**Project 2**

**The Business Scenario**:

A supermarket is offering a new line of organic products. The supermarket’s management wants to determine which customers are most likely to purchase these products and develop a profile of the typical customer that purchases organic products.

As a side issue, they would also like to understand whether customers that purchase organic products spend more (or less) on average than other customers. If they find that customers who purchase organic products are also highly profitable customers, that makes the additional cost of stocking organic products more palatable for the management.

**The Data:**

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 (organics.csv).

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** | **Measurement Level** | **Description** |
| ID | ID | Nominal | Customer loyalty identification number |
| DemAffl | Input | Interval | Affluence grade on a scale from 1 to 30 |
| DemAge | Input | Interval | Age, in years |
| DemCluster | Rejected | Nominal | Type of residential neighborhood |
| DemClusterGroup | Input | Nominal | Neighborhood group |
| DemGender | Input | Nominal | M = male, F = female, U = unknown |
| DemRegion | Input | Nominal | Geographic region |
| DemTVReg | Input | Nominal | Television region |
| PromClass | Input | Nominal | Loyalty status: tin, silver, gold, or platinum |
| PromSpend | Input | Interval | Total amount spent in the store this year |
| PromTime | Input | Interval | Time as loyalty card member |
| TargetBuy | Target | Binary | Organics purchased? 1 = Yes, 0 = No |
| TargetAmt | Rejected\*\* | Interval\*\* | Number of organic products purchased\*\*  Note that you could use this as an input for exploratory analysis – You just don’t need to consider it as a Target variable for this exercise (which is why it was rejected in the original analysis of these data) |

**Tasks and Questions**:

Your primary analytical task is to build a model (or models) that will enable the supermarket to identify customers who are most likely to purchase organic products and that will enable you to develop a profile of the “typical” organic products purchaser.

Note that there are some missing values in this data set, so be sure to have a strategy for dealing with those.

Some questions to consider as you explore and analyze the data:

1. Are there any noticeable differences between customers that purchase organic products versus those who don’t purchase organic products?
2. Continuing along a similar path, are there any noticeable differences in the percentage of customers who purchase organic products across the different loyalty status groups (for example, is the percentage of platinum customers who purchase organic products higher than the percentage of tin customers who purchase organic products)?
3. What factors seem to have the most impact on a customer’s likelihood to purchase organic products?

**Deliverables:**

**Report:**

Once you have completed your analysis, you should produce a report for the management of the supermarket. This report should be 1-3 pages in length and may include embedded tables and figures as needed. The audience for the report is non-technical. You can assume that they understand the business problem but they are not familiar with the data or with the analytical methods that you apply. The report should include:

* a brief introduction to the problem
* a brief summary of the final model (for a business audience)
* a short description of the typical organic products customer
* a set of recommendations that the supermarket could implement in order to increase the sales of organic products
* a file that contains the scores (probabilities) for the new customers (contained in the file new\_organics.csv)
* a technical appendix that outlines your modeling strategy and includes the detailed output for the model (or models) that you used to address the business questions above.