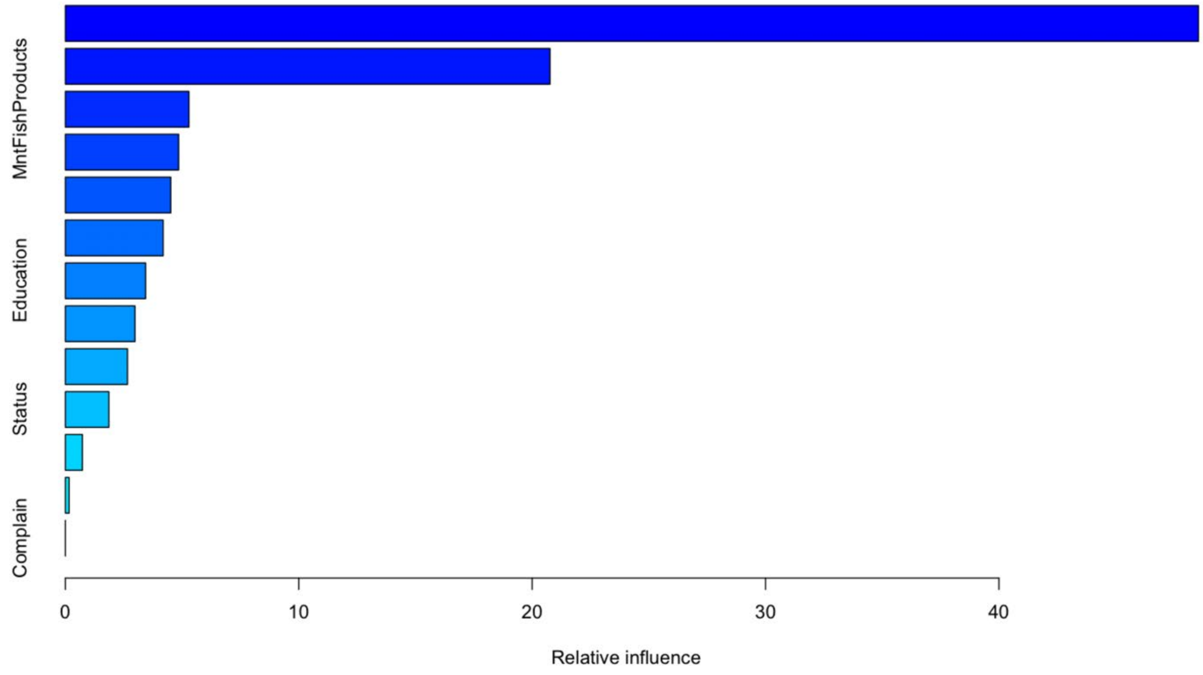
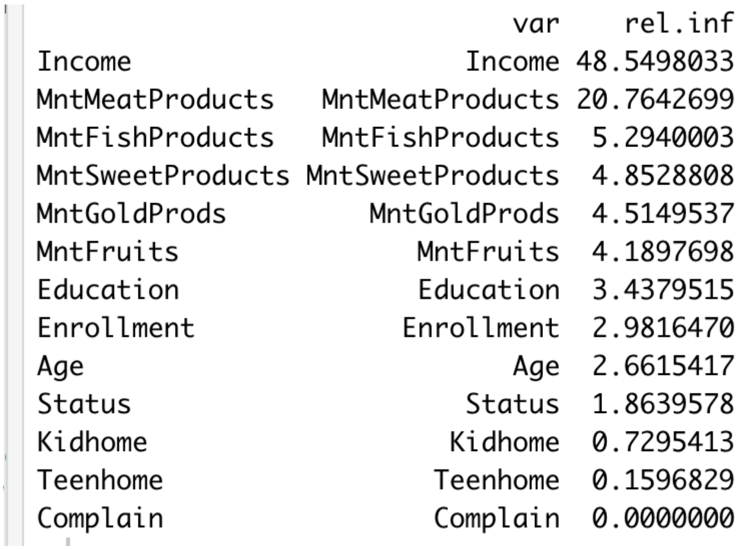


Figura 17 Importanza delle variabili del boosting per MntWines



```
                                  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 2 Importanza variabili boosting per MntWines

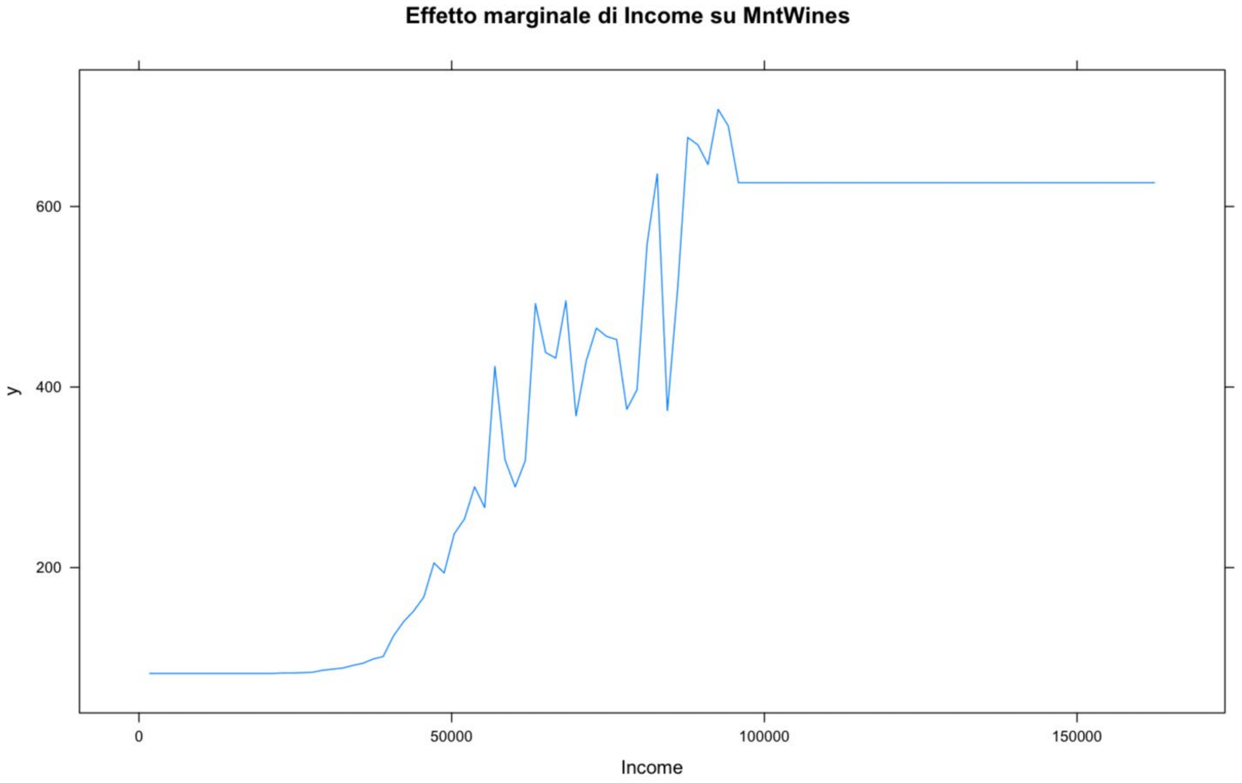


Figura 18

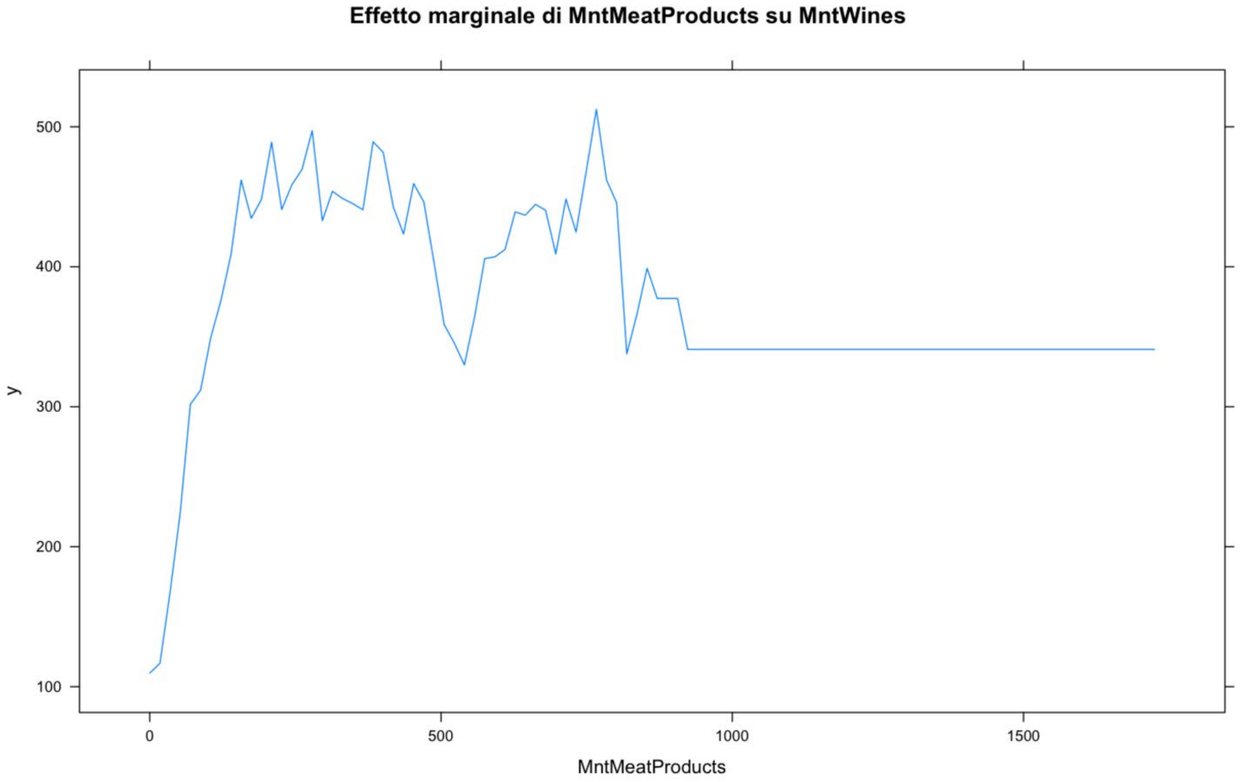


Figura 19

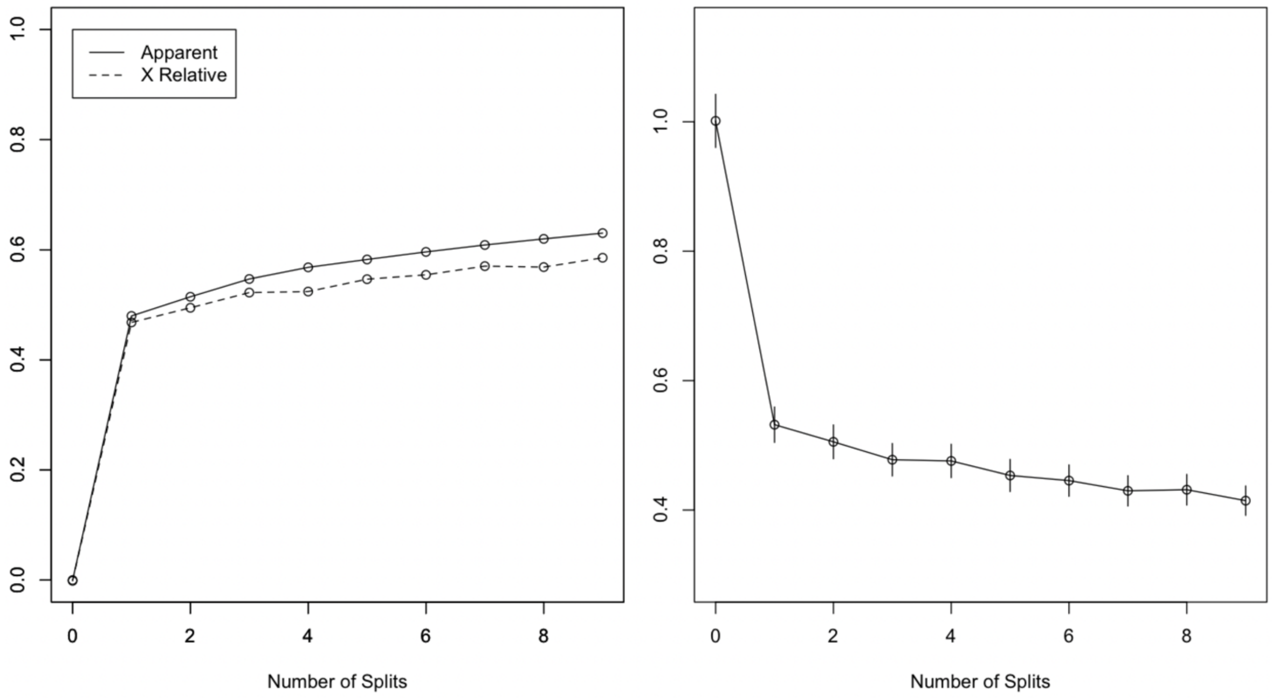


Figura 20 R^2 e cp dell'albero di regressione per MntWines

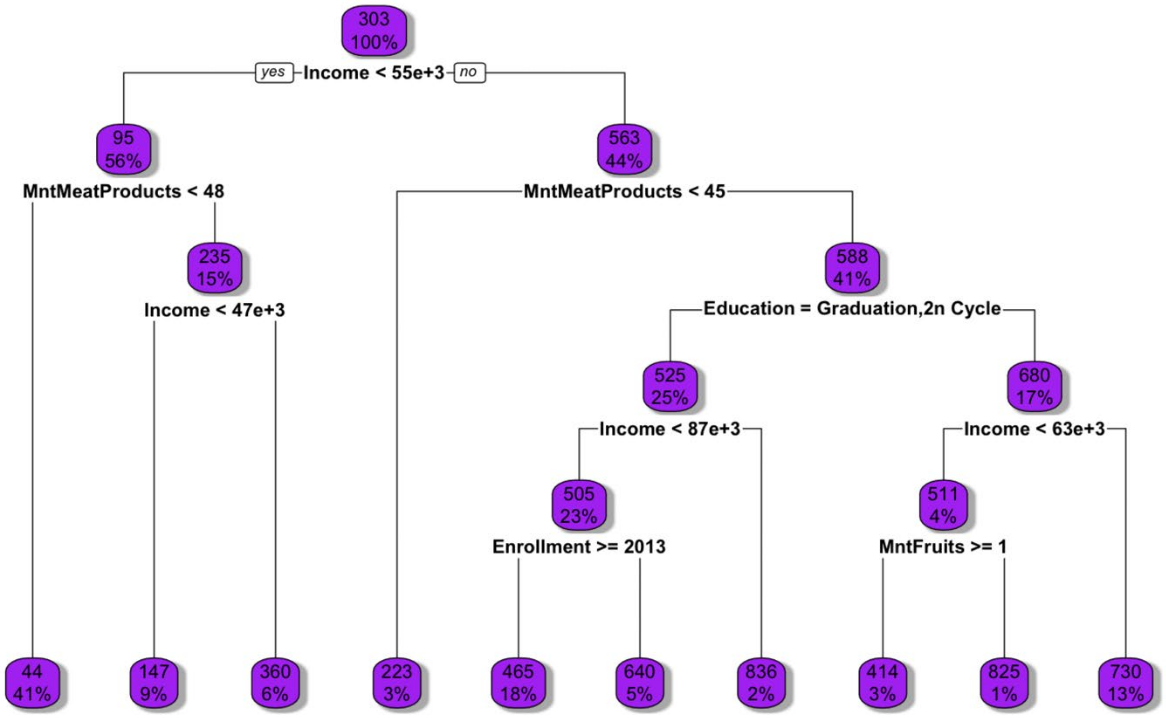


Figura 21 Albero di regressione per MntWines

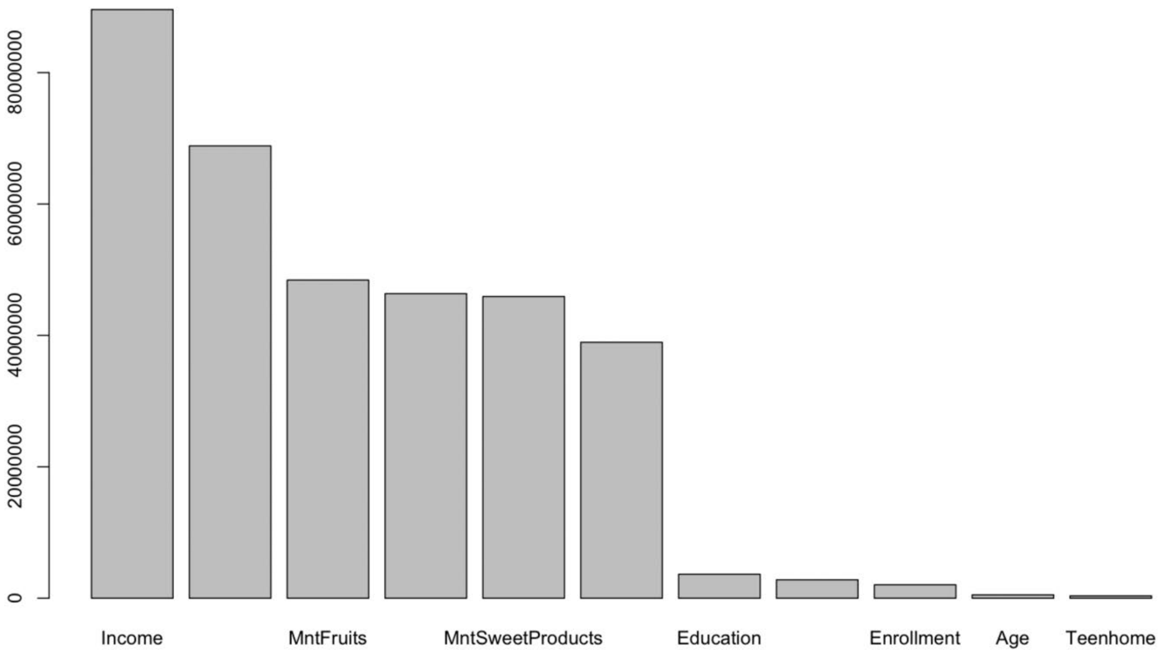
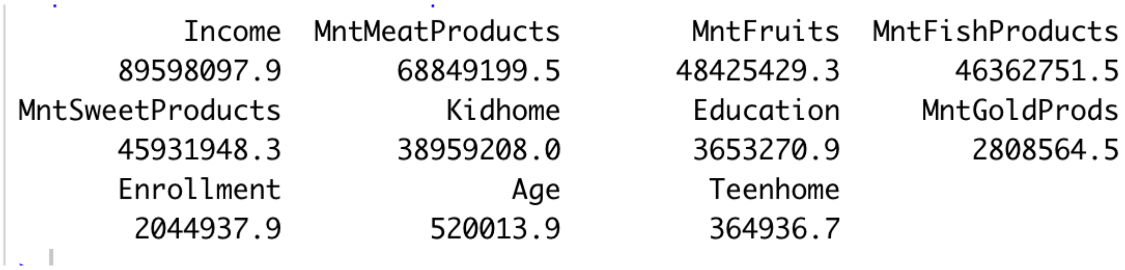


Figura 22 Importanza delle variabili dell'albero di regressione per MntWines



```
          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 3

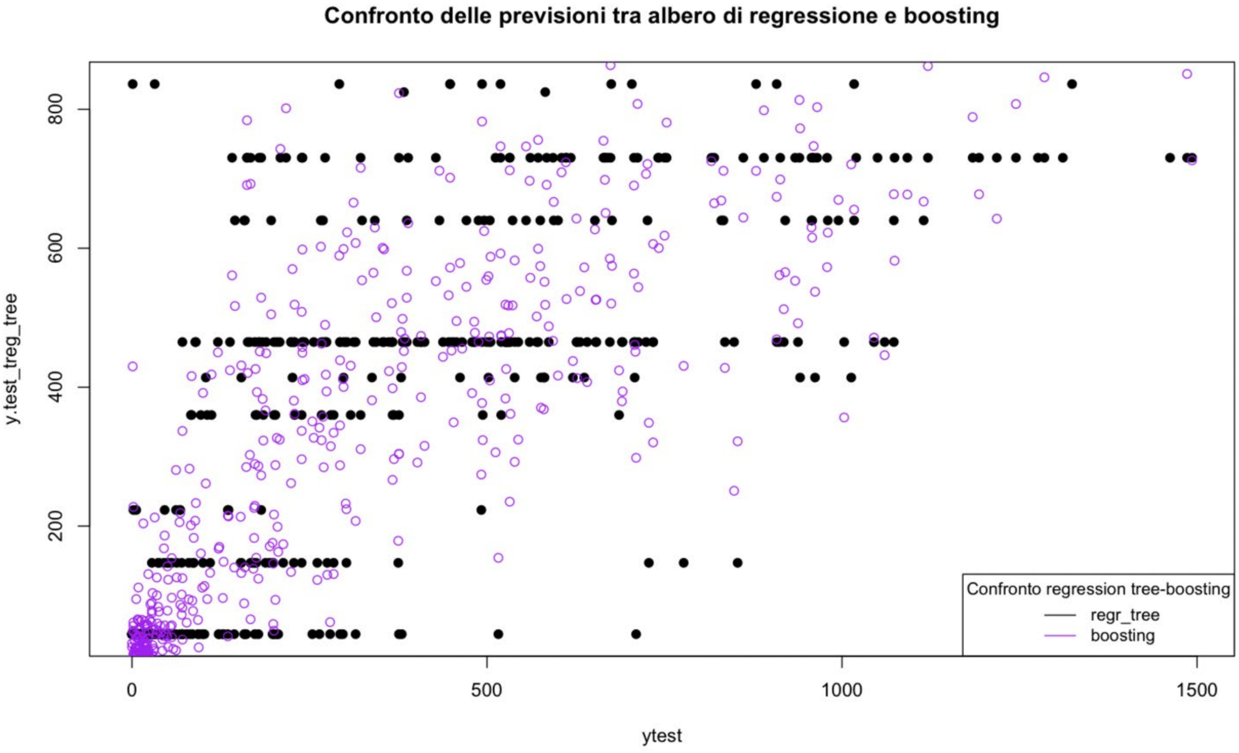


Figura 23

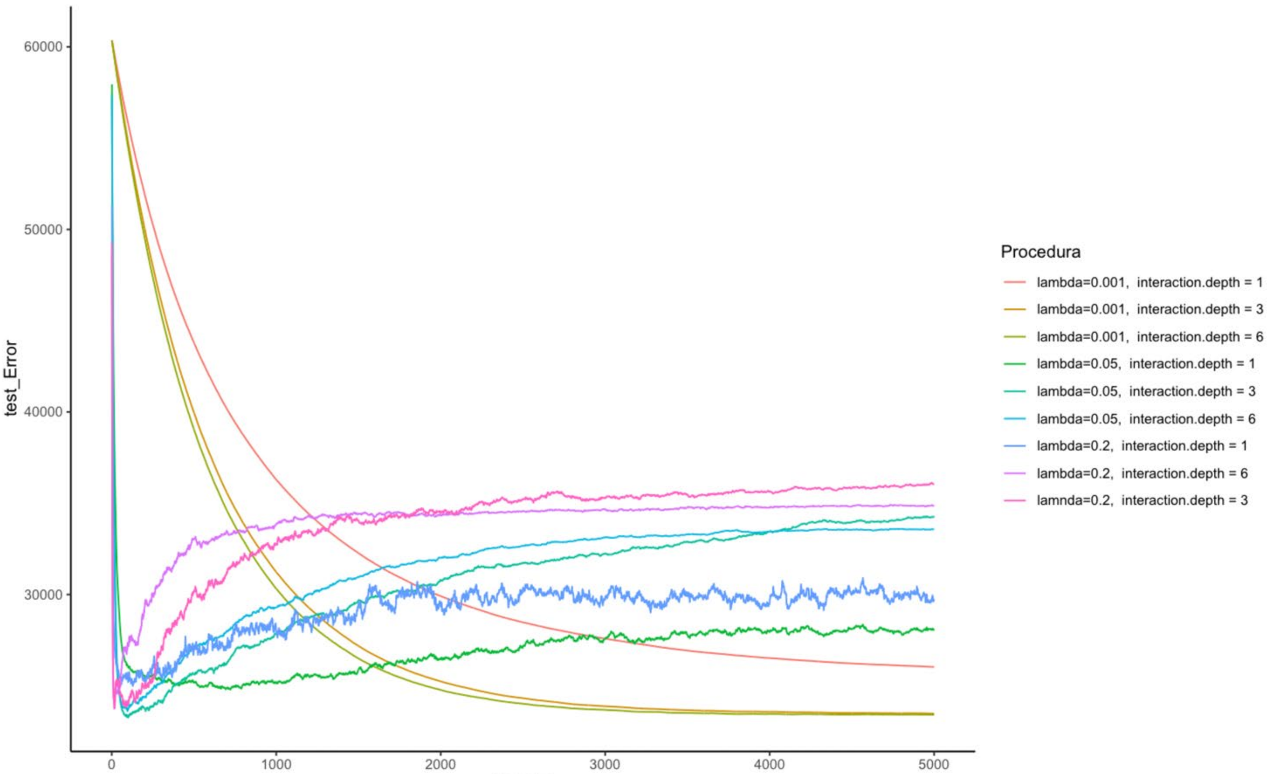


Figura 24 Test error per MntMeatProducts

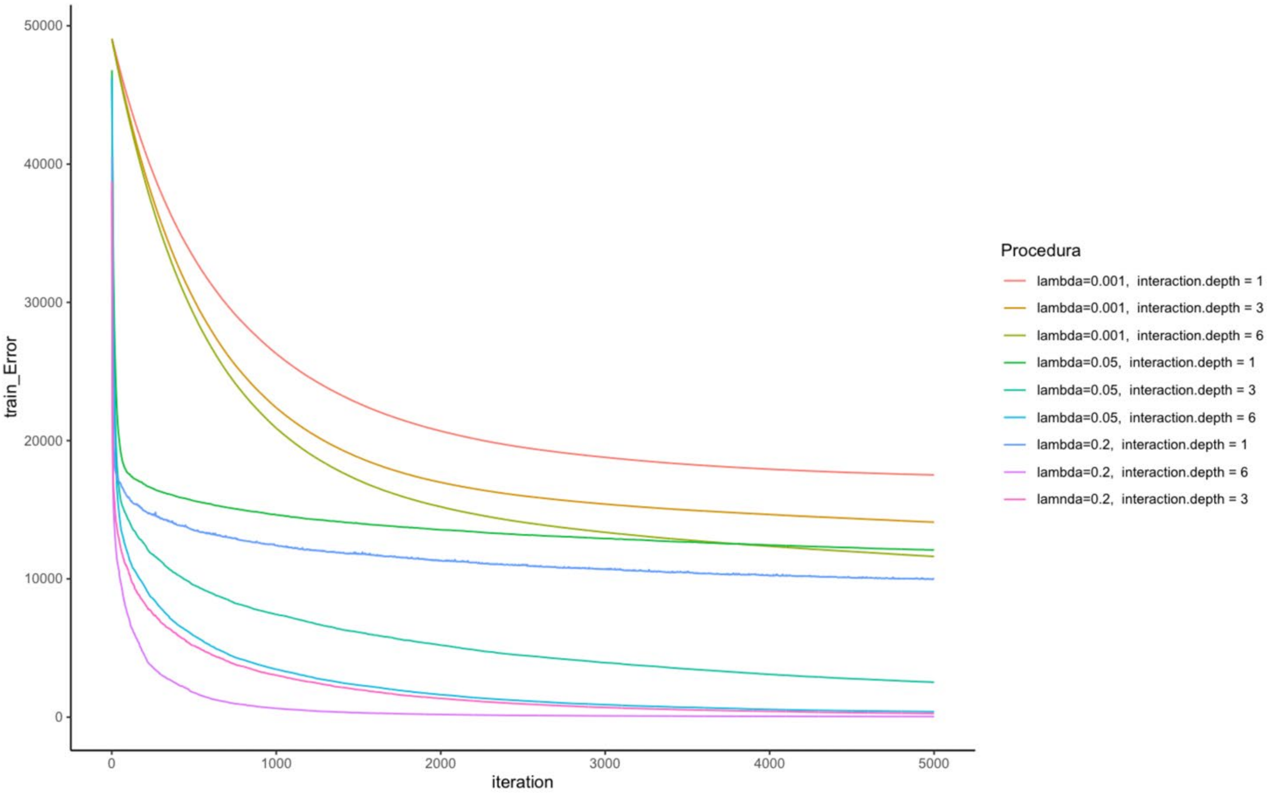


Figura 25 Train error per MntMeatProducts

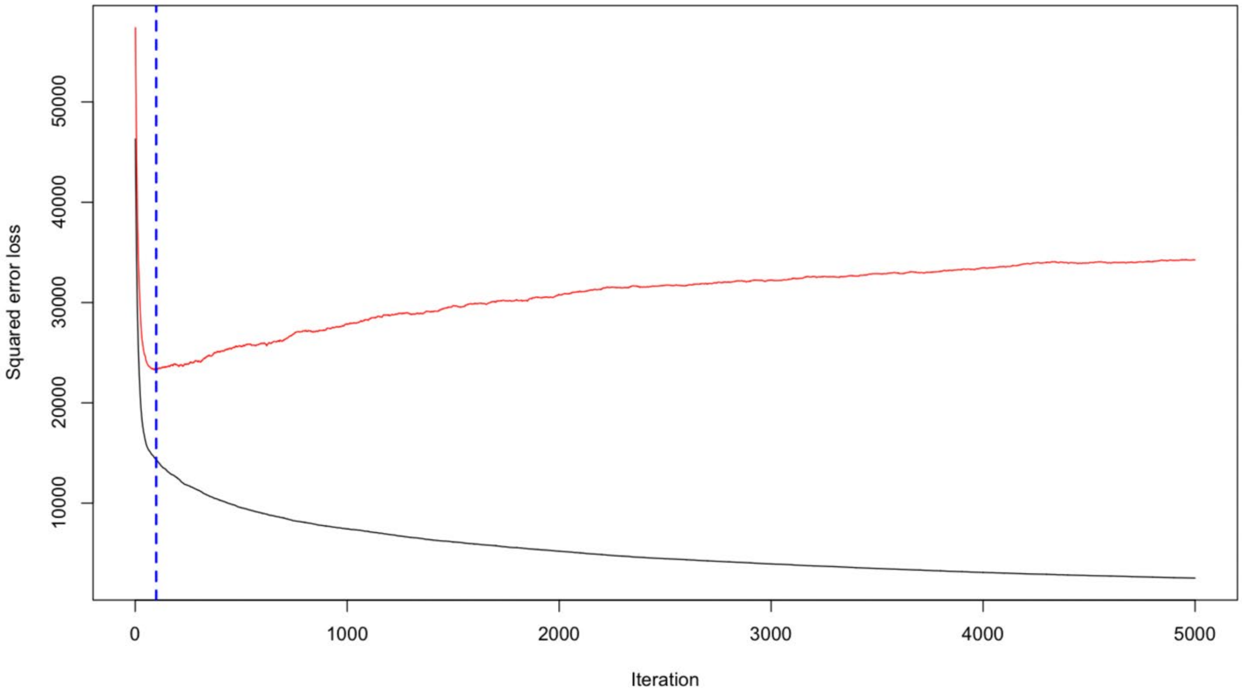


Figura 26 Numero ottimo di alberi la procedura pori a 99

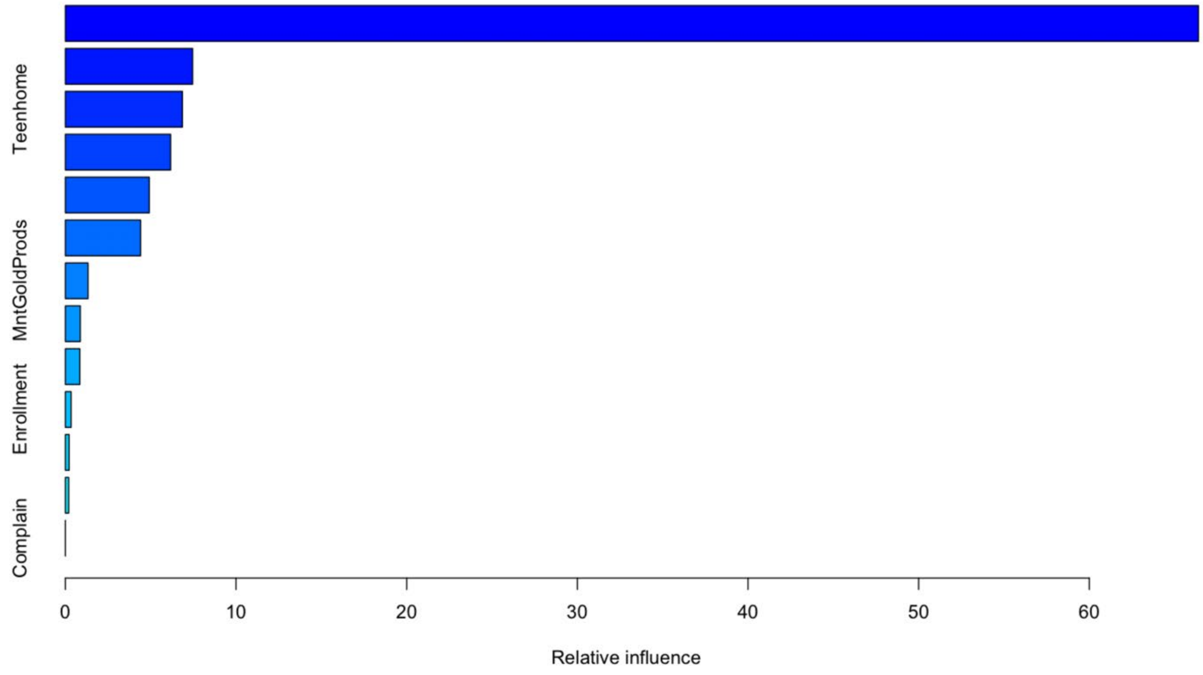
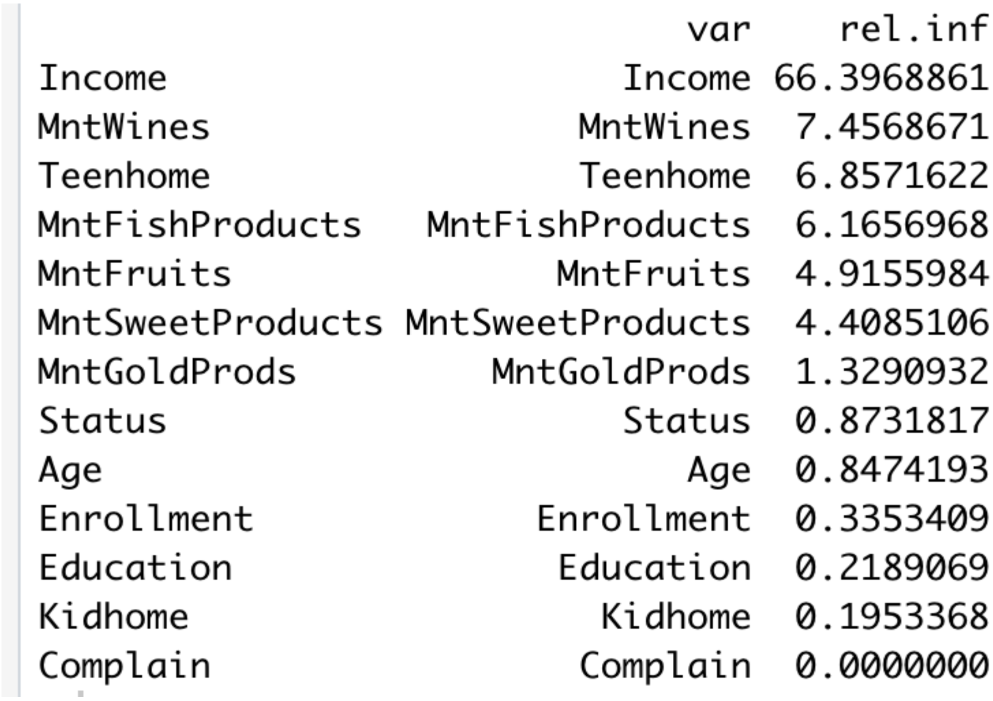


Figura 27 Importanza delle variabili del boosting per MntMeatProducts



```
                                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 4

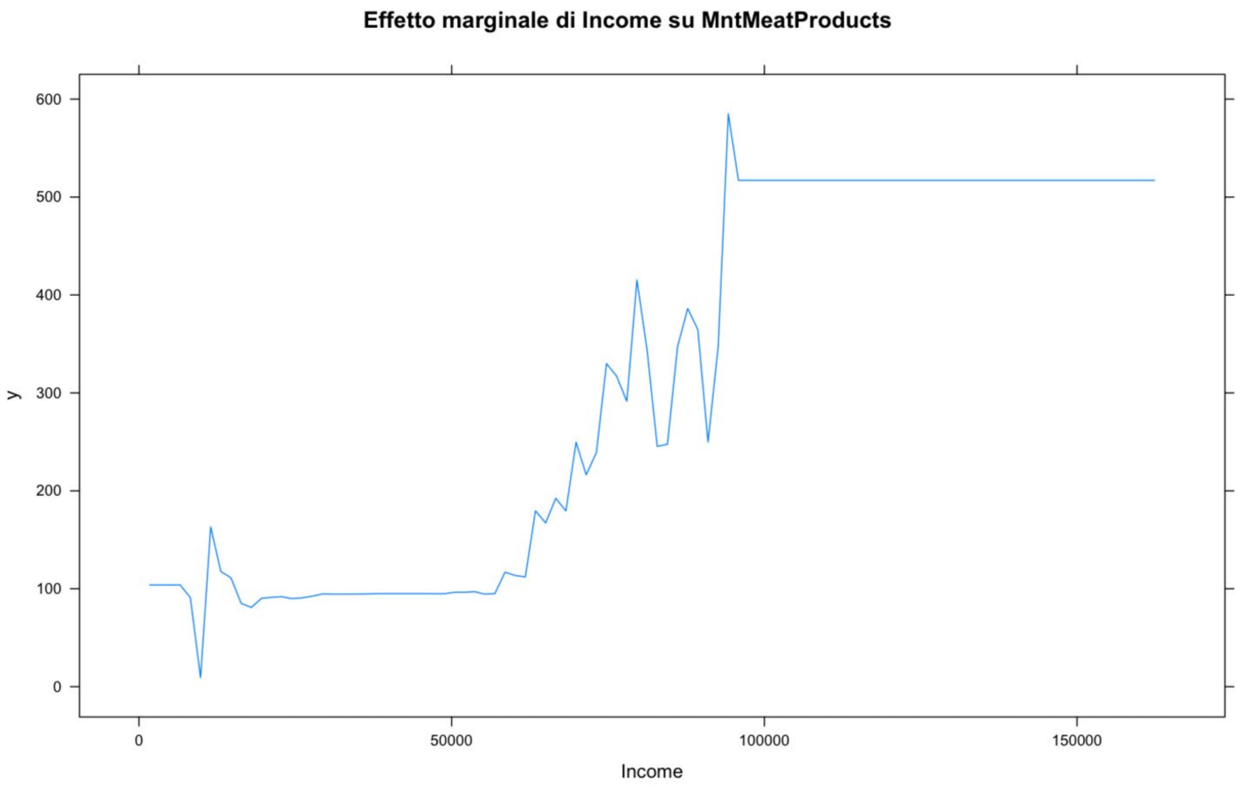


Figura 28

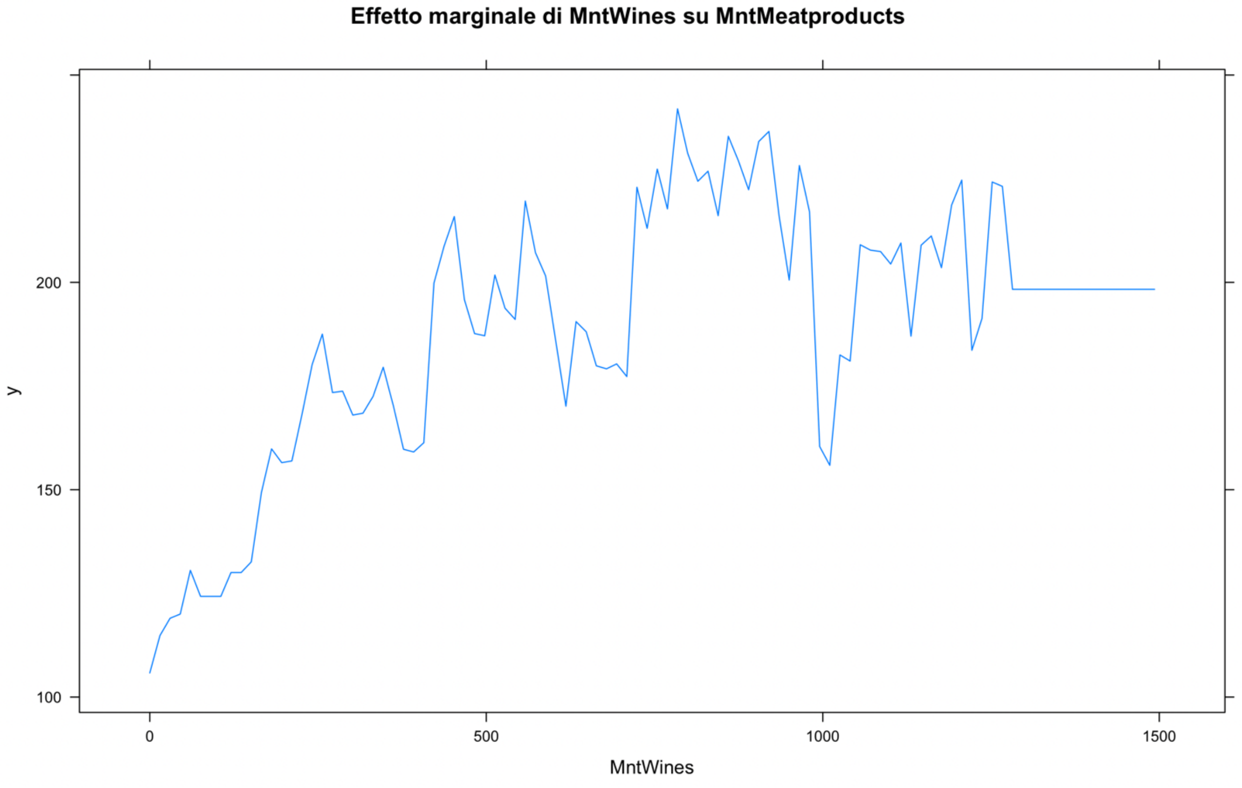


Figura 29 Effetto marginale di MntWines su MntMeatProducts

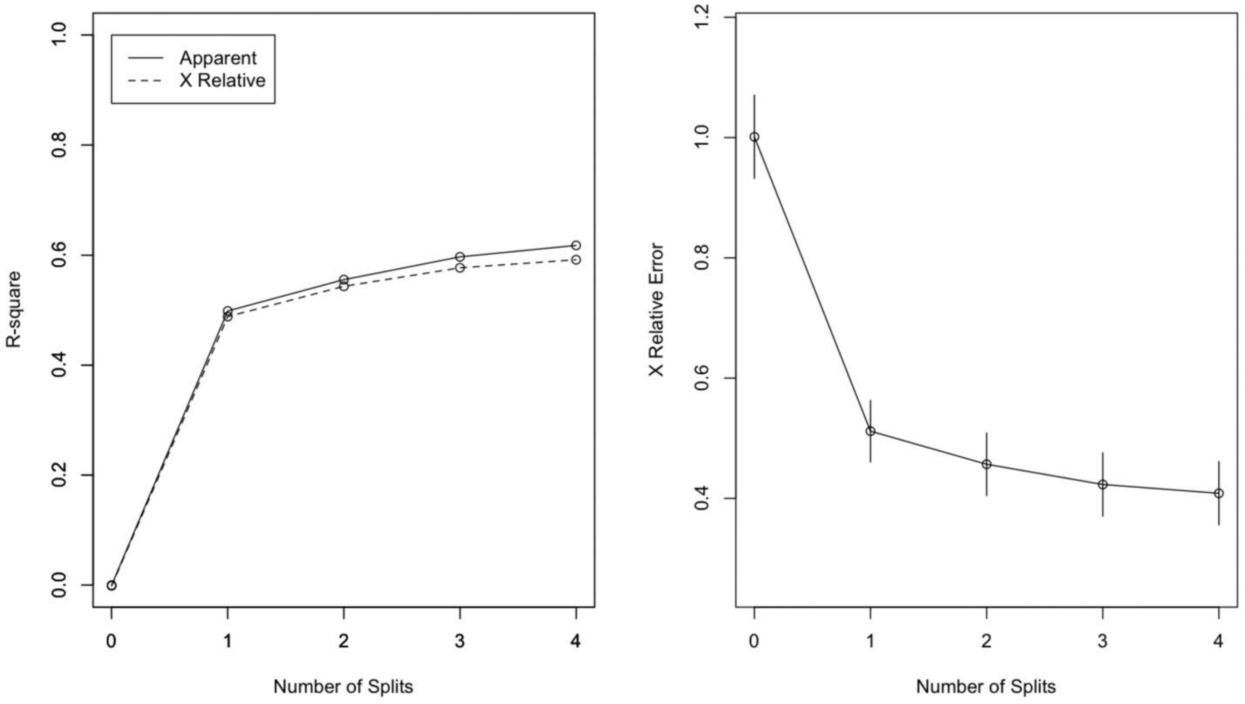


Figura 30 R^2 e cp dell'albero di regressione per MntMeatProducts

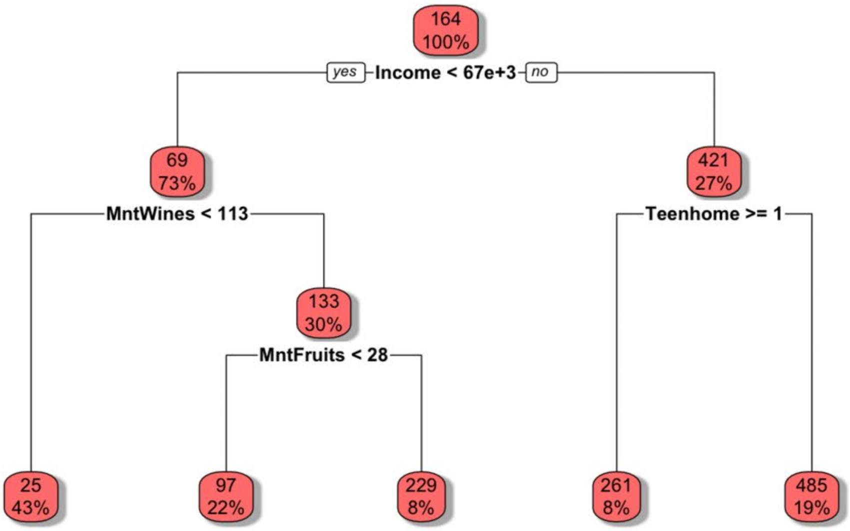


Figura 31 Albero di regressione per MntMeatProducts

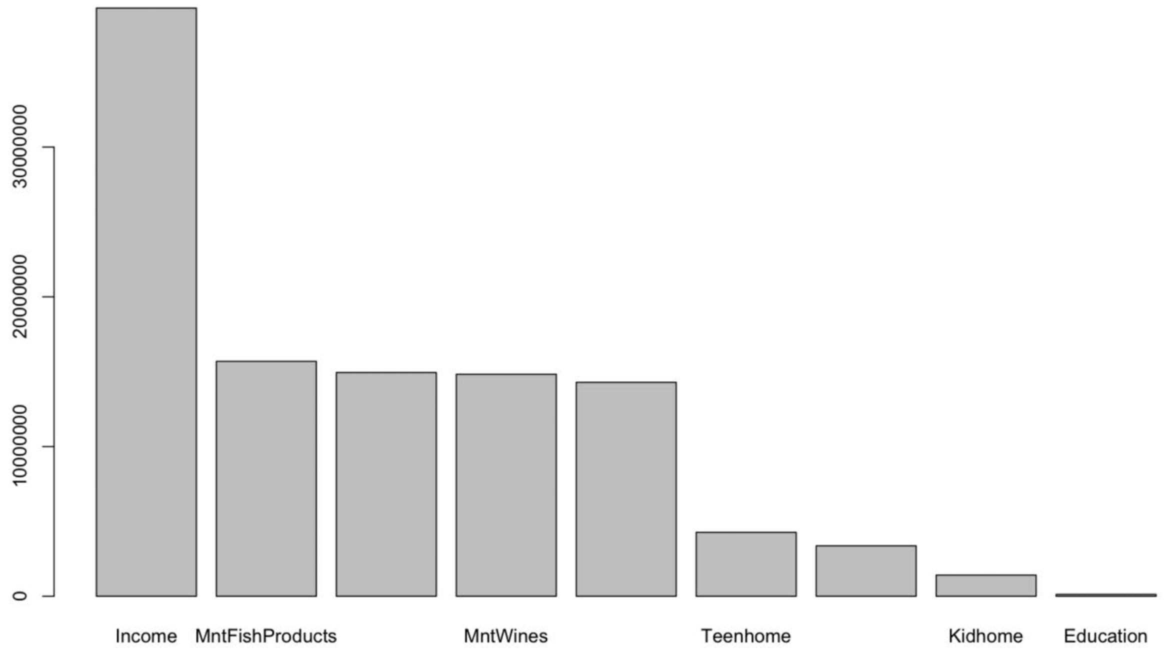
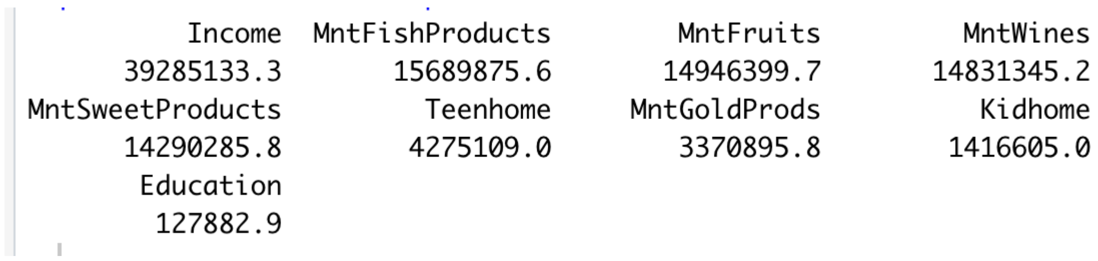


Figura 32 Importanza delle variabili dell'albero di regressione per MntMeatProducts



```
         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 5

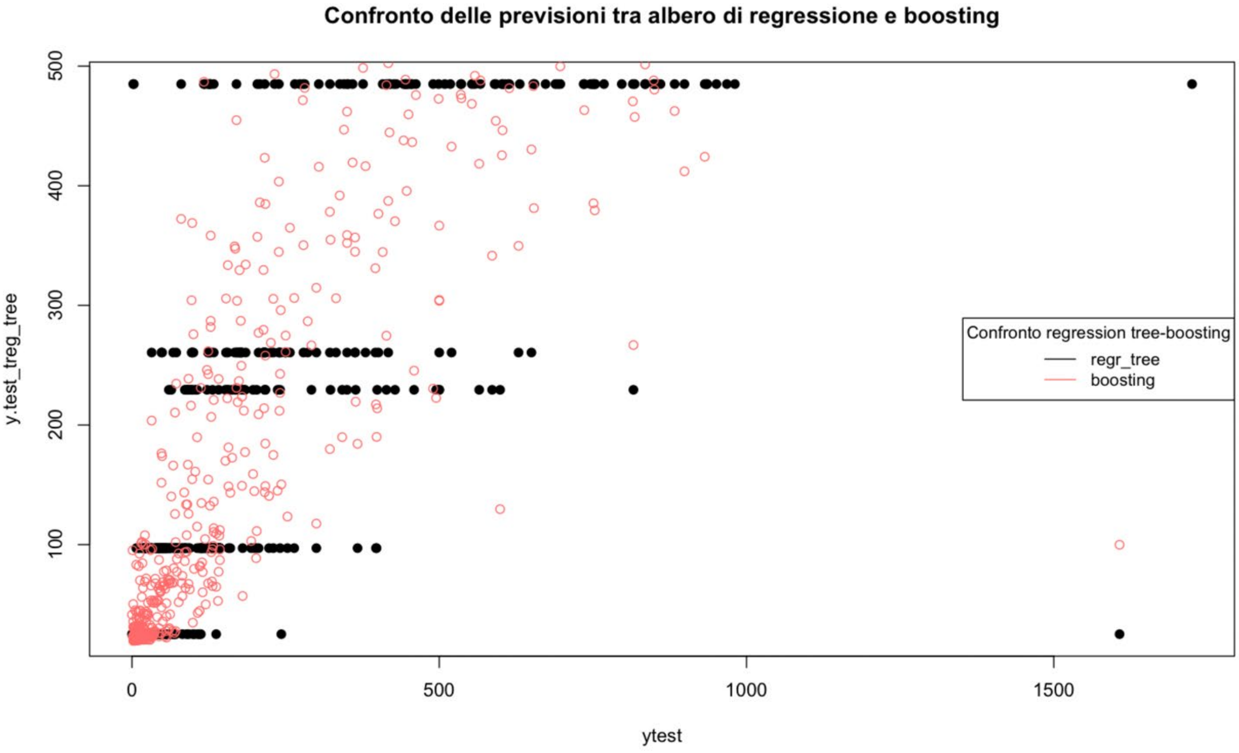

Confronto delle previsioni tra albero di regressione e boosting
y.test_treg_tree
100
200
300
400
500
0
500
1000
1500
ytest
Confronto regression tree-boosting
regr_tree
boosting

Figura 33