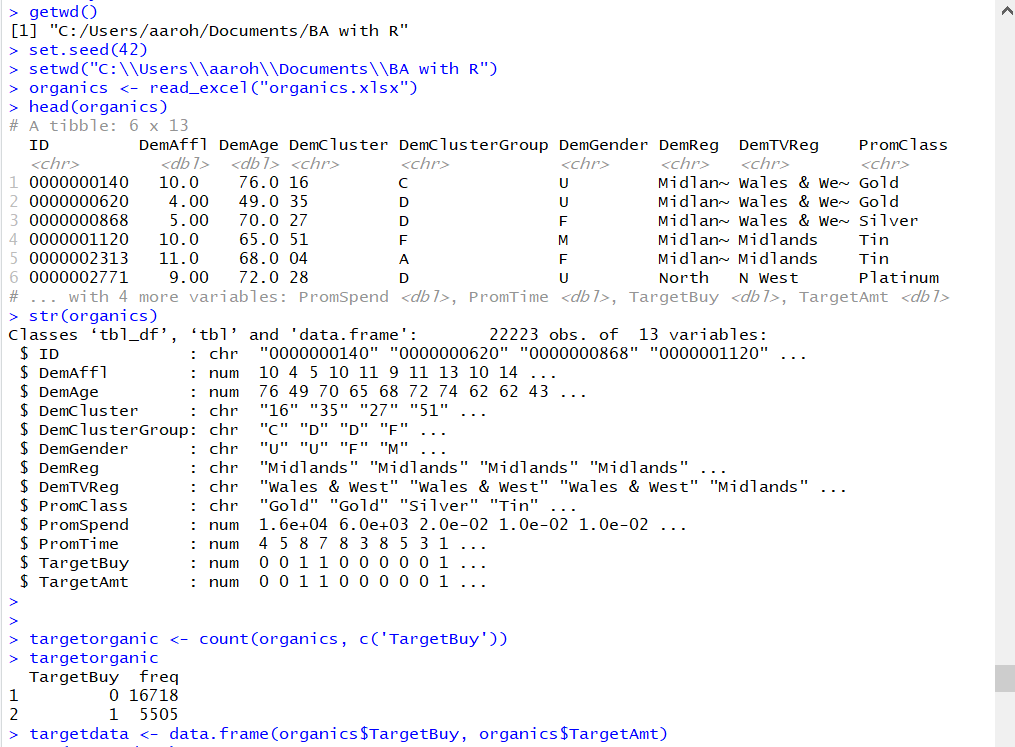
**Classification Techniques**

**Customer Classification**

A supermarket is offering a new line of organic products. The supermarket's management wants to determine which customers are likely to purchase these products. The supermarket has a customer loyalty program. As an initial buyer incentive plan, the supermarket provided coupons for the organic products to all of the loyalty program participants and collected data that includes whether these customers purchased any of the organic products.

The ORGANICS data set contains 13 variables and over 22,000 observations. The variables in the data set are shown below with the appropriate roles and levels:

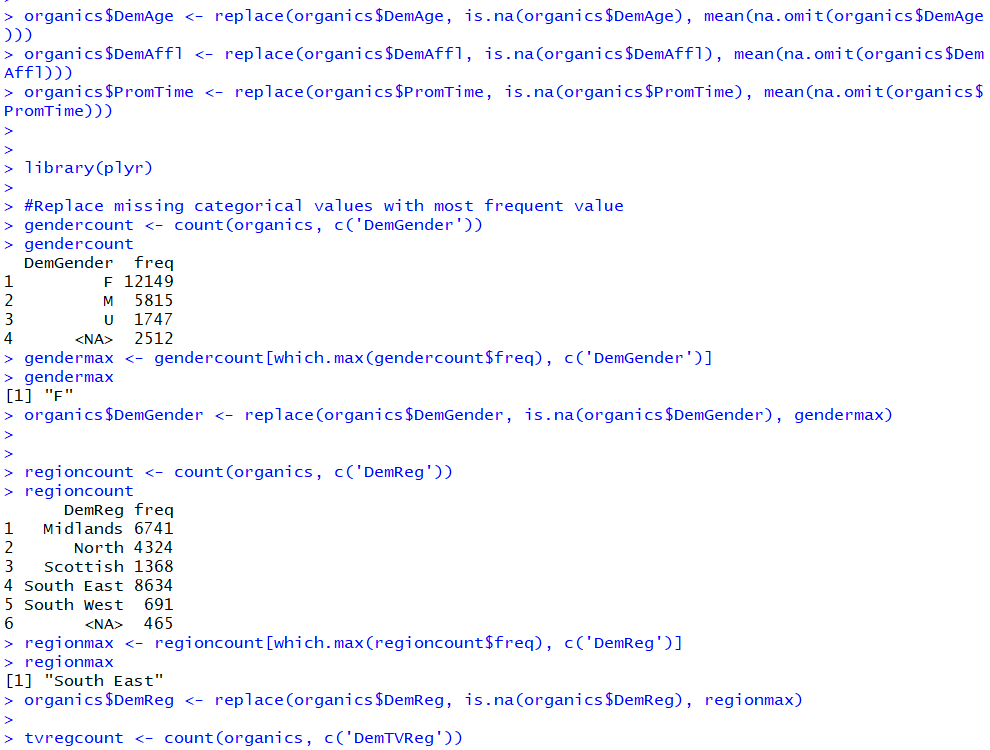
|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Model Role** | **Data Type** | **Description** |
| ID | ID | Categoric | Customer loyalty identification number |
| DemAffl | Input | Numeric | Affluence grade on a scale from 1 to 30 |
| DemAge | Input | Numeric | Age, in years |
| DemCluster | Rejected | Categoric | Type of residential neighborhood |
| DemClusterGroup | Input | Categoric | Neighborhood group |
| DemGender | Input | Categoric | M = male, F = female, U = unknown |
| DemRegion | Input | Categoric | Geographic region |
| DemTVReg | Input | Categoric | Television region |
| PromClass | Input | Categoric | Loyalty status: tin, silver, gold, or platinum |
| PromSpend | Input | Numeric | Total amount spent |
| PromTime | Input | Numeric | Time as loyalty card member |
| TargetBuy | Target | Numeric | Organics purchased? 1 = Yes, 0 = No |
| TargetAmt | Rejected | Numeric | Number of organic products purchased |
| Although two target variables are listed, these exercises concentrate on the binary target variable TargetBuy. | | | |

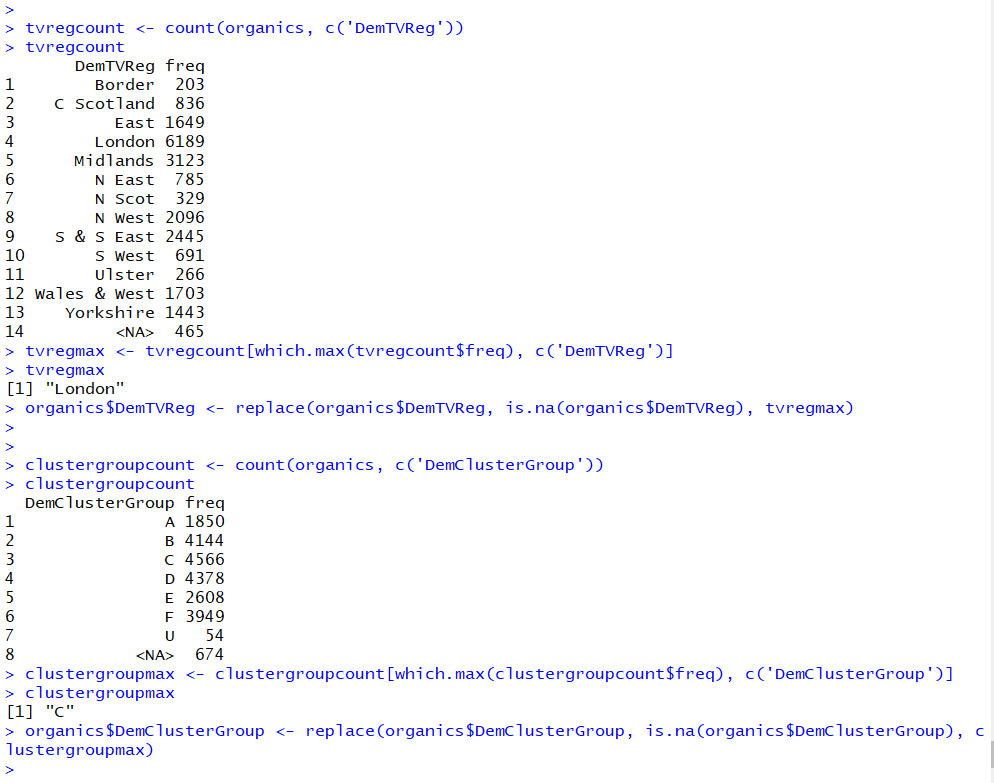


Missing values are checked for all the columns

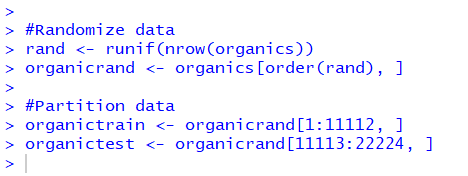
There are 2 types of variables that we are dealing here: continuous and categorical.

For missing categorical values, we have replace it with the maximum frequency value in the column and for the missing continuous variable we have replaced it with the mean of the column.

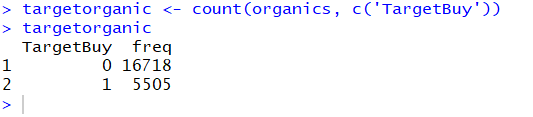




Once all the missing values are imputed, the dataset is randomized and partitioned equally in training and test datasets.

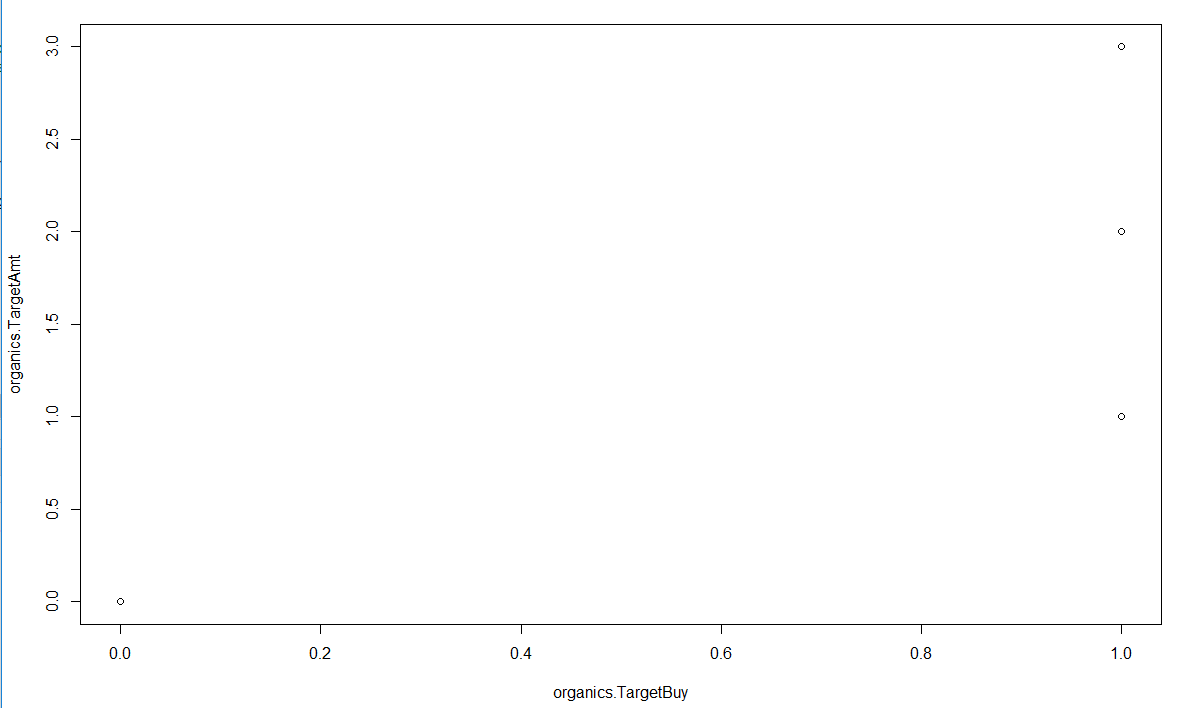


The target variable is a binary variable with values 0 and 1.

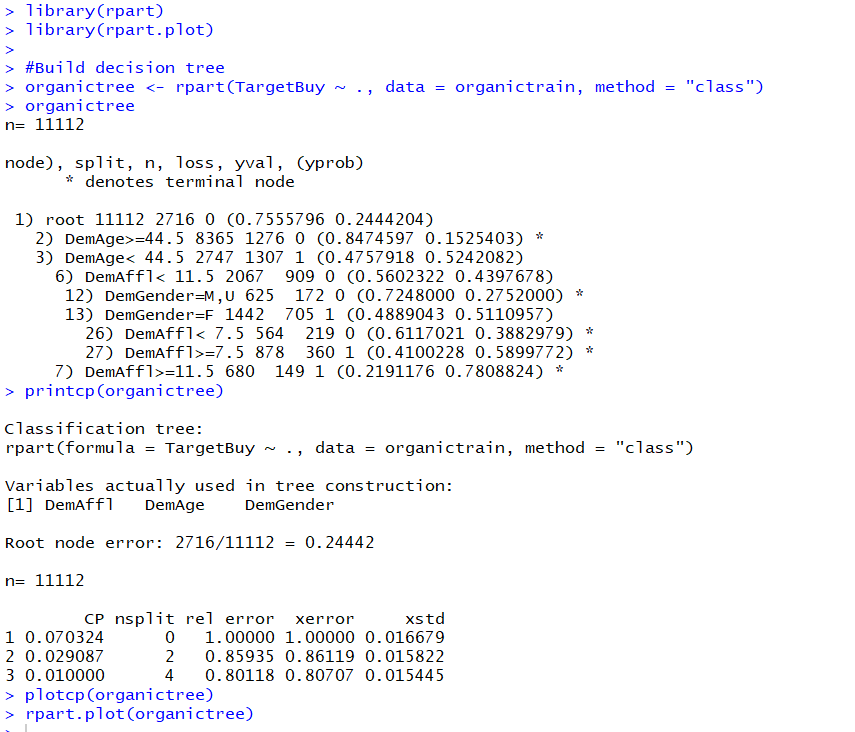


From the above we observe that approximately 1/4th of the total count purchased organic products.

No, TargetAmt cannot be used as an input for a model used to predict TargetBuy since the independent and dependent variables need to have a linear relationship between them. However, both the variables in discussion here do not show linear relationship.



“TargetBuy” status



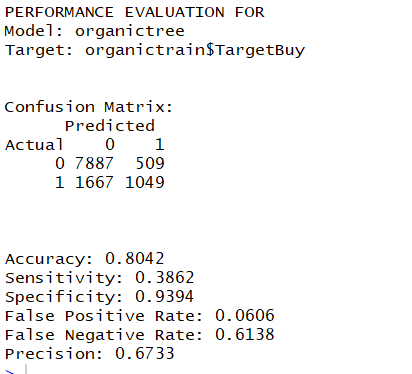
5 leaves are observed and there are no pure nodes.

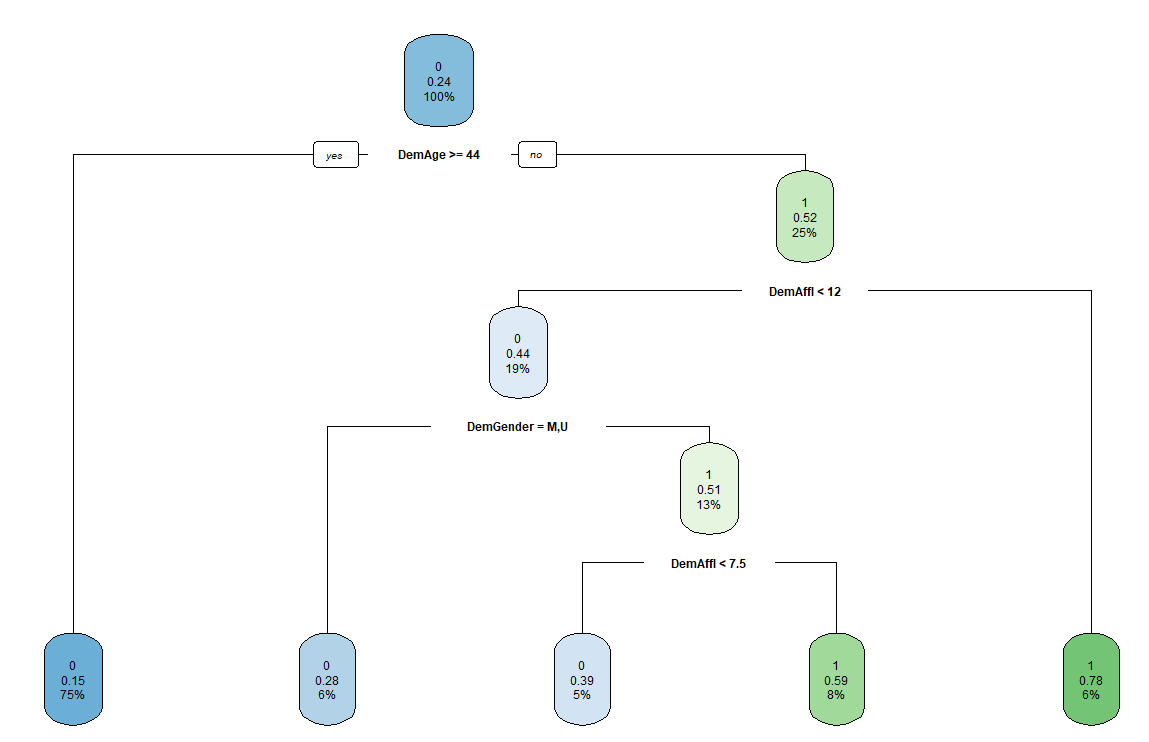
DemAge is the first variable used for split

A function is created named testModelPerformance which draws the performance matrix for the training dataset along with showing various factors like accuracy, sensitivity, specificity.

From the below results we see that 7887 zeroes were predicted correctly and 1049 1’s were predicted correctly. We see 509 0’s were predicted as 1’s and 1667 1’s were predicted as 0’s. We observe from the stats below the accuracy of the model is 80.42%

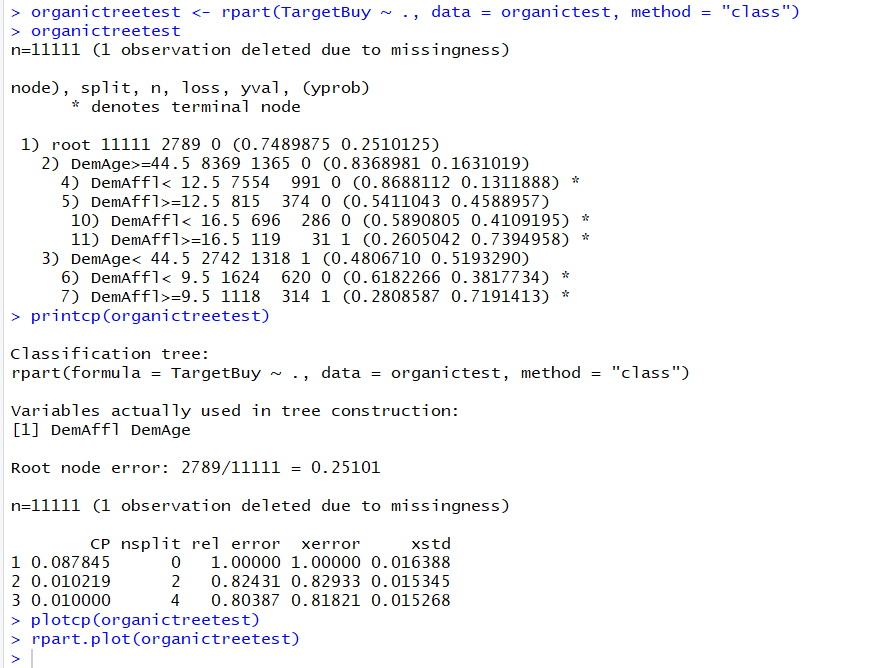


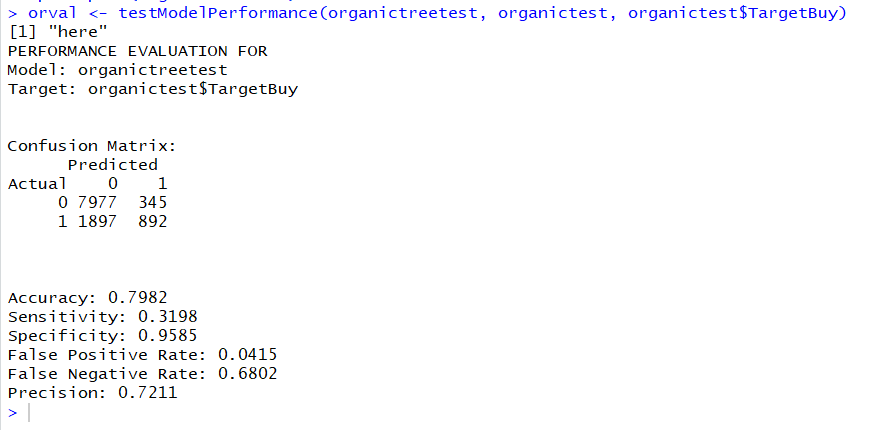


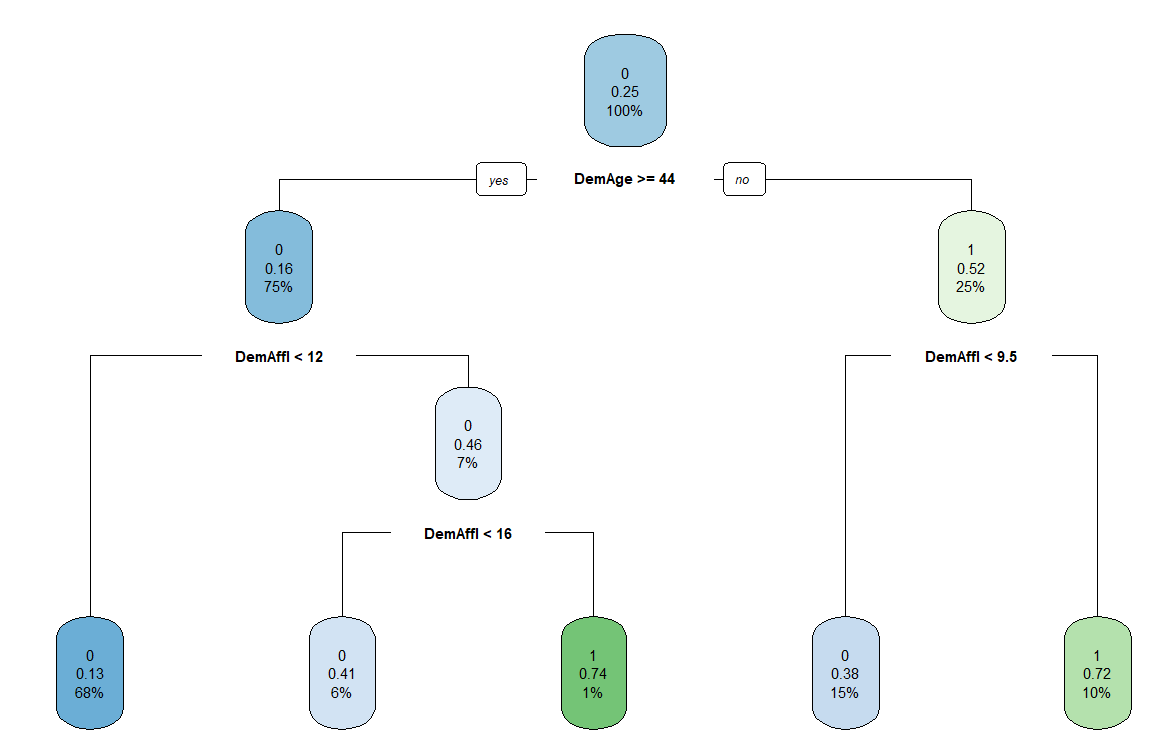


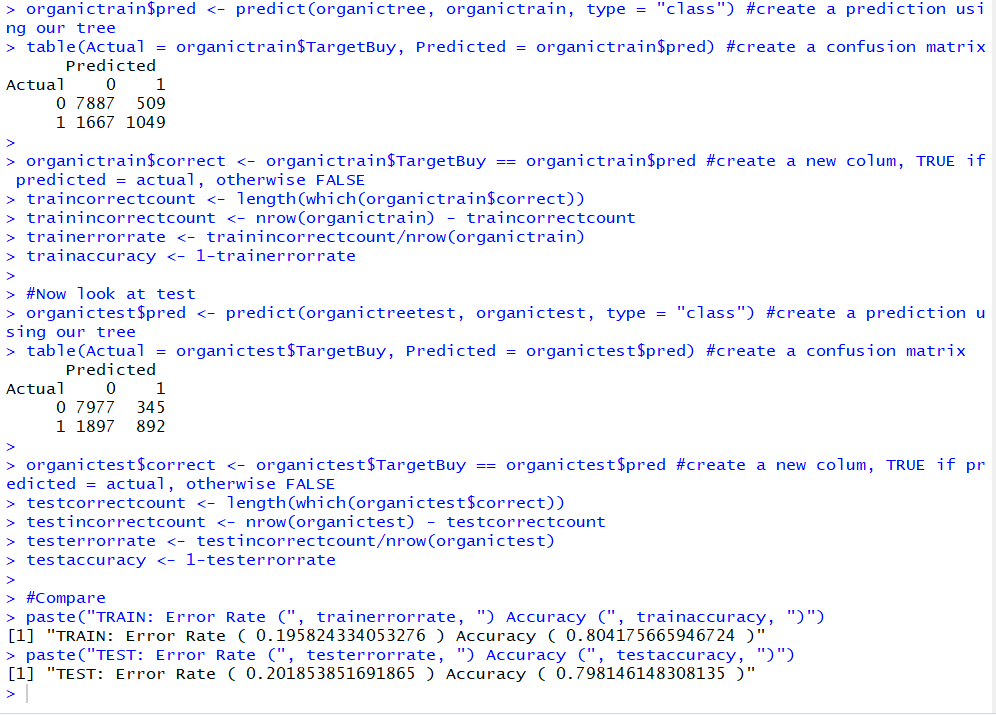
We see that the first split is done on the variable DemAge similar to that of the training dataset.

From the below results we see that 7977 zeroes were predicted correctly and 892 1’s were predicted correctly. We see 345 0’s were predicted as 1’s and 1897 1’s were predicted as 0’s. We observe from the stats below the accuracy of the model is 79.82%



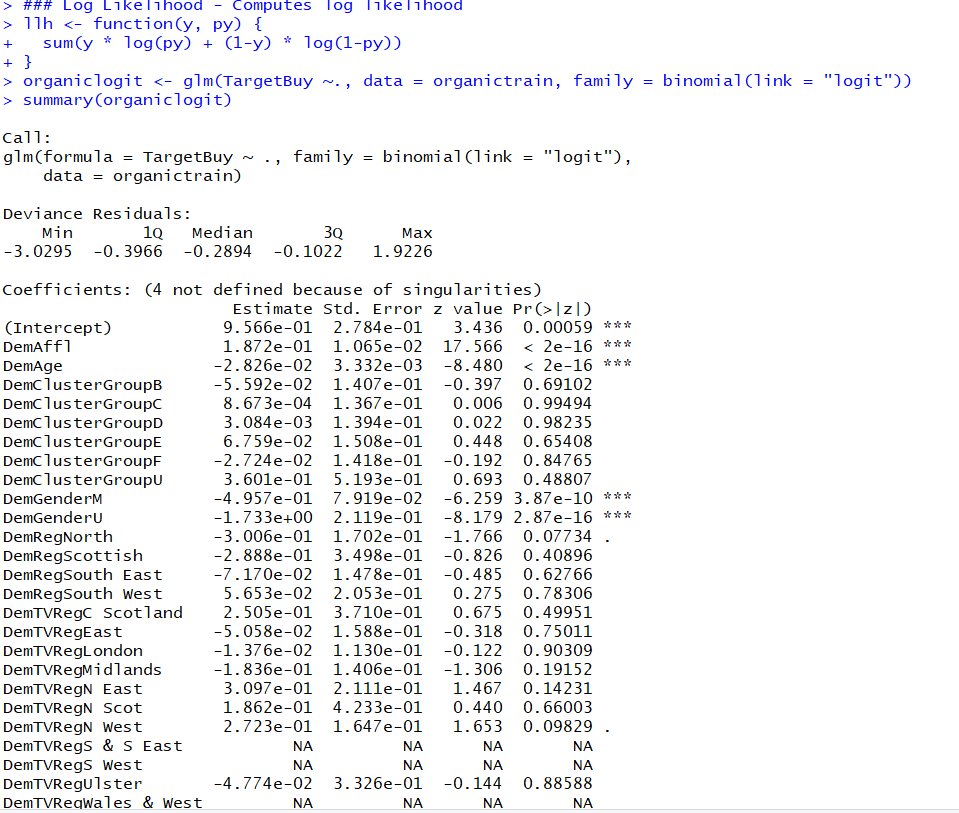




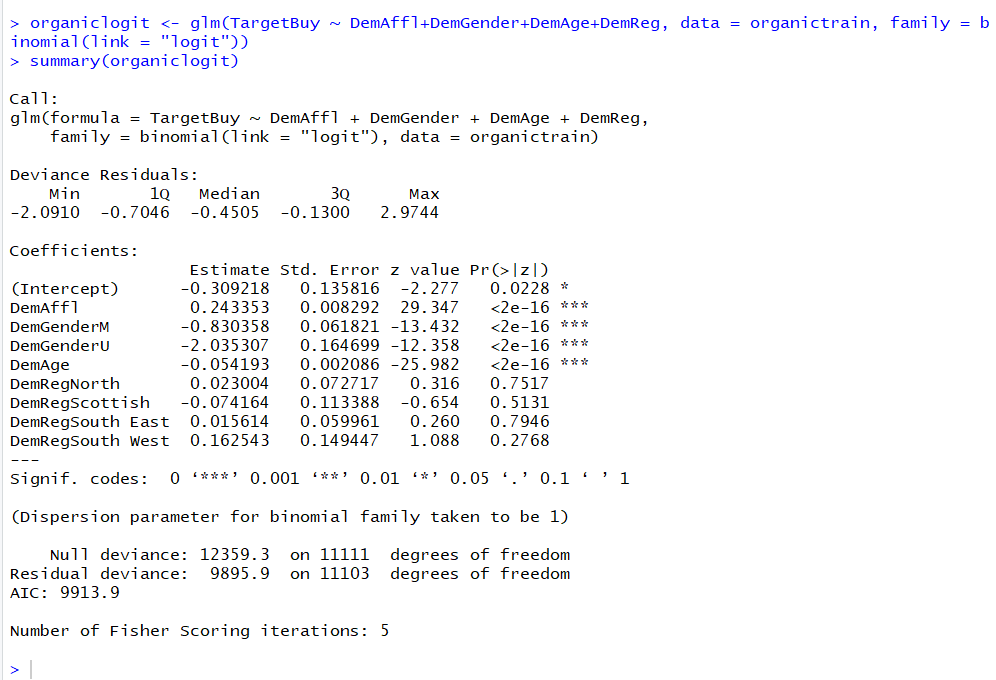


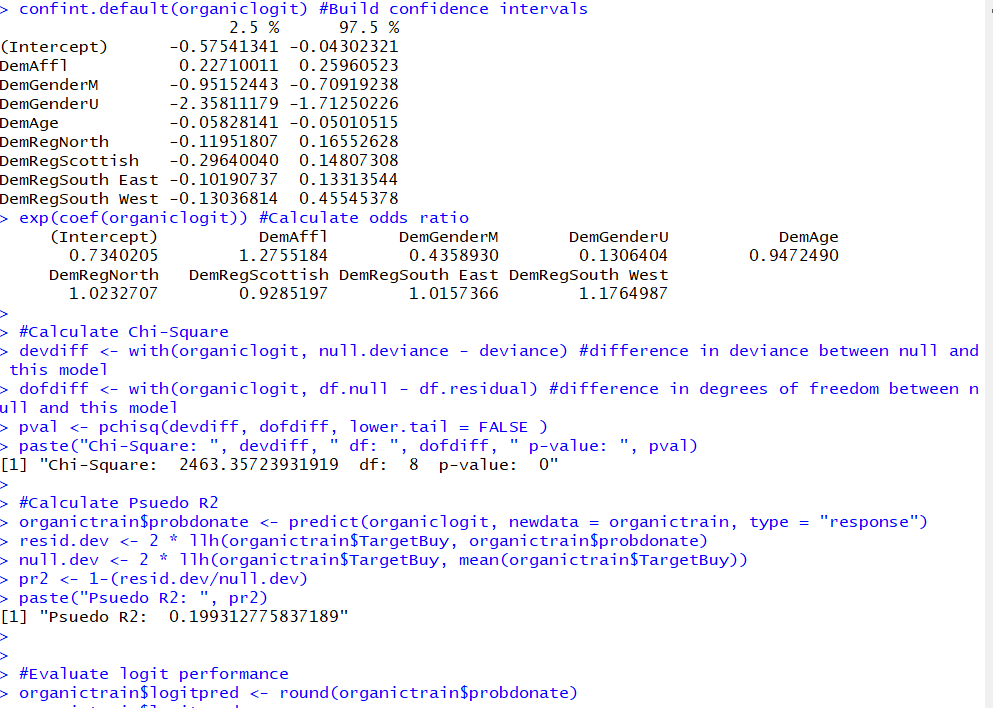
From the above confusion matrix for the training dataset we see that 7887 zeroes were predicted correctly and 1049 1’s were predicted correctly. We see 509 0’s were predicted as 1’s and 1667 1’s were predicted as 0’s. We observe from the stats below the accuracy of the model is 80.42% while the error rate was 19.5%

For the test dataset, we see that 7977 zeroes were predicted correctly and 892 1’s were predicted correctly. We see 345 0’s were predicted as 1’s and 1897 1’s were predicted as 0’s. We observe from the stats below the accuracy of the model is 79.82% error rate being 20.18%

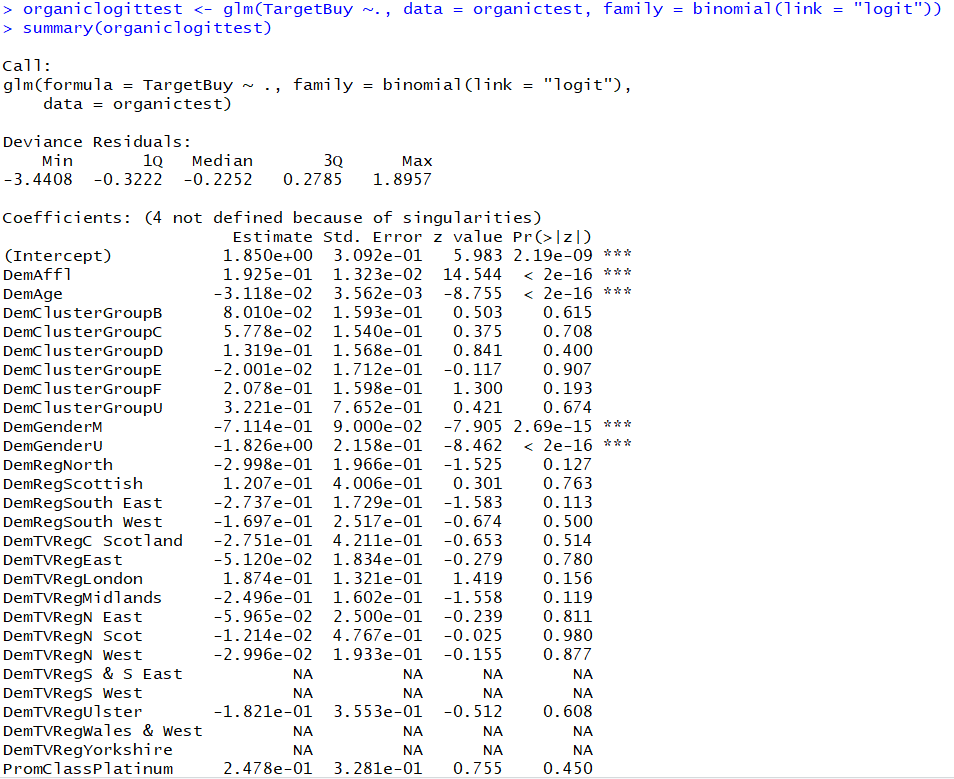


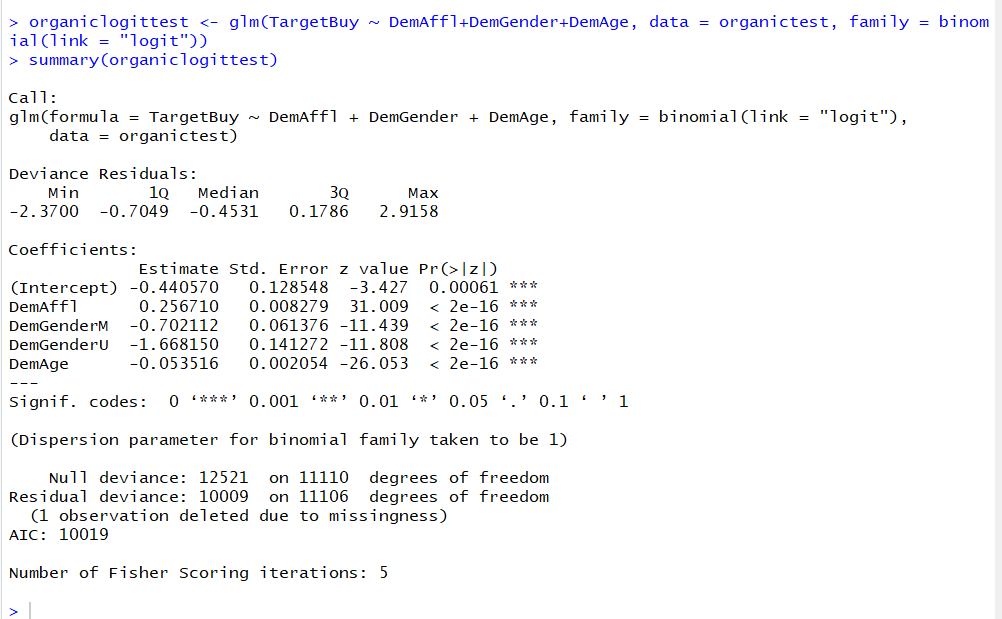
We see that the variables DemAffl, DemAge and DemGender are significant, hence we run a logistic regression using these variables.

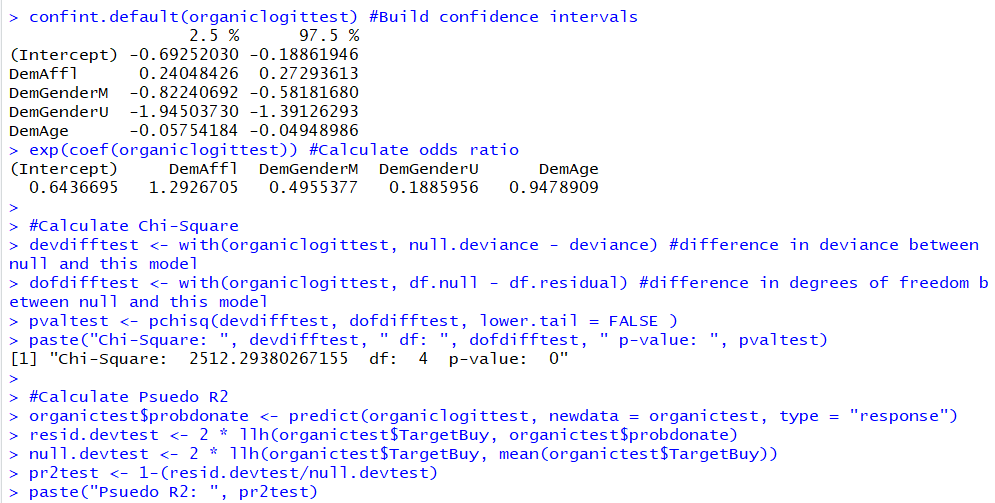




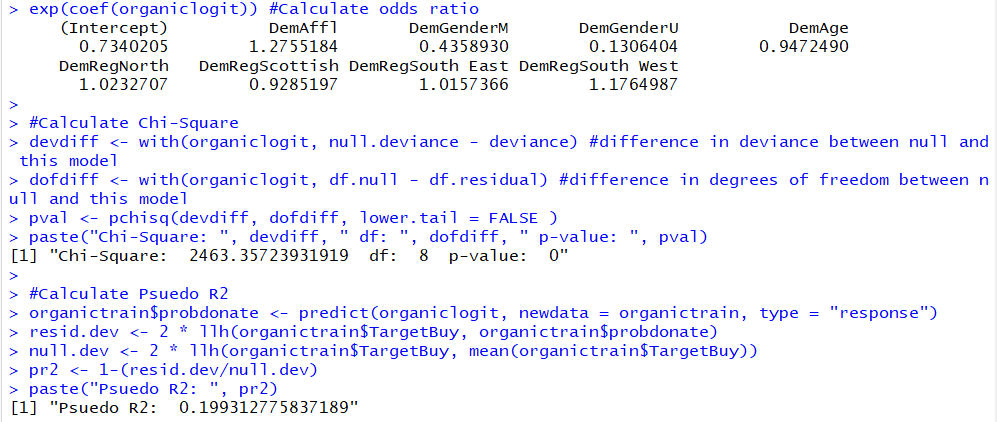
Below is logistic regression on test data. Here we see that variables DemAffl, DemAge and DemGender are significant so we run another logistic regression using these variables



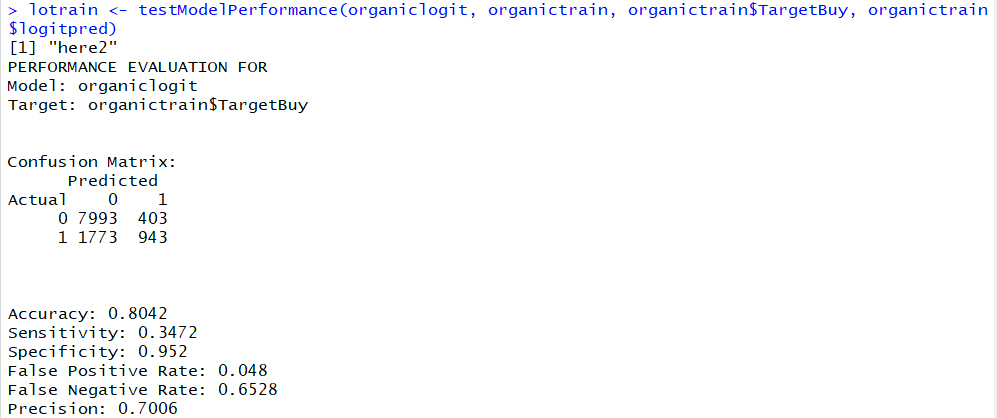




Training Dataset

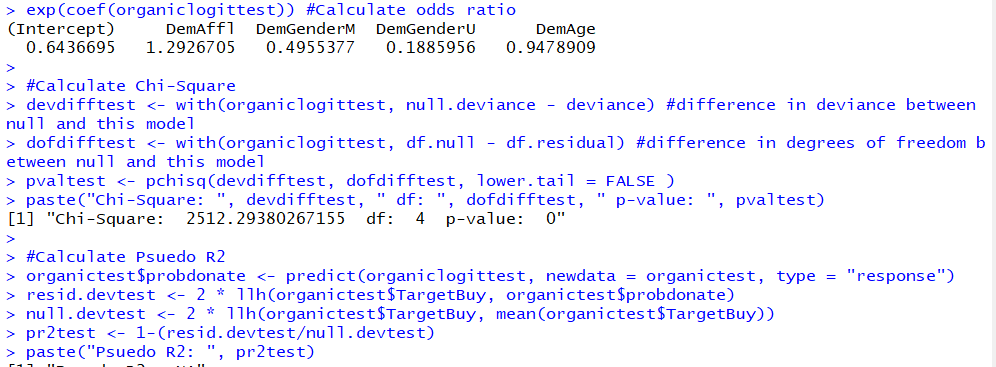


We observe that for the training dataset the Chi-Square is 2463.357 and the pseudo R-squared is 0.199

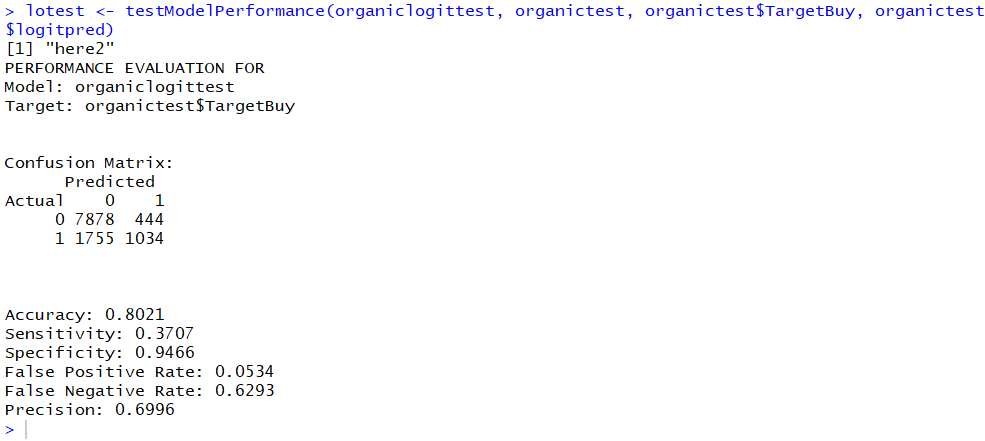


From the above confusion matrix for the training dataset we see that 7993 zeroes were predicted correctly and 943 1’s were predicted correctly. We see 403 0’s were predicted as 1’s and 1773 1’s were predicted as 0’s. We observe from the stats below the accuracy of the model is 80.42% and sensitivity 34.72%

Testing Data



The Chi-square for the test dataset is 2512.293



From the above confusion matrix for the test dataset we see that 7878 zeroes were predicted correctly and 1034 1’s were predicted correctly. We see 444 0’s were predicted as 1’s and 1755 1’s were predicted as 0’s. We observe from the stats below the accuracy of the model is 80.21%

and sensitivity 37.07%

For the decision tree approach, we see that for the training dataset the accuracy is 80.42% and sensitivity is 38.62% while for the testing dataset accuracy being 79.82% and sensitivity is 31.98%.

When we see the logistic regression approach for classification, for training dataset the accuracy is 80.42% and sensitivity is 34.72% and for test dataset accuracy is 80.21% and sensitivity is 37.07%. Thus, it is observed that the logistics regression classification is better than decision tree approach since the accuracy and sensitivity is better and gives accurate results for testing dataset.