Project 2: Organics

Ziwan Peng

1. Introduction to the problem

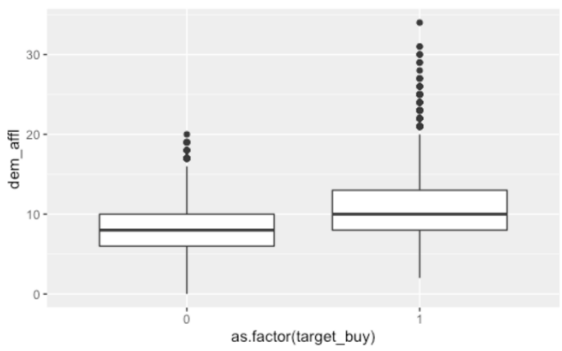
Along with offering a new line of organic products, the supermarket cares about the probability of customers to purchase these products. Based on the information collected from customer loyalty program, the management team will determine a model to predict which customers are most likely to purchase and develop a profile of the typical customer who purchases organic products. The model and profile will direct the supermarket how to implement the next steps so that to increase the sales of organic products.

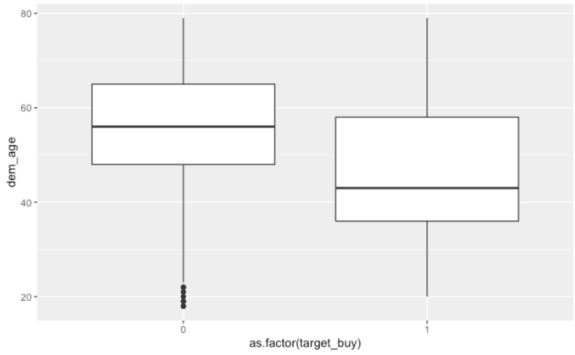
2. Explore the data

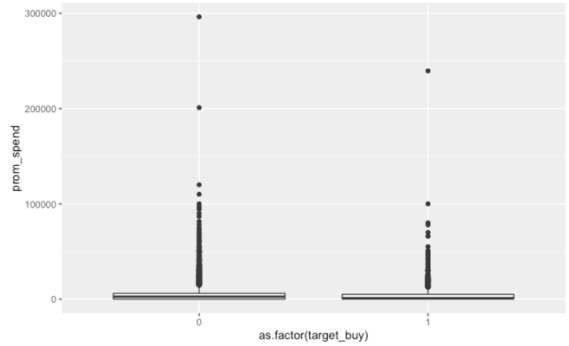
1) The data set contains 13 variables and over 22,000 observations. The variables that matter to our problem are as shown below:

Name	Model Role	Measurement Level	Description
dem_affl	Input	Interval	Affluence grade on a scale from 1 to 30
dem_age	Input	Interval	Age, in years
dem_cluster_group	Input	Nominal	Neighborhood group
dem_gender	Input	Nominal	M = male, F = female, U = unknown
dem_reg	Input	Nominal	Geographic region
dem_tv_reg	Input	Nominal	Television region
prom_class	Input	Nominal	Loyalty status: tin, silver, gold, or platinum
prom_spend	Input	Interval	Total amount spent in the store this year
prom_time	Input	Interval	Time as loyalty card member
target_buy	Target	Binary	Organics purchased? 1 = Yes, 0 = No

2) Explore the numeric variables

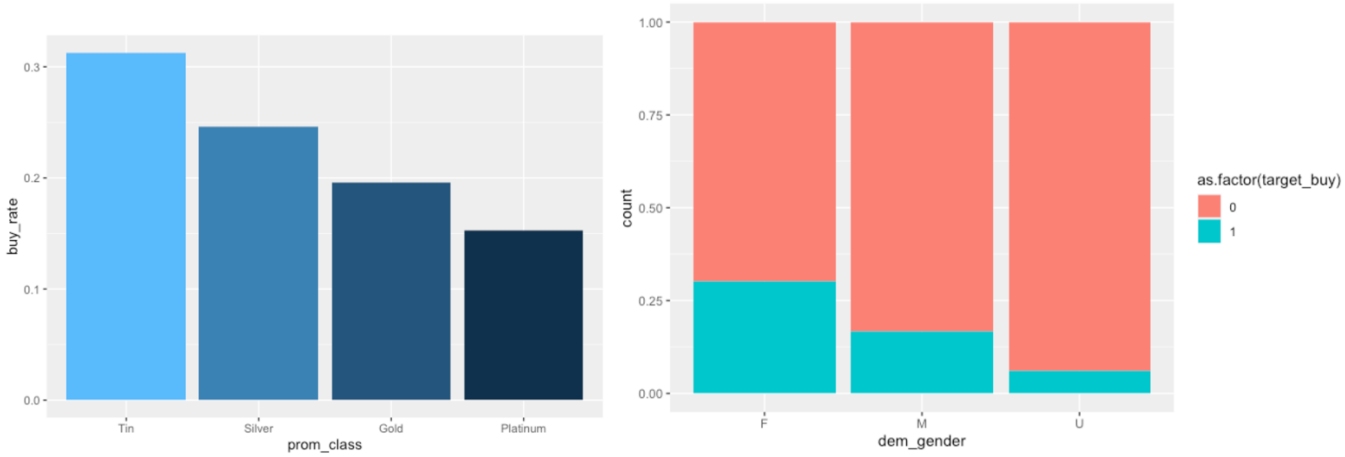






From the boxplots, it can be observed that people with higher affluence grade and younger age are more likely to purchase organic products. But there's no significant evidence showing that the people buying organic products would spend more in the store.

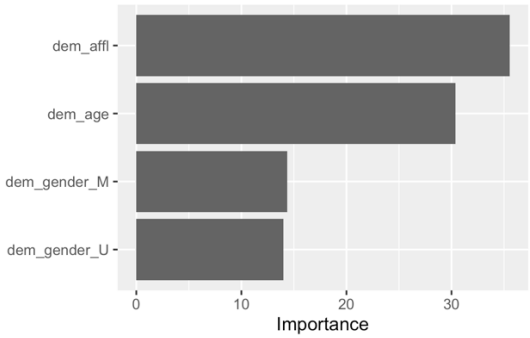
3) Explore the nominal variables



There's a descending trend in organic products buying rate with higher loyalty status. In other words, the customer is more loyalty to the supermarket, the less likely he/she would purchase organic products. When it comes to gender, we can see that the female group has the highest purchasing will in organic products and the unknown gender group (the people don't want to identify themselves) has the lowest rate in organic products purchasing.

3. Final model: logistic regression

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	-0.3850	0.1096	-3.5134	4e-04
dem_affl	0.2475	0.0070	35.5275	0e+00
dem_age	-0.0531	0.0017	-30.3653	0e+00
dem_gender_M	-0.7459	0.0519	-14.3779	0e+00
dem_gender_U	-1.7585	0.1256	-14.0019	0e+00



The final model includes significant variables affluence, age, gender_male and gender_unknown. With higher affluence grade and younger age, the probability of purchasing organic products will increase. In the meantime, the customer's gender defined as male and unknown would have a negative impact on organic products purchases. From the importance graph of variables, the affluence grade is the most important factor that affects the customer's behaviors on organic products.

.metric <chr>	.estimator <chr>	.estimate <dbl>	part <chr>
accuracy	binary	0.8034842	training
accuracy	binary	0.8012599	testing
roc_auc	binary	0.7846957	training
roc_auc	binary	0.7892176	testing

The final model performed the best in the practices with 0.8 accuracy rate and 0.78 roc_auc score (the higher, the better), which means it's a reliable model to determine which customers are most likely to purchase these products. Both high performance in train set and test set would benefit in generalizing the model to predict the probability of new customers purchasing organic products.
Therefore, based on our final model, the profile of typical organic products customer is female with high affluence grade and young age.

4. Recommendations

1) Instead of focusing on increasing the customer loyalty, the supermarket team should develop more affluent customers to increase the sales of organic products because the affluence grade contributes most to the purchasing.
2) Age does matters. The supermarket team can implement ads like organics benefits sports to attract the youth.

3) In the following promotion activities, the supermarket can use female-friendly strategies like pretty packaging to attract female customers since they are prone to purchasing organic products.

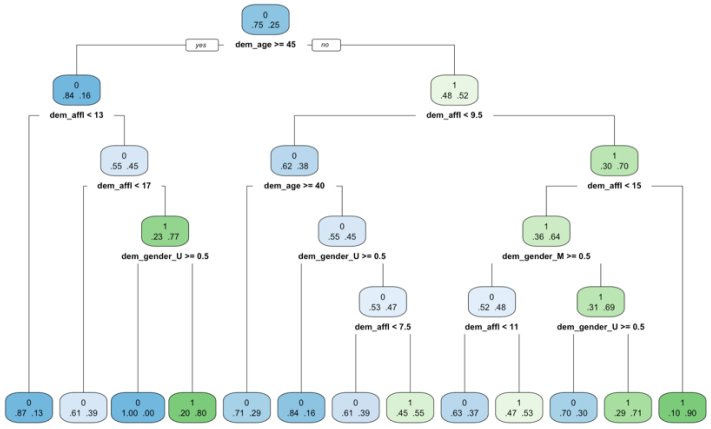
5. Technical appendix

I employed three models: logistic regression and decision tree model with engine "rpart" and "C5.0".
For logistic regression: I first performed the full model and then used stepwise method to remove insignificant variables. Here is the final logistic regression model and it's performance:

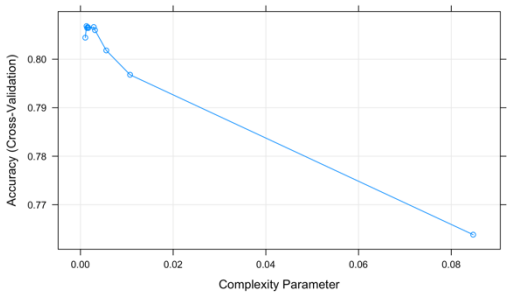
term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	-0.3850	0.1096	-3.5134	4e-04
dem_affl	0.2475	0.0070	35.5275	0e+00
dem_age	-0.0531	0.0017	-30.3653	0e+00
dem_gender_M	-0.7459	0.0519	-14.3779	0e+00
dem_gender_U	-1.7585	0.1256	-14.0019	0e+00

.metric	.estimator	.estimate	part
<chr>	<chr>	<dbl>	<chr>
accuracy	binary	0.8034842	training
accuracy	binary	0.8012599	testing
roc_auc	binary	0.7846957	training
roc_auc	binary	0.7892176	testing

For decision tree with engine "rpart": I first performed the model and tuned the cost complexity to find the best model. Here is the final decision tree model with engine rpart and the most appropriate cost complexity 0.0013.

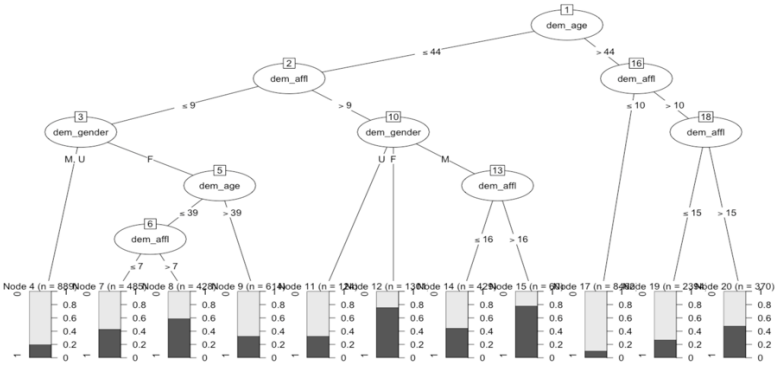


.metric	.estimator	.estimate	part
<chr>	<chr>	<dbl>	<chr>
accuracy	binary	0.8082412	training
roc_auc	binary	0.7474122	training
accuracy	binary	0.8042598	testing
roc_auc	binary	0.7504702	testing



	cp
	<dbl>
2	0.001305483

For decision tree with engine "C5.0": Here is the final decision tree model with engine C5.0.



.metric	.estimator	.estimate	part
<chr>	<chr>	<dbl>	<chr>
accuracy	binary	0.8109411	training
roc_auc	binary	0.7646722	training
accuracy	binary	0.8054597	testing
roc_auc	binary	0.7664385	testing

Comparing with the two decision tree models, the final logistic regression model has the highest roc_auc score with close accuracy rate. Therefore, I chose it as the final model to predict which customers are most likely to purchase organic products.