INFLUENTIAL MARKETING: A NEW DIRECT MARKETING STRATEGY ADDRESSING THE EXISTENCE OF VOLUNTARY BUYERS

by

Lily Yi-Ting Lai
B.Sc., University of British Columbia, 2004

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

In the School
of
Computing Science

SIMON FRASER UNIVERSITY

Fall 2006

APPROVAL

Name: **Lily Yi-Ting Lai**

Degree: **Master of Science**

Title of Thesis: **Influential Marketing: A New Direct Marketing Strategy Addressing the Existence of Voluntary Buyers**

Examining Committee:

Chair: **Dr. Martin Ester**
Associate Professor of Computing Science

Dr. Ke Wang
Senior Supervisor
Professor of Computing Science

Dr. Jian Pei
Supervisor
Assistant Professor of Computing Science

Dr. S. Cenk Sahinalp
Internal Examiner
Associate Professor of Computing Science

Date Approved: Dec. 4th, 2006

ABSTRACT

The traditional direct marketing paradigm implicitly assumes that there is no possibility of a customer purchasing the product unless he receives the direct promotion. In real business environments, however, there are "voluntary buyers" who will make the purchase even without marketing contact. While no direct promotion is needed for voluntary buyers, the traditional response-driven paradigm tends to target such customers.

In this thesis, the traditional paradigm is examined in detail. We argue that it cannot maximize the net profit. Therefore, we introduce a new direct marketing strategy, called "influential marketing." To achieve the maximum net profit, influential marketing targets only the customers who can be positively influenced by the campaign. Nevertheless, targeting such customers is not a trivial task. We present a novel and practical solution to this problem which requires no major changes to standard practices. The evaluation of our approach on real data provides promising results.

Keywords: classification; direct marketing; supervised learning; data mining application

Subject Terms: Data mining; Business – Data processing; Database marketing; Direct marketing – Data processing

ACKNOWLEDGEMENTS

I would like to express my gratitude to my senior supervisor Dr. Ke Wang for his continuous guidance, patience, and support. He has shown me on many occasions the importance of bridging research and real world applications, for which I am grateful. In addition, I want to thank my supervisor Dr. Jian Pei for his insightful commentary and valuable input.

I am thankful to Daymond Ling, Jason Zhang, and Hua Shi who represent CIBC. Their expertise in direct marketing has helped this research tremendously. It was rewarding and intriguing to have the opportunity to learn the science behind direct marketing; it has certainly enriched my horizons.

Finally, I want to thank my family, and James. Without their continuous support, I would not be here today.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

Direct marketing is a marketing strategy where companies promote their products to potential customers via a direct channel of communication, such as telephone or mail. Unlike mass marketing, companies employing direct marketing target only a selected group of customers. For instance, a bank may decide to directly promote their first-time home buyer mortgage program to only newlywed customers. In accordance with the general principle of marketing, a direct marketing campaign strikes for the maximum net profit. Nevertheless, how does a campaign select which customers to contact so that it can achieve the maximum net profit?

Over the last decade, data mining has established itself as a solid research field. Its application spans across multiple disciplines, including economics, genetics, fraud detection, and so forth. Data mining focuses on the discovery of hidden patterns in data. This fits the purpose of direct marketing where companies need to study the underlying patterns of customers' purchasing behaviors based on a large set of historical data. As a result, data mining techniques have been extensively applied in direct marketing to determine the ideal targeting groups. Traditionally, such process involves three main steps:

1. Collect historical data from a previous campaign. Each historical customer sample is associated with a number of individual characteristics (e.g. age, income, marital status) and a response variable. The response variable indicates whether a customer responded after receiving the direct promotion.

2. Construct a data mining model based on the historical data. The objective is to estimate how likely a customer will respond to the direct promotion. Often, the response rate is low; for example, less than 3% is not unusual. Such a low response

rate imposes a certain degree of difficulty in the modeling process, often referred to as the *class imbalance problem.*

3. Deploy the model to rank all potential customers in the current campaign according to their estimated probability of responding. Contact only the highest ranked customers (i.e. those who are most likely to respond) in an attempt to achieve the maximum net profit.

Since the goal of the traditional direct marketing model is to identify customers who are most likely to respond to the promotion, it follows that the effectiveness of such a model, or campaign, is determined by the response rate of contacted customers. This evaluation criterion has long been adopted by numerous works in both academic and commercial settings [LL98, KDD98, Bha00, PKP02, DR94]. Intuitively, it seems that the more responders that exist among those contacted customers, the better — in other words, as long as a contacted customer responds, it is considered to be a positive result. However, is this really the case? Remember that ultimately, the goal of a direct marketing campaign is to maximize the net profit.

An implicit assumption made by the traditional direct marketing paradigm is that *profit can only be generated by a direct promotion.* In other words, it has been assumed that a customer would not make the purchase unless being contacted by the campaign. As such, how one would behave without the direct promotion is of no concern. However, we have to wonder if such an assumption holds in real life. It is not unrealistic to believe that some customers will make the purchase on their own without receiving the contact.

1.1 Motivation

The following example shows that if customers have decided to buy the product *before* the product is directly marketed to them, then the traditional objective does not address the right problem.

Example 1. John is 25 years old and recently got married. He and his wife have a joint account at Bank X. John, a newlywed, is planning to buy a house soon. He has decided to apply for a mortgage at his home bank Bank X after hearing great things about it from a good friend.

Applying traditional direct marketing strategies, Bank X discovered that young newlyweds are more likely to respond to the direct promotion on the bank's mortgage program. Therefore, the bank sent John a brochure about its mortgage program. Though it is true that John will respond to the direct promotion (brochure), he would have done so even without it. Therefore, from the bank's point of view, contacting John does not add any new value to the campaign — doing nothing will produce the same response from John. ■

There are two important observations from the above example. First, certain customers buy the product based on factors other than the direct promotion. Customers may voluntarily purchase due to prior knowledge about the product and/or the effect of *word-of-mouth* or *viral marketing* [DR01, KKT03]. We call such customers "voluntary buyers." For instance, John from Example 1 is a voluntary buyer who has a high natural response rate; he is a newlywed and has decided to apply for Bank X's mortgage program due to good word-of-mouth. Rather than contacting John, Bank X's promotion should have contacted customers with low natural response rates instead. This would have been more meaningful as those customers would only have considered purchasing *after* contact, unlike John. A classic example of viral marketing is Hotmail (http://www.hotmail.com). This free emailing service attaches an advertisement with every outgoing email message sent. Upon seeing the advertisement, recipients who do not use Hotmail may be influenced to sign up, further spreading the promotional message.

The second observation is that the traditional paradigm is response-driven and hence has the tendency to target voluntary buyers. As voluntary buyers always respond regardless of a contact, they have the highest response rates. Yet, this is a waste of resources because *no direct marketing is required to generate a positive response from such buyers.*

Therefore, in addition to avoiding non-buyers as in the traditional strategy, we advocate the significance of avoiding voluntary buyers. Essentially, a campaign should focus solely on those who will buy if and only if they are contacted — we believe that this is the right objective of a direct marketing campaign.

One question that arises is the following: how significant in practice is the portion of voluntary buyers? If it is insignificant, it may be acceptable to "push voluntary buyers through" to close the deal while focusing on avoiding non-buyers. To answer this question, a real campaign was carried out (see details in Chapter 7). Instead of contacting all selected customers, a random subset of those selected customers was withheld from contact. It turns out that while the contacted group had a response rate of 5.4%, the not-contacted group had a response rate of 4.3%. In other words, 80% of the responders contacted would have responded even without the contact! Aside from cost considerations, unnecessary promotions can potentially annoy customers and project a negative image of the company. In the worst case, they may lead customers to switch to a competing product or company. Clearly, unnecessary contacts to voluntary buyers incur both economic and social costs.

For direct marketing, the assumption that a purchase can only be generated by a direct promotion is too simplistic and does not reflect real world phenomena. Following such assumption will lead a campaign to be response-driven and, consequently, waste resources on voluntary buyers. In this thesis, we recognize the implications such an unrealistic assumption has on the field of direct marketing. Our research first conducts experiments on real campaign data following the traditional strategy. Then, we introduce a new strategy for direct marketing, called *influential marketing*. We will discuss our proposed solution to influential marketing in detail. Ultimately, the goal of influential marketing is still maximizing the net profit, except that now the existence of voluntary buyers is taken into consideration.

1.2 Contribution

The contributions of this thesis are outlined as follows.

1. Before introducing influential marketing, we first go through the traditional direct marketing paradigm to understand its principles firsthand. In the context of the traditional strategy, we examine the performances of two classifiers, the association-rule based algorithm (ARC) [WZYY05] developed by SFU and the decision tree in SAS Enterprise Miner [SAS]. The experiment was done using a real data set, as provided by our collaborative partner, the Canadian Imperial Bank of Commerce (CIBC). This part is also considered as an extension to ARC in which we compare the performance of ARC to other classifiers. CIBC produced their result as well on the same data set. The results produced by the three models are compared.

2. Based on purchasing behaviours, a new classification scheme for customers is introduced. All customers are classified into three classes: decided, undecided, and non. While decided and non customers have made up their minds on whether to buy the product, undecided customers will buy if and only if they are contacted. We argue that direct marketing should target only undecided customers. *Influential marketing* refers to this objective.

3. The major challenge is that undecided customers are not explicitly labeled. Therefore, standard supervised learning is not directly applicable. Our novel solution addresses this challenge while requiring no major changes to the standard campaign practice.

4. Using real campaign data, we compare our proposed solution with related work. The study shows that our approach is the most effective in terms of maximizing the net profit.

1.3 Thesis Organization

The remainder of the thesis is organized as follows.

In Chapter 2, we provide background information related to the work in this thesis.

In Chapter 3, we present the result obtained on real campaign data following the traditional direct marketing paradigm. We discuss the traditional approach in detail.

In Chapter 4, we introduce the new classification scheme for customers. We discuss in detail why the traditional paradigm does not solve the right problem. The definition of influential marketing is formally stated, with arguments given on why influential marketing has the correct objective to direct marketing.

In Chapter 5, we present our proposed solution to influential marketing. The solution covers data collection, model construction, and model evaluation. How to determine the optimal number of customers to contact is also discussed.

In Chapter 6, we compare our work with related work in the literature.

In Chapter 7, we compare three different approaches on a real campaign data. We show that our approach is the best at targeting undecided customers.

In Chapter 8, we provide suggestions for possible future work and summarize the work in this thesis.

CHAPTER 2
BACKGROUND

This Chapter provides background information on the important concepts related to the work presented in this thesis. Chapter 2.1 discusses classification in data mining. Chapter 2.2 looks at the class imbalance problem. An overview on the standard campaign practice is given in Chapter 2.3. In Chapter 2.4, the two classification algorithms used in the thesis are discussed.

2.1 Classification in Data Mining

Data mining is the process of extracting useful patterns or relationships from large data sets. Major sub areas of data mining include association rules, classification and prediction, and cluster analysis [HK01]. In this section, we give an overview on classification, which is the data mining technique most widely used for direct marketing. Our solution to influential marketing also relies on classification.

In classification, we wish to construct a model from a set of historical data. The model should describe a predetermined set of data *classes*. For example, in traditional direct marketing there are usually two predetermined classes, namely the "responder" class and the "non-responder" class. An *observation* or *sample* is a record representing an entity, e.g. a customer. Each observation is associated with a certain number of characteristic attributes and belongs to exactly one of the predetermined classes. The classification model aims to correctly assign each observation to its class. Classification is an example of *supervised learning* because the class label of each observation is known during the modelling process.

Two main steps are involved in classification [HK01]. First, in the *Learning* stage, a model is learnt from a subset of the historical data, called the *training set*. By analyzing

each sample in the training set, the model attempts to extract the patterns that differentiate the different classes. Many different techniques have been proposed for constructing such a classification model, including decision trees, Bayesian networks, neural networks, and so forth [HK01]. The second stage is *Classification*. In this stage, a subset of the data independent of the training samples, usually referred to as the *testing set*, is used to estimate the future performance of the model constructed in the first stage. It is imperative that the future performance of a model is estimated using a set of unseen data, as is the case in a real campaign. For each unseen observation, the class label predicted by the model is compared to the label as given in the data. The effectiveness of the model is judged by the evaluation criterion selected, e.g. accuracy of correct class prediction.

For a more reliable assessment of future performance, the k-fold cross validation [Sto74] is often applied. An advantage of this technique is that all samples in the data set are fully utilized. In a k-fold cross validation, the data is randomly separated into k partitions of equal size. In each of the k runs, $(k - 1)$ partitions are combined to form the training set and the remaining partition is held out as the testing set. This process repeats k times, each time with a different partition of training and testing sets. The average performance of the model on all k testing sets provides a more reliable evaluation than a single, random testing set.

In direct marketing, only a limited number of all potential customers will be selected for the direct promotion. As a result, the classification model is required to *rank* customers by how likely they belong to the class initiating the contact. Exactly how many will be selected in order to achieve the highest net profit depends on the performance of the model. For this reason, a classifier adopted for direct marketing should not only classify, but also classify with a confidence measurement for ranking observations. Most supervised learning algorithms are capable of such ranking or can be easily modified to do so.

2.2 Standard Campaign Practice for Direct Marketing

Generally, there are three main steps in the standard campaign practice for direct marketing regardless of the supervised learning algorithm or the evaluation criterion used. Below we describe the three steps.

1. Data Collection: No interesting patterns can be validly discovered without a set of historical data that is representative of the population of interest. Each observation in the historical data set should belong to exactly one of the predetermined classes. In direct marketing, such historical data is collected by observing the purchasing behaviours of customers from a previous campaign. Customers in the previous campaign may or may not have received the direct promotion. Whether a customer was to receive the direct promotion may have been randomly decided or by a data mining model. Each observation is associated with a number of attributes (e.g. age, income) plus a response variable. The response variable indicates whether one had responded in the previous campaign.

   A company that conducts a direct marketing campaign will set an "observation window," usually in the range of three to four months. Customers selected for observation will either receive or not receive the contact from the company at the beginning of the observation period. Customers that respond within the observation window will count as respondents for the campaign.

2. Model Construction and Evaluation: Once the historical data has been collected, the next step is model construction. While the actual construction of the model may differ by the supervised learning algorithm and evaluation criterion used, the general purpose is to predict the purchasing behaviours of customers. When more than one models are constructed, the model with the best performance during evaluation is selected for the campaign.

Model construction may also involve several data preprocessing steps such as treating missing values and noisy data, and reducing the number of attributes [HK01]. In this thesis, we do not consider the details of data preprocessing.

3. Campaign Execution (Model Deployment): Once the model is ready, the next step is to deploy the model in the current campaign. The model is applied to rank all potential customers by the predicted probability of belonging to the class initiating the contact. Only the top x% of the ranked list will receive the promotion (if the majority of potential customers are contacted, then direct marketing would not differ much from mass marketing). The selection of an optimal x, i.e. the x that produces the highest net profit, depends on the (predicted) performance of the model. In addition, if budget constraints apply, the selection of x should be realizable within the budget constraint. See more discussion on the optimal selection of x in Chapters 3.4 and 5.4.

2.3 The Class Imbalance Problem

Typically, the response rate in a direct marketing campaign is low. It is not unusual to see a response rate of less than 5%. As a result, the size of the "responder" class tends to be much smaller than the size of the "non-responder" class. Such situation where the class distribution is significantly skewed toward one of the classes is commonly known as the *class imbalance problem* [Jap00]. The more interesting class is usually the smaller class. Other examples of classification applications where class imbalance is common include the detection of oil spills in satellite images [KHM98], and the detection of various fraudulent behaviors [CS98, FP97, ESN96].

Research has shown that the issue of class imbalance hinders the performance of many classification algorithms [Jap00, Wei04, JAK01]. For instance, the decision tree C4.5 [Qui93] attempts to maximize the accuracy on a set of training samples. When the class distribution is skewed, simply classifying all samples into the majority class (e.g. the "non-responder" class) can achieve high accuracy. Typical solutions to the class

imbalance problem include under-sampling, over-sampling, and classification costs/benefits.

In under-sampling, instead of using all observations of the majority class to train the model, only a random subset of the majority class is used in addition to the minority class. Training samples of the majority class are randomly eliminated until the ratio of the majority and minority classes reach a preset value, usually close to 1. A disadvantage of under-sampling is that it reduces the data available for training. In over-sampling, training samples of the minority class is over-sampled at random until the relative size of the minority and majority classes is more balanced. Note that over-sampling may increase classification costs as it increases the size of the training set.

Another solution for the class imbalance problem considers the costs of misclassifications or similarly, the benefits of correct classifications. For example, MetaCost [Dom99] is a general framework for making error-based classifiers cost-sensitive, avoiding the tedious process of creating a cost-sensitive version for each individual algorithm. It incorporates a cost matrix $C(i, j)$, which specifies the cost of classifying a sample of true class j into class i. Instead of considering all samples as equal, a sample of the class of interest is assigned a higher value, i.e. the cost of misclassifying becomes higher. For a sample s, the optimal prediction is the class i that leads to the minimum expected cost

$$\sum_j P(j \mid s)C(i, j).$$

[ZE01] examines a more general case in which the cost of classification is dependent on each sample. The optimal predicted label for s is the class i that maximizes

$$\sum_j P(j \mid \mathrm{s})B(i, j, s),$$

where $B(i, j, s)$ represents the benefit of classifying s to class i when the true class is j.

2.4 The Supervised Learning Algorithms

In the experiment conducted for our work, two supervised learning algorithms are used. We discuss the two algorithms in this section.

2.4.1 The Association Rule Classifier (ARC)

The first supervised learning algorithm used is the association rule based classifier, or ARC, as proposed in [WZYY05]. ARC has been specially designed with the consideration of class imbalance and high dimensionality in mind (a data set incurs a high dimensionality when there are a large number attributes associated with the data, e.g. hundreds of attributes). Both issues are widespread in direct marketing. As suggested by its name, ARC first makes use of the association rule [AS94] to summarize the characteristics of the class of interest. Then it constructs a *covering tree* and performs pruning based on *pessimistic estimation* as in C4.5 [Qui93].

Since association rule mining is only applicable with categorical attributes, ARC requires all independent attributes that are continuous to be discretized. Then for each independent attribute A, there are a finite number of m categorical values or *items*, denoted $a_1, \ldots, a_m$ associated with A. The "positive class" refers to the class of interest (e.g. the "responders") and the "negative class" refers to the class with a low ranking (e.g. the "non-responders"). All observations should belong to either the positive class or the negative class.

ARC constructs the classification model first by generating a set of *focused association rules* (*FAR*). An *item* $A = a_i$ (item a_i of attribute A) is said to be "focused" if $A = a_i$ appears in at least p% of the positive class and no more than n% of the negative class. A FAR is a rule of the following form, where we use *f-item* to denote a focused item:

$$f\text{-}item_1, \ldots, f\text{-}item_k \rightarrow \text{positive}.$$

Only focused items can constitute the left-hand side of a FAR. At least p% of the positive samples should have all the items on the left-hand side; in other words, the *support* of a FAR in the positive class is at least p%. Essentially, the focused association rules concentrate on the common characteristics of the positive class which are rare in the negative class. This makes sense since the objective of the model is to identify characteristics exclusive to the positive class so that positive samples can be ranked higher than negative samples.

Let r denotes a FAR. $Supp(r)$ denotes the percentage of all observations containing both sides of the rule r. $lhs(r)$ denotes the set of *f-items* on the left-hand side of r, and $|lhs(r)|$ denotes the number of items in $lhs(r)$. A rule r is said to be *more general* than another rule r' if $lhs(r) \subseteq lhs(r')$.

Given the set of FARs based on the training set, the next step is to rank all FARs in order to construct a *covering tree*. In the order as described below, r is ranked higher than r' if,

- $O\_avg(r) > O\_avg(r')$, or
- $O\_avg(r) = O\_avg(r')$, but $Supp(r) > Supp(r')$, or
- $Supp(r) = Supp(r')$, but $|lhs(r)| < |lhs(r')|$, or
- $|lhs(r)| < |lhs(r')|$, but r is created before r'.

$O\_avg(r)$ is the average profit generated by all samples matching r. ARC thus is capable of handling direct marketing tasks where the amount of profit varies from customer to customer.

While a sample s may match many FARs, it has only one *covering rule* — the r that has the highest rank among all matching FARs of s. A rule r is useless and should be disregarded if it has no chance of covering any samples.

Once the set of rules is ranked, a covering tree can be constructed. In the covering tree, r is the "parent" of r' if r is more general than r' and has the highest rank. A child rule always has a higher rank than its parent; otherwise, the parent rule will cover all the samples matched by the child rule and the child rule is useless. The root of the tree represents the default rule,

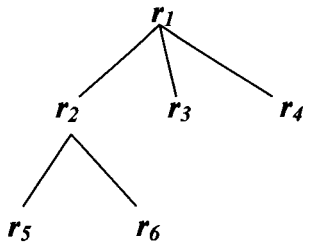
$$\phi \rightarrow \text{negative}.$$

An example of a covering tree is given in Figure 2.1. $A = a_1$, $B = b_2$, and $C = c_3$ are the focused items. To find the parent of r_5, we look at all the rules that are more general than

r_5, which are r_1, r_2, and r_3. Of the three rules, r_2 has the highest rank and therefore is the parent rule of r_5. Similarly, r_2 is the parent rule of r_6.

ID	Rules	Rank
r_1	$\phi \rightarrow$ negative	6
r_2	$A = a_1 \rightarrow$ positive	3
r_3	$B = b_2 \rightarrow$ positive	4
r_4	$C = c_3 \rightarrow$ positive	5
r_5	$A = a_1, B = b_2 \rightarrow$ positive	1
r_6	$A = a_1, C = c_3 \rightarrow$ positive	2

(a) The set of FARs.



(b) The covering tree based on (a)

Figure 2.1 Example of a covering tree.

To avoid overfitting, the covering tree is pruned. Suppose for each r (excluding the default rule), r covers M samples and E of them belong to the negative class. Then the estimated profit of r, denoted *Estimate*(r), is calculated as follows:

$$Estimated(r) = M \times (1 - U_{CF}(M, E)) \times O\_avg(r) - M \times U_{CF}(M, E) \times (cost\ per\ contact)$$

For the default rule, *Estimate*(r) = 0.

The estimated average profit for a non-default rule r, denoted $E\_avg(r)$, is $Estimated(r)/M$. The exact computation of $U_{CF}(M, E)$ can be found as part of the C4.5 code.

The pruning is done in a bottom-up fashion. At a tree node r, we compute the estimated profit for the entire subtree, $E\_tree(r)$; $E\_tree(r)$ is calculated by $\sum Estimated(u)$ over all nodes u within the subtree of r (including r). In addition, we compute $E\_leaf(r)$, the estimated profit of r after pruning the subtree; this can be done by assuming that r covers all the samples in its subtree. If $E\_tree(r) \leq E\_leaf(r)$, the subtree is pruned; otherwise, the subtree remains intact.

Take the example in Figure 2.1. $Estimated(r) = 0$, and $E\_leaf(r_1) = 0$. Suppose that $E\_tree(r_2) < E\_leaf(r_2)$, where $E\_leaf(r_2) = Estimated(r_2) + Estimated(r_5) + Estimated(r_6)$. Then the tree after pruning is shown in Figure 2.2.

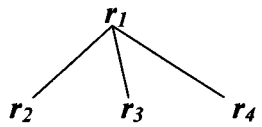


Figure 2.2 The covering tree after pruning.

Then the final model is given by the set of FARs remaining after pruning. In the above example, the final rules are $\{r_1, r_2, r_3, r_4\}$. To make prediction on an observation, the model returns the covering rule of the observation. If no positive rule is matched, then the default rule is returned as the covering rule. Observations can be ranked by the rank of the covering rule.

The remaining issue is on the selection of the minimum support for the positive class, p, and also the selection of the maximum support for the negative class, n. [WZYY05] recommends choosing n based on available computational resources. A smaller n will filter out more items, which allows the choice of a smaller p. Initially, n is set to the percentage of the positive class in data and p is set to 1% (as suggested by [WZYY05]). The optimal values of the two parameters are determined in a trial-and-error fashion. A random subset of the training data is withheld for tuning. Each run helps fine-tune the parameters until the best result, e.g. the highest net profit as produced by the tuning data, is reached.

2.4.2 The Decision Tree in SAS Enterprise Miner

Another supervised learning algorithm chosen for our experimentation is the decision tree option available in SAS Enterprise Miner [SAS]. We will refer to this algorithm as *SAS EM Tree*. With its comprehensive tools for data analysis, SAS is the leader in business intelligence solutions across various industries. SAS Enterprise Miner is one of the many

software packages available in SAS and offers tools that support the complete data mining process, ranging from data preparation, model construction/evaluation, to model deployment. In particular, our collaborative partner, the Canadian Imperial Bank of Canada (CIBC), uses SAS as their only business intelligence software for all aspects of data analysis. In this section, we discuss the basics of a decision tree.

A decision tree employs the divide-and-conquer approach by recursively partitioning the data into smaller subsets. With each partition, an input attribute A is chosen as the *test* and the current set of training samples is divided into subsets T_1, T_2..., T_n by the possible outcomes a_1, a_2, ..., a_n of A. For each partition to select the best *test*, the concept of *information gain* is used.

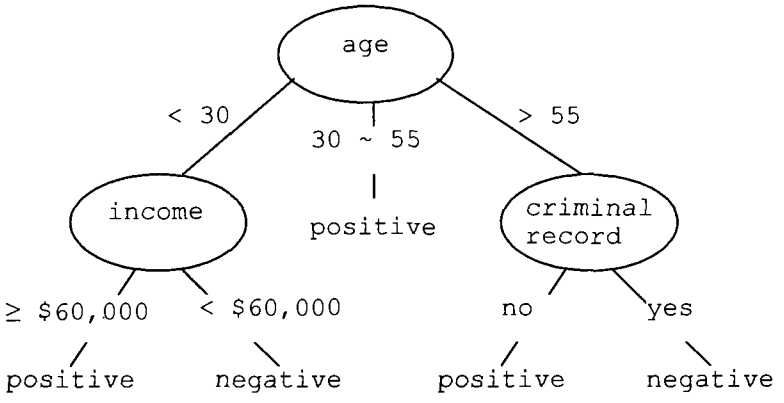


Figure 2.3 An example of a decision tree.

Suppose there are m distinct classes C_1, C_2, ..., C_m in the data. Let T denote the set of data at the current partition. The *entropy* of T, which measures the average amount of information needed to identify the class of a sample in T, is defined as follows:

$$Entropy(T) = \sum_{i=1}^{m} - p_i \cdot \log_2(p_i)$$

where p_i is the probability of an arbitrary sample belonging to class C_i.

Now, consider a similar measurement after T has been partitioned by the n possible outcomes of an independent attribute A. Then the entropy of T partitioned by A can be found as the weighted sum over the subsets, as

$$Entropy_A(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \times Entropy(T_i)$$, where $|T_i|$ is the size of set T_i.

Then the following quantity gives the information gain by partitioning T with the outcomes of A:

$$Gain(A) = Entropy(T) - Entropy_A(T).$$

In other words, $Gain(A)$ is the expected reduction in entropy if T is partitioned by A. For any partition, the attribute which maximizes the information gain is selected as the *test*. The recursive partitioning continues until all the subsets consist of samples belonging to a single class (or other stopping criterion as specifically set by the algorithm, e.g. stop when the number of observations in the current partition is less than n).

Overfitting can happen when branches of the tree reflect anomalies in the training data due to noise or outliers. To avoid overfitting, pruning is applied. Suppose a leaf covers N samples and E of them are classified incorrectly. For a given confidence level CF, the upper limit of the probability of an error in this leaf can be found from the confidence limits for the binomial distribution, written as $U_{CF}(E, N)$.

The predicted error rate at a leaf is given by $U_{CF}(E, N)$. Then the predicted number of errors at a leaf covering N training examples is then given by $N \times U_{CF}(E, N)$. The predicted number of errors for a subtree is the sum of its predicted errors of its branches. A node will be pruned if removing the node leads to a smaller predicted number of errors; otherwise, it is kept. The set of logic statements, or rules, derived from the pruned decision tree gives the final classification model. The predicted label/class for a leaf is the class covering the majority of the samples.

A simple modification can be made for a decision tree to perform ranking. Specifically, observations can be ranked by the confidence of the matched rule, which is usually computed as the percentage of positive samples in the matched leaf. Such ranking is available in SAS EM Tree.

CHAPTER 3
THE TRADITIONAL DIRECT MARKETING PARADIGM

Before introducing influential marketing, we first focus on the traditional direct marketing paradigm which is still widely adopted in the industry. As part of our research, we experimented with two supervised learning algorithms, ARC and SAS EM Tree, following the traditional strategy. We applied the two algorithms on a real data set provided by the Canadian Imperial Bank of Canada (CIBC). In addition, a third model based on the same data set was constructed "in-house" by CIBC.

Recall that a traditional direct marketing model has the objective of identifying the customers who are most likely to respond upon contact; as a result, the models are evaluated by the response rate of contacted customers.

3.1 The Data Set

The data set we received from CIBC has 304,698 observations. Each observation represents a customer and is described by 407 variables, both categorical and numerical. The variable "index" gives the ID of the customer and the variable "target" describes whether the customer had responded to the previous campaign. There are a lot of missing values. Detailed descriptions on the meaning of each variable were not provided. All of the observations had received the direct promotion.

Based on the responses, there are two classes — the positive class consists of the responders, and the negative class consists of the non-responders. The response rate is very low, at a mere 1.13%. Thus, the class distribution is extremely skewed, suggesting that many supervised learning algorithms may not perform well without the training data being sampled.

3.2 The Supervised Learning Algorithms

In Chapter 2.4, we discussed in detail the approach of each of the two supervised learning algorithms. We now look at the reasoning for our choices. For ARC, we also discuss the modification made for our experiment. Essential information on the third model developed by CIBC is also presented.

3.2.1 The Association Rule Classifier (ARC)

One of the main reasons for choosing ARC is for its superior ability at handling imbalanced class distributions. It utilizes the association rule mining, making sampling unnecessary in many cases otherwise requiring sampling. In [WZYY05], ARC has been shown to produce the best result among many algorithms on the data set used for KDD-98 [Kdd98], which has a skewed class distribution. In addition, ARC can handle high dimensionality (the data set has more than 400 variables) without a considerably long running time.

Yet another significant advantage of ARC lies in the expressiveness of the model constructed. Each final rule is expressive in itself. For instance, a rule may indicate that "*if* a person makes more than $50,000/year and is between the age of 25 and 40, *then* he is a good candidate to contact for insurance policy No.23007." For a direct marketing campaign (and for many other purposes where data mining is useful), it is very important that the constructed model can be easily interpreted by human. In the banking industry, for example, a data mining model not only can be used for the particular marketing campaign, but other aspects of the customer relationship management (CRM), e.g. in-person consultation in the bank, can also benefit from understating the patterns of customer behaviors.

Other supervised learning algorithms, such as SVM [Joa99] and multilayer neural networks, may also perform well; however, it is difficult for a person to interpret the models produced by these algorithms. While the models may be deemed "effective" based on the evaluation criterion, not being expressive in the form of association rules is a great disadvantage in direct marketing.

In our experiment, we have made two slight adjustments to ARC as it was originally presented. In the original context where ARC was developed, the objective of the model was to target customers with high expected profits. It was pointed out that there are often many responders that bring small revenues, whereas few responders bring large revenues. By ranking the rules using $O\_avg(r)$, ARC is able to identify the small group of more precious customers associated with larger profits. In our experiment, rather, the focus is on whether a customer will respond upon contact. Essentially, all responders are considered equal and each is associated with the same amount of expected profit. Therefore, the first adjustment is to uniformly set the expected profit of each responder to \$1.00 (the expected profit of each non-responder remains at \$0.00). The cost per contact is set at \$0.00. Then it follows that for a rule r, the original $O\_avg(r)$ now gives the confidence of r.

The *confidence* of a rule r is defined as:

$$\frac{\#\ of\ appearances\ of\ lhs(r)\ in\ the\ positive\ class}{\#\ of\ appearances\ of\ lhs(r)\ in\ the\ data}.$$

Recall that $lhs(r)$ denotes the set of focused items on the left-hand side of r.

In other words, a rule with a higher confidence is ranked higher. This matches the purpose of the traditional paradigm as the model targets the customers with a higher chance of responding, i.e. those matching rules with higher confidences.

The second adjustment made to ARC deals with the initial filtering of items. Recall that n is the maximum support for the negative class, and that an item with a support more than n in the negative class is removed in the first stage of the algorithm. In our experiment, we replace n with c, in which an item is filtered in the initial stage *if the confidence of the item is less than c% in the data.* For an item, its confidence is calculated as the following:

$$\frac{\#\ of\ appearences\ of\ the\ item\ in\ the\ positive\ class}{\#\ of\ appearences\ of\ the\ item\ in\ the\ data}.$$

Note that the parameter c takes into account the occurrence of an item in the negative class relative to the positive class, whereas n only considers the number of appearances in the negative class. As an example, suppose positive samples consist of 5% of the entire data set. An item, $A = a_1$, has appeared in 8% of the negative class. If n is used and set to 5%, which is the recommended value to start with, then $A = a_1$ will be filtered out since it appears in more than 5% of the negative class. However, we note that $A = a_1$ actually has a confidence of 12%, more than double of the positive rate in the data — this suggests that the item should help discriminate between the classes. The use of c instead of n will not filter out a potentially useful item such as $A = a_1$. We suggest initially setting c to about twice the percentage of the positive rate in the data and gradually decreasing its value during the tuning process until the result does not improve significantly.

3.2.2 The Decision Tree in SAS Enterprise Miner (SAS EM Tree)

SAS is the only data analysis software used by our collaborative partner CIBC. For this reason, we wanted to apply SAS Enterprise Miner in our experiment. We use the decision tree algorithm, one of the most recognized supervised learning algorithms in data mining, from SAS Enterprise Miner to construct the model. The model constructed has great interpretability, expressing itself in terms of association rules.

3.2.3 The Model Constructed by CIBC

The third model is constructed by CIBC, following their usual data mining procedure for a direct marketing campaign. The modeller at CIBC trained the model, and then passed on to us the validation result. The supervised learning algorithm of their choice is linear logistic regression. A considerable amount of time is spent on data preparation, including variable selection, variable transformation, and variable imputation.

The first step in their modelling procedure is to significantly reduce the number of input variables by performing exploratory analysis on the data. This is done by having the modeller examine the relationships between the different variables and the target variable using the analysis tools provided by SAS. The modeller is also responsible for truncating

outliers, imputing missing values, adding indicators to represent missing values, transform variables, and so forth.

Knowledge on the meaning of each variable also plays a part in the modelling process. Particularly, for this data set the modeller at CIBC has modelled "visa-only customers" and "non visa-only customers" separately based on her knowledge in the banking business.

When working with ARC and SAS EM Tree, we do not manually perform variable selection, variable imputation, etc.

3.3 Experimental Results

Using ARC and SAS EM Tree, we performed the experiments using 5-fold cross validation. The original data set was randomly partitioned into five disjoint subsets. In each of the five runs, four of the subsets are combined to form the training set while the remaining subset forms the testing set. The result reported for each of ARC and SAS EM Tree is the average result of the five runs. In according with the traditional direct marketing paradigm, the positive class consists of the responders and the negative class consists of the non-responders.

The result from each of the three models is shown in Figure 3.1. The x-axis represents the different percentiles x, as mentioned in Chapter 2.2. The y-axis represents the response rate. For example, a point (30%, 2%) in Figure 3.1 means that the response rate of the top 30% customers ranked by the model is 2%.

"ARC" denotes the model constructed using ARC. "SAS EM Tree" denotes the model constructed using the decision tree in SAS Enterprise Minder, and "CIBC" denotes the model constructed in-house by CIBC.

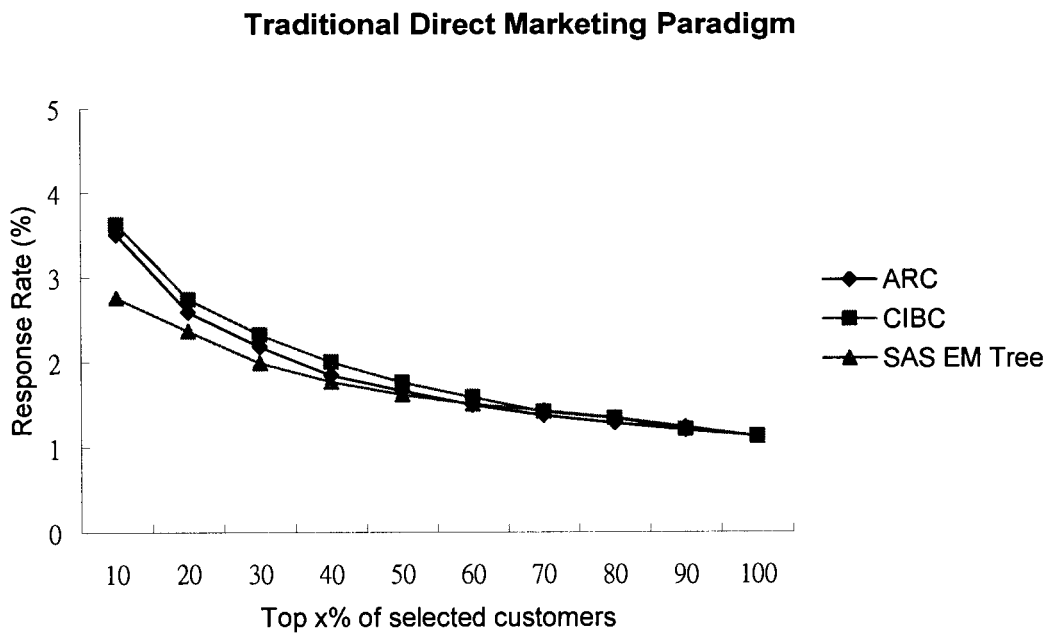


Figure 3.1 Comparison of Models – The Traditional Paradigm

3.3.1 Model: ARC

We implemented all stages of ARC using the programming language Java. Categorical attributes are first discretized using the public software MLC++ [KSD97]. A random 30% of the training set is held for tuning while the model is trained using the remaining 70%. We take the model that produces the highest lift index [LL98] in the tuning data as the best model. The lift index calculates the weighted average response rate of each decile, where the top deciles weigh more than the bottom deciles.

A systematic tuning is performed starting with c at 2.5%, which is slightly larger than twice the rate of the positive class in the data. The best result obtained on the tuning set was when c equals 2% and s, the minimum support of a rule in the positive class, equals 0.3%. The model is then applied to the testing set. The average result on all five testing sets is reported by "ARC" in Figure 3.1.

3.3.2 Model: SAS EM Tree

Unlike ARC, the decision tree does not perform well when the class distribution is highly skewed. The response rate of 1.13% in the original data set is too small for the decision tree to handle. Without any sampling, the classifier will simply classify all samples as negative and at the same time achieve an accuracy of nearly 100%.

In order to apply SAS EM Tree on our data set, we performed under-sampling on the negative class. The under-sampling was done at different rates so that the positive class is at 10, 20, 30, 40, and 50% of the entire training data (instead of the original 1.13%). The best result was obtained at the rate of 30%, as shown by "SAS EM Tree" in Figure 3.1.

3.3.3 The Reported Result from CIBC

CIBC also reports their result following the modelling procedure described in Chapter 3.2.3. The result is reported by "CIBC" in Figure 3.1.

3.4 Discussion

From Figure 3.1, we observe that the models produced by ARC and CIBC have similar performances. They both are able to target responders, as observed by the decreasing trend of the response rate along the x-axis. As an example, the predicted response rates for both models at the 10%-percentile is more than 3.5%, which is more than triple the random response rate of 1.13%. On the other hand, SAS EM Tree performed less favourably, achieving a response rate of less than 3% in the 10%-percentile.

One thing to note here is the complexity of the modelling procedure. While both ARC and SAS EM Tree run autonomously, the procedure adopted by CIBC is more time-consuming and requires extensive manual labour. When considering this, the fact that the result produced by ARC is comparable to the result produced by CIBC is rather impressive.

Recall that a direct marketing campaign needs to identify the optimal percentile x which maximizes the net profit within the budget constraint. Let C be the cost per contact, and R

be the revenue per purchase. Let P be the set of potential customers and $|P|$ be the number in of customers in P. We use RR_x to represent the response rate at percentile x. For example, according to Figure 3.1, $RR_{10\%}$ for "ARC" is 3.5%. Then at percentile x, $|P|*x$ is the number of customers to contact and $|P|*x*RR_x$ is the number of responders. Following the traditional direct marketing paradigm, the net profit at x is calculated as follows.

$Net\ profit(x)$ = total revenue – total cost = $|P|*x*RR_x*R - |P|*x*C$.

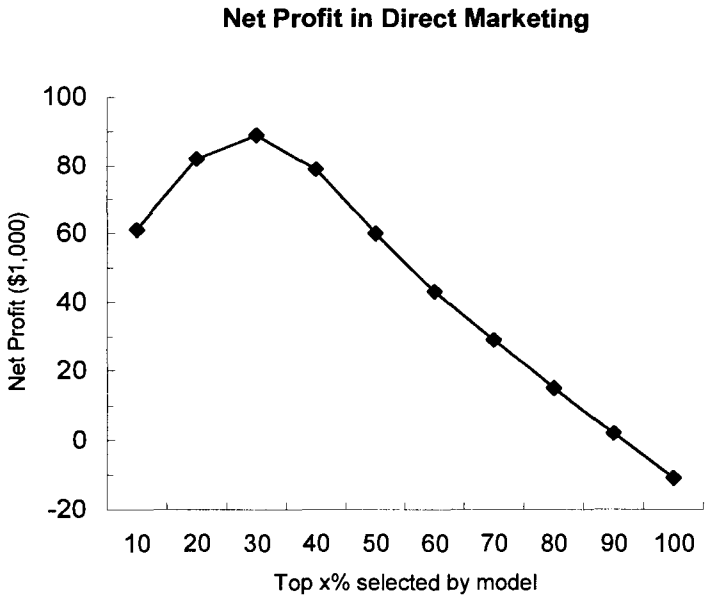


Figure 3.2 Net profit in direct marketing.

Based on the response rates at the different percentiles, we can produce a plot by having the different percentiles in the x-axis and their corresponding calculated net profits in the y-axis; an example of such a plot is shown in Figure 3.2. In Figure 3.2, we observe that the highest net profit is achieved at x = 30%. Therefore, the campaign shall contact 30% of the potential customers to achieve the maximum net profit.

We have experimented with the traditional direct marketing paradigm by applying supervised learning algorithms on a real set of data. We have also discussed the selection

of the optimal x percentile following the traditional strategy. The key observations from this Chapter are summarized below.

1. In the traditional direct marketing paradigm, a contact is deemed profitable if the contacted customer responds. In other words, it assumes that all positive responses are generated by the direct campaign.

2. Based on the assumption in 1, a model of the traditional paradigm targets the customers who are most likely to respond to the direct campaign. Therefore, how the customers would behave in the absence of the contact is not considered. This perception is reflected in every aspect of the modelling procedure. For example, the historical data is often collected on only those who have been contacted. Or, during model construction, the positive class consists solely of the customers who responded to the direct promotion.

3. The model with the highest response rate *among contacted customers*, especially in the top percentiles, is selected to be the best model. Similarly, the optimal marketing percentile for a campaign is selected as the one that achieves the highest net profit *among contacted customers*.

For the remainder of the thesis, we will examine the validity of the traditional approach. In particular, we attempt to answer the following question, "Can traditional direct marketing really maximize the net profit?"

CHAPTER 4
INFLUENTIAL MARKETING

Consider a pool P of potential customers. Ultimately, a direct marketing campaign aims to maximize the net profit over P. As is the case of many campaigns, we assume that each customer purchase generates a fixed revenue R and that contacting each customer incurs a fixed cost C, where $R > C$. Then the net profit of each purchase is $R - C$. The *set of buyers over P* consists of all the customers who eventually buy the product, either voluntarily or due to the direct promotion. Clearly, to maximize the net profit for a fixed number of contacts, we need to maximize the set of buyers over P.

4.1 The Three Classes of Customers

In real business environments, almost certainly there will be customers who base their purchasing decisions on factors other than a direct promotion. Traditionally, all customers are classified into either one of the following two categories: the "responder" who responds to the direct promotion, and the "non-responder" who does not respond to the direct promotion. The implicit assumption is that all responses were solicited by the campaign.

However, many of the responders may have in fact responded voluntarily. This observation prompts us to propose a new classification of customers that is based on their purchasing behaviors. The focus here is to differentiate between those who voluntarily buy the product from those who will only purchase if contacted. We categorize all customers into three classes, as described below.

- **Decided** – the customers who voluntarily buy the product, regardless of a direct promotion.
- **Undecided** – the customers who buy the product if and only if the product is directly promoted to them. This is the only group of customers whose purchasing decision can be positively influenced by a direct marketing campaign.
- **Non** – the customers who will not buy the product, regardless of a direct promotion.

Each customer belongs to exactly one of these classes. Although we refer to the classes of customers, our solution does not require the explicit determination of the class of each customer.

4.2 Influential Marketing

Based on the three classes of customers, the set of buyers for a campaign now consists of the following two parts: (i) All decided customers in P (decided customers always make the purchase), and (ii) The undecided customers in P who are contacted by the campaign (undecided customers who are contacted will also make the purchase).

Note that the number of customers in (i) is the same across all campaigns. However, the number of customers in (ii) depends on whom the campaign selects to contact.

We define the following variables.

- ***S***: the set of customers contacted by the campaign,
- ***D***: the set of decided customers in S,
- ***U***: the set of undecided customers in S.

Additionally, let $|X|$ denotes the number of customers in set X. For example, $|D|$ represents the number of decided customers in S. We now introduce three important definitions that will be used throughout this thesis.

Definition 1 (*RR*) *RR* is the *response rate* of S, calculated by $\frac{|D|+|U|}{|S|}$.

Definition 2 (*DBR*) *DBR* is the *decided buyer rate* of S, calculated by $\frac{|D|}{|S|}$.

Definition 3 (*UBR*) *UBR* is the *undecided buyer rate* of S, obtained by $\frac{|U|}{|S|}$.

Note that *RR*, *DBR*, and *UBR* refer to only the buyers who are contacted by the campaign, and that $RR = DBR + UBR$.

As we have discussed, the traditional strategy of direct marketing aims to maximize *RR*. Yet, we argue that maximizing *RR* does not maximize the net profit as doing so cannot maximize the set of buyers over P. The following example illustrates this.

Example 2. Let P be a pool of 100 customers, of which 20 customers are decided, 15 customers are undecided, and 65 customers are non. We consider two direct marketing campaigns, M1 and M2. Each campaign selects and contacts 20 customers from P. Therefore, $|S| = 20$ for both campaigns. In M1, of the 20 customers contacted, 7 are decided, 2 are undecided, and 11 are non, i.e. $|D| = 7$ and $|U| = 2$. Then we have $RR = \frac{7+2}{20} = \frac{9}{20} = 45\%$, $DBR = \frac{7}{20} = 35\%$, and $UBR = \frac{2}{20} = 10\%$. The set of buyers over P consists of the 20 decided customers in P and 2 undecided customers contacted by the campaign, for a total of 22.

In M2, of the 20 customers contacted, 6 responded. Among the 6 responders, 1 is decided and 5 are undecided. The other 14 contacted customers are non and therefore did not respond. Then $RR = \frac{6}{20} = 30\%$, $DBR = \frac{1}{20} = 5\%$, and $UBR = \frac{5}{20} = 25\%$. The set of buyers over P consists of the 20 decided customers in P and 5 undecided customers contacted by the campaign, for a total of 25. ■

Figures 4.1 and 4.2 illustrate each of the two campaigns M1 and M2. In the figures, the largest rectangle represents the set of potential customers *P*, which consists of a fixed number of decided, undecided, and non customers. The dashed triangle represents *S*, the set of customers contacted by each campaign. Recall that both campaigns contact the same number of customers.

In Figure 4.1, the gray area highlights the set of customers that contributes to *RR*. We see that M1 has a higher *RR* than M2 (45% vs. 30%, as in Example 2; also shown by the larger highlighted area).

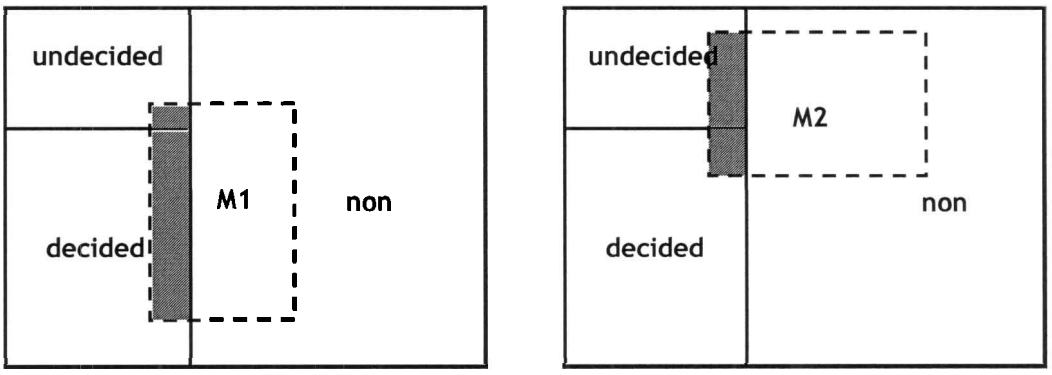


Figure 4.1 Illustration of the set of buyers over *S* for M1 and M2.

However, a closer examination indicates that M1 actually produces a smaller set of buyers over *P*. This is illustrated in Figure 4.2, where the set of buyers is highlighted in gray.

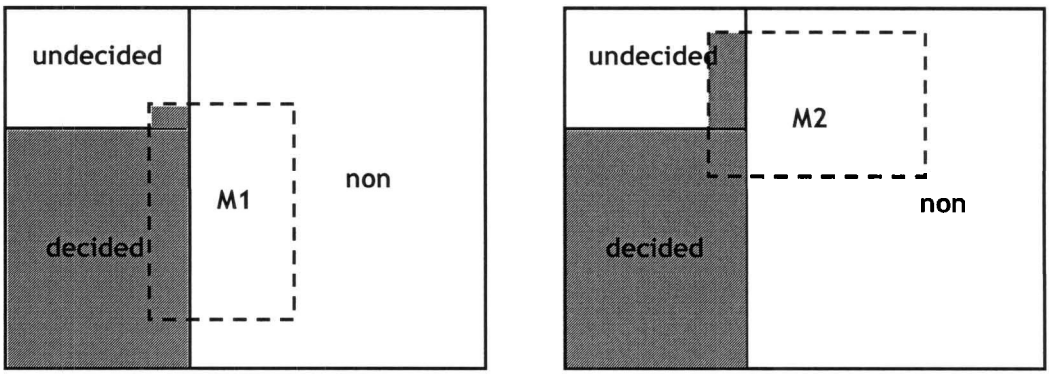


Figure 4.2 Illustration of the set of buyers over *P* for M1 and M2.

Why is it that a campaign can achieve a higher *RR* but a smaller set of buyers over *P* (and hence, a lower net profit)? The key point lies in the number of undecided customers targeted. Note that the majority of responders targeted by M1 are actually decided customers whose purchasing decision is positive regardless of the direct promotion — contacting such customers does not increase the total number of buyers. On the other hand, M2 has targeted more undecided customers who can actually be positively influenced.

If we were to follow the traditional objective of maximizing *RR*, then M1 is preferable to M2. However, M2 actually produces a larger set of buyers, and hence, a larger net profit. In fact, M2 should be preferable to M1.

This analysis clearly illustrates a key point: for a fixed number of contacts, the only way to generate a higher net profit is to contact undecided customers. Instead of indiscriminately targeting all likely responders, we want to target only those who can be

positively influenced. This establishes the core concept of *influential marketing*, as defined below.

Definition 4 (Influential Marketing) For a given number of contacts, *influential marketing* aims to maximize *UBR* by targeting undecided customers. ■

As the traditional paradigm targets decided customers to maximize *RR*, it misses the opportunity of converting undecided customers into buyers. The traditional paradigm will only maximize the number of buyers over *P* if there are no decided customers (i.e. the only possible buyers are the undecided customers). Essentially, influential marketing generalizes traditional direct marketing in the presence of voluntary buyers.

4.3 The Challenges

The main challenge of influential marketing is that undecided customers were not explicitly labeled in the historical campaign data. When a contacted customer responds, she could be either a decided or undecided customer; we do not know which. Therefore, supervised learning where the positive and negative classes are explicitly labeled is not directly applicable. Another challenge is that given their existing investments, campaign managers generally do not wish to switch to a completely different set of marketing methodologies. This consideration is particularly important in the business environment where deliveries have a priority over publications. A solution that addresses these requirements must be innovative in that it not only identifies undecided buyers, but also requires little changes to the standard campaign practice. We present a solution with this goal in mind.

CHAPTER 5
PROPOSED SOLUTION

We now present our solution to influential marketing. We discuss in detail how historical data should be collected, and how to construct and evaluate a model that ranks undecided customers.

5.1 Data Collection

How to collect data so that patterns of undecided customers can be learned presents certain challenges. Since the traditional strategy focuses on contacted customers, historical data often had been collected on only contacted customers. This does not allow us to study how the customers would behave in the absence of the direct marketing campaign.

Recall from Chapter 4.2, for the set of contacted customers S, $RR = UBR + DBR$. Unfortunately, while RR can be directly computed from S, the direct computation of UBR and DBR is not possible because customers are not explicitly labeled by the three classes. However, if we can estimate DBR, then we will be able to compute UBR by $RR - DBR$. Suppose we have a set of customers *similar* to S, denoted S_{sim}. While the customers in S are contacted, the customers in S_{sim} are not contacted. It follows that RR of S_{sim} can be used to approximate DBR of S because (a) similar sets of customers tend to behave similarly [Mon91], meaning that the two sets should have similar DBR, and (b) for S_{sim}, $RR = DBR$ (all purchases must have been voluntary since none of the customers were contacted).

The idea is similar to a clinical trial. For example, the testing of a new drug is commonly conducted with two groups. The two groups are usually denoted as "Treatment" and "Control," where individuals in both groups have similar backgrounds, e.g. similar

medical histories. Those in Treatment are subjected to the new drug while those in Control are not (they may receive a placebo instead). The effectiveness of the drug is then measured on the Treatment group, while the Control group establishes a baseline for comparison.

In the context of our work, the *treatment* received by a subject (customer) is the direct promotion. Based on this observation, we collect two disjoint sets of data, Treatment and Control, from the previous campaign.

- **Treatment**: a set of customers who were contacted.
- **Control**: a set of customers who were not contacted. The purchasing behaviours of Control are used to approximate those of Treatment under the alternative marketing decision that Treatment was not contacted.

In order to have the purchasing behaviours of Control to approximate those of Treatment, the two groups need to have a similar distribution. This can be done by ensuring that Treatment and Control came from *the same underlying population.* Suppose the training data came from a previous campaign that had selected some customers. Instead of contacting all the customers selected, the campaign could reserve a subset of the selected customers for Control. In particular, for each customer c selected by the campaign, toss a $\rho/(1-\rho)$-sided coin to determine whether c was added to Treatment (with probability ρ) or added to Control (with probability $1-\rho$). ρ is the *treatment/control rate* that specifies the relative size ratio of Treatment and Control. Since Treatment/Control is a random split of all selected customers, they come from the same underlying population. As usual, the customers in Treatment were contacted and their purchasing behaviours were recorded. The customers in Control were not contacted but their purchasing behaviours were also recorded.

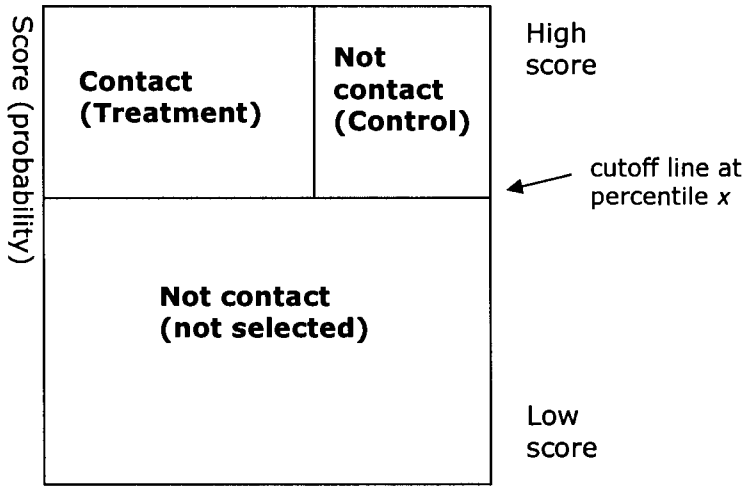


Figure 5.1 Illustration of data collection.

With the same split ratio, we randomly split Treatment into two sets, *T1* and *T2*, and randomly split Control into two sets, *C1* and *C2*. <*T1*, *C1*> forms the *training set* and is used for model training. <*T2*, *C2*> forms the *validation set* and is used for model evaluation.

5.2 Model Construction

We use the training data <*T1*, *C1*> to construct a model for targeting influential marketing. As stated, the major challenge is that undecided customers are not explicitly labeled. Nevertheless, if we know whether a customer received the promotion and made the purchase, we can deduce the possible classes the customer belongs to from a total of three down to one or two. For example, if we know that someone has received the promotion but did not respond to it, then she must belong to the non class.

Based such observation, the first step of our proposed solution is to partition the samples in the training set <*T1*, *C1*> into four groups, as shown by the *learning matrix* in Table 5.1:

Table 5.1 The learning matrix.

		Response	
		Yes	**No**
Treatment (*T1*)		decided + undecided (1)	non (2)
Control (*C1*)		decided (3)	non + undecided (4)

- Group (1): The buyers in *T1*, i.e. the customers who were contacted and responded. These customers can be either decided or undecided; we do not know which.

- Group (2): The non-buyers in *T1*, i.e. the customers who were contacted but did not respond. These customers must be non.

- Group (3): The buyers in *C1*, i.e. the customers who purchased the product without being contacted. These customers must be decided.

- Group (4): The non-buyers in *C1*, i.e. the customers who received no contact and made no purchase. These customers can be either non or undecided; we do not know which.

For influential marketing, the task of the data mining model is to recognize the characteristics of undecided customers. However, at first glance it is not clear how such a model can be constructed as the group of undecided customers is not clearly separated from the rest of the non and decided customers.

Based on the learning matrix and its four groups, our insight for the model construction is as follows. First, we observe that undecided customers are present only in groups (1) and (4). This motivates us to combine (1) and (4) to form the positive class, denoted *PC*, which covers all undecided customers. Nonetheless, *PC* also includes some decided and non customers. To isolate the characteristics of undecided customers (or to "remove" the characteristics of decided and non customers) in *PC*, we combine (2) and (3) to form the negative class, denoted *NC*. It follows that the only type of customer that appears in *PC* but not *NC* is the undecided customer. Therefore, if we apply standard supervised learning with *PC* and *NC*, the algorithm shall have the tendency *to pick out the characteristics exclusive of PC*. Such characteristics are likely those of undecided customers.

This approach is summarized in Figure 5.2. It takes a supervised learning algorithm *Alg* as a parameter. *Alg* should be able to rank samples by their probability of belonging to *PC*.

Model Construction

Input: The training data <*T1*, *C1*>

Parameter: A supervised learning algorithm *Alg*.

Output: A model that ranks samples by the probability of belonging to the undecided class.

1. Combine (1) and (4) into the positive class *PC*;
2. Combine (2) and (3) into the negative class *NC*;
3. Feed *PC* and *NC* into *Alg*;
4. Return the model learnt by *Alg*.

Figure 5.2 Model construction.

Alg will rank a customer *c* higher if *c* has a higher probability of belonging to *PC*. From the above analysis, we know that the only class of customer belonging to *PC* but not *NC* is the undecided. Therefore, a higher rank given by *Alg* indicates a higher probability of being undecided.

Heuristic 1. Assume that the supervised learning algorithm *Alg* is capable of ranking customers by their probability of belonging to the positive class. Then for the model returned by Figure 5.2, if customer *c1* is ranked higher than customer *c2*, *c1* has a higher probability of being undecided than *c2*. ■

Our construction of *PC* and *NC* has an extra benefit: it lessens the impact of the class imbalance problem. The typically low response rate in a direct marketing campaign means that the size of group (1) tends to be much smaller than the size of group (2), and the size of group (3) tends to be much smaller than group (4). Our approach combines the small (1) with the large (4) to classify against the combination of another small (3) and large (2), effectively lessening the situation of class imbalance. This is experimentally shown in Chapter 7.

Furthermore, the learning matrix provides a flexible way for the campaign to focus on the different class of customers. Typically, the size of (3) is much smaller than the size of (4) due to the vast majority of non customers. When targeting undecided customers as in influential marketing, we may want to over-sample (3) to emphasize the characteristics of decided customers in *NC* — this should help the "removal" of decided customers in *PC*. On the other hand, when the focus is on all responders, either undecided or decided as in the traditional paradigm, we may want to under-sample (4) to limit the number of non customers in *PC*.

5.3 Model Evaluation

Recall that a direct marketing campaign contacts the top *x*% customers as ranked by the model. We refer to *x* as the *marketing percentile*. The choice of *x* should maximize the net profit while staying within the budget constraint.

Suppose a model M has been constructed from the training data <*T1*, *C1*>. To evaluate the effectiveness of M, we will apply M on the validation data <*T2*, *C2*>. Recall that while *T2* and *C2* came from the same underlying population, *T2* was contacted and *C2* was not contacted.

To perform the evaluation, we first need to select an evaluation criterion. In the traditional paradigm, the most common evaluation criterion is the response rate of contacted customers (as in the experiment conducted in Chapter 3). However, we have shown in Chapter 4 that *RR* is, in fact, inadequate for evaluation; this is because that *RR* does not measure how well a model can target undecided customers, which is the only class of customers that can increase the net profit. Instead, in influential marketing we use *UBR* as the evaluation criterion (as stated in Definition 4). Since *UBR* cannot be directly computed, we will indirectly estimate its value.

For a specified marketing percentile x, let *T2x* and *C2x* denote the top x-percentile of the ranked list of *T2* and *C2*, respectively. Let

- *MT* denote *RR* in *T2x*, (equivalent to RR_x in Chapter 3);
- *MC* denote *RR* in *C2x*.

Recall that $MT = UBR + DBR$, where *UBR* and *DBR* are the percentages of undecided and decided customers in *T2x*, respectively. For an estimated *DBR*, we can compute *UBR* by $MT - DBR$. As discussed in Chapter 4.1, we can use *RR* of *C2x*, i.e. *MC*, to approximate *DBR* of *T2x* because (a) *T2x* and *C2x* are ranked by applying the same model to a similar population, and (b) all the responses of *C2x* are generated from decided customers. Then it follows that for *T2x*, $UBR = MT - DBR = MT - MC$.

Theorem 1. For *T2x*, *UBR* is given by $MT - MC$. ■

Essentially, we use the response rate of voluntary buyers observed from the non-contacted group to approximate the response rate of voluntary buyers of the contacted group, given that the two groups are similar.

An effective direct marketing campaign should perform better than random marketing, where customers are randomly selected for contact. Let Random denote the selection model for random marketing. Note that the expected *RR* of Random is the same across all marketing percentiles x. Hence, we take *MT* and *MC* of Random to be those at $x = 100\%$, denoted *RT* and *RC*, respectively. *UBR* of Random is then given by $RT - RC$.

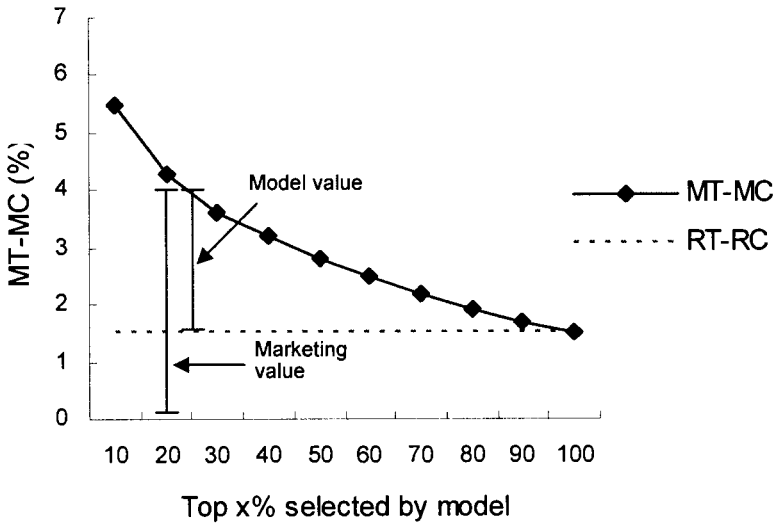


Figure 5.3 The positive influence curve (PIC).

The evaluation of a model can be summarized using the *positive influence curve* (*PIC*), as shown in Figure 5.3. *PIC* indicates how effective a model is at ranking undecided customers. The x-axis represents the marketing percentile and the y-axis represents $MT - MC$, which is the equivalent of *UBR*. The baseline performance of Random, or $RT - RC$, is illustrated using the dashed line.

MT – MC measures the *marketing value* of the campaign relative to the purchases made by decided customers only. On the other hand, (*MT – MC*) – (*RT – RC*) measures the *model value* of M relative to the marketing value of random marketing. A model of a large marketing value has a high *PIC*. A model of a large model value has a steep decreasing trend as x increases.

5.4 Optimal Marketing Percentile

We should keep in mind that the ultimate goal in direct marketing is to maximize the net profit. Hence, a very important task as part of the model evaluation is to determine the optimal marketing percentile x that produces the highest net profit. Recall that R is the revenue per purchase and C is the cost per contact, where $R > C$. Suppose we are selecting customers from the pool P, with $|P|$ denoting the number of customers in P.

For each point (x, y) on *PIC*, the total net profit is equal to $r1 + r2 - tc$, where

- $r1$ is the revenue generated by all the decided customers in P,
- $r2$ is the revenue generated by the undecided customers within the top x percentile,
- tc is the total cost for contacting the top x percentile of P.

To compute the total number of decided customers in P, we need to determine the percentage of buyers in P who voluntarily make the purchase. This percentage is measured by the voluntary response rate over the entire pool of customers, i.e. when x = 100%; such is the equivalence of *MC* at x = 100%. As previously mentioned, *RC* also represents *MC* at x = 100%. It then follows that the number of decided customers in P is given by $|P|*RC$. Since $|P|*RC$ is fixed for all campaigns, $r1$ is fixed across all points on *PIC*. Maximizing the net profit then becomes maximizing $r2 - tc$. The number of customers within the top x percentile of P is $|P|*x$, of which $|P|*x*y$ are undecided customers. Therefore, we have

$$r2 - tc = |P|*x*y*R - |P|*x*C = |P|*x(y*R - C).$$

To determine the optimal x, we compute $r2 - tc$ at each point (x, y) on PIC and choose the x that produces the highest $r2 - tc$. Clearly, the optimal x depends on PIC (i.e. the model), R, and C, but not on the fixed pool size $|P|$. In the presence of budget constraints, x must be selected so that the cost $|P|*x*C$ falls within the campaign budget.

Often, a campaign may wish to reserve some selected customers for Control, as described in Chapter 4.1. In this case, only ρ percent of those selected will be contacted (i.e. Treatment) and the remaining $1-\rho$ percent are watched for purchase behaviors without contact (i.e. Control). Then,

$$r2 - tc = |P|*x*\rho(y*R - C).$$

For the pre-determined treatment/control rate ρ, the optimal x is the same as discussed above and does not depend on ρ.

Model Evaluation

Input: A model M, the validation data <*T2*, *C2*>, and the marketing percentile x.

Output: *PIC*

1. Apply M to rank the records in *T2*;
2. Apply M to rank the records in *C2*;
3. Let *T2x* be the top x-percentile of the ranked list of *T2*;
4. Let *C2x* be the top x-percentile of the ranked list of *C2*;
5. Let *MT* be *RR* of *T2x* and let *MC* be *RR* of *C2x;*
6. Return *MT* – *MC*.

* To produce *PIC*, repeat steps 3–6 with different marketing percentiles x.

Figure 5.4 Model evaluation.

CHAPTER 6
RELATED WORK

6.1 Traditional Approaches

For years, the fundamental principle of direct marketing research has been to differentiate responders from non-responders, or valuable customers from less valuable customers. This is based on the belief that a small subset of all customers is responsible for the majority of profits. The idea was coined as *differential marketing* by Garth Hallberg in [Hal95], with the objective of maximizing the hit rate of responders or valuable customers.

Whereas the traditional approach indiscriminately targets all responders, our approach further categorizes responders into decided and undecided. Not only we consider how a customer would behave if contacted by the campaign, but we also take into account how a customer would have behaved if not contacted. Most importantly, we have shown that when there are decided (voluntary) customers, the net profit cannot be maximized if we do not differentiate between decided and undecided customers.

When maximizing *RR*, a model is driven to target the most likely responders, which tends to focus incorrectly on the decided customers. Recall that the optimal marketing percentile x of the traditional paradigm is the one that maximizes $|P|*x*MT*R - |P|*x*C$, whereas influential marketing looks at the x that maximizes $|P|*x*\{MT - MC\}*R - |P|*x*C$. Clearly, the traditional strategy makes no attempt to isolate the undecided customers.

Some works [Giu03, P-SM99, YFB05] use the accumulative lift curve (ALC) as the performance criterion. Each point (x, y) on the ALC indicates that the top x percent of the ranked list contains y percent of all responders. The more responders near the top of the ranked list the better. Similar to the measure of response rate, ALC drives a model to identify most likely responders. Other works use the receiver operating characteristic

(ROC) curves [Giu03, HM82, HL05] to measure the relative trade-offs between true positives and false positives. Alternatively, cost-sensitive learning [Dom99, ZE01] recognizes that identifying a buyer as a non-buyer (false negative) incurs a higher cost than identifying a non-buyer as a buyer (false positive) and attempts to reduce the first type of error. Essentially, when following any of these evaluation criteria, a model tends to target decided customers instead of undecided customers.

6.2 Lo's Approach

To our knowledge, the only previously published work that addresses the same problem as considered in this thesis is Lo's work [Lo02]. For every customer i, their proposed solution predicts two probabilities: the probability of i responding if contacted, and the probability of i responding if not contacted. Customers are then ranked by the difference between the two predicted probabilities (a customer is ranked higher if the predicted difference is greater).

In order to construct a model that can predict the two probabilities for every customer i, a treatment variable T is introduced in the training data. $T_i = 1$ indicates that i was contacted and $T_i = 0$ indicates that i was not contacted. X represents the set of independent variables. A supervised learning algorithm is then used to train a model for the response variable, where X, T, and $X*T$ (the interaction variables) are the independent variables. Then for every customer, the model computes the two probabilities by assuming $T_i = 1$ and $T_i = 0$.

Expressed in terms of the learning matrix, Lo's approach essentially combines (1) and (3) as the positive class, and combines (2) and (4) as the negative class. As in the traditional approach, the response variable decides whether a sample belongs to the positive or negative class.

For a supervised learning algorithm to pick up the treatment variable T in the training data, $T = 1$ needs to be more strongly associated with the positive class than the negative class. This implies that the model or campaign generating such data has successfully

targeted undecided customers. In other words, Lo's approach depends on a successful model for producing the training data, which defeats the purpose of constructing such models from the training data. In fact, their approach has only been evaluated using a simulated data set with five attributes with an unusually high response rate, e.g. MT = 90% at x = 10%. Looking at the real campaign data used in our experiments, we observed that T = 1 is not significantly more associated with the positive class than with the negative class (see Chapter 7). Fundamentally, the limitation of Lo's approach is the lack of attempt in isolating undecided customers. In fact, all undecided customers in Control, i.e. group (4), are labelled as negative.

CHAPTER 7
EXPERIMENTAL EVALUATION

In this Chapter we present the experimental results.

7.1 The Data Set and Experimental Settings

We experiment on a real campaign data set from a loan product promotion, provided by the Canadian Imperial Bank of Commerce (CIBC). It contains 24,506 records. Each observation is described by 608 independent attributes, about 1/6 of which are categorical and the remaining numerical.

A detailed breakdown of the data is shown in Table 7.1. From the table, we see that the random response rate of Treatment, i.e. RT, is $\frac{1{,}182}{1{,}182+20{,}816} = 5.4\%$ and the random response rate of Control, i.e. RC, is $\frac{108}{108+2400} = 4.3\%$. Thus, the marketing value of Random, i.e. the marketing value of contacting customers randomly relative to the purchases made by decided customers, is only 1.1% ($RT - RC$). If the direct promotion can raise the percentage of undecided customers by a "small" amount, e.g. $MT - MC$ = 3%, this will be a significant increase relative to the marketing value of Random. We also note that the distribution between the response and treatment variables is statistically significant ($p < 0.025$ following the chi square test). This means that the observed difference in response rate between Treatment and Control is unlikely to have happened by chance. Nonetheless, as we have discussed, this difference is small (1.1%).

Note that the two groups, "Response = Yes" and "Response = No," have a very similar distribution — 92% and 90% of the records received the treatment (i.e. T = 1),

respectively. This suggests that Lo's approach of using the treatment variable T will not be effective on this campaign data set.

Table 7.1 Breakdown of the campaign data.

	Response	
	Yes	**No**
Treatment	(1) 1,182	(2) 20,816
Control	(3) 108	(4) 2,400

All experimental results were obtained using 3-fold cross validation. We chose 3-fold due to the small presence of group (3). Treatment and Control were randomly split into 3 equal-sized partitions, respectively. In each of the three runs, we construct the validation data $<T2, C2>$ by taking one of the partitions from each of Treatment and Control; the union of the remaining two partitions in Treatment and Control constitute the training data, $T1$ and $C1$, respectively. For each percentile x, we collected UBR of $T2x$, which is given by $MT - MC$ as stated in Theorem 1. The UBR reported is the average of the three runs based on the 3-fold cross validation. We also report the standard error of this average UBR. Three different approaches are compared, as described below:

- **Traditional approach**. This refers to the traditional objective of maximizing RR of contacted customers.
- **Lo's approach**. To our knowledge, this is the only other work that considers a similar problem.
- **Proposed approach**. This refers to the solution proposed in this paper.

The experiments were conducted using the supervised learning algorithms ARC and SAS EM Tree, as discussed in Chapter 2. ARC is chosen for its superior ability at handling imbalanced class distributions. When using ARC, the numerical attributes are first

discretized using the machine learning library MLC++ [KSD97]. Since this research is a collaborative work with CIBC, demonstrating that our proposed solution can be applied within the SAS environment is important. We therefore choose SAS EM Tree as one of the supervised learning algorithms.

Both algorithms are capable of ranking the samples according to their probability of belong to the positive class. We first present the result from each approach; then we collectively compare all approaches using *PIC*.

7.2 Traditional Approach

Following the traditional direct marketing paradigm, we use only the observations of the contacted customers, i.e. *T1*, to construct the model since how one would behave in the absence of the contact is of no concern. Then the positive class consists of the responders in *T1* and the negative class consists of the non-responders in *T1*. Figure 7.1 plots *MT* and *MC* against the marketing percentile x, with ARC being the supervised learning algorithm. Recall that *MT* is the response rate of the top x-percentile of *T2* (i.e. *T2x*) and *MC* is the response rate of the top x-percentile of *C2* (i.e., *C2x*).

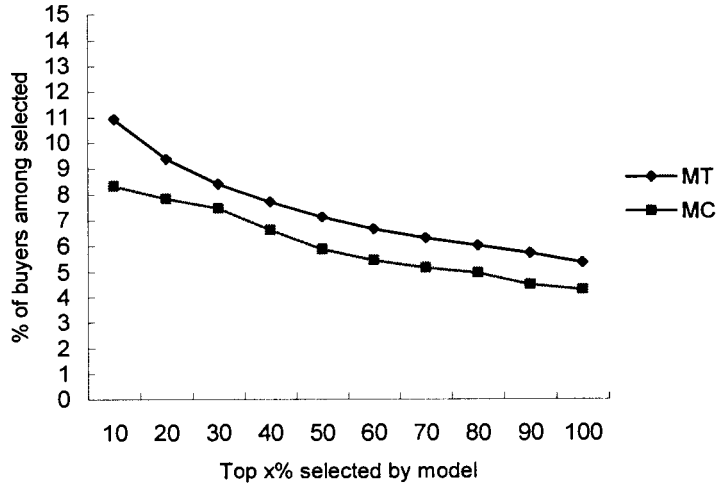


Figure 7.1 Traditional approach using ARC.

Observe that in Figure 7.1, the curves for *MT* and *MC* are highly correlated. Consequently, the difference *MT* – *MC* is small. Recall that *T2x* and *C2x* are similar, except that *T2x* was contacted and *C2x* was not. The high correlation between the two curves indicates that the responders in *T2x* are also likely to respond voluntarily in *C2x*. In other words, the majority of responders in *T2x* are decided customers.

We also experimented with SAS EM Tree. Unlike ARC, the decision tree performs poorly when the class distribution is highly imbalanced (in *T1*, only 5.4% is labeled as positive). Therefore, we under-sample the negative class in *T1* at different rates and construct a model for each rate. The model with the best lift index [LL98] is selected as the final model. The result of the final model is similar to what was obtained with ARC, in which *MT* and *MC* are highly correlated.

The experiment clearly shows the following: while the models built by the traditional direct marketing paradigm indeed produce high response rates among those contacted (e.g. high *MT*), they favour decided customers (e.g. high *MC* as well; hence the small *MT* – *MC*). An effective model should target undecided customers instead, so that more buyers can be generated with fewer contacts. Evidently, *MT* is an ineffective evaluation criterion. The only criterion we should really consider is *MT* – *MC*, the rate of undecided buyers.

7.3 Lo's Approach

The result following Lo's approach with ARC is shown in Figure 7.2. For the typical top four *x* marketing percentiles, the standard errors of averaged *UBR*s obtained from the 3 folds are in the range of 0.13 ~ 0.61. Compared to the traditional approach, *MT* – *MC* has improved by a slight margin. Looking more closely, we observe that the treatment variable *T* is not a good indicator on whether an observation belongs to the positive or negative class. Specifically, from 7.1 we see that $T = 1$ is similarly associated with both the positive and negative classes. As a result, the model constructed is likely to ignore *T*. When this occurs, quite often the predicted probability given $T_i = 1$ and the predicted

probability given $T_i = 0$ is the same. This renders it ineffective to rank customers based on the difference between the two probabilities. We have also experimented with SAS EM Tree following Lo's approach. The result is worse in that it is even more similar to the result of the traditional approach in Figure 7.1.

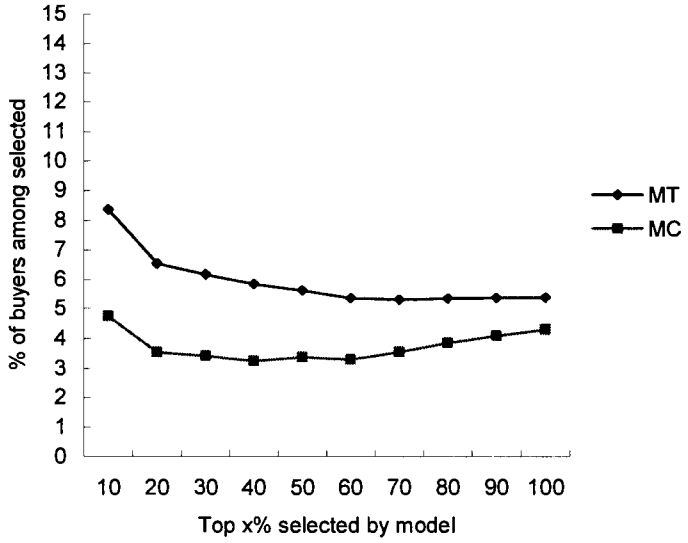


Figure 7.2 Lo's approach using ARC.

7.4 Proposed Approach

We now present the result obtained following our proposed approach. Figure 7.3 gives the result with ARC. The result with SAS EM Tree is similar. Recall that our approach combines (1) and (4) into the positive class *PC* and (2) and (3) into the negative class *NC*. As a result, the percentage of positive records in the training data has increased to 14.6% from 5.4% of the traditional approach (and 5.3% of Lo's approach). Since the class distribution has become more balanced, it was not necessary to under-sample the majority class when using SAS EM Tree.

Unlike Figures 7.1 and 7.2, the curves for *MT* and *MC* in Figure 7.3 have a reverse trend up to the top 50% percentile — *MT* decreases whereas *MC* increases, while *MT* is always above *MC*. The standard errors of averaged *UBR*s for the top four x percentiles are in the range of 0.25 ~ 1.2. This reverse trend results in a large $MT - MC$, which in turns means

that our model is effective at targeting undecided customers. A deeper analysis is as follows. Note that the purchasing decisions of the decided and non customers are the same in both *T2x* and *C2x*. However, while the undecided customers in *T2x* contribute to *MT* (because they were contacted), the undecided customers in *C2x* do not contribute to *MC* (because they were not contacted). Therefore, the difference *MT* – *MC* is caused solely by the presence of undecided customers.

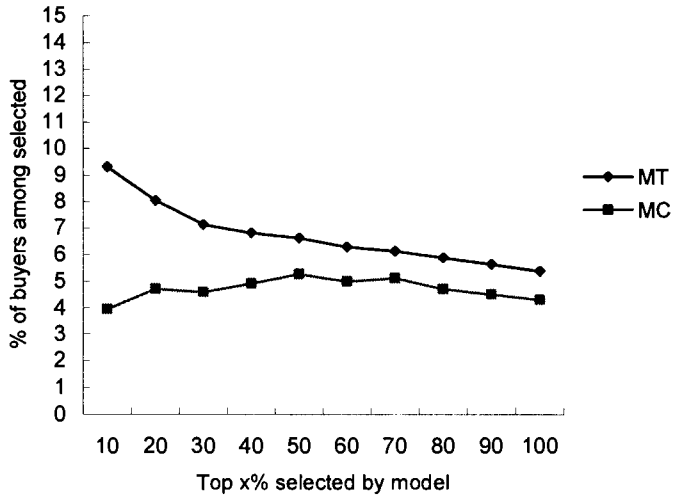


Figure 7.3 Proposed approach using ARC.

As a model focuses on undecided customers, fewer decided customers are selected. It is our intention to not contact decided customers. This explains the decrease in both *MT* and *MC* compared to Figure 7.1.

As Table 7.1 shows, the size of group (3) is quite small. Recall that (3) is included in *NC* to "remove" the characteristics of the decided customers from *PC*. Due to the small size of (3), this effect is quite limited. To increase this effect, we over-sample (3) up to 5 times, 10 times, and 20 times in the training data. The models built from these over-sampled training sets have all shown a further increase in *MT* – *MC* when compared to Figure 7.3. For example, Figure 7.4 shows the result when the over-sampling rate is 10 times (a similar result was observed for the other over-sampling rates) where the standard

errors of averaged *UBR*s for the top four x percentiles are in the range of 0.23 ~ 0.9. Observe that the increase in *MT* – *MC* is mainly due to the further decline in *MC*, indicating that even fewer **decided** customers have been selected. This effect is exactly what we want to achieve by over-sampling (3). A similar improvement with such over-sampling was observed with SAS EM Tree.

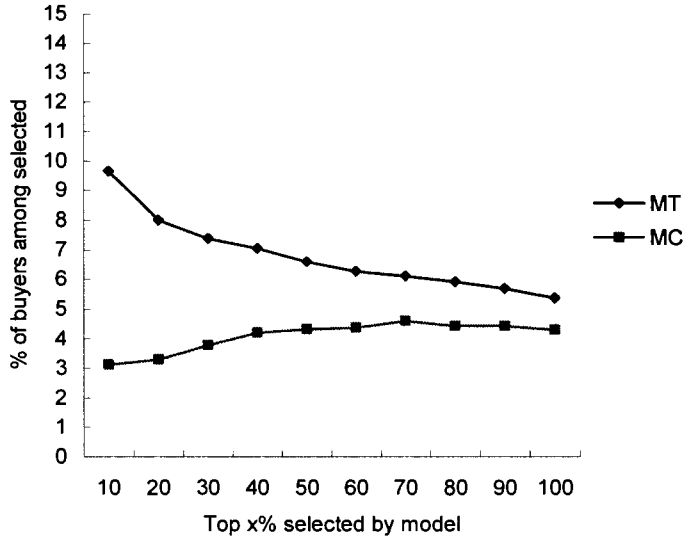


Figure 7.4 Proposed approach using ARC. 10 times over-sampling of (3).

Besides evaluating our proposed solution on the validation data <*T2*, *C2*>, we also performed the evaluation on an independent testing set whose response variable was withheld by CIBC, i.e. the testing data received by us contains no response variable. After we ranked the customers in this testing set using the model constructed following our proposed solution, the result was sent back to CIBC for verification. The result obtained is similar to Figure 7.3, again showing a large *MT* – *MC*.

7.5 Summary of Comparison

We summarize the results using *PIC* in Figures 7.5 and 7.6. "Traditional" denotes the traditional approach. "Lo's" denotes Lo's approach. "Influential-10" denotes the

proposed approach with group (3) over-sampled 10 times, and "Influential" denotes the proposed approach with no over-sampling.

We can see that in both figures, Influential-10 produces the highest *MT* – *MC*. The decreasing trend of both Influential and Influential-10 indicates that the models built by our approach do in fact focus on undecided customers. On the other hand, the nearly flat trend produced by the traditional approach suggests that the method fails to target undecided customers.

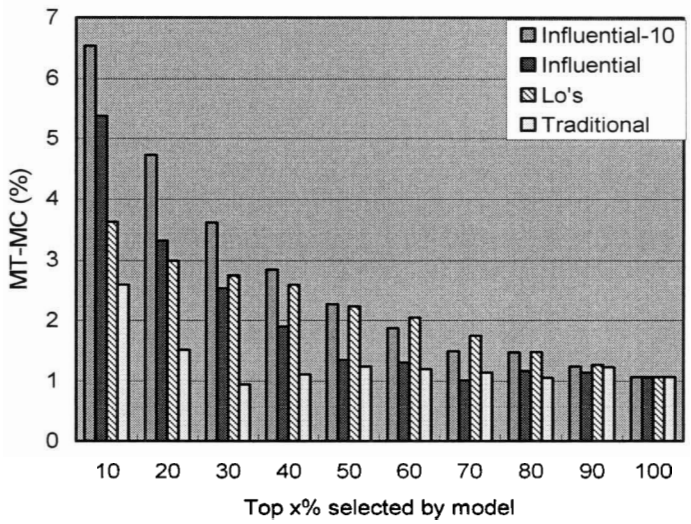


Figure 7.5 Comparisons using *PIC* (ARC).

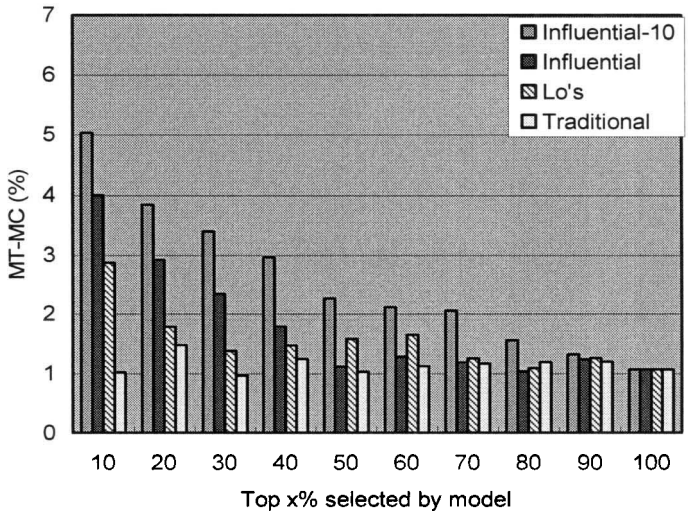


Figure 7.6 Comparisons using *PIC* (SAS EM Tree).

Our approach also performs better than Lo's approach. The improvement is significant in the top 20 percentile for the ARC option. For SAS EM Tree, the improvement extends to the top 40 percentile. These improvements are significant considering the generally very low response rates in direct marketing. Since a typical direct marketing campaign has a small marketing percentile, e.g. 10–30%, it shall benefit from our approach.

CHAPTER 8
DISCUSSION AND CONCLUSIONS

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

1. In rare but still possible cases, there may exist a fourth class of customers who are "negatively affected" by the direct promotion. This may occur when the direct promotion is ineffective and thus leaves a negative impression of the product or company. For example, Jane was initially a voluntary buyer but she felt annoyed by the attitude of the person that contacted her about the product. As a result of her feelings, she eventually found an alternative product and did not purchase the product she originally intended to.

   In this thesis, we did not consider the possibility that a direct contact may actually have a negative influence on a person's purchasing behavior. Nonetheless, if the frequency of such scenario happening in real life is not negligible, it will be an interesting topic for future exploration.

2. The size of Control should be sufficiently large for the training of an influential model. In the real data set used in our work, there were only 108 observations in group (3) of the learning matrix. This sample size is rather small, which prompts us to over-sample group (3) during model training in an attempt to reinforce the features of decided customers in the negative class. In the past, the Control group, if available, has been used for comparative purposes only. Campaign managers tend to limit the size of Control based on the belief that they would incur a loss of profit if they "give up" (not contact) a potential customer identified as a likely responder. However, for the purpose of model training in influential marketing, it is best that the size of Control is not too small.

3. In direct marketing, the response, or purchasing behavior, of a customer can be influenced by the campaign. In other words, the "state" of a customer is not absolutely fixed. Generally, the influential paradigm becomes necessary when the state of the subject can be influenced by an external action (e.g. a direct contact). On the other hand, in areas such as fraud or defective detections, no external action resulted from data mining can change the state of the subject.

In summary, this thesis recognizes the presence of voluntary buyers in the real business environments. Analysis showed that the traditional direct marketing paradigm tends to incorrectly target decided customers which does not maximize the net profit. To tackle this issue, we presented *influential marketing* to target only undecided customers who can be positively influenced. Our innovative solution addresses the two major challenges: (a) without the explicit labelling of customers by the new taxonomy, it is still possible to construct a model capable of ranking undecided customers within the standard supervised learning procedure; (b) the solution introduces no major changes to the standard campaign practice. These properties make our solution immediately deployable. We presented an in-depth comparison with previous approaches both analytically and experimentally. The study showed great promises in influential marketing.

BIBLIOGRAPHY

[AS94] R. Agrawal and R. Srikant, Fast algorithms for mining association rules, In *Proc. of the 20th Int'l Conference on Very Large Databases*, 1994.

[BL97] M.J.A. Berry and G.S. Linoff, Data Mining Techniques: For Marketing, Sales, and Customer Support, John Wiley & Sons, 1997.

[Bha00] S. Bhattacharyya, Evolutionary algorithms in data mining: Multi-objective performance modeling for direct marketing, In *Proc. of ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining*, 2000.

[CS98] P. Chan and S. Stolfo, Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection, In *Proc. of ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining*, 1998.

[DR94] W. Desarbo and V. Ramaswamy, Crisp: customer response based iterative segmentation procedures for response modeling in direct marketing, *Journal of Direct Marketing*, 8(3), 1994, pp. 7-20.

[Dom99] P. Domingos, Metacost: A general method for making classifiers cost sensitive, In *Proc. of ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining*, 1999.

[DR01] P. Domingos and M. Richardson, Mining the network value of customers, In *Proc. of ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining*, 2001.

[ESN96] K. Ezawa, M. Singh, and S. Norton, Learning goal oriented Bayesian networks for telecommunications risk management. In *Proc. Of the International Conference on Machine Learning*, 1996.

[FP97] T. Fawcett and F. Provost, Adaptive fraud detection, *Data Mining and Knowledge Discovery*, 1(3), 1997.

[Giu03] P. Giudici, Applied Data Mining, John Wiley & Sons, 2003.

[Hal95] G. Hallberg, *All Consumers Are Not Created Equal: Differential Marketing Strategy for Brand Growth and Profits*, John Wiley & Sons, 1995.

[HK01] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2001.

[HL05] J. Huang and C.X. Ling, Using AUC and accuracy in evaluating learning algorithms, *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 17(3), 2005, pp. 299-310.

[HM82] J.A. Hanley and B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology*, 143(1), 1982, pp. 29-36.

[Hug96] A.M. Hughes, The Complete Database Marketer: Second-generation Strategies and Techniques For Tapping the Power of Your Customer Database, Irwin Professional, Chicago, 1996.

[Jap00] N. Japkowicz, The class imbalance problem: Significance and strategies, In *Proc. of the Int'l Conf. on Artificial Intelligence*, 2000.

[JAK01] M.V. Joshi, R.C. Agarwal, and V. Kumar, Mining needles in a haystack: Classifying rare classes via two-phase rule induction, In *Proc. of ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining*, 2001.

[Joa99] T. Joachims, *Making Large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning*, MIT-Press, 1999.

[KDD98] KDD-Cup-98. *http://www.kdnuggets.com/meetings/kdd98/kdd-cup-98.html.*

[KHM98] M. Kubat, R. Holte, and S. Matwin, Machine learning for the detection of oil spills in satellite radar images, *Machine Learning*, 30(2-3) 1998.

[KKT03] D. Kempe, J. Kleinberg, and E. Tardos, Maximizing the spread of influence through a social network, In *Proc. of ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining*, 2003.

[KM97] M. Kubat and S. Matwin, (1997), Addressing the curse of imbalanced data sets: One sided sampling, In *Proc. of the Int'l Conf. on Machine Learning*, 1997.

[KSD97] R. Kohavi, D. Sommerfield, and J. Dougherty. Data mining using MLC++: A machine learning library in C++. *International Journal on Artificial Intelligence Tools*, 6(4), 1997, pp. 537-566. (http://www.sgi.com/tech/mlc/.)

[LL98] C.X. Ling and C. Li, Data mining for direct marketing: Problems and solutions. In *Proc. of ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining*, 1998.

[Lo02] V.S.Y. Lo, The true lift model – A novel data mining approach to response modeling in database marketing, *SIGKDD Explorations*, 4(2), 2002, pp. 78-86.

[Mon91] D.C. Montgomery, *Design and Analysis of Experiments, 3rd Edition.* John Wiley & Sons, 1991.

[P-SM99] G. Piatetsky-Shapiro and B. Masand, Estimating campaign benefits and modeling lift, In *Proc. of ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining*, 1999.

[PKP02] R. Potharst, U. Kaymak, and W. Pijls, Neural networks for target selection in direct marketing, *Neural Networks in Business: Techniques and Applications*, 2002, pp. 89-100.

[Qui93] J.R. Quinlan, *C4.5: Programs for machine learning*, Morgan Kaufmann, 1993.

[SAS] SAS Institute. The SAS system. (http://www.sas.com/.)

[Sto74] M. Stone, 1974. Cross-validatory choice and assessment of statistical predictions, *Journal of the Royal Statistical Society*, Series B(36), 1974, pp. 111-147.

[Wei04] G.M. Weiss, Mining with rarity: A unifying framework, *SIGKDD Explorations*, 6(1), 2004.

[WZYY05] K. Wang, S. Zhou, Q. Yang and J.M.S. Yeung, Mining customer value: From association rules to direct marketing, *Journal of Data Mining and Knowledge Discovery*, 11(1), 2005, pp. 57-80.

[YFB05] L. Yan, M. Fassino, and P. Baldasare, Enhancing the lift under budget constraints: An application in the mutual fund industry, In *Proc. of ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining*, 2005.

[ZE01] B. Zadrozny and C. Elkan, Learning and making decisions when costs and probabilities are both unknown, In *Proc. of ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining*, 2001.