**Organic Products Case Study**

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

**Tasks and Questions**:

Your primary analytical task is to develop a set of models that will enable the supermarket to identify customers who are most likely to purchase organic products, identify key factors that increase the likelihood of purchasing organic products, and 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?
4. Can you identify certain groups of customers who are more likely to purchase organic products? How would you describe those customers?

**Modeling requirements:**

* You must create at least four different models (not 4 different methods of modeling, but maybe something like a default value and an elastic net type of classification) including at least one logistic regression and at least one classification tree. This is the minimum requirement, and so you may choose to explore more than four models during the course of your analysis. Your slides should include a table comparing the accuracy, precision, recall, and area under the ROC curve (for both the training and test data sets) for ALL models that you consider.

**Deliverables:**

**Annotated Power Point Presentation:**

Once you have completed your analysis, you should create a presentation directed towards the management of the supermarket. These managers are not technical and are primarily concerned with what this analysis means for the supermarket. While they need to be reassured that your analysis is sound, they do not need to understand the details of how you arrived at your final models. The presentation should include:

* a brief introduction to the business problem
* an overview of any findings from your exploratory analysis of the data
* a brief summary of your model(s) (for a business audience)
* a short description of the typical organic products customer(s)
* a set of recommendations that the supermarket could implement in order to increase the sales of organic products

**NOTE:** Your slides should meet all of the requirements for a good business presentation. In particular, they should be professional and appropriate for a business audience, and they should not be wordy. Every slide should have a notes section that details what you would say about the slide if you were presenting the slide to an audience. These notes should be written in complete sentences using proper grammar.

**Scored Data Set:**

* a file that contains the scores (probabilities) for the new customers (contained in the file new\_organics.csv) (Helpful notebook is the end of the Telco churn example notebook)

**Python Code and Results:**

* an html file containing your code and the results from your analyses