

Tayko Software Reseller

Tayko Software is a software catalog firm that sells games and educational software. It started out as a software manufacturer, and added third party titles to its offerings. It has recently put together a revised collection of items in a new catalog, which it is preparing to roll out in a mailing.

In addition to its own software titles, Tayko's customer list is a key asset. In an attempt to grow its customer base, it has recently joined a consortium of catalog firms that specialize in computer and software products.

The consortium affords members the opportunity to mail catalogs to names drawn from a pooled list of customers. Members supply their own customer lists to the pool, and can "withdraw" an equivalent number of names each quarter. Members are allowed to do predictive modeling on the records in the pool so they can do a better job of selecting names from the pool.

Tayko has supplied its customer list of 200,000 names to the pool, which totals over 5,000,000 names, so it is now entitled to draw 200,000 names for a mailing. Tayko would like to select the names that have the best chance of performing well, so it conducts a test – it draws 20,000 names from the pool and does a test mailing of the new catalog to them.

This mailing yielded 1065 purchasers – a response rate of 0.05325. To optimize the performance of the data mining techniques, it was decided to work with a stratified sample that contained equal numbers of purchasers and non-purchasers. For ease of presentation, the data set for this case includes just 1000 purchasers and 1000 non-purchasers, an apparent response rate of 0.5. Therefore, after using the data set to predict who will be a purchaser, we must adjust the purchase rate back down by multiplying each case's "probability of purchase" by $0.05325/0.5$ or **0.1065**.

There are two response variables in this case: "purch" indicates whether or not a prospect responded to the test mailing and purchased something, while "spend" indicates, for those who made a purchase, how much they spent. The overall procedure in this case will be to develop two models. One will be used to classify records as "purchase" or "no purchase." The other will be used for those cases that are classified as "purchase," and will predict the amount they will spend.

The following table provides a description of the variables available in this case. A partition variable is used because we will be developing two different models in this case and we want to preserve the same partition structure for assessing each model.

Codelist				
Var. #	Variable Name	Description	Variable Type	Code Description
1.	usa	Is it a US address?	binary	1: yes 0: no
2 - 16	s_*	Source catalog for the record (15 possible sources)	binary	1: yes 0: no
17.	freq	Number of transactions in last year at source catalog	numeric	
18.	last	How many days ago was last update to cust. record	numeric	
19.	first	How many days ago was 1st update to cust. record	numeric	
20.	web	Customer placed at least 1 order via web	binary	1: yes 0: no
21.	male	Customer is male	binary	1: yes 0: no
22.	resid	Address is a residence	binary	1: yes 0: no
23.	purch	Person made purchase in test mailing	binary	1: yes 0: no
24.	spend	Amount spent by customer in test mailing (\$)	numeric	
25.	part	Variable indicating which partition the record will be assigned to	alpha	t: training v: validation s:test

The following shows the first few rows of data:

num	usa	s_a	s_b	s_c	s_d	s_e	s_h	s_m	s_o	s_p	s_r	s_s	s_t	s_u	s_w	s_x	freq	last	first	web	male	res	purch	spend	part
1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	3662	3662	1	0	1	1	128	s
2	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2900	2900	1	1	0	0	0	s
3	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	2	3883	3914	0	0	0	1	127	t
4	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	829	829	0	1	0	0	0	s
5	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	869	869	0	0	0	0	0	t
6	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1995	2002	0	0	1	0	0	s
7	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	1498	1529	0	0	1	0	0	s

Analysis

The following gives an overview of the analysis for this case which we will develop both in class and for homework. Further details will be available as we work through it (i.e., do not follow these directions explicitly).

- 1) Develop a model for predicting spending among the purchasers
 - a) Make a copy of the data sheet, sort by the "Purchase" variable, and remove the records where Purchase = "0" (the resulting spreadsheet will contain only purchasers) – this spreadsheet is called “Tayko.xls.”
 - b) Partition this data set into training and validation partitions on the basis of the partition variable.
 - c) Develop models for predicting spending using
 - i) Multiple linear regression (use best subset selection) – homeworks 2 and 3.
 - ii) Regression trees – we will cover this in class during week 5.
 - d) Choose one model on the basis of its performance with the validation data.
- 2) Develop a model for classifying a customer as a purchaser or non-purchaser.

- a) Partition the data (Tayko_all.xls) on the basis of the partition variable, which has 800 "t's," 700 "v's" and 500 "s's" (standing for training data, validation data and test data, respectively) randomly assigned to cases.
 - b) Develop models for classifying customers:
 - i) Using the "best subset" option in logistic regression, implement the full logistic regression model, select the best subset of variables, then implement a regression model with just those variables to classify the data into purchasers and non-purchasers (logistic regression is used because it yields an estimated "probability of purchase," which is required later in the analysis) – we will cover this in class during week 4.
 - ii) Classification trees – we will cover this in class during week 5.
- 3) Consider the test data partition that includes both purchasers and non-purchasers. Note that we have not used this partition in any of our analysis to this point, so it will give an unbiased estimate of the performance of our models. It is best to make a copy of the test data portion of this sheet to work with, since we will be adding analysis to it – we will cover this in class during week 6.
- a) Copy the "predicted probability of success" (success = purchase) column from the classification of test data (from step 2) to this sheet.
 - b) Score the chosen prediction model (from step 1) to this data sheet.
 - c) Arrange the following columns so they are adjacent:
 - i) Predicted probability of purchase (a)
 - ii) Predicted spending \$ (b)
 - iii) Actual spending \$.
 - d) Add a column for expected spending [= probability of purchase * predicted spending].
 - e) Sort all records on the "expected spending" column.
 - f) Calculate cumulative lift (= cumulative "actual spending" divided by the average spending that would result from random selection) – note that total spending in the test data partition was \$46951 from 500 customers.
- 4) Each catalog costs approximately \$2 to mail (including printing, postage and mailing costs). Estimate the net profit that the firm could expect from its remaining 180,000 names if it randomly selected them from the pool – we will cover this in class during week 6.

$$([\mathbf{46951}/500 * \mathbf{0.1065}] - 2) * 180000 = \$8.00 * 180000 = \$1.44m$$
- 5) Using the cumulative lift from 3(f), estimate the net profit that would result from mailing to the 180,000 names selected using your data mining models – we will cover this in class during week 6.
We are mailing to 180,000/5,000,000 or 3.6% of the total remaining list. To estimate the lift that would result from mailing to the top 3.6%, we proceed 3.6% of the way down the cumulative lift table, in other words to the 18th of the 500 test records. The lift here is 5.449565826 so we can expect average revenue per catalog mailed to be

$$(5.449565826 * [\mathbf{46951}/500 * \mathbf{0.1065}] - 2) * 180000 = \$52.50 * 180000 = \$9.45m.$$