* **Problem Introduction:**

This project is to develop a profile of the typical customer that purchases organic products to help the supermarket decide which customers are most likely to purchase these products and explore whether customers that purchase organic products spend more (or less) on average than other customers so that the supermarket can judge whether the additional cost of stocking organic products more palatable for the management.

* **Final Model:**

My final model is a decision tree shown below:

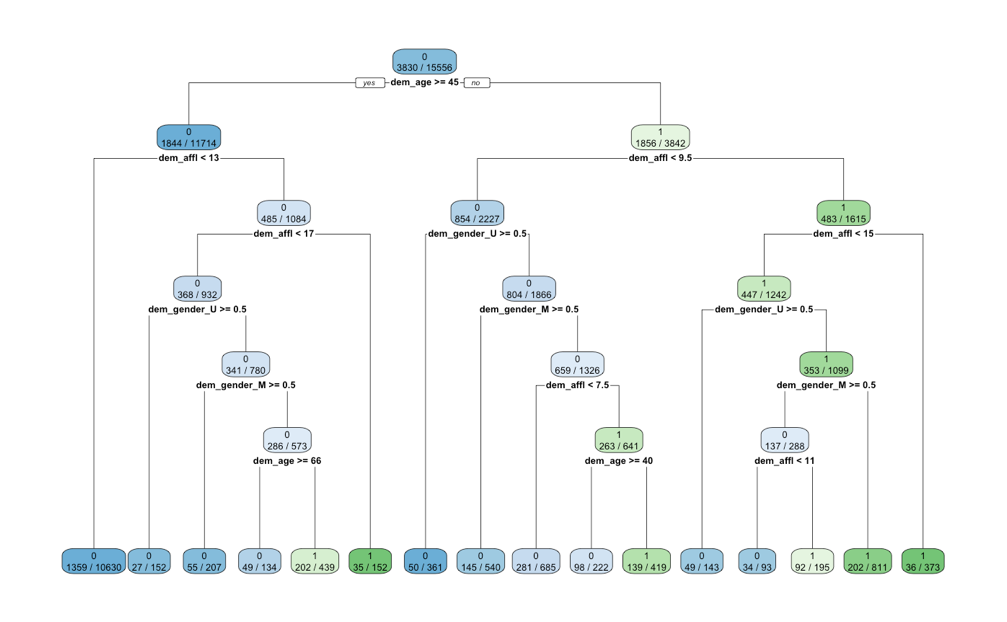


Figure 1 Final Model

The green nodes mean that most customers of that cluster have bought organic products while the blue nodes mean that most customers of that cluster haven’t bought organic products. The deeper the color is, the more proportion of that cluster has customers that have (or have not) bought organic products.

For example, we can conclude from this model that most of the customers whose age are less than 45 with affluence grade >= 15 have bought organic products while most of the customers whose age are greater than or equal to 45 with affluence grade less than 13 haven’t bought organic products before.

* **Organic Products Customer Characteristics**

Base on the final model, we can find that age, affluence grade and gender are the three most important factors that can affect customers’ choice of organic products.

Among all the customers, here are 6 types of customers who are most likely to buy organic products:

Table 1 Organic Products Customer Characteristics

|  |  |  |  |
| --- | --- | --- | --- |
|  | Age | Affluence Grade | Gender |
| 1 | A<45 | AG>=15 | / |
| 2 | A<45 | 9.5<=AG<15 | Female |
| 3 | A<45 | 9.5<AG<11 | Male |
| 4 | A<45 | 11<=AG<15 | Female |
| 5 | A<40 | 7.5<=AG<9.5 | Female |
| 6 | 45<=A<66 | 13<=AG<17 | Female |

Take #6 as an example, we can see that females who have ages between 45 and 66, affluence grades between 13 and 17 are likely to buy organic products.

* **Recommendations**

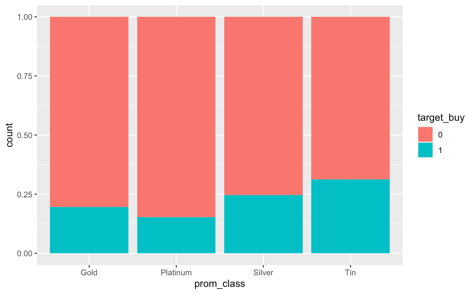
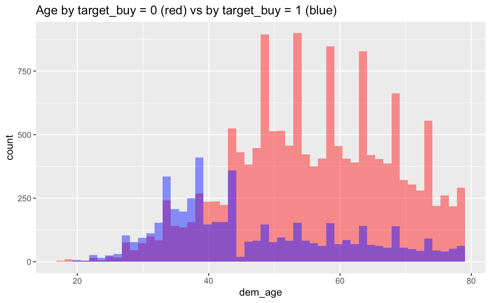
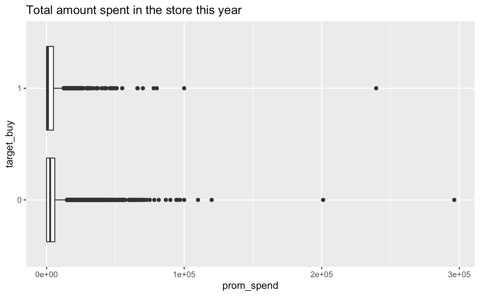


Figure 2.1 Amount Spent vs. Organics Purchased Figure 2.2 Age vs. Organics Purchased Figure 2.3 Loyalty class vs. Organics Purchased

- Customers that purchase organic products spent less than those who don’t purchase organic products on average. Thus, I suggest that the supermarket shouldn’t introduce so many kinds of organic products to reduce the additional cost of stocking organic products since organic consumers don't have much purchasing power.

- According to the final model, age is the most important factor affecting whether or not to buy organic products. Young people are more likely to be the target buyers for organic products. Thus, we can launch some activities targeted at young users to increase their purchasing power whereas launching some preferential or trial purchase activities to make more middle-aged and elderly people accept and willing to buy organic products.

- People in different loyalty statuses have different preferences. The willingness of buying organic products for each loyalty group has the sequence of Tin>Silver>Gold>Platinum. Therefore, we can advertise a large number of organic products to the users of the Tin group to meet their interests and to make them choose this supermarket and increase loyalty to this supermarket.

- The supermarket should target users based on the criteria of age, gender and affluence grade. Based on the user portraits we generated, we can further analyze the preferences of our target users and promote some products they are interested in, thus increasing their purchasing power.

* **Technical Appendix**

Explore data:

-The skim result shows that the complete rate for each factor is larger than 0.8.

- boxplot, histogram, barplot for target/numeric/characters.

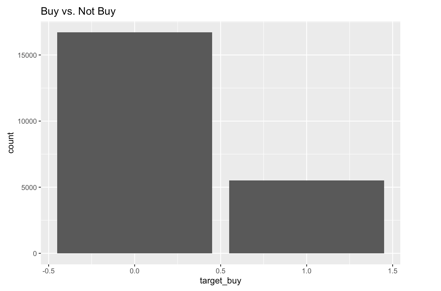
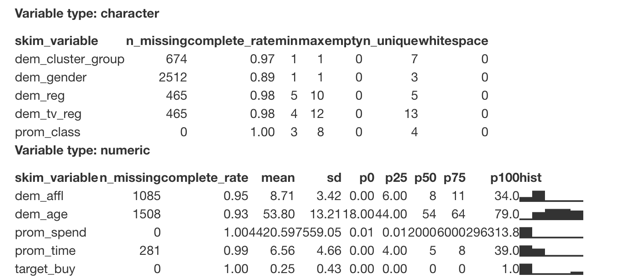


Figure 3 skim result Figure 4 Target Exploration

Make factors, Data set split

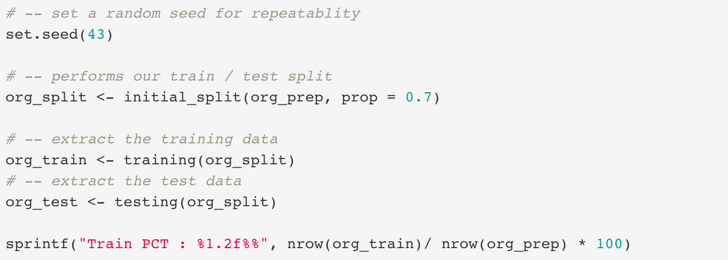
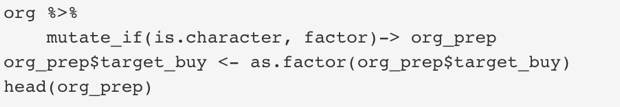


Figure 5 Make Factors Figure 6 Data Set Split

Generate Models:

- 4 models: default decision tree; decision tree with the best complexity; logistic full model; logistic reduced model.

- Best complexity is 0.0011.

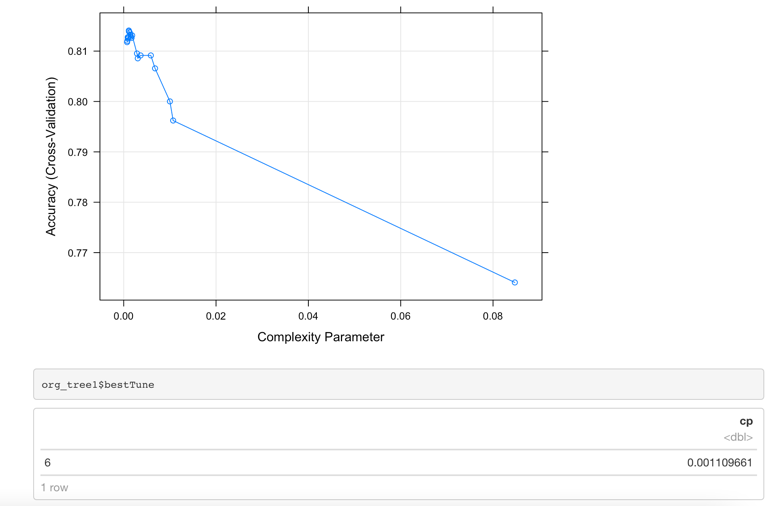
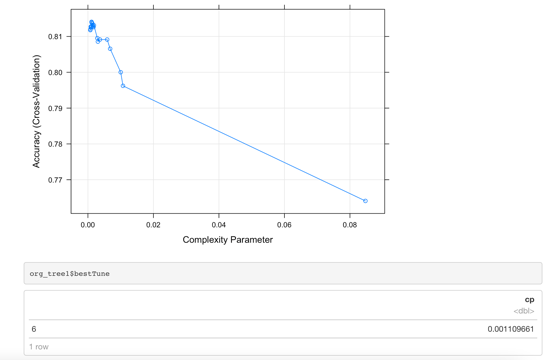


Figure 7.1 Complexity Plot Figure 7.2 Best Complexity

- Models are shown below.

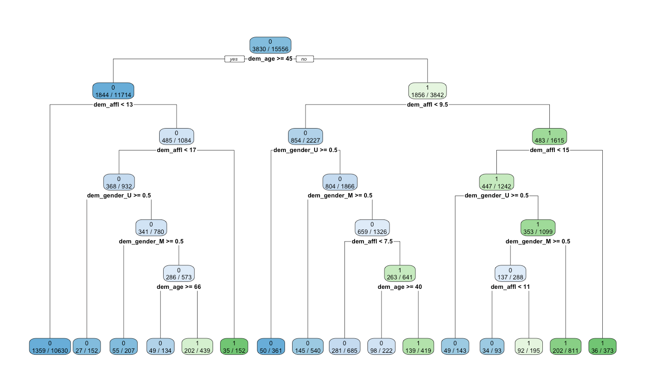
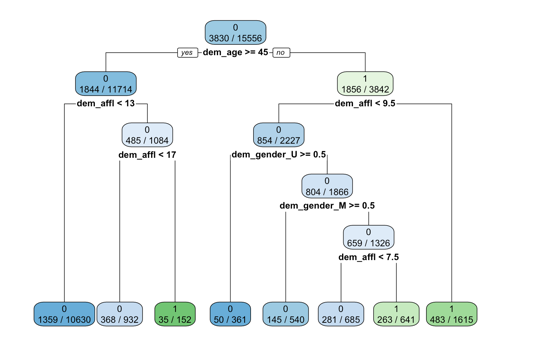


Figure 8.1 Default Decision Tree Figure 8.2 Best Complexity Decision Tree

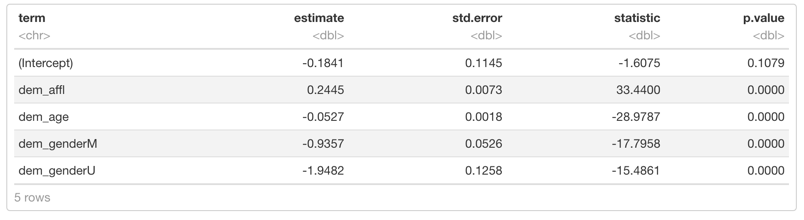
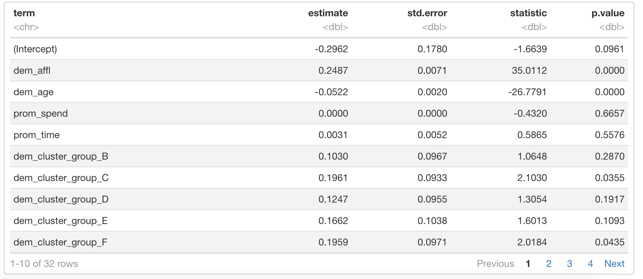


Figure 9.1 Full Logistic Model Figure 9.2 Reduced Logistic Model

Models comparison:

Table 2 Models Comparison

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy(train) | AUC(train) | Accuracy(test) | AUC(test) |
| Default Decision Tree | 0.808 | 0.746 | 0.807 | 0.751 |
| Decision Tree (cp=0.0011) | 0.817 | 0.753 | 0.812 | 0.758 |
| Full Logistic | 0.810 | 0.802 | 0.807 | 0.809 |
| Reduced Logistic | 0.796 | 0.793 | 0.791 | 0.800 |

The second decision tree is the best model since it has the largest accuracy for both train and test sets. And the accuracy difference between train and test is 0.005, which is acceptable.

Fit in the new chart:  
- input chart, subset, bake and predict

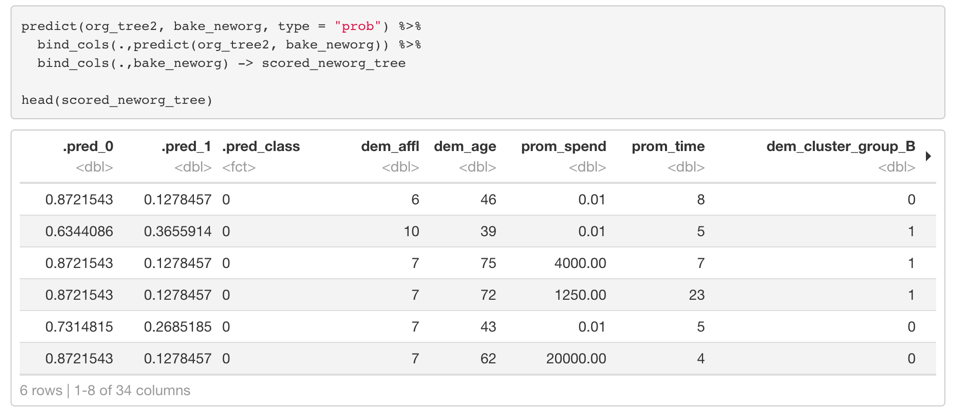


Figure 10 Prediction for New Model