Homework 2: Identifying Target Customers

Yurong Jiang

Case & Data: The case and the Excel data file for this homework are available at the Decision Pros website. You can also find them on Blackboard. There are two data files related to this homework assignment, one called “Estimation Sample” and the other called “Holdout Sample”.

BBB Club’s management wants to know whom to send a specially produced brochure promoting a new art book The Art History of Florence. Thus, it runs an experiment, as we discussed in class. You will work through the analysis.

Question 1: Estimation using the Estimation Sample

Using ME-XL Excel Add-In and the data in the Estimation Sample, estimate a logit model predicting the probability of response as a function of all the available variables:

- Gender
- Amt_purchased
- Frequency
- Last_Purchase
- First_purchase
- P_Child
- P_Youth
- P_Cook
- P_DIY
- P_Art

To do this, select ME-XL→ Customer Choice (Logit). Once you’ve filled in your information properly, you should be able to run a logit model.

Look at your logit model results. **Report only the items that are requested below.**

(a) (5 points**) Report:** Logit model coefficients and t-statistics. Interpret the sign of each coefficient. Specifically, note which variables increase the probability that the customer makes a purchase and which decreases this probability. Highlight which factors most influenced the customers’ decision to buy or not to buy. *Hint: you do not need to compute any probabilities because higher coefficients imply higher probabilities.*

Table 1 shows the coefficient estimates of the logit model and the corresponding t-statistics:

Variables / Coefficient estimates	Coefficient estimates	Standard deviation	t-statistic
Gender	**-0.8632319**	0.13744997	-6.2803354
Amount purchased	**0.00186414**	0.00079182	2.35424651
Frequency	**-0.0755142**	0.01659374	-4.5507628
Last purchase	**0.61177129**	0.09381274	6.52119642
First purchase	-0.0147792	0.01280267	-1.1543851
P_Child	**-0.8112489**	0.11670677	-6.9511726
P_Youth	**-0.6370422**	0.14337788	-4.4430994
P_Cook	**-0.9230066**	0.11948153	-7.7250988
P_DIY	**-0.9058697**	0.14370263	-6.30378
P_Art	**0.68611236**	0.12701763	5.40170961
Const-1	-0.3515281	0.2143841	-1.6397116
Baseline		n/a	n/a

Table 1

The sign of each coefficient means the effect it will bring to the dependent variable. A positive sign means a positive effect and will increase the probability of response; a negative sign means a negative effect and will decrease the probability of response. Besides, the bigger the number, the larger the effect. Based on t-statistic, variables Gender, Amount purchased, Frequency, Last_purchase, P_Child, P_Youth, P_Cook, P_DIY, and P_Art have statistical significance in predicting whether a customer will respond to brochure promoting. On the other hand, First_purchase has no significance.

Therefore,

- Factors increase the probability: Amt_purchased, Last_purchase, and P_Art
- Factors decrease this probability: Gender, Frequency, P_Child, P_Youth, P_Cook, and P_DIY
- Factors most influenced the customers' decision to buy: P_Art
- Factors most influenced the customers' decision not to buy: P_Cook

Question 2: Applying the scoring model to the Holdout Sample

These steps follow the same process we used in our in-class RFM exercise. Unless a particular step says "report", you do not need to report anything in your write-up.
Hint: Please refer to the Excel file "Bookbinders Book Club Student Worksheets" in Blackboard for useful formulae

(c) Compute a logit score for each person in the Holdout Sample. *Hint: the instructor already provided the logit score calculation for the first respondent for your reference.*

(d) Sort the 2300 prospects in the Holdout Sample in decreasing order of logit scores.

(e) (5 points) Compute the Decile Scores based on logit scores. Create a plot with the deciles on the horizontal axis and the percent cumulative response on the vertical axis. This is the Lift

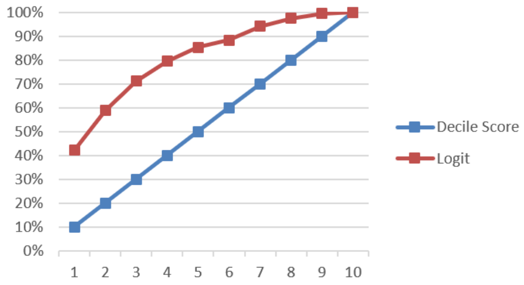
Curve.
Report: The decile scores and the lift curve plot. Comment on this plot. What conclusions can you draw about the quality of your scoring rule? Is it a good rule? How does it compare with the lift curve from the RFM exercise we did in class? *Hint: see instructor's formulae in the "RFM sorted" worksheet as guidelines*

Decile Score	Logit
10%	42.2%
20%	58.8%
30%	71.1%
40%	79.4%
50%	85.3%
60%	88.2%
70%	94.1%
80%	97.5%
90%	99.5%
100%	100.0%

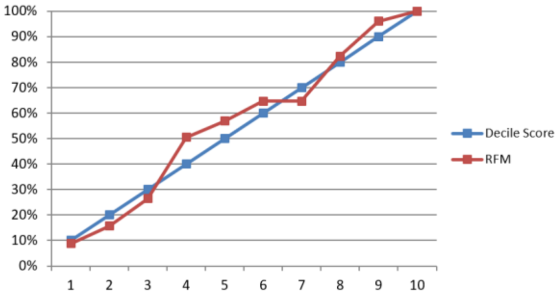
Table 2

Based on Graph 1, the quality of the scoring rule of logit model is pretty good because the lift curve is always above and away from the decile scores. The decile scores line represents random guesses. Besides, by only contacting 20% consumer, the company can contact 59% of people who will respond.

Graph 2 is the lift curve from the RFM method. It is pretty bad and the lift curve of the logit model is much better than it. In the RFM graph, sometimes the accuracy is even lower than randomly guess.

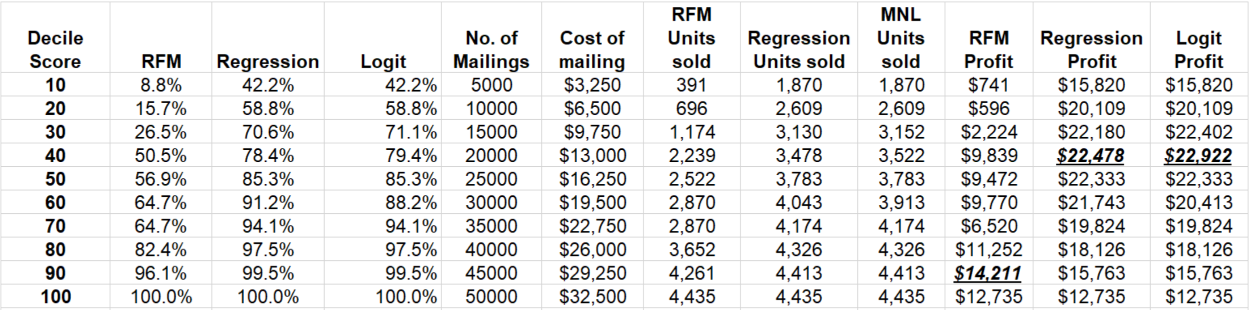


Graph 1 Lift Curve of the Logit Model



Graph 2 Lift Curve of the RFM Method

(f) (10 points) Bookbinders is considering a similar mail campaign in the Midwest where it has data for 50,000 customers. Such mailings typically promote several books. The allocated cost of the mailing is $0.65/addressee (including postage) for the art book, and the book costs Bookbinders $15 to purchase and mail. The company allocates overhead to each book at 45 percent of cost. The selling price of the book is $31.95. Based on the logit model, how many percentiles of customers should Bookbinders target? How would Bookbinders' profit be at this percentile? How much more profit would you expect the company to generate using this model as compared to sending the mail offer to the entire list? *Hint: see instructor's formulae in the "Model Evaluator" worksheet as guidelines*



Decile Score	**RFM**	**Regression**	**Logit**	**No. of Mailings**	**Cost of mailing**	**RFM Units sold**	**Regression Units sold**	**MNL Units sold**	**RFM Profit**	**Regression Profit**	**Logit Profit**
10	8.8%	42.2%	42.2%	5000	$3,250	391	1,870	1,870	$741	$15,820	$15,820
20	15.7%	58.8%	58.8%	10000	$6,500	696	2,609	2,609	$596	$20,109	$20,109
30	26.5%	70.6%	71.1%	15000	$9,750	1,174	3,130	3,152	$2,224	$22,180	$22,402
40	50.5%	78.4%	79.4%	20000	$13,000	2,239	3,478	3,522	$9,839	***$22,478***	***$22,922***
50	56.9%	85.3%	85.3%	25000	$16,250	2,522	3,783	3,783	$9,472	$22,333	$22,333
60	64.7%	91.2%	88.2%	30000	$19,500	2,870	4,043	3,913	$9,770	$21,743	$20,413
70	64.7%	94.1%	94.1%	35000	$22,750	2,870	4,174	4,174	$6,520	$19,824	$19,824
80	82.4%	97.5%	97.5%	40000	$26,000	3,652	4,326	4,326	$11,252	$18,126	$18,126
90	96.1%	99.5%	99.5%	45000	$29,250	4,261	4,413	4,413	***$14,211***	$15,763	$15,763
100	100.0%	100.0%	100.0%	50000	$32,500	4,435	4,435	4,435	$12,735	$12,735	$12,735

Table 3

Based on the logit model, Bookbinders should target 40% percentiles of customers and would get $22,922 profit at this percentile. Compared to sending the mail offer to the entire list, the company will generate $10,185($22,922 – $12,735) more profit by using this model.

Question 3: Repeat analyses listed in Q1 and Q2 above using a regression model.

Answer the following questions:
(l) (5points) **Report:** regression analysis results (be sure to report both coefficients and t-statistics); interpret the sign of each coefficient; and the decile scores and the lift curve plot. How many percentiles of customers should Bookbinders target? How much profit will the company make based on the regression model?

Table 4 shows the coefficient estimates of the regression analysis and the corresponding t-statistics:

	Coefficients	Standard Error	t Stat
Intercept	**0.364228**	0.030741	11.84824
Gender	**-0.13092**	0.02003	-6.53612
Amount purchased	**0.000274**	0.000111	2.464059
Frequency	**-0.00909**	0.002179	-4.17005
Last purchase	**0.097029**	0.013559	7.156089
First purchase	-0.002	0.001816	-1.10263
P_Child	**-0.12626**	0.016401	-7.69817
P_Youth	**-0.09636**	0.02011	-4.79153
P_Cook	**-0.14149**	0.016606	-8.52024
P_DIY	**-0.13523**	0.019787	-6.83425
P_Art	**0.117849**	0.019443	6.061375

Table 4

The sign of each coefficient means the effect it will bring to the dependent variable. A positive sign means a positive effect and will increase the probability of response; a negative sign means a negative effect and will decrease the probability of response. Besides, the bigger the number, the larger the effect. Based on t-statistic, variables Gender, Amount purchased, Frequency, Last_purchase, P_Child, P_Youth, P_Cook, P_DIY, and P_Art have statistical significance in predicting whether a customer will respond to brochure promoting. On the other hand, First_purchase has no significance.

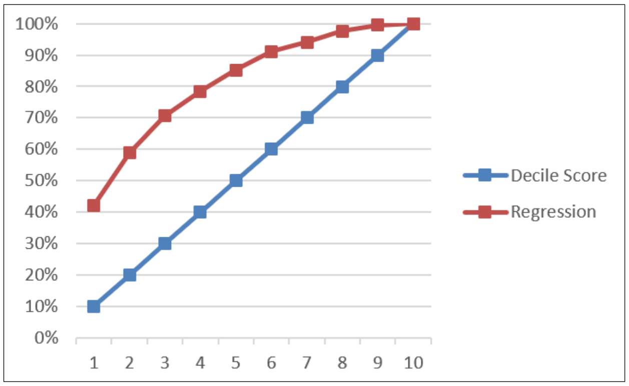
Therefore,

- Factors increase the probability: Amt_purchased, Last_purchase, and P_Art
- Factors decrease this probability: Gender, Frequency, P_Child, P_Youth, P_Cook, and P_DIY
- Factors most influenced the customers' decision to buy: P_Art
- Factors most influenced the customers' decision not to buy: P_Cook

The decile scores and the lift curve plot:

Decile Score	**Regression**
10%	42.2%
20%	58.8%
30%	70.6%
40%	78.4%
50%	85.3%
60%	91.2%
70%	94.1%
80%	97.5%
90%	99.5%
100%	100.0%

Table 5



Graph 3 Lift Curve of the Regression Model

Based on the regression model, Bookbinders should target 40% percentiles of customers and would get $22,478 profit at this percentile (see in Table 3). Compared to sending the mail offer to the entire list, the company will generate $9,743 ($22,478 – $12,735) more profit by using this model.

(m) (5 points) compare and contrast with results from the logit; the regression; and the RFM analysis. Summarize the advantages and limitations of each of the modeling approaches.

Comparing three models, the logit model performs the best, regression model is the second-best model, while RFM analysis performs the worst. From Table 3 we can know that, both regression model and logit model can maximize their profit by targeting first 40% percentile customers. This means that their cost of mailing is the same. However, the profit of logit model is larger than regression model. In addition, the logit model can cover about 85% of ‘best’ customer by just contacting top 50% of the list. When using RFM, we can reach the largest profit by contacting 90% of the list, but earn a profit of $14,211 which is much lower than Logit model ($22,922).

For logit model, one advantage is that it’s very good at predicting a binary outcome. It can capture target customers more accurately and get more profit. Besides, we can easily interpret the result of logit model. However, the limitation is we cannot use it to predict continuous amount. We also cannot use it to solve non-linear problems.

For regression model, it performs well in a lot of situations. Also, this model is very easy to use and we can easily interpret the result of regression model to others. Regression model is still one of the most popular models in the data analysis field. However, the model assumes that variables are independent with each other which is not very realistic in some situations.

For RFM model, one of the biggest advantages is that we don’t need to do experiments and we can save a lot of money and human resource. RFM model is very popular in banks. However, it’s a very arbitrary model. Because it only takes account R(recency), F(frequency), M(monetary), it does not perform well under some business models.