The True Lift Model - A Novel Data Mining Approach to Response Modeling in Database Marketing

Victor S.Y. Lo
Fidelity Investments
82 Devonshire Street R3D
Boston, MA 02109-3614
Victor.lo@fmr.com

ABSTRACT

In database marketing, data mining has been used extensively to find the optimal customer targets so as to maximize return on investment. In particular, using marketing campaign data, models are typically developed to identify characteristics of customers who are most likely to respond. While these models are helpful in identifying the likely responders, they may be targeting customers who have decided to take the desirable action or not regardless of whether they receive the campaign contact (e.g. mail, call). Based on many years of business experience, we identify the appropriate business objective and its associated mathematical objective function. We point out that the current approach is not directly designed to solve the appropriate business objective. We then propose a new methodology to identify the customers whose decisions will be positively influenced by campaigns. The proposed methodology is easy to implement and can be used in conjunction with most commonly used supervised learning algorithms. An example using simulated data is used to illustrate the proposed methodology. This paper may provide the database marketing industry with a simple but significant methodological improvement and open a new area for further research and development.

Keywords

Database marketing, data mining, knowledge discovery, predictive modeling, response modeling, customer relationship management, customer development, upselling and cross-selling, treatment effect, true lift, interaction effect

1. INTRODUCTION

Among the many applications of data mining and knowledge discovery, database marketing is a key area where scientific methods are often applied to analyze a massive amount of business data (see e.g. [17;19;22;32]). In the past decade, many industries have adopted a data warehousing project to capture customer profile and interaction data (e.g. [18;9]). This is the critical requirement for maximizing learnings of customers so that companies can provide the best offers and messages to the right customers through the right channels. Maximizing the "learning relationship" is a key step in customer relationship management or one-to-one marketing [29;30]. Applications of data mining in database marketing often require professionals with multi-disciplinary skills (e.g. [4], p.196-197) and thus have drawn interests from professionals in many areas.

In order to maximize campaign return on investment, predictive modeling is often applied to determine the characteristics of customers (or prospects) who are likely to respond using data from a previous campaign. Then prior to launching the next similar campaign, the model can be used to identify likely responders. This reduces the number of contacts (mails, calls, etc.) required to reach the desirable number of responders (e.g. [2;3;33]). Other applications include using predictive modeling to determine the right offer, right message, and right channel for each customer (if multiple offers are tested in the previous campaign e.g. [1]). For simplicity, we will focus on targeting the right customers (or prospects) in this paper while determining the right treatment (offer, message, and channel) may be considered as an extension.

Common types of marketing campaigns where response modeling can be applied include (e.g. [4], chapters 3-5):

(1) Acquisition – which prospects are most likely to become customers; this also includes win-back campaigns where attrited customers are targeted;

(2) Development – which customers are most likely to buy additional products (cross-selling) or to increase monetary values (up-selling);

(3) Retention – which customers are most likely to be 'saved' by a retention campaign; this essentially identifies who have more 'controllable' risks as opposed to those who will attrite regardless of the retention effort.

In this paper, we point out that the current methodology widely published in the literature and commonly used in business is not directly designed to meet the desired business objective. We then propose a new and simple response modeling methodology that can be used with commonly used statistical and data mining techniques.

This paper is organized as follows. In section 2, we discuss the appropriate campaign measurement of model effectiveness and then provide the appropriate business objective for response modeling. Section 3 outlines the current response modeling methodology. Section 4 first provides the mathematical objective function to meet the appropriate business objective and then proposes a new methodology to achieve the objection function. Section 5 uses a simulated example to illustrate the proposed methodology. Section 6 provides recommendations for future research, followed by conclusions in section 7.

2. BUSINESS OBJECTIVE

To derive the appropriate business objective for response modeling, we first discuss the measurement of model

effectiveness in marketing campaigns. For campaign-specific response modeling, the main purpose is to improve future campaign return on investment. This can be accomplished by identifying the customers (or prospects) who are most likely to respond where a response can be accepting an offer, increasing revenue or sales, etc.

Table 1: Campaign measurement of model effectiveness

	Treatment (e.g. mail)	Control (e.g. no-mail)	Incremental (treatment minus control)
Model	A	B	A-B
Random	C	D	C-D
Model minus random	A-C	B-D	(A-B)-(C-D)

After a model is used for a campaign, analysts may measure the effectiveness of the model. Table 1 shows a typical example. In this table, 'Model' is the group of customers identified as 'good' (e.g. likely responders) by the model. It may be the top decile or top 2 or 3 deciles predicted by the model. 'Random' is the group of customers targeted randomly. 'Treatment' is the group that received a treatment (an offer) through direct mail, e-mail, web, or live channel. 'Control' is the group similar to the treatment group but no treatment is offered and their data is captured for comparison purpose (e.g. [27]). A, B, C, and D are observed aggregate values of a campaign. For examples, they may represent the mean number of sales generated, the mean new revenue generated, or simply the mean response rates, i.e. they can be continuous or proportions. If A > C (statistically significantly), it means that the model is in the right direction of targeting customers who are likely to respond. However, we would like to see what would have happened if the customers did not receive the treatment and that is the role of the control group.

Now let's look at the difference between 'Treatment' and 'Control.' If A > C and B > D (assumed all statistically significant) but A – B = 0 (i.e. assumed statistically insignificant), it means that while the model is able to pick the likely responders, it does not add any 'value' to the campaign. It is because A – B = 0, or A=B, implies that the customers in the model-treatment cell (where A is) received the same value as those in the model-control cell (where B is) who did not receive the treatment. Note that A-B represents the treatment and control difference (e.g. with and without mail) for those identified by the model. Similarly, C-D represents the treatment and control difference for the random group. To claim that the model 'works' (i.e. 'Model' is better than 'Random'), we not only require that A-B > 0 but also that A-B > C-D. We propose that the appropriate measure of the gain (in response rate, revenue, or sales, etc.) due to the treatment is (A-B) – (C-D). Thus, we suggest that the quantitative business objective of a model, measured by its effectiveness for campaigns, is to maximize (A-B) – (C-D), which is hereafter referred to as the *true lift*.

3. CURRENT METHODOLOGY

The current methodology typically uses the treatment data to identify characteristics of customers who are likely to respond. In other words, for individual customer (or prospect) $i \in W$,

$$E(Y_i \mid X_i; \text{treatment}) = f(X_i), \quad (1)$$

where W = set of all customers (or prospects), Y_i = dependent variable (e.g. respond or not, sales, revenue, profit), $X_i = (X_{i1},\ldots, X_{ip})'$ is a vector of p independent variables that represent individual characteristics (e.g. age, income, past transaction behavior), $f(.)$ is the functional form of the model. For examples, in linear regression, Y_i is continuous, and $f(.)$ is a linear function of X_i; in logistic regression, Y_i is binary (e.g. respond or not), expected Y_i becomes the probability that $Y_I = 1$ (i.e. response rate), and $f(.)$ is a logistic function of X_i ([16]). In fact, equation (1) is a general *supervised learning* model form as $f(.)$ may be nonlinear or other complicated functions such as step-functions (e.g. decision trees such as CART and CHAID, see [6;28]), splines ([36;38]), composite functions (e.g. multi-layer perception in neural networks [5;12]), other neural network models (e.g. [23;35]), mixture models (e.g. [37;12]), Bayesian models (e.g. [13;20]), or hybrids (e.g. [11;14;26]).

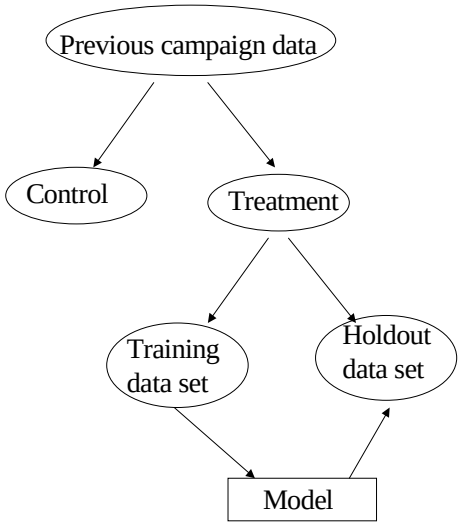
In data mining, as summarized in Fig 1, the entire data set from a previous campaign is usually divided into training (or learning) and holdout (or validation) samples. $f(.)$ is then fitted (or trained) using the training sample. Then the fitted model is applied to the holdout sample for validation purposes. Examples of validation include plotting the observed average values by model-based decile (see e.g. [3;33]). Unlike in traditional statistical modeling where the objective is to find the best 'overall' fitness measured by the sum of squares of residuals or joint log-likelihood function of *all* individuals (e.g. [8]), database marketers are often interested in targeting only the 'best' customers or prospects such as top 1-3 deciles. Other related measures are mentioned in [31].

Mathematically, the current methodology uses model (1) to predict Y_i for each customer i in a new data set and then select the set of customers S ($\subseteq W$) that have the highest predicted Y_i given treatment, i.e.

$$\text{Maximize} \sum_{i \in S} E(Y_i \mid X_i; \text{treatment}) \quad (2)$$

where i denotes customer i, subject to budget or other resource constraints. If Y_i denotes customer profitability, this is equivalent to maximizing the overall profit. If Y_i denotes response rate, this implies maximizing the sum of response rates. This approach has been mentioned in vast amount of literature such as [2;3;33].

Fig 1. Current Methodology



4. PROPOSED METHODOLOGY

4.1 Objective Function

In section 2, we have proposed that the appropriate model measure in campaigns is the difference between model and random incremental (treatment minus control) differences, i.e. (A-B) – (C-D) in Table 1, or the *true lift.*

Note that the current methodology for objective (2) is essentially maximizing the difference between the model-treatment and random-treatment, i.e. A (with model) and C (random) under the treatment column in Table 1. In order to achieve the objective of maximizing the *true lift*, (A-B) – (C-D), we should select the set of customers S ($\subseteq W$) that have the highest incremental differences between treatment and control, i.e.

$$Maximize \sum_{i \in S} \{E(Y_i \mid X_i; treatment) - E(Y_i \mid X_i; control)\} \quad (3)$$

subject to budget or other resource constraints.

Objective (3) is equivalent to identifying the customers (or prospects) whose responses will be positively influenced by the treatment. Consider a simplified example, in the automobile industry, a company selling high-end expensive cars uses direct mails to target individuals who are wealthy and currently have an expensive car. However, if their current car is already very old, they may decide to replace the car prior to receiving a direct mail (some decide to buy the same car the company is selling and others decide to buy other cars). On the other hand, if their current car is not old enough, they may not consider buying a new one. In either case, the treatment response rate is not expected to be better than the natural response rate in the control group. The challenge is to target individuals who own an expensive car of right age so that they will respond to the direct mail (treatment) before making a *natural* decision (control). That is, we need to find the characteristics of individuals whose decisions will be positively influenced by the direct mail.

4.2 Customer Development versus Acquisition

To achieve (3), we need to estimate both $E(Y_i|X_i;treatment)$ and $E(Y_i|X_i;control)$. Note that (3) reduces to (2) if $E(Y_i|X_i;control)$ is close to zero. The latter condition may be valid for some acquisition problems. For example, in the credit card industry, prospects usually will not take the initiative to apply for a credit card unless a direct mail is received (typically with a good offer such as low interest rate, low rate of balance transfer, or cash rebate). That means, the response rate for the control (no mail) is close to zero. However, in customer development campaigns including cross-selling and upselling, customers may more likely take the initiative regardless of whether they receive a direct mail. For instance, in retail banking, prior to the maturity of a customer's CD (certificate of deposit), the customer may initiate a transfer to a money market mutual fund to preserve the relatively high interest income. This is not surprising given the internal competition between the CD and mutual fund products (e.g. [21]). Therefore, her decision to transfer money to another product may be made before receiving a direct mail and thus, the treatment and control groups may not give any different campaign result. Nevertheless, (3) is the appropriate general objective for all types of campaigns and is especially important for customer development.

4.3 Example of Logistic Regression

We now use a special case of binary response variable to describe our proposed methodology. If Y_i is a binary response variable representing whether customer i responds to a campaign, we consider the following set of independent variables: X_i, T_i, and X_i*T_i where $T_i = 1$ if i is in the treatment group and = 0 if i is in the control group. We may model the response rate using a linear logistic regression [16]:

$$P_i = E(Y_i \mid X_i) = \frac{exp(\alpha + \beta' X_i + \gamma T_i + \delta' X_i T_i)}{1 + exp(\alpha + \beta' X_i + \gamma T_i + \delta' X_i T_i)} \quad (4)$$

where $\alpha, \beta, \upsilon, \delta$ are parameters to be estimated.

In equation (4), α denotes the intercept, β is a vector of parameters measuring the main effects of the independent variables, γ denotes the main treatment effect, and δ measures additional effects of the independent variables due to treatment. In reality, some variable reduction procedure will usually be applied to narrow down the list of X_i, T_i, and X_i*T_i first before estimating the parameters in equation (4).

Then,

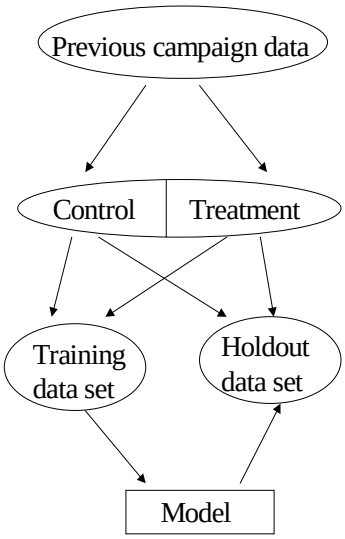
$$P_i \mid treatment - P_i \mid control = \frac{exp(\alpha + \gamma + \beta' X_i + \delta' X_i)}{1 + exp(\alpha + \gamma + \beta' X_i + \delta' X_i)} - \frac{exp(\alpha + \beta' X_i)}{1 + exp(\alpha + \beta' X_i)} \quad (5)$$

That is, the parameter estimates obtained in equation (4) can be used in equation (5) to predict the treatment and control difference in response rate for each i in a new data set. Then the data set can be sorted by the predicted difference between treatment and control. Those with high positive predicted differences will be selected for next campaign targeting.

While equation (4) may be new in database marketing, a similar practice of modeling and analysis has been applied widely in the biomedical clinical trial literature (e.g. [7;10;15;25;34]). For example, [25] concludes that a new HIV treatment is more effective than the traditional treatment (ddI alone) for HIV-infected children under 3 years old but not for older children. In clinical trials, the objective is usually to measure the treatment effect given the presence of other factors (baseline characteristics such as age at treatment initiation) and also possibly determine whether the treatment effect is the same across various groups (e.g. whether gender * treatment is a significant interaction factor

[34]). The latter appears to be similar to our objective in database marketing. However, in clinical trials, the objective is to identify risk factors that may affect treatment-response relationship while in database marketing, it is to explicitly predict, at the individual level, who will be more positively influenced by the treatment effect. Another major difference is that the number of independent variables and the number of observations are normally much smaller in clinical trial studies.

Fig 2. Proposed Methodology



4.4 Estimation and Validation Procedures

For model estimation, we use both the treatment and control groups. Fig 2 indicates that both the training and holdout data sets include the treatment and control groups as opposed to the current methodology in Fig 1 where only the treatment group is used.

More generally, we propose the following procedure for estimating $E(Y_i|X_i;treatment)$ and $E(Y_i|X_i;control)$ where Y_i can be continuous or binary; and the model functional form can be linear, linear logistic, or nonlinear:

1. Include data, $\{Y_i,X_i\}$ from both the treatment and control groups in the analysis data set;
2. Assign a dummy variable T_i to 1 for the treatment group and 0 for the control group;
3. Divide the data set into training and hold-out samples;
4. Further divide the training sample into two sub-samples by T_i, i.e. one is treatment and the other is control;
5. Choose a variable selection method (or called feature extraction). In each sub-sample (treatment and control), use the method to narrow down your list of independent variables, X_i (often an essential step in data mining as there are normally hundreds of independent variables);
6. Take the union of the two reduced sets of independent variables from 5 and thus, the new X_i has only q elements, where q<original number of independent variables, p;
7. Multiply all independent variables, X_i, (from step 6) by T_i to form the interaction effects, X_i*T_i;
8. Choose a data mining or statistical technique for supervised learning;
9. Fit a model using Y_i as the dependent variable and X_i, T_i, and X_i*T_i as independent variables;
10. Use stepwise procedure (or similar model selection procedure) to determine the best parsimonious model.

After the best model is selected, we propose the following procedure for validation using the holdout sample:

1. For each individual in the hold-out sample, compute the predicted values of expected Y_i for both the treatment and control, i.e. predict $E(Y_i|X_i;treatment)$ and $E(Y_i|X_i;control)$;
2. Subtract the control value from the treatment value to estimate the treatment and control difference (in order to achieve objective (3));
3. Rank and decile the entire hold-out sample by the predicted difference;
4. In each decile, compute the observed mean value of Y_i's in the treatment group and the observed mean value of Y_i's in the control group and then take the observed difference;
5. Plot the observed difference between treatment and control by decile to validate the model;
6. The expected *true lift* can be measured by how much the top decile(s) perform better than random using the observed treatment and control difference from step 6.

We will illustrative the above procedures in Section 5.

The general form of (4) and (5) is simply as follows:

$$E(Y_i \mid X_i) = f(X_i, T_i, X_i * T_i) \qquad (6)$$

$$E(Y_i \mid X_i;treatment) - E(Y_i \mid X_i;control) \qquad (7)$$
$$= f_1(X_i) - f_2(X_i),$$
$$where\ f_1(X_i) = f(X_i, T_i, X_i * T_i)\ when T_i = 1, and$$
$$f_2(X_i) = f(X_i, T_i, X_i * T_i)\ when T_i = 0.$$

For example, if Y_i is a continuous variable representing revenue, we may use a multi-layer perceptron neural network with a single hidden layer:

$$Y_i = w_{20} + \sum_{j=1}^{J} \frac{w_{2j}}{1 + exp(-z_{ji})},$$

$$where\ z_{ji} = w_{1,0j} + \sum_{k=1}^{K} w_{1,kj} x_{ik}^{*},$$

where $w_{20}, w_{21}, ..., w_{2J}$ are the parameters (weights) linking the bias and the J hidden units on the hidden layer to the dependent variable on the output layer,

$w_{1,0j}, w_{1,1j}, ..., w_{1,Kj}$ are the parameters linking the bias and the K input variables on the input layer to jth hidden unit, and

$x_{i1}^{*}, ..., x_{iK}^{*}$ are the input variables selected from the original set of independent variables X_i, T_i, and $X_i * T_i$.

If the model function cannot be written in closed form in equations (6) and (7), a simple modification will be required. See the Appendix for the case when Naïve Bayes is used.

5. A SIMULATED EXAMPLE

We now illustrate the methodology with a simulation study.

5.1 The Simulated Data

We assume the following logistic models for treatment and control response rates:

Control response rate

$$= \frac{exp(-7.5 + 0.02\,age + 0.005\,wealth + 0.00155\,asset)}{1 + exp(-7.5 + 0.02\,age + 0.005\,wealth + 0.00155\,asset)}, \quad (8)$$

Treatment response rate

$$= \frac{exp(-8.0 + 0.04\,age + 0.005\,wealth + 0.00165\,asset)}{1 + exp(-8.0 + 0.04\,age + 0.005\,wealth + 0.00165\,asset)} \quad (9).$$

where the independent variables are generated as follows:

$age \sim N(45, 13^2)$, $trade \sim Bin(n=10, p=0.15)$,

$wealth \sim N(800+3age, 150^2)$, $asset \sim N(400+0.3wealth, 150^2)$,

and $home\ value \sim N(0.7wealth, 70^2)$.

To reflect reality, these variables are created to be correlated with each other. Other than *trade*, they can be jointly expressed in the following multivariate normal form (*wealth*, *asset*, and *home value* are all in $000):

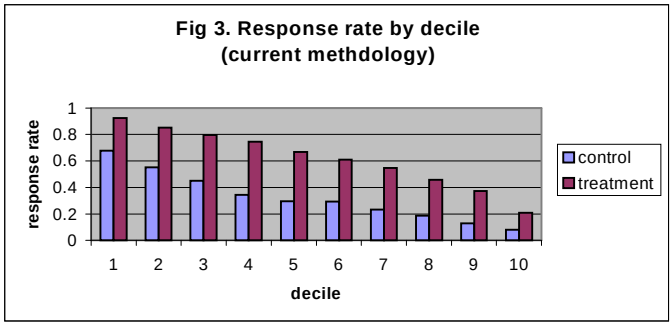
$$\begin{pmatrix} age \\ wealth \\ asset \\ home\ value \end{pmatrix} \sim MVN\left(\begin{pmatrix} 45 \\ 935 \\ 680.5 \\ 654.5 \end{pmatrix}, \begin{pmatrix} 13^2 & 507 & 152 & 355 \\ 507 & 150^2 & 6{,}750 & 15{,}750 \\ 152 & 6{,}750 & 150^2 & 4{,}725 \\ 355 & 15{,}750 & 4{,}725 & 70^2 \end{pmatrix} \right)$$

Note that while *home value* is available for modeling, it does not appear in the true models in (8) and (9). Equations (8) and (9) indicate that both treatment and control response rates are affected by age, wealth, and asset. Additionally, the fact that the coefficients of age and asset in (9) are higher than those in (8) indicate that the differential effects between treatment and control appear in age and asset. In particular, older people and people with higher assets tend to respond to the treatment more likely than people in the control group.

Using the above true models, we randomly generated 100,000 observations with 80,000 from the treatment group and 20,000 from the control group.

5.2 Applying the Current Methodology

We apply the current methodology in this section. First, we randomly divide the set of 100,000 simulated observations into training and holdout samples with 50,000 each. Using the current methodology, we only focus on the treatment data. Then, we attempt to narrow down the set of independent variables by generating a simple correlation matrix between the response rate and all five independent variables using the treatment data. Due to the high correlation (empirical coefficient>0.85) between *wealth* and *home value* and the fact that *wealth* has a higher correlation with the response, we drop the *home value* variable.



Next, we run a stepwise linear logistic regression of the response rate on *age*, *trade*, *wealth*, and *asset* in SAS, resulting in the following model (subscript *i* is omitted for convenience):

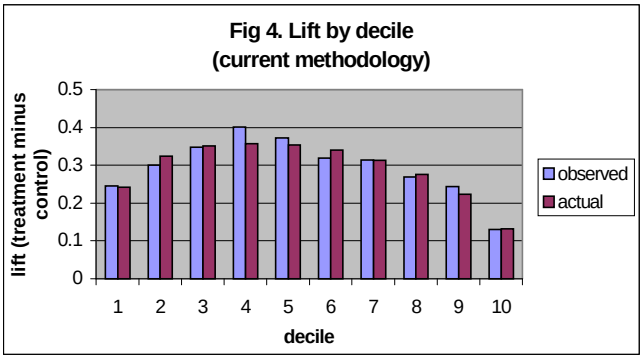
Estimated treatment response rate

$$= \frac{exp(-6.82 + 0.0407\,age + 0.00484 wealth + 0.00158\,asset)}{1 + exp(-6.82 + 0.0407\,age + 0.00484\,wealth + 0.00158\,asset)} \quad (10)$$

Then we use the above estimated model (10) to rank and decile the holdout sample. Fig 3 shows the performance in both the treatment and control groups. The horizontal axis is the decile (predicted) based on the above estimated model and the vertical axis shows the observed response rate by decile. The model does perform 'well' in the sense of the declining response rate by decile. In particular, the top decile generates a treatment response rate of 0.93 compared to the random treatment response rate of 0.62. Note that both treatment and control rates decline by decile. To evaluate the *true lift*, we subtract the control rate from the treatment rate in each decile as shown in Fig 4. In Fig 4, in

addition to the observed difference between treatment and control rates, we compute the 'actual' lift that is simply the response rate difference between the actual treatment and control rates in (8) and (9). The actual lift is only available through simulations but will not be available in practice.

Fig 4 shows that, first, the actual and observed have similar patterns and, second, both indicate that the treatment minus control lift does not decline by decile. In fact, the fourth and fifth deciles have the highest lift.



5.3 Applying the Proposed Methodology

To apply the proposed methodology, we follow the estimation procedure outlined in Section 4.4. In the training sample, we now include both the treatment and control data for modeling. Following the estimation steps 1-4, we have T_i=1 for the treatment group and T_i=0 for the control group.

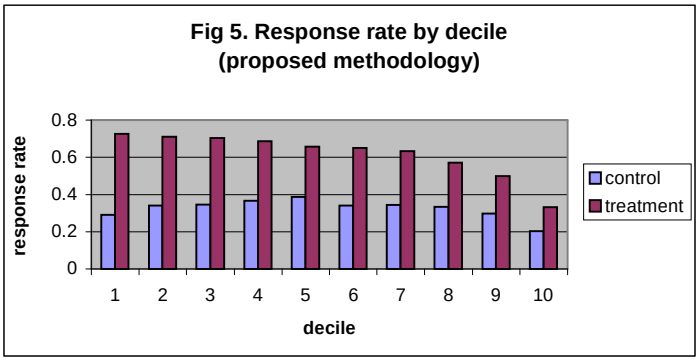
Following the estimation steps 5 and 6 in section 4.4, we perform variable selection by investigating the correlation matrix of the response rate and all five independent variables by T_i. Once again, due to the high correlations (empirical coefficients>0.85) between *wealth* and *home value* in both the treatment and control groups and the fact that *wealth* has a higher correlation with the response rate, we drop the *home value* variable in both the treatment and control groups.

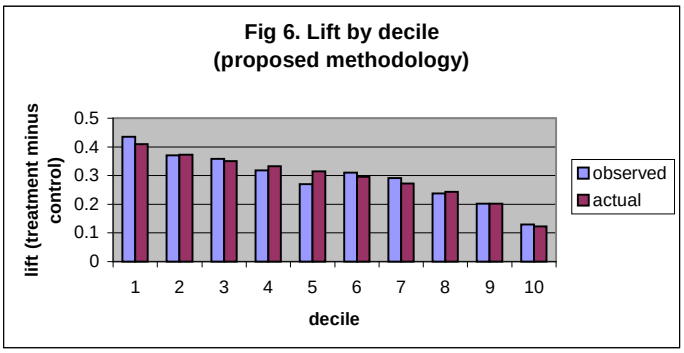
We then run a stepwise linear logistic regression using the model form in equation (4) in SAS, resulting in the following estimated model (subscript *i* is omitted for convenience):

$$\text{Estimated response rate} = \frac{exp(-7.60 + 0.0236\ age + 0.0048\ 6\ wealth + 0.001\ 6\ asset + 0.739\ T + 0.017\ age * T)}{1 + exp(-7.60 + 0.0236\ age + 0.0048\ 6\ wealth + 0.001\ 6\ asset + 0.739\ T + 0.017\ age * T)} \quad (11)$$

(11) can be rewritten as:

$$\text{Estimated control response rate} = \frac{exp(-7.60 + 0.0236\ age + 0.00486\ wealth + 0.0016\ asset)}{1 + exp(-7.60 + 0.0236\ age + 0.00486\ wealth + 0.0016\ asset)}, \quad (12)$$





Estimated treatment response rate

$$= \frac{exp(-6.86 + 0.0406\ age + 0.00486\ wealth + 0.0016\ asset)}{1 + exp(-6.86 + 0.0406\ age + 0.00486\ wealth + 0.0016\ asset)} \quad (13).$$

Note that the estimated model in (12) and (13) is very close to the actual model in (8) and (9) except that the estimated model is not able to capture the treatment and control differential effect of *asset.*

Next, in the holdout sample, we follow the validation steps 1-3 to estimate the treatment and control response rates using (12)-(13) and then take the difference to estimate the lift for each individual. We then rank and decile the holdout sample with the estimated treatment minus control lift.

Fig 5 shows the observed response rate by decile for treatment and control, respectively. This does not show any declining pattern. However, following the validation steps 4-6, Fig 6 shows that the observed difference by decile has a declining pattern. Similarly for the 'actual' difference which is again the response rate difference between the actual treatment and control rates in (8) and (9). In the top decile, the actual lift is 0.41 compared to the average (random) lift of 0.29. Again, in reality, 'actual' is not available and 'observed' can be used to approximate the 'actual.'

Comparing Fig 6 to Fig 4, we can recognize the significant improvement using the proposed methodology, which aims at solving the appropriate business problem by maximizing the expected *true lift.*

6. FUTURE RESEARCH

We hope to open an area for academics and industrial participants to improve or extend the proposed methodology. This will further benefit the industry of database marketing.

We have considered the following areas for potential improvement or extension:

1. Control size determination: The control group has traditionally been used for comparison purpose only. Marketing managers often attempt to minimize the control group size so as to maximize campaign return. On the other hand, statisticians need to set the appropriate size level such that the treatment and control differences, if exist, can be detected with a high likelihood. The proposed methodology, however, adds another requirement to the control size, i.e. not only the treatment and control differences need to be detectable, but both treatment and control groups need to be sufficiently large for model development. Hence, the optimal control size will need to be determined given the modeling requirement and campaign objective.
2. Variable reduction (feature extraction): Many variable reduction techniques have been proposed for data mining (e.g. [24]). In this paper, our modeling objective of capturing the differential effect of treatment versus control may be a unique problem in the area of variable reduction. This is equivalent to capturing the interaction effects of treatment and other independent variables in addition to their own main effects. For instance, one may consider using interaction detection algorithms such as decision trees (e.g. [6;28]).
3. Multiple treatments: When more than one treatment (i.e. multiple offers, channels, messages, etc.) are available, the proposed methodology can be easily extended. However, more interaction variables will be generated which will complicate the estimation procedure.
4. Estimation procedure: Large-scale simulation studies may be used to understand the robustness of the proposed estimation procedure with respect to the number of independent variables, the scale and variability of independent variables, and correlations among independent variables. Other similar procedures (e.g. fitting models on treatment and control data separately instead of using interaction effects to differentiate between treatment and control) may also be studied with rigorous statistical theories and empirical simulations.
5. Validation procedure: Similar to the estimation procedure, it would be useful to use simulation studies to examine the sensitivity of the proposed validation procedure to various factors such as the scale and variability of independent variables.
6. Variability of estimates: Interval estimates can be provided for the treatment and control differences using asymptotic properties or simulations so as to assess the variability.

7. CONCLUSION

We have presented a novel approach to response modeling in database marketing. This has improved the current methodology because it directly addresses the objective of maximizing the *true lift*. Specifically, the current methodology may find the customers (or prospects) who will take the desirable action regardless of the treatment. The proposed methodology, however, identifies the customers (or prospects) such that the incremental difference between treatment and control responses is maximized. In other words, we aim at finding the characteristics of customers (or prospects) whose campaign response decisions can be positively influenced by the treatment. This is particularly important for customer development campaigns (upselling and cross-selling).

The proposed methodology is simple and can be readily applied in conjunction with most commonly used linear or nonlinear modeling techniques for supervised learning such as linear regression, logistic regression, decision tree, spline regression, and neural network.

We hope that this paper will open a new line of research and thus will further benefit the database marketing industry.

8. ACKNOWLEDGMENTS

Many thanks to Johnny Shi and Jane Zheng for brainstorming ideas and implementing the proposed methodology, Kathleen Kane for pointing out that the current methodology may not solve the right business problem, and Florence H Yong for her valuable comments on a draft of this article.

9. APPENDIX – NAÏVE BAYES CASE

Naïve Bayes is a relatively simple data mining technique. However, by definition, it does not have a closed functional form. As a result, equation (6) cannot be explicitly written and thus equation (7) cannot be derived.

Assume that Y_i is a binary response variable representing whether customer i responds to a campaign, we consider the following set of p independent variables: $X_i = (X_{i1},\ldots, X_{ip})$' and T_i where $T_i = 1$ if i is in the treatment group and = 0 if i is in the control group. Additionally, $P(.)$ denotes a probability or probability distribution and $P(.|.)$ denotes a conditional probability or conditional probability distribution.

When no treatment effect is considered (subscript i is omitted for convenience), by Bayes' theorem,

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)} \propto \prod_{j=1}^{p} P(X_j \mid Y)\, P(Y), \qquad (14)$$

since $X = (X_1,\ldots, X_p)$', assuming conditional independence of $X_1,\ldots, X_p$ given Y, the key assumption of Naïve Bayes. Each component in equation (14) can then be estimated empirically from the training data set by generating cross-tabulations of independent variables (X_j's) by response variable (Y).

If the objective is only to classify new observations into $Y = 0$ or 1, then using (14), the predicted outcome is simply the one associated with the greater of $P(Y=1|X)$ and $P(Y=0|X)$, because the original dominator $P(X)$ does not depend on Y. In database marketing, we often would like to estimate the response probabilities, $P(Y=1|X)$. This can be achieved by normalizing:

$$P(Y=1 \mid X) = \frac{P(Y=1 \mid X)}{P(Y=1 \mid X)+P(Y=0 \mid X)} = \frac{\prod_{j=1}^{p} P(X_j \mid Y=1)\,P(Y=1)}{\prod_{j=1}^{p} P(X_j \mid Y=1)\,P(Y=1)+\prod_{j=1}^{p} P(X_j \mid Y=0)\,P(Y=0)}. \quad (15)$$

With the treatment effect, we aim at selecting the set of customers S that have the highest incremental differences in response probability between treatment and control (from objective function (3)):

$$Maximize \sum_{i \in S} \{P(Y_i=1 \mid X_i;T_i=1) - P(Y_i=1 \mid X_i;T_i=0)\}$$

Similar to (15), to estimate the probabilities in the above objective function, we have (omitting subscript i for convenience):

$$P(Y=1 \mid X;T=1) = \frac{J_1}{J_1+J_2}, \quad (16)$$

where $J_1 = \prod_{j=1}^{p} P(X_j \mid Y=1;T=1)P(Y=1 \mid T=1)$, and

$$J_2 = \prod_{j=1}^{p} P(X_j \mid Y=0;T=1)P(Y=0 \mid T=1);$$

$$P(Y=1 \mid X;T=0) = \frac{J_3}{J_3+J_4}, \quad (17)$$

where $J_3 = \prod_{j=1}^{p} P(X_j \mid Y=1;T=0)P(Y=1 \mid T=0)$, and

$$J_4 = \prod_{j=1}^{p} P(X_j \mid Y=0;T=0)P(Y=0 \mid T=0).$$

Each component in equations (16) and (17) can then be estimated empirically from the training data set, conditional on treatment or control. That is, instead of introducing a treatment/control dummy, T, we estimate the conditional response probabilities by generating '3-way cross-tabulations' of independent variables (X_j's) by response variable (Y) and by treatment versus control.

10. REFERENCES

[1] Almquist, E. and Wyner G. Boost your marketing ROI with experimental design. Harvard Business Review, Oct, 2001, p.135-141.

[2] Berry, M.J.A. and Linoff, G.S. Data Mining Techniques: For Marketing, Sales, and Customer Support, Wiley, 1997.

[3] Berry, M.J.A. and Linoff, G.S. Mastering Data Mining. Wiley, 2000.

[4] Blattberg, R.C., Getz, G., and Thomas J.S. Customer Equity: Building and Managing Relationships As Valuable Assets. Harvard Business School Press, 2001.

[5] Bose, N.K. and Liang P. Neural Network Fundamentals with Graphs, Algorithms, and Applications. McGrawHill, 1996.

[6] Breiman, L., Olson, R., Friedman, J., and Stone, C. Classification and Regression Trees. Chapman & Hall, 1984.

[7] Collett, D. Modeling Survival Data in Medical Research. Chapman & Hall, 1997, p.241-249.

[8] Dudewicz, E.J. and Mishra, S.N. Modern Mathematical Statistics. Wiley, 1988.

[9] Fabris, P. Advanced navigation: Marketing secrets from the financial sector show how data mining charts a profitable course to customer management. CIO magazine, 11, No.15, 1998, p.50-55.

[10] Goetghebeur, E. and Lapp, K. The effect of treatment compliance in a placebo-controlled trial: regression with unpaired data. Applied Statistics, 46, No.3, 1997, p.351-364.

[11] Guo, H. and Gelfand, S.B. Classification trees with neural network feature extraction. IEEE Transactions on Neural Networks, 3, No.6, 1992, p.923-933.

[12] Haykin, S. Neural Networks: A Comprehensive Foundation. Prentice Hall, 1999.

[13] Heckerman, D. Bayesian networks for data mining. Data Mining and Knowledge Discovery, 1, No.1, 1997, p.79-119.

[14] Henery, R.J. Combining classification procedures. In Machine Learning and Statistics, The Interface, by Nakhaeizadeh, G. and Taylor, C.C. (ed.). Wiley, 1997, p.153-177.

[15] Holmgren, E.B. and Koch, G.G. Statistical modeling of dose-response relationships in clinical trials – a case study. Journal of Biopharmaceutical Statistics, 7, No.2, 1997, p.301-312.

[16] Hosmer, D.W. and Lemeshow, S. Applied Logistic Regression. Wiley, 1989.

[17] Hughes, A.M. Strategic Database Marketing. McGrawHill, 1994.

[18] Inmon, W.H. and Hackathorn, R.D. Using the Data Warehouse. Wiley, 1994.

[19] Jackson, R. and Wang, P. Strategic Database Marketing. NTC Publishing, 1996.

[20] Jensen, F.V. An Introduction to Bayesian Networks. Springer, 1996.

[21] Koch, T.W. Bank Management, 3rd edition. Dryden, 1995, p.352.

[22] Kotler, P. Marketing Management, 8th edition. Prentice-Hall, 1994, chapter 24.

[23] Leonard, J.A., Kramer, M.A., and Ungar, L.H. Using radial basis functions to approximate a function and its error bounds. IEEE Transactions on Neural Networks, 3, No.4, 1992, p.624-627.

[24] Lerner, B., Guterman, H., Aladjem, M., Dinstein I. A comparative study of neural network based feature extraction paradigms. Pattern Recognition Letters 20, 1999, p.7-14.

[25] McKinney, R.E., Johnson, G.M., Stanley, K., Yong, F., Keller, A., O'Donnel, K.J., Brouwers, P., Mitchell, W.G., Yogev, R., Wara, D.W., Wiznia, A., Mofenson, L., McNamara, J., Spector, S.A., and the Pediatric AIDS Clinical Trials Group Protocol 300 Study Team. A randomized study of combined zidovudine-lamivudine versus didanosine monotherapy in children with sympotomatic therapy-naïve HIV-1 infection. Journal of Pediatrics, 133, No.4, 1998, p. 500-508.

[26] Mojirsheibani, M. Combining classifers via discretization. Journal of the American Statistical Association, 94, No.446, 1999, p.600-609.

[27] Montgomery, D.C. Design and Analysis of Experiments, 3rd edition. Wiley, 1991, p.79-80.

[28] Murthy, S.K. Automatic construction of decision trees from data: a multi-disciplinary survey. Data Mining and Knowledge Discovery, 2, 1998, p.345-389.

[29] Peppers, D. and Rogers, M. Enterprise One-to-One. Doubleday, 1999.

[30] Peppers, D. and Rogers, M. The One-to-One Fieldbook. Doubleday, 1999.

[31] Piatetsky-Shapiro, G. and Steingold, S. Measuring lift quality in database marketing. SIGKDD Explorations, 2, Issue 2, 2000, p.76-80.

[32] Roberts, M.L. and Berger P.D. Direct Marketing Management. Prentice-Hall, 1999.

[33] Rud, O.P. Data Mining Cookbook. Wiley, 2001.

[34] Russek-Cohen, E. and Simon, R.M. Evaluating treatments when a gender by treatment interaction may exist. Statistics in Medicine, 16, issue 4, 1997, p.455-464.

[35] Specht, D.F. A general regression neural network. IEEE Transactions on Neural Network, 2, No.6, 1991, p.568-576.

[36] Wahba, G. Spline Models for Observational Data. Society for Industrial and Applied Mathematics, 1992.

[37] Wang, P. and Puterman, M.L. Mixed logistic regression models. Journal of Agricultural, Biological, and Environmental Statistics, 3, No.2, 1998, p.175-200.

[38] Zhang, H. and Singer, B. Recursive Partitioning in the Health Sciences. Springer, 1999.

About the author:

Victor Lo is Vice President of Modeling at Fidelity Investments where he manages a team of data miners to support the retail marketing group. Previously, he was VP and Manager of Modeling and Analysis at FleetBoston Financial, a VP/Lead Analytic Consultant at Fleet Bank, and a Senior Associate at Mercer Management Consulting. In addition to analytics and management, his work includes bridging the gap between data miners, business analysts, and marketers by recommending and applying novel techniques and state-of-the-art tools to improve targeting and tailoring strategies.

Throughout Victor's career, he has applied techniques such as complex experimental design for conjoint-based surveys and direct marketing campaigns; cross-sectional time series regression for measuring advertising effectiveness; correspondence analysis for perceptual mapping and brand positioning; cluster analysis for segmentation using survey and behavioral data; simulation and sensitivity analysis for financial modeling; advanced statistical modeling for brand, pricing, and feature optimization using discrete choice analysis; hybrid modeling for customer long-term valuation; survival analysis for employee retention; and data mining techniques such as decision tree and neural network for database marketing.

Victor has a master degree in Operational Research from University of Lancaster, UK, and a PhD in Statistics from the University of Hong Kong. He also received postdoctoral training in Management Science from University of British Columbia, Canada. His previous academic research included applications of probability, statistical, and nonlinear optimization models in gambling strategies and quality engineering. Further, he co-authored a graduate level econometric book in horse-racing and published articles in Management Science and The Statistician.