ADS-505 Applied Data Science for Business

Module 2 Assignment 2.1: Use Case - Tayko Software Cataloger (Using Python) (80 Points)

Tayko.csv is the dataset for this case study (linked in the assignment prompt in Blackboard).

For this assignment, you will use a Jupyter notebook to answer the questions in the "Assignment" section of this document. Be sure to include the number and question that you are answering in your notebook.

BACKGROUND

Tayko is a software catalog firm that sells games and educational software. It started out as a software manufacturer and later 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 expand 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.

THE MAILING EXPERIMENT

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.

This mailing yielded 1065 purchasers, a response rate of 0.053. 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 nonpurchasers. For ease of presentation, the dataset for this case includes just 1000 purchasers and 1000 non purchasers, an apparent response rate of 0.5. Therefore, after using the dataset to predict who will be a purchaser, we must adjust the purchase rate back down by multiplying each case's "probability of purchase" by 0.053/0.5, or 0.107.

DATA

There are two outcome variables in this case. Purchase indicates whether or not a prospect responded to the test mailing and purchased something. Spending 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 second will be used for those cases that are classified as purchase and will predict the amount they will spend.

Table 21.6 shows the first few rows of data. Table 21.7 provides a description of the variables available in this case.

Table 21.6 First 10 records from Tayko dataset

sequence_number	US	source_a	source_c	source_b	source_d	source_e	source_m	source_o	source_h	source_r	source_s	source_t	source_u	source_p	source_x	source_w	Freq	last_update_days_ago	1st_update_days_ago	Web order	Gender=male	Address_is_res	Purchase	Spending
1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	2	3662	3662	1	0	1	1	127.87
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
3	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	2	3883	3914	0	0	0	1	127.48
4	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	829	829	0	1	0	0	0
5	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	869	869	0	0	0	0	0
6	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1995	2002	0	0	1	0	0.06
7	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	1498	1529	0	0	1	0	0.06
8	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	3397	3397	0	1	0	0	0.08
9	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	525	2914	1	1	0	1	488.5
10	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3215	3215	0	0	0	1	173.5

(*Source*: Reproduced with permission from Datastats, LLC, ©2020 and Cytel Corp)

Table 21.7 Description of Variables for Tayko Dataset

				Code
Variable number	**Variable name**	**Description**	**Variable type**	**description**
1	US	Is it a US address?	Binary	1: Yes

				0: No
2–16	Source_*	Source catalog	Binary	1: Yes
		for the record		0: No
		(15 possible sources)		
17	Freq.	Number of transactions	Numerical	
		in last year at		
		source catalog		
18	last_update_days_ago	How many days ago	Numerical	
		last update was made		
		to customer record		
19	1st_update_days_ago	How many days	Numerical	
		ago first update		
		to customer record was made		
20	RFM%	Recency–frequency–	Numerical	
		monetary percentile,		

		as reported by		
		source catalog		
		(see Section 21.2)		
21	Web_order	Customer placed at	Binary	1: Yes
		least one order		0: No
		via web		
22	Gender=mal	Customer is male	Binary	1: Yes
				0: No
23	Address_is_res	Address is	Binary	1: Yes
		a residence		0: No
24	Purchase	Person made purchase	Binary	1: Yes
		in test mailing		0: No
25	Spending	Amount (dollars) spent	Numerical	
		by customer in		
		test mailing		

University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

ASSIGNMENT

1. Each catalog costs approximately $2 to mail (including printing, postage, and mailing costs). Estimate the gross profit that the firm could expect from the remaining 180,000 names if it selects them randomly from the pool. **(6 Points)**
2. Develop a model for classifying a customer as a purchaser or nonpurchaser.
   2.1. Partition the data randomly into a training set (800 records), validation set (700 records), and test set (500 records). **(6 Points)**
   2.2. Run logistic regression with L2 penalty (solver='lbfgs', cv=5, max_iter=500), using method LogisticRegressionCV, to select the best subset of variables, then use this model to classify the data into purchasers and nonpurchasers. Use only the training set for running the model. (Logistic regression is used because it yields an estimated "probability of purchase," which is required later in the analysis.) **(6 Points)**
3. Develop a model for predicting spending among the purchasers.
   3.1. Create subsets of the training and validation sets for only purchasers' records by filtering for Purchase = 1. **(6 Points)**
   3.2. Develop models for predicting spending with the filtered datasets, using:
      3.2.1. Multiple linear regression (use stepwise regression) **(6 Points)**
      3.2.2. Regression trees **(6 Points)**
      3.2.3. Choose one model on the basis of its performance on the validation data and explain your reasoning for selecting it **(6 Points)**
4. Return to the original test data partition. Note that this test data partition includes both purchasers and nonpurchasers. Create a new data frame called Score Analysis that contains the test data portion of this dataset.
   4.1. Add a column to the data frame with the predicted scores from the logistic regression. **(4 Points)**
   4.2. Add another column with the predicted spending amount from the prediction model chosen. **(4 Points)**
   4.3. *Let's skip this question.*

4.4. Add a column for expected spending: adjusted probability of purchase × predicted spending. **(4 Points)**

4.5. Plot the cumulative gains chart of the expected spending (cumulative expected spending as a function of number of records targeted). **(6 Points)**

4.6. Using this cumulative gains curve (not directly!), estimate the gross profit that would result from mailing to the 180,000 names on the basis of your data mining models. **(10 Points)**
Hint: For this part, you need to calculate average spending based on the model and then the expected profit.

5. Briefly explain, in two to three paragraphs, the business objective, the data mining models used, why they were used, the model results, and your recommendations to your non-technical stakeholder team. **(10 Points)**

Note: Although Tayko is a hypothetical company, the data in this case (modified slightly for illustrative purposes) were supplied by a real company that sells software through direct sales. The concept of a catalog consortium is based on the Abacus Catalog Alliance.