# Active Learning for Decision-Making

Maytal Saar-Tsechansky • Foster Provost

*Department of Management Science and Information Systems, Red McCombs School of Business, The University of Texas at Austin, Austin, Texas 78712, USA*
*Department of Information, Operations, and Management Sciences, Leonard N. Stern School of Business, New York University,*
*44 West Fourth Street, New York, NY 10012, USA*
maytal.saar-tsechansky@mccombs.utexas.edu • fprovost@stern.nyu.edu

Working Paper CeDER-04-06, Stern School of Business, New York University, NY, NY 10012

**November, 2004**

---

## Abstract

This paper addresses focused information acquisition for predictive data mining. As businesses strive to cater to the preferences of individual consumers, they often employ predictive models to customize marketing efforts. Building accurate models requires information about consumer preferences that often is costly to acquire. Prior research has introduced many "active learning" policies for identifying information that is particularly useful for model induction, the goal being to reduce the acquisition cost necessary to induce a model with a given accuracy. However, predictive models often are used as part of a decision-making process, and costly improvements in model accuracy do not always result in better decisions. This paper develops a new approach for active information acquisition that targets decision-making specifically. The method we introduce departs from the traditional error-reducing paradigm and places emphasis on acquisitions that are more likely to affect decision-making. Empirical evaluations with direct marketing data demonstrate that for a fixed information acquisition cost the method significantly improves the targeting decisions. The method is designed to be generic—not based on a single model or induction algorithm—and we show that it can be applied effectively to various predictive modeling techniques.

**Key words**: active learning, information acquisition, decision-making, class probability estimation, cost-sensitive learning.

## 1. Introduction

Because of advances in computing power, network reach, availability of data, and the maturity of induction algorithms, businesses increasingly take advantage of automated predictive modeling, or predictive "data mining," to influence repetitive decisions. Consider an example: Telecommunications companies face severe customer retention problems, as customers switch back and forth between carriers (the problem of "churn"). For each customer, at each point in time, the company faces a decision between doing nothing and intervening in an attempt to retain the customer. Increasingly, decision-making is based on predictive models built from data on the effectiveness of offers and of inaction. For this example, ideally the predictive model would estimate the probability of imminent loss of the customer; the probability estimate would then be combined with utility information to maximize expected profit.[1]

Acquiring additional customer feedback can improve modeling, but this information comes at a cost. Firms collect information about customers directly via solicitations, e.g., surveys of the customers themselves, or direct acquisition from a third party. For example, Acxiom provides detailed consumer demographic and lifestyle data to a variety of firms, including Fortune 500 firms, in support of their marketing efforts; other direct marketing firms such as Abacus Direct maintain and sell specialized consumer purchase information (New York Times, 1999). Firms collect information indirectly, via interactions initiated by the firm for the purpose of collecting relevant data, and also via normal business interactions (e.g., Amazon's acquisition of customer preferences via purchases and product ratings). All these acquisitions involve costs to the firm.

For this paper, we consider the acquisition of a particular kind of information. Following the terminology used by Hastie, et al., (2001) we refer to the data used to induce models as *training* data. Importantly for this paper, in the usual ("supervised learning") scenario training data must be *labeled*, meaning that the value of the target variable is known (e.g., whether or not a particular customer would respond positively to the current offer). However, acquiring labels may be costly. For example, obtaining preference

[1] For this paper we ignore issues pertinent to this example like calculations of lifetime value, but see (Rosset et al. 2003) for a treatment from the data-mining perspective.

information for individual customers involves solicitation costs, incentives required for revealing their preferences, negative reactions to solicitations, etc. Firms also incur opportunity costs when information is acquired over time through normal business interactions. For example, making a particular offer to a random sample of web site visitors, for the purpose of acquiring training labels, may preclude making another offer already known to be profitable.

Because of the inefficiencies imposed by these label-acquisition costs, researchers have studied the *selective* acquisition of labels for inductive model building (e.g., optimal experimental design (Kiefer, 1959; Fedorov, 1972) and active learning (Cohn et al., 1994)). The motivation is that focused selection of cases for label acquisition should result in better models for a given acquisition budget, as compared to the standard approach where labels are acquired for cases sampled uniformly at random, and therefore should reduce the cost of inducing a model with a particular level of performance. Research to date offers various label-acquisition strategies for inducing statistically accurate predictive models (e.g., Fedorov, 1972; Cohn et al., 1994; Lewis and Gale, 1994; Roy and McCallum, 2001).

However, business applications employ predictive models to help make particular business decisions. Of course, a more accurate model may lead to better decisions, but concentrating on the decisions themselves has the potential to produce a more economical allocation of the information acquisition budget. Prior work has not addressed how labels should be acquired to facilitate decision-making directly.

We consider the decision of whether or not to initiate a business action. A characteristic of such decisions is that they require an estimation of the expected utility for each action, hence (in the presence of uncertainty) an estimation of the probabilities of different outcomes. Consider for example a model for predicting whether a mobile service customer will terminate her contract, where the model supports the decision of whether or not to offer the customer incentives to renew her contract. Prior work provides label acquisition strategies to improve (for a given budget) the estimation of the probability of renewal. However, such a strategy may not be best for improving intervention decisions. As we show later, the ability to identify potentially wasteful investments can result in considerable economies.

The contribution of this paper is the development and demonstration of a new method for selecting cases for label acquisition that (1) targets decision-making directly (it is *decision-centric*), and (2) can be applied to various predictive modeling techniques (it is *generic*). The goal is to allow the induction of better models for decision-making, given a fixed label-acquisition budget. We demonstrate the method using data from a direct-marketing campaign, with the objective of acquiring customer feedback that will increase profits from future customer solicitations. The decision-centric approach results in significantly higher decision-making accuracy and profit (for a given number of acquisitions) compared to the usual strategy: sampling cases for label acquisition uniformly at random. Moreover, the decision-making accuracy and profit obtained with the new method are significantly higher than those obtained by acquiring labels to reduce model error. Notably, even though the decision-centric method does result in superior decision-making, the average statistical accuracy of the model induced is lower than that obtain with the error-reducing method. Each method is better at the task for which it was designed. To demonstrate the generic nature of the method, we apply the method to three different model induction algorithms and show that the decision-centric method consistently results in superior performance compared to the error-reduction method.

The rest of the paper is organized as follows. Section 2 discusses the current paradigm for selective label acquisition for induction (*active learning*). In Section 3 we analyze the impact of traditional active learning on decision-making efficacy and lay the theoretical foundation for the new decision-centric approach. The new method is presented in Section 4. Then, in Section 5 we demonstrate the proposed approach. We estimate the costs and benefits of direct mailing decisions and analyze the performance of the proposed and existing label acquisition methods. We present some limitations to the work in Section 6, and we discuss managerial implications and conclude in Section 7.

## 2. Active Learning: Terminology, Framework and Prior Work

We first introduce the notation and the terminology we employ. A firm wants to induce a probabilistic classification model to estimate the probability of alternative outcomes. A categorical classification model is a mapping of an input vector $x \in X$ to a label $y \in Y$ from a set of discrete labels or classes $Y$. The model is constructed through induction where a training set of labeled "examples"—$(x, y)$ pairs—are generalized into a concise model $M : X \rightarrow Y$. Differently from a categorical classification model, a probabilistic classification model also assigns a probability distribution over $Y$ for a given input vector $x$. A maximum *a posteriori* decision rule would then map $x$ to the label $y$ with the highest estimated probability. Examples of probabilistic classifiers are the Naïve Bayes classifier (Mitchell, 1997), suitably designed tree-based models (Breiman et al., 1984; Provost & Domingos, 2003) and logistic regression. For brevity, we refer to the estimation of the probability that an input vector $x$ belongs to a class $y$, $\hat{p}(y \mid x)$, as *class probability estimation* (*CPE*).

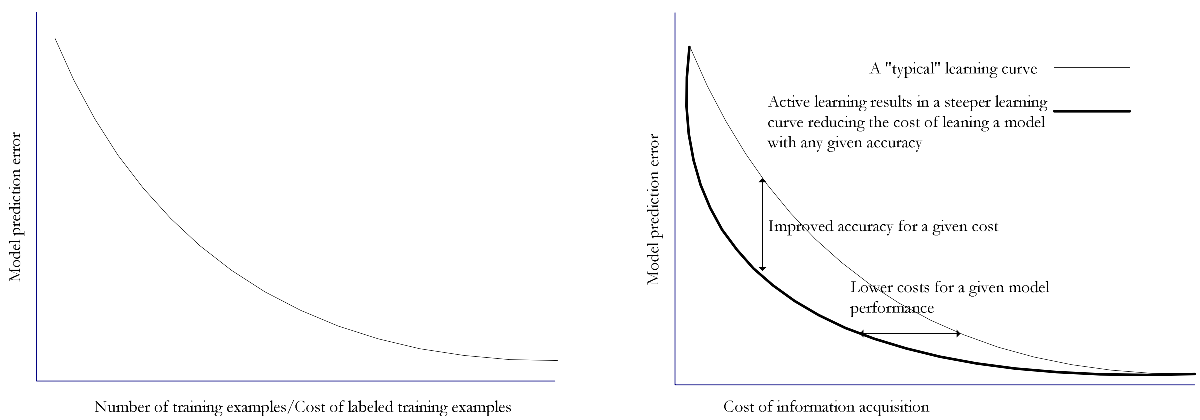
## 2.1 Active Learning

A signature of a modeling technique's predictive performance for a particular domain of application is captured by its *learning curve*, depicting the model's predictive accuracy[2] as a function of the number of training data used for its induction. A prototypical learning curve is shown in Figure 1 where the model improves with the number of training examples available for induction, steeply first, but with decreasing marginal improvement (cf., Cortes, et al., 1994). For this paper we assume that the cost of acquiring data is uniform, so the learning curve also shows the cost of learning a model with any given accuracy. Consider for example a company modeling consumer preferences to predict the probability of response to various offers. The customer preference model can be improved as more feedback on various offers is acquired, resulting in more effective product recommendations and a potential increase in profit. The cost of acquiring customer feedback corresponds to the graph's x-axis, and hence the learning curve characterizes the model's performance as a function of the information acquisition cost.

Consider a typical setting where there are many potential training examples for which labels can be acquired at a cost; for example, customers to whom we can send an offer to determine whether they will

[2] Model accuracy here refers to a model's predictive performance on out-of-sample data. This measure is sometimes referred to as generalization performance.

respond. Let us refer to examples whose labels are not (yet) acquired as *unlabeled* examples, and to examples whose labels already have been acquired as *labeled* examples. The goal of active learning is acquire the labels of unlabeled examples in order to produce a better model. Specifically, for a given number of acquisitions, we would like the model's generalization performance to be better than if we had used the alternative strategy of acquiring labels for a representative sample of examples (via uniform random sampling).




a. A learning curve describes a model performance as a function of the number of training examples or information-acquisition cost

b. Active learning economizes on information-acquisition cost for a particular model accuracy

Figure 1: The learning curve and the effect of active learning

Let us examine the learning curve that results from traditional active learning. The thin learning curve in Figure 1b corresponds to acquiring the labels of examples that were sampled randomly and then using these labeled examples for model induction. The thick-lined curve in Figure 1b is an idealized learning curve resulting from active learning, where fewer labeled training examples are needed to induce a model with any given accuracy. Active learning attempts to label examples that are particularly informative for reducing the model error, so ideally it results in a steeper learning curve. Similarly, for a given acquisition budget (point on the x-axis), the acquisitions directed by active learning produce a model with lower prediction error.

Active learning methods operate iteratively. At each phase they (i) estimate the expected contribution of potential unlabeled examples if their labels were to be acquired,[3] (ii) label some examples, and (iii) add these examples to the training set. Figure 2 describes an algorithm framework outlining the prevailing active learning paradigm. Specifically, an induction method first is applied to an initial set *L* of labeled examples (usually selected at random or provided by an expert). Subsequently, sets of $M \geq 1$ examples are selected by the active learning method in phases from the set of unlabeled examples *UL*, until a predefined condition is met (e.g., the labeling budget is exhausted). To select the best examples for labeling in the next phase, each candidate unlabeled example $x_i \in UL$ is assigned an effectiveness score $ES_i$ based on an objective function, reflecting its estimated contribution to subsequent induction. Examples then are selected based on their effectiveness scores and their labels are acquired before being added to the training set *L*. (And the process iterates.)

**Input**: an initial labeled set *L*, an unlabeled set *UL*, a model induction algorithm *I,* a stopping criterion, and an integer *M* specifying the number of actively selected examples in each phase.

1 While stopping criterion not met

/* perform next phase: */

2 Apply inducer *I* to *L* to induce model $E$

3 For each example $\{x_i \mid x_i \in UL\}$ compute $ES_i$, the effectiveness score

4 Select a subset *S* of size M from *UL* based on $ES_i$

5 Remove *S* from *UL*, label examples in *S*, and add *S* to *L*

**Output:** Model *E* induced with *I* from the final labeled set *L*

Figure 2: A Generic Active Learning Algorithm

[3] An often-tacit assumption of active learning methods is that acquiring labels for certain training examples will affect similar examples when the model is used. We will revisit this below.

The framework in Figure 2 highlights the challenge of an active learning method: to estimate the relative contribution of possible training examples (the effectiveness score) prior to acquiring their labels. Most existing methods compute effectiveness scores based on some notion of the uncertainty of the currently held model. For example, *uncertainty sampling* [Lewis and Gale, 1994] is a generic active learning method designed for inducing binary classifiers. Uncertainty sampling defines the most informative examples (whose labels should be acquired) as those examples for which the current classification model assigns a CPE that is closest to 0.5. The rationale is that the classification model is most uncertain regarding the class membership of these examples, and so the estimation of the classification boundary can be improved most by acquiring their labels for training.

## 2.2 Prior Work

The role of information acquisition in decision-support has been studied by many in the management literature. For example, Allen and Gale (1999) examine the role of increasing information costs on the emergence of financial intermediary institutions, demonstrating that the rising cost of information necessary to successfully participate in sophisticated financial markets was the key factor in the formation of intermediaries. Makadok and Barney (2001) examine the creation of informational advantages by firms through the acquisition of information about their competition. They focus on the acquisition of information for supporting strategy-formulation decisions. This paper concentrates on information acquisition to improve operational decisions that must be made repeatedly, so even modest marginal improvements can have a large cumulative effect on profit.

Many organizations employ predictive models effectively, often as key tools for extracting customer, competitor and market intelligence (Wall Street Journal, 1997; Resnik and Varian, 1997; New York Times 2003a; New York Times, 2003b). Research on predictive models for business intelligence has focused primarily on modeling techniques (e.g., West et al., 1997; Moe and Fader, 2004). However, such intelligence relies on information that requires significant time and/or money to obtain. Therefore, it is important to understand what are the fundamental properties of information that will be particularly effective for inducing accurate predictive models, so as to direct the acquisition of such information. Mookerjee and Mannino (1997) consider the cost of retrieving examples for inductive learning. They argue for the important role of information cost for learning and aim to reduce the cost of attribute specification required to retrieve cases relevant for classification. They demonstrate that the incorporation of such considerations can reduce information acquisition costs.

Another related stream of research builds on the (classic) multi-armed bandit problem originally proposed by Robbins (1952). Given $k$ slot machines with different rates of return, a gambler has to decide which to play in a sequence of trials. There are various formulations of the goal, but generally the gambler wants to maximize the overall reward. An important difference between the multi-armed bandit settings and that of active learning is that for the gambler it is sufficient to estimate the success probability of each machine, whereas an active learner must induce a predictive model over the dependent-variable domain space.

The challenge of data acquisition specifically for modeling has been studied extensively in the statistical community. In particular, the problem of optimal experimental design (Kiefer, 1959; Fedorov, 1972) or *OED* examines the choice of observations for inducing parametric statistical models when observations are costly to acquire. The objective is to devise a distribution over the independent variables reflecting the contribution of label acquisition for these examples. Although there are substantial similarities between work calling itself active learning and work on optimum experimental design (and not many cross-references), there is an important difference between the two. OED studies parametric statistical modeling, whereas active learning is concerned primarily with non-parametric machine-learning modeling or with generic methods that apply (in principle) to a variety of modeling methods. This is an important distinction because methods for OED depend upon closed-form formulations of the objective function that cannot be derived for non-parametric models.

The fundamental notion of active learning has a considerable history in the literature. Simon and Lea (1974) describe conceptually how induction involves simultaneous search of two spaces: the hypothesis space and the example space. The results of searching the hypothesis space can affect how the example space will be sampled. In the context of acquiring examples for classification problems, Winston (1975) suggests that the best examples to select for learning are "near misses," instances that miss being class members for only a few reasons. This notion underlies most active learning methods, which address classification models (e.g., Seung et al., 1992; Lewis and Gale, 1994; Roy and McCallum, 2001) and are designed to improve classification accuracy (rather than the accuracy of the probability estimations, to which we will return presently).

As mentioned previously, most existing active-learning methods address categorical classification problems and compute effectiveness scores based on some notion of the uncertainty of the currently held model. To our knowledge, this idea was introduced in the active-learning literature by the Query By Committee (QBC) algorithm (Seung et al., 1992). In the QBC algorithm each potential example is sampled at random, generating a "stream" of training examples, and an example is considered informative and its label is acquired if classification models sampled from the current version space[4] (Mitchell, 1997) disagree regarding its class membership. The QBC algorithm employs disagreement among different classification models as a binary effectiveness score—capturing uncertainty in current predictions of each unlabeled example's class membership. Subsequently, authors proposed a variety of alternative effectiveness scores for this uncertainty (e.g., Lewis and Gale, 1994; Roy and McCallum, 2001).

A different approach to active learning for categorical classification problems attempts to estimate directly the expected improvement in accuracy if an example's label were acquired. Roy and McCallum (2001) present an active-learning approach for acquiring labeled documents and subsequently using them for inducing a Naïve Bayes document classifier. Their method estimates the expected improvement in class entropy obtained from acquiring the label of each potential learning example; it acquires the example that brings about the greatest estimated expected reduction in entropy.

Decision-making situations often require more than categorical classification. In particular, for evaluating different courses of action under uncertainty it is necessary to estimate the probability distribution over possible outcomes, which enables the decision-making procedure to incorporate the costs and benefits

[4] The version space refers to the set of all hypotheses or models that predict the correct class of all the examples in the training set.

associated with different actions. In targeted marketing, for example, the estimated probability that a customer will respond to an offer is combined with the corresponding costs and revenues to estimate the expected profits from alternative offers. More generally, accurate estimations of response probabilities enable a decision maker to rank alternatives correctly, to identify the actions with the highest expected benefits, and to maximize utility over multiple courses of action. To our knowledge there is only one study of generic active learning methods for inducing accurate class probability estimation (CPE) models (Saar-Tsechansky and Provost 2004), in which the effectiveness score is based on uncertainty in the CPEs rather than in the classifications (we return to this below). However, as we discuss in more detail next, improving the CPEs generally may not be as effective as focusing on the particular decision-making task.

## 3. Active Learning for Decision-Making

The objective of all prior active learning methods has been to lower the cost of learning *accurate* models, be they accurate models for categorical classification or accurate models for class probability estimation. Therefore, these methods employ strategies that identify and acquire labels for training examples that are estimated to produce the largest reductions in the model's prediction error. From a management perspective, it is important to ask whether these strategies are best when the learned models will be used in a particular decision-making context. In particular, an error-reducing strategy may waste resources on acquisitions that improve model accuracy, but produce little or no improvement in decision-making. More accurate CPEs do not necessarily imply better decision-making.

How should active learning strategies be designed to avoid such wasteful investments? We next analyze the relationship between costly label acquisitions and decision-making efficacy, deriving the fundamentals for new active learning approaches designed specifically for decision support.

### 3.1 The Impact of Label Acquisition on Decision Making Quality

As described above, we consider the decision of whether or not to initiate a business action, such as mailing a direct marketing solicitation, or offering a costly incentive for contract renewal. We would like to estimate whether the expected utility from action would exceed that of inaction. Let $x_i$ be an example (e.g., the description of a customer) and let $f_i$ denote the (unknown) probability that the action with respect to $x_i$ will be successful (e.g., customer $x_i$ will respond to the marketing campaign, or will renew her contract). Given that action is taken, let the utility of success and the utility of failure with respect to instance $x_i$ be $U_i^S$ and $U_i^F$ respectively. Let the corresponding utility of inaction be $\Psi_i$. Finally let C denote the cost of action. To maximize utility, action should be initiated if $f_i \cdot U_i^S + (1 - f_i) \cdot U_i^F - C \geq \Psi_i$, or equivalently, if the probability of a successful outcome exceeds the threshold $f_i^{Th}$ given by

$$f_i^{Th} = \frac{\Psi_i + C - U_i^F}{U_i^S - U_i^F} \qquad (1)$$

For a decision maker to act optimally it is necessary to estimate the probability of success. Because training information is costly, we would like to reduce the cost of inducing an estimation model that will render decisions of a given quality. One approach to reducing the cost of learning accurate CPEs is via traditional active learning methods, which are designed to improve the model's average performance over the instance space.

However, improvement of class probability estimations may not always be justified. Consider the case in which the actual probability of success exceeds the threshold $f_i^{Th}$ (suggesting action is better than inaction). For the induced model to allow a decision maker to act optimally it is sufficient that the estimated probability of success $\hat{f}_i$ exceed the threshold as well, even if it is highly inaccurate. Improvement of the probability estimation when the current estimation already specifies the correct decision would not affect decision-making, and therefore the cost of the improvement would be wasted. In fact, as we will illustrate, if the true probability is just above the threshold and the estimate has a non-negligible variance, improving the probability may adversely affect decision-making (cf., Friedman 1997).

Since a model is induced from a sample, the model's probability estimation $\hat{f}_i$ can be treated as a random variable. Let $\Gamma_i$ be the best "action" and let $\hat{\Gamma}_i$ be the estimated best action derived using the model's probability estimation. Similarly to Friedman's analysis of incorrect classification decisions (Friedman, 1997), the probability of making a "wrong decision"–-i.e., a decision that is inconsistent with the decision derived using the true probability of success—is given by:

$$P(\hat{\Gamma} \neq \Gamma) = I\left(f < f_i^{Th}\right)\int_{f_i^{Th}}^{\infty} p(\hat{f})d\hat{f} + I\left(f \geq f_i^{Th}\right)\int_{-\infty}^{f_i^{Th}} p(\hat{f})d\hat{f} \qquad (2)$$

where the indicator function $I(L)$ is 1 if $L$ is true and 0 otherwise. For example, if the actual probability were smaller than the threshold $f_i^{Th}$, the expected utility of action would not exceed that of inaction; a sub-optimal decision would result if the estimated probability were larger than the threshold.

In order to reduce the cost of inducing a probability estimation model that will allow for satisfactory decision-making, it is important to understand the circumstances under which costly improvements in CPE accuracy should be avoided. If we approximate $p(\hat{f})$ with a normal distribution, the probability of making an inconsistent decision is given by

$$P(\hat{\Gamma} \neq \Gamma) = \Phi\left[sign\left(f - f^{Th}\right)\frac{E\hat{f} - f^{Th}}{\sqrt{\operatorname{var}\hat{f}}}\right] \qquad (3)$$

where $\Phi$ denotes the right-hand-side tail of the standard normal distribution, E denotes an expectation and *var* denotes the variance of a random variable. Assume for illustration that a learner is used to induce a model from a training sample for estimating the probability that customers would respond to a certain offer. For a given customer $x_i$ the model produces (on average over different samples) a CPE such that the expected profit from an offer solicitation to $x_i$ is higher than the expected profit of not making the offer, i.e., $\hat{f}_i > f_i^{Th}$. Also assume that the true probability of response suggests the same ($f_i > f_i^{Th}$). So we expect the model to lead to the correct decision: make the offer. Under such circumstances it may not be cost-effective to acquire additional labels (customer feedback) to improve the estimation; improving the estimation may increase the chance of decision-making error! From (3) we see that indeed the larger the average CPE produced by the learner and hence the more biased the estimates are, the more likely it is that the model would produce the correct decision. This is because the larger bias reduces the chance that—due to estimation variance—the estimated expected profit from action would (mistakenly) fail to exceed that of inaction.

There is an incentive, however, to remove such CPE bias when the estimated probability and the true probability of response lead to different decisions.[5] For example, assume that the true expected benefits from inaction exceed those of action, but that on average the learner induces a model that suggests otherwise. In this scenario improving the CPEs reduces the likelihood of decision-making error. Existing active learning methods employ a greedy strategy, acquiring examples for model improvement whenever such improvement is deemed possible. The above analysis suggests that for cost-effective acquisition of examples to support decision-making, an active learner is well advised to take a decision-centric strategy, acquiring labels for training examples when an improvement in the estimation is likely to lead to better decisions, and avoiding acquisitions otherwise even if they might produce a more-accurate model. Unfortunately, the true probability and thus the right decision are unknown, and therefore it is impossible to determine whether or not an improvement is called for.

In summary, active label acquisition targeting improved accuracy generally may not be best for cost-effective decision-making. In fact, somewhat counter-intuitively, in certain cases improving CPEs can be detrimental to decision-making. Ideally, we would like to improve CPEs only when the decision is wrong; however because $f_i$ is unknown we cannot determine whether or not the model's prediction is correct. In the following section we develop an approach for cost-effective acquisition of examples that offers an alternative.

## 4. Goal-Oriented Active Learning

Instead of estimating directly which decisions are erroneous, we propose an alternative method based on a related property that avoids the need to know $f_i$. We propose acquiring labels for examples where a

[5] There is a incentive whenever the estimated probability $\hat{f}_i$ and the true probability $f_i$ are on different sides of the threshold $f_i^{Th}$.

relatively small change in probability estimation can affect decision-making, and avoiding acquiring labels otherwise. Specifically, we will prefer acquisitions when $\hat{f}_i$ is closer to $f_i^{Th}$. For example, consider two scenarios concerning a given decision. In scenario A the estimated class probability is considerably higher than the threshold probability. In scenario B the estimated probability of response is only marginally greater than the threshold probability. In scenario A the evidence in the training data is strongly in favor of action. As a result a more substantial change in the estimated probabilities is necessary to affect the decision in Scenario A as compared to Scenario B, requiring more training examples to sway the estimation in favor of inaction (all else being equal). The approach we propose here acquires labeled examples pertaining to decisions that are likely to be less costly to affect; i.e., decisions for which a relatively small change in the estimation can change the preference order of choice. Of course, although the design is suggested by the theoretical development above, this is a heuristic method.

The new method we propose operates within the active learning framework presented in Figure 2. At each phase, $M \geq 1$ examples are selected from the set of unlabeled examples $UL$; their labels are acquired and the examples are added to the set of labeled training examples $L$. The effectiveness score is calculated as follows. Each example $x_i \in UL$ is assigned a score that reflects the relative effect the example is expected to have on decision-making if its label were acquired and the example added to the training set. In particular, the score is inversely proportional to the minimum absolute change in the probability estimation that would result in a decision different from the decision implied by the current estimation, i.e., the score of example $x_i$ is inversely proportional to $\left|\hat{f}_i - f_i^{Th}\right|$.

For selection, rather than selecting the examples with the highest scores ("direct selection," as is common in active learning), a sampling distribution is created. Specifically, the effectiveness scores are considered to be weights on the examples, and examples are drawn from a distribution where the probability of an example to be selected for labeling is proportional to its weight. Earlier work (Iyengar *et al.*, 2000; Saar-Tsechansky and Provost 2004) has demonstrated that sampling from a distribution of effectiveness scores is preferable to direct selection. It reduces the chance of acquiring labels of outliers or other atypical examples (Saar-Tsechansky and Provost 2004).

Figure 3: The Goal Oriented Active Learning (GOAL) Algorithm

**Input**: Set of unlabeled examples $UL$, initial set of labeled examples $L$, Inducer (learner) $I$, stopping criteria.

**While** (stopping criterion)

1 Apply inducer $I$ to $L$, resulting in estimator $E$

2 Apply estimator $E$ to $UL$

3 For all examples $\{x_i \mid x_i \in UL\}$ compute $D(x_i) = 1 / \lambda \cdot \left(\beta + \left|\hat{f}_i - f_i^{Th}\right|\right)$, where $\lambda = \sum_{i=1}^{size(UL)} 1 / \left(\beta + \left|\hat{f}_i - f_i^{Th}\right|\right)$, such that $D$ is a distribution

3 Sample from the probability distribution $D$, a subset $S$ of $M$ examples from $UL$ without replacement

5 Remove $S$ from $UL$, label examples in $S$, and add them to $L$

6 Select the top $M$ instances from $UL$, remove them from $UL$, label them and add them to $L$

**End While**

**Output**: Estimator $E$ induced from $L$

Formally, the sampling-distribution weight $D(x_i)$ assigned to example $x_i \in UL$ is given by

$D(x_i) = \dfrac{1}{\lambda \cdot \left(\beta + \left|\hat{f}_i - f_i^{Th}\right|\right)}$, where $\beta$ is some small real number to avoid division by zero (in the empirical evaluation that follows $\beta = 0.001$), $f_i^{Th} = \dfrac{\Psi_i + C - U_i^F}{U_i^S - U_i^F}$ as above, and $\lambda$ is a normalizing factor given by $\lambda = \sum_{i=1}^{size(UL)} 1 / \left(\beta + \left|\hat{f}_i - f_i^{Th}\right|\right)$, to make $D$ a distribution. Figure 3 presents pseudocode of the method, which we call Goal-Oriented Active Learning (GOAL).

## 5. A Direct Marketing Campaign Case Study

We evaluate GOAL using data from a direct-marketing campaign. We use these data for evaluation because they comprise real consumer interactions along with information-acquisition costs. The data pertain to a charity's periodic solicitations to potential donors and are publicly available through the University of California at Irvine repository [Blake et al., 1998]. The challenge of direct marketing stems from the non-negligible cost of solicitation; hence the organization seeks to maximize campaigns' profits by better targeting potential donors. For these data, each solicitation costs 68 cents (printing, mailing costs, etc.) and response amounts range from $1 to $200, with an average of $15. The average response rate is approximately 5%. Because of the low response rate and the cost of solicitation, informed decisions that minimize wasteful solicitations are critical to the success of the campaigns. Importantly for this paper, the estimated probabilities come from an induced predictive model that in turn requires costly acquisitions of customer responses. For a cost-effective utilization of the charity's donor base, it is important to reduce the number of solicitations necessary to allow for effective targeting, or alternatively, to increase the effectiveness of targeting for a given solicitation budget.

### 5.1 Acquisition Strategies for the Direct Marketing Problem

Let us first describe the context in which active acquisition of consumer responses takes place. Given an acquisition budget, an acquisition strategy solicits potential donors and acquires their responses (i.e., whether or not a given consumer responded to the solicitation, and if so, in what amount). These become the labeled training data. Once the labels are acquired, a targeting model is induced from the training data and is subsequently employed to target potential donors for a new campaign. In the new campaign, a successful solicitation is one that results in a contribution that exceeds the solicitation cost. So, the objective is to reduce the acquisition cost necessary to achieve a particular level of profit, or alternatively to increase the profit for a particular acquisition investment.

We will compare three label-acquisition strategies:

(1) Acquisition of responses from a representative set of donors, using random sampling from a uniform distribution. Uniform random sampling is the most widely applied practice for the acquisition of labels based on a set of unlabeled training examples. In spite of its simplicity of implementation, random sampling is remarkably effective because it attempts (implicitly) to obtain a representative sample of the example space. We will refer to this label-acquisition strategy as RANDOM.

(2) An active-learning method that focuses on error reduction: BOOTSTRAP-LV. Because probability estimates are used to evaluate the expected profitability of alternative solicitations, an acquisition strategy that improves these estimations is also likely to improve targeting decisions. To our knowledge BOOTSTRAP-LV is the only generic method designed specifically to reduce class probability estimation error. BOOTSTRAP-LV follows the traditional paradigm of using uncertainty in estimations to calculate effectiveness scores. Specifically, BOOTSTRAP-LV estimates the variance in learned models' response probabilities, for each potential example, and assigns a higher score to the acquisition of responses from examples with higher variance. BOOTSTRAP-LV was shown to result in lower probability estimation error for a given acquisition cost compared both to random acquisition of responses, and to active learning designed for improving categorical classification accuracy (Saar-Tsechansky and Provost 2004).

(3) GOAL. Let the estimated probability that a potential donor $x_i$ would respond to a mailing be $\hat{f}_i$, the estimated contribution amount (described below) be $\hat{U}_i^S$ and the mailing cost be $C$. The profit from inaction is zero; hence a solicitation is initiated if $f_i \cdot \hat{U}_i^S - C \geq 0$ and the threshold probability is $f_i^{Th} = \frac{C}{\hat{U}_i^S}$.

Therefore, the weight assigned to acquiring donor $i$'s response in GOAL is given by $1/\lambda\left(\beta + \left|\hat{f}_i - f_i^{Th}\right|\right) = 1/\lambda\left\{\beta + \left|\hat{f}_i - \frac{C}{\hat{U}_i^S}\right|\right\}$.

## 5.2 Experimental Setting

In order to evaluate the three acquisition strategies, we compare the decision-making efficacy and profits generated from solicitation decisions derived from the models induced with each. We now describe the induction methods examined, the data partitioning, and the method for calculating generated profits.

For estimating the probability of response, we use three induction methods.[6] Our first experiments focus on Probability Estimation Trees (PETs)—unpruned C4.5 classification trees (Quinlan, 1993) for which the Laplace correction [Cestnik 1990] is applied at the leaves. Not pruning and using the Laplace correction has been shown to improve the CPEs (Provost and Domingos 2003; Perlich et al. 2003). Subsequently, in order to demonstrate the generic nature of the methods, we also compare the three acquisition strategies using logistic regression and Naïve Bayes (Mitchell 1997). For this application, revenues from successful solicitations are not known in advance and therefore also must be estimated from the data. We use a linear regression model based on a set of predictors that was identified in earlier studies.[7]

On a separate (holdout) set of potential donors, we compare the profits generated by each method for an increasing number of acquired, labeled training examples. More specifically, at each phase the responses of $M$ additional donors are acquired by each method and added to its respective set of training examples. Each point on each curve shown hereafter is an average over 10 independent experiments. For each experiment,

---

[6] The predictors are: household income range, date of first donation, date of most recent donation, number of donations per solicitation, number of donations given in the last 18 months, amount of last donation, and whether the donor responded to three consecutive promotions.

[7] The predictors for the regression model are: the amount of the most recent gift, the number of donations per solicitation, average donation amount in response to the last 22 promotions, and the estimated probability of donation as estimated by the CPE model. Following Zadrozny and Elkan (2001) the CPE estimation is incorporated as a predictor in the linear regression model to remove a sample selection bias. Because large gifts are rare there exists a selection bias towards one group of frequent donors who donate small amounts resulting in the regression model underestimating gifts by donors who contribute large amounts infrequently. To alleviate such a bias, Heckman (1979) recommends incorporating the probability of belonging to either group (i.e., the probability of making a donation) as a predictor in the regression model.

the data are randomly partitioned into: an initial set of labeled training examples selected at random (used to build the first model) $L$; an unlabeled pool of donors $UL$ from which the three strategies acquire additional responses, which then are added to $L$ (cumulatively as the curves progress); and an independent out-of-sample test set $T$ of potential donors whose responses and donations are known, for evaluating the three methods. To reduce variance, the same data partitions are used by all methods.

The profit for each method is calculated via the following simulated process (recall, responses are known to the experimenters for the entire test set). For each potential donor in the test set, either a solicitation is mailed or no action is taken. The solicitation is mailed if the expected revenue exceeds the solicitation cost. The cost of mailing is subtracted from the total profit whenever a solicitation is made; if a donor responds to a solicitation the actual donated amount is added to the overall profit. This profit calculation is depicted in Figure 4.

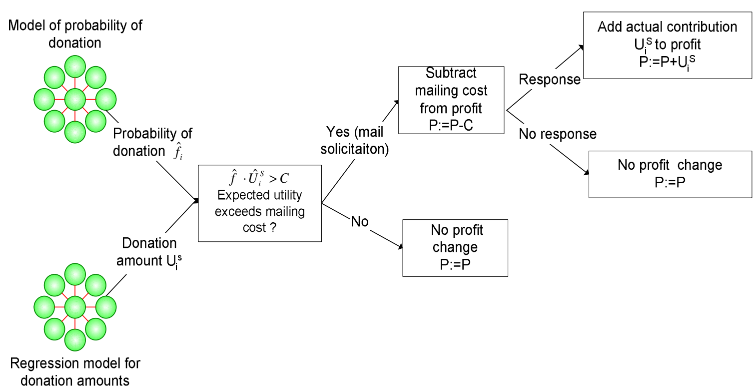



Figure 4: Decision-making profitability calculation from charity solicitations

## 5.3 Results

In order to evaluate the effectiveness of the GOAL acquisition policy we first measure the accuracy of mailing decisions enabled by each acquisition strategy. Specifically, we measure decision-making efficacy as the proportion of targeting decisions made correctly by each model. Ultimately, the capability to avoid non-profitable mailings as well as to identify profitable ones is critical for a campaign's success. Figure 5 shows the proportion of mailing decisions made correctly when GOAL, BOOTSTRAP-LV and RANDOM are used to acquire donor responses for model induction. Mailing decision accuracy is shown for an increasing cost of label (response) acquisitions. Each method was given (the same) 2000 initial training examples. At each phase 10 responses were acquired by each method and in total 5000 responses were acquired actively by each.

Note that initially all methods have access to the same, small set of responses. Therefore, the same probability estimation model is induced by all methods resulting in the same performance. As additional donors' responses are acquired by each method, the sets of responses available for training begin to differ in composition, resulting in different learned models. As Figure 5 shows, as more responses are acquired and the composition of the training sets diverges, the relative advantage of GOAL becomes more apparent. GOAL improves (on average) more decisions per acquisition than either of the other methods. For a given cost, a model trained with donor responses acquired by GOAL obtains a higher proportion of correct targeting decisions when compared with BOOTSTRAP-LV's CPE-error-reduction policy or with the acquisition of responses uniformly at random. Similarly, RANDOM acquisitions are clearly inferior to those obtained with BOOTSTRAP-LV. GOAL's superiority with respect to BOOTSTRAP-LV is statistically significant (p=0.05) once 200 responses are acquired by each method.

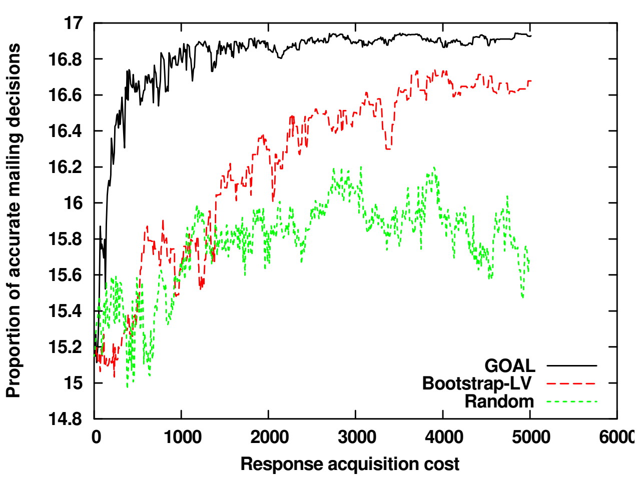
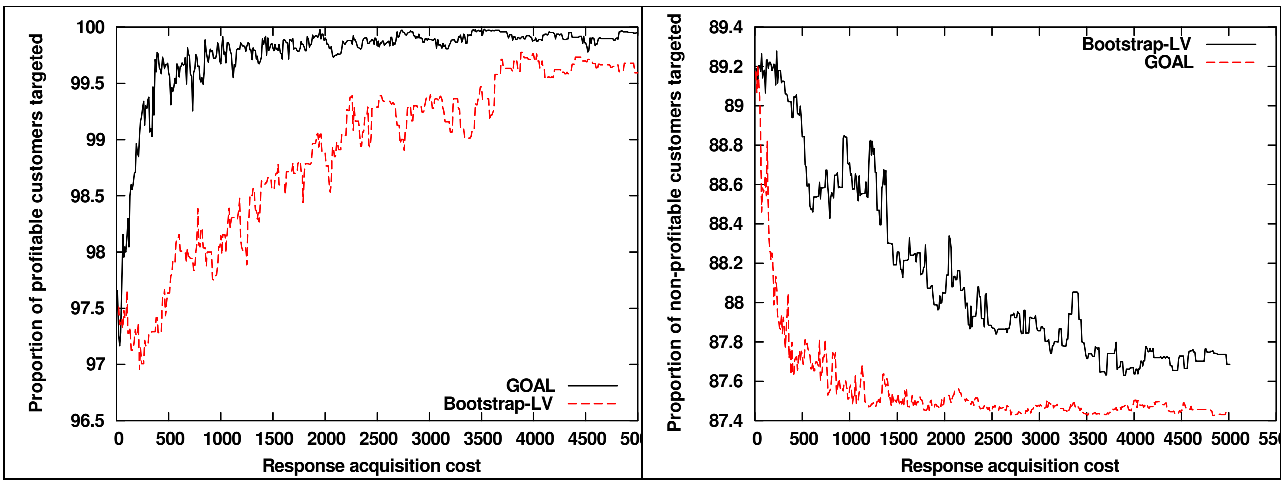



Figure 5: Mailing accuracy rate as a function of response acquisition cost

| Number of response acquisitions | Percentage Improvement |
|---|---|
| 200 | 7.84 |
| 300 | 8.63 |
| 400 | 9.57 |
| 500 | 7.74 |
| 1000 | 6.92 |
| 1500 | 4.63 |
| 2000 | 3.40 |
| 2500 | 2.45 |
| 3000 | 2.32 |

Table 1: Percent improvement in mailing effectiveness obtained using GOAL with respect to BOOTSTRAP-LV (p=0.05).

The difficulty of improving model accuracy sufficiently to improve targeting decisions is well demonstrated by the number of acquisitions required in order to obtain a given improvement in mailing decision performance. For example, BOOTSTRAP-LV must acquire more than 2000 responses in order to increase the mailing-decision accuracy from 15.1% to 16.4%; GOAL must acquire only about 300 responses to exhibit the same improvement in performance. On average over all acquisition phases, GOAL increases the mailing-decision accuracy rate by 3.66%. The largest improvements are exhibited in the early acquisitions phases, where GOAL results in more than 9% improvement compared to BOOTSTRAP-LV. Table 1 shows the improvements in the proportion of correct mailing decisions obtained with GOAL, over BOOTSTRAP-LV, for an increasing number of response acquisitions. The reported improvements are significant according to a paired t-test (p=0.05).




a. Proportion of profitable donors targeted

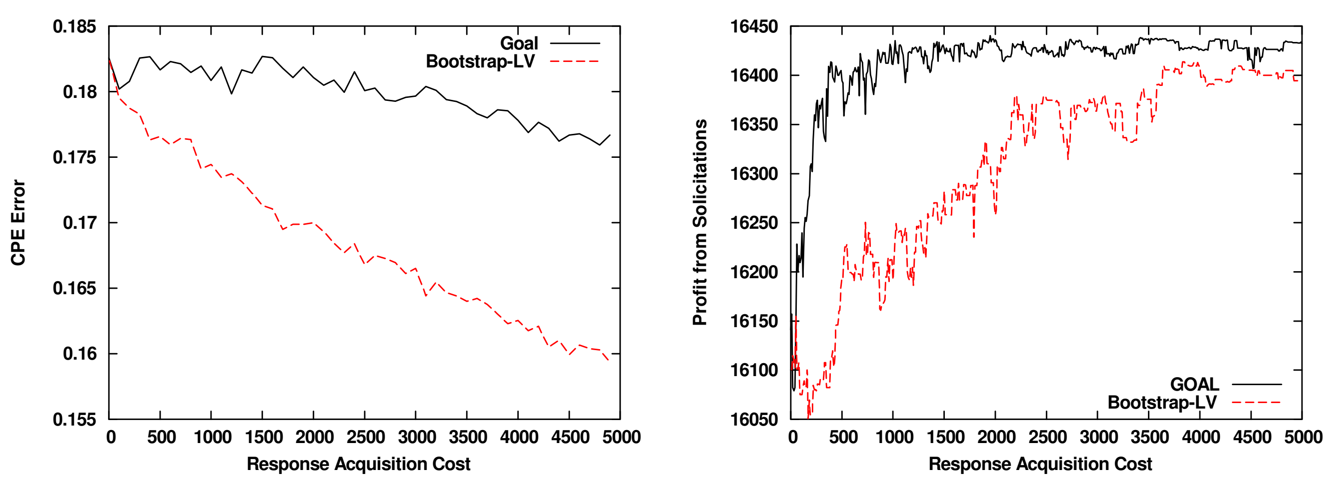
b. Proportion of non-profitable donors targeted

Figure 6: Proportion of profitable donors targeted with GOAL and BOOTSTRAP-LV.

One element of campaign profitability is the model's effectiveness at targeting *profitable* donors, and similarly avoiding targeting non-profitable ones. Figure 6a reports, for an increasing number of response acquisitions, the proportion of the set of profitable donors targeted with GOAL and BOOTSTRAP-LV. Figure 6b reports the proportion of non-profitable donors targeted by each. Clearly, the training responses acquired with GOAL produce a model that identifies more of the profitable donors and that avoids targeting more of the non-profitable donors than do the training responses acquired with the error-reducing approach. GOAL's performance is already statistically significantly superior (p=0.05) before 200 responses are acquired.

Taken together, these results strongly support our contention that for this problem, GOAL's decision-centric response acquisitions are more informative and effective (on average) compared to acquiring training responses to improve CPE accuracy generally (using BOOTSTRAP-LV).

Of course it is possible that the improved decision accuracy afforded by GOAL simply is a result of improved class probability estimation. GOAL is designed to improve decisions directly, while BOOTSTRAP-LV's acquisition strategy is designed to improve the model's CPEs. However, perhaps BOOTSTRAP-LV is not effective at its intended purpose. Figure 7a compares the error of the probability estimates produced by GOAL with those generated by BOOTSTRAP-LV (as always, on out-of-sample test sets). Probability estimation accuracy is measured with *BMAE* (Best-estimate Mean Absolute Error), computed as $BMAE = \frac{\sum_{i=1}^{N} |p_{Best}(x_i) - p(x_i)|}{N}$, where $p(x_i)$ is the probability estimated by the model under evaluation (and that was induced from the selected subset of the available examples); $N$ is the number of test examples for which the models are evaluated; $p_{Best}$ is a surrogate to the best estimated probability and is estimated by a "best" model induced using the entire set of available examples $L \cup UL$ (and using a more complicated modeling approach, a Bagged-PET which generally produces superior CPEs as compared to a PET (Provost and Domingos, 2003; Perlich et al. 2003)).




a. Error rates of estimated probabilities of response.

b. Profits from direct mailing.

Figure 7: Comparison of mailing profitability and CPE accuracy using a PET model.

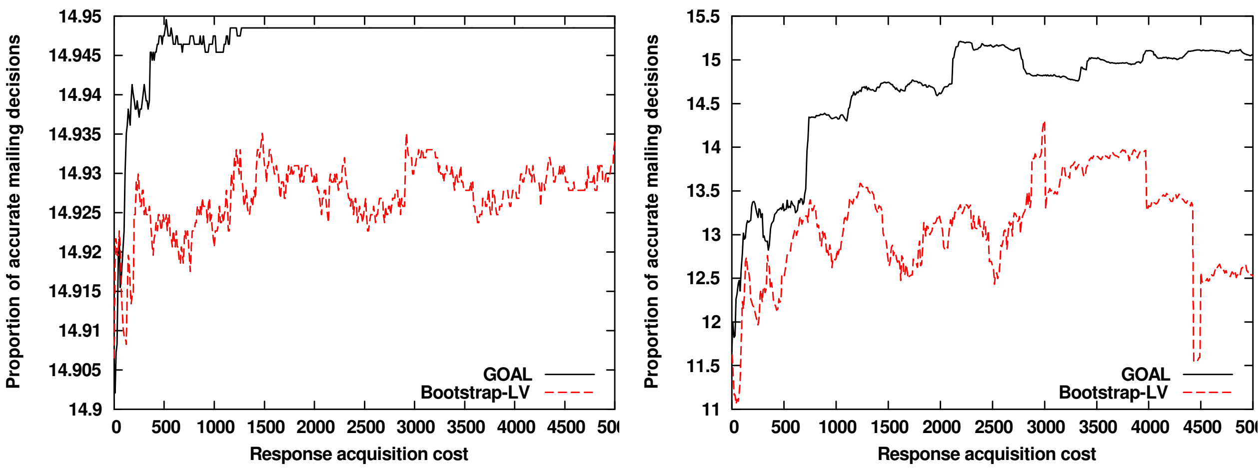
In contrast to the pattern shown in Figure 5, on average the class probability estimations obtained with GOAL for a given acquisition cost are considerably worse than those obtained with BOOTSTRAP-LV. BOOTSTRAP-LV's improved average error is statistically significant (p=0.05) after both strategies have acquired 600 examples. Bootstrap-LV acquires responses that improve the accuracy of response probability estimation, regardless of the subsequent impact on decision-making. As the discussion in Section 3 suggests, some improvements in CPE accuracy may not impact decision-making. Because these improvements come at a cost, they result in wasteful solicitations that are not rewarded with improved mailing decisions. GOAL is designed to avoid such acquisitions, and it is able to exhibit improved decision making for a given cost; however the average probability estimations it produces are inferior.

Figure 7b explores whether the improved decision accuracy also results in superior profitability. The graph shows the profits generated from direct marketing mailings for increasing cost of response acquisitions. From 200 response acquisitions onward and until 2500 responses are acquired, GOAL results in statistically significantly higher profits than BOOTSTRAP-LV, according to a paired t-test (p=0.05)—again, GOAL produces models yielding better targeting decisions.

In summary, Figures 6 and 7a show that GOAL and BOOTSTRAP-LV each excels at the task for which it was designed: BOOTSTRAP-LV to improve the average CPEs and GOAL to improve decision-making. In particular, while BOOTSTRAP-LV obtains considerably better average probability estimations for a given cost, these improvements often do not result in more accurate targeting. GOAL, on the other hand, avoids many costly CPE improvements that are not likely to alter decisions, and thereby reduces the cost of obtaining a given level of decision-making efficacy.

We designed GOAL to be generic: it does not depend on the form of the model or on the induction algorithm. Hence, it can be applied with any model for estimating class probabilities. Whether it will be effective for various models must be demonstrated empirically. Figure 8a compares the accuracy of targeting decisions using logistic regression models induced with GOAL and with BOOTSTRAP-LV. Again, GOAL is able to acquire more-informative responses for inducing donor response models. GOAL's superior performance is statistically significant (p=0.05) once 200 donor responses are acquired.

Figure 8b shows the direct-mailing decision accuracy for GOAL and BOOTSTRAP-LV when the base model is a Naïve Bayes classifier. For this model as well, GOAL results in better decision-making for a given investment in response acquisition. For the Naïve Bayes model, GOAL's superiority is statistically significant (p=0.05) once more than 750 donor responses have been acquired by each policy.

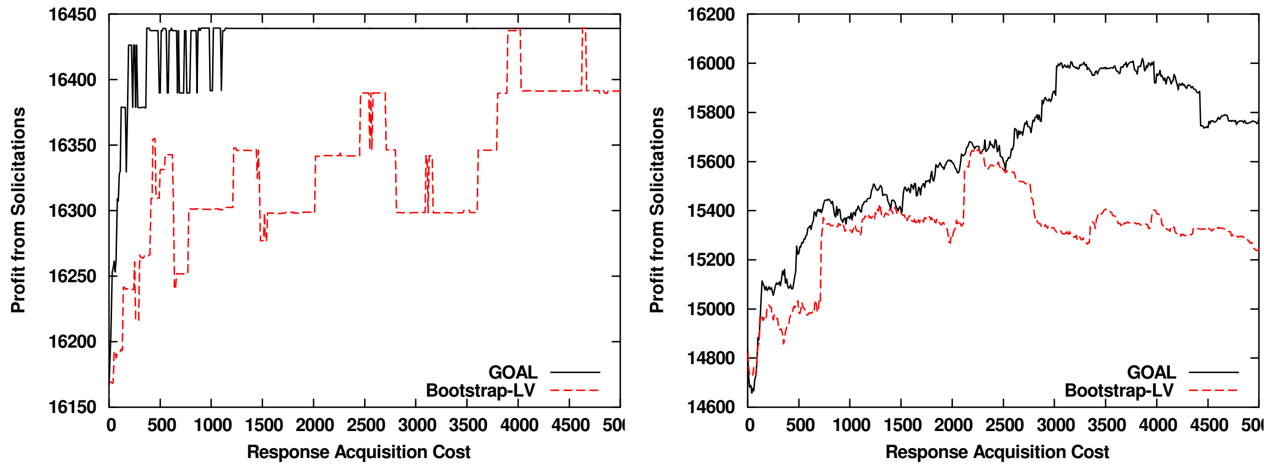



a. Mailing accuracy rate with logistic regression

b. Mailing accuracy rate with a Naïve Bayes model

Figure 8: Mailing decision accuracy rate using GOAL and BOOTSTRAP-LV

Figure 9 compares the profitability resulting from GOAL's acquisitions to the profitability obtained with BOOTSTRAP-LV. For the logistic regression model, GOAL's improvement is significant (p=0.05) once 200 acquisitions are made by each method. Once 4000 responses are acquired the two acquisition policies exhibit comparable performance again. GOAL results in statistically superior profitability with the Naïve Bayes model once more than 2600 responses have been acquired.




a. Logistic regression model.

b. Naïve Bayes model

Figure 9: Direct marketing profitability with GOAL and BOOTSTRAP-LV

## 6. Limitations

Above we mentioned that although GOAL's technique is informed by the theoretical analysis, it is essentially a heuristic (as are all existing active learning methods, except arguably OED, which has its own drawbacks). Thus, there is room for alternative methods that may perform even better. For example, we have ignored that the population to which the method will be applied (e.g., the potential customer base) may be available when the model is built. It may be that the similarities between "training" customers and to-be-targeted customers could be taken into account to improve the potential decision impact of response acquisitions. We rely on the weighted sampling to take these similarities into account implicitly.

We have not taken into account any information about the expected distribution of CPEs, beyond the point value for each case. As with Bootstrap-LV, it may be possible to improve GOAL by looking more carefully at the distribution. For example, a case with a high-variance probability estimate might be preferred over a low-variance estimate with the same point value or mean. Also, in this study we focused the active acquisitions to benefit the CPE model. However, we also induce a regression model for donation amounts. It may be that a method could actively solicit to optimize the two models simultaneously. In addition, we assumed that the cost of acquiring labels is uniform across examples (e.g., customers). For applications with non-uniform acquisition costs, the method would benefit from an extension that integrated the acquisition costs with the selection criteria. Ideally, the method would maximize the expected value of the information acquired.

Finally, the employment of active interactions with consumers also gives rise to a new challenge of balancing between the benefits of offering consumers product recommendations intended to increase sales at present versus acquisition of information that will benefit future actions. The tradeoff between activities initiated to support learning and actions intended to exploit what is already known has been explored in the robotics literature where robots balance between initiating actions for accomplishing a given task learning actions that attempt to improve the robot's ability to predict the results of its future actions. A similar framework may prove useful for businesses as well, as they proactively initiate actions from which they can both benefit immediately and use to improve future actions (Pednault et al, 2002).

## 7. Conclusion and Managerial Implications

Because the information required for effective modeling often is costly to obtain, it is beneficial to devise mechanisms to direct the acquisition of data for cost-effective improvements to decision-making. In this paper we examined the induction of class probability estimation models used for comparing alternative courses of action in the presence of uncertainty. For reducing information costs, traditional active learning can be applied and it performs significantly better than the standard approach of acquiring information from a representative sample using uniform random sampling. However, label acquisition costs can be reduced even more if an acquisition strategy is designed to improve decision-making specifically.

The GOAL acquisition method is derived from theoretical observations regarding the conditions under which class probability estimation error is more likely to undermine decision-making. When applied to data from direct-marketing campaigns, GOAL identifies donors whose responses significantly improve models for donor targeting, as measured by decision accuracy and campaign profitability. GOAL's decision-centric strategy is superior to the alternatives. Examining the relationship between error reduction and decision-making efficacy reveals that the economies exhibited by GOAL indeed are derived from acquiring labels that will affect solicitation decisions, sometimes at the expense of CPE error reduction. An additional advantage of GOAL is that it has a simple, straightforward formulation, and that it does not require considerable computation (unlike alternatives like Bootstrap-LV).

This paper introduces the notion that decision-centric, active acquisition of modeling information can lead to more profitable model building and use. We have motivated the use of such techniques with examples of modeling consumer preferences and loyalty, which recently have been high-profile modeling applications (e.g., on-line recommendation mechanisms, churn prediction for banks and telecommunications companies, and other customer relationship management applications). These are applications for which the acquisition of labels for training data carries clear costs. There are many other uses of predictive modeling in business (e.g., West et al, 1997; Moe and Fader, 2004), most of which have associated data acquisition costs. A further implication of this work is that decision-centric data acquisition strategies should be considered elsewhere as well.[8]

The notion of decision-centric *active* information acquisition also suggests that businesses should consider modifying their strategies for acquiring information through normal business transactions. A firm, such as Amazon.com, that models consumer preferences for customized marketing of products can accelerate learning about consumers by proactively offering recommendations—not merely to induce immediate sales, but for the purpose of improving recommendations in the future. The decision-centric acquisition approach presented here suggests that such active acquisition of information may well result in better decisions in the future. For firms such as Amazon.com, such capacity potentially could be employed to accelerate induction of consumer preferences and to offer more accurate and effective recommendations, earlier.

[8] From the research perspective there is a striking lack of publicly available data sets that include all the cost/benefit information required to evaluate methods such as this. This should be less of a limitation from a managerial point of view, where sources of both data and cost/benefit information should be available more readily.

## 6. References

Allen F. and Gale, D. 1999. Innovations in financial services, relationships and risk sharing, Management Science 45, 1239-1253.

Blake, C.L. & Merz, C.J. 1998. UCI Repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science, [http://www.ics.uci.edu/~mlearn/MLRepository.html].

Breiman, L; Friedman, J. H.; Olshen, R. A.; and Stone, C. J. 1984. Classification and regression trees. Belmont, CA: Wadsworth International Group.

Cestnik, B. 1990. Estimating probabilities: A crucial task in machine learning. In Proceedings of the Ninth European Conference on Artificial Intelligence, 147-149.

Cohn, D., Atlas, L. and Ladner, R. 1994. Improved generalization with active learning. Machine Learning, 15:201-221.

Cortes, C., L. Jackel, S. Solla, V. Vapnik, and J. Denker, 1995. Learning curves: asymptotic values and rate of convergence. Advances in Neural Information Processing Systems 6, 327-334. Morgan Kaufmann.

Friedman. J. H. 1997. On bias, variance, 0/1-loss, and the curse-of-dimensionality. Journal of Knowledge Discovery and Data Mining, 55-77.

Hastie, T., R. Tibshirani, and J. Friedman, 2001. The Elements of Statistical Learning. New York: Springer-Verlag.

Iyengar, V. S., Apte, C., and Zhang T. 2000. Active learning using adaptive resampling. In Sixth ACM Special Interest Group on Knowledge Discovery in Databases International Conference on Knowledge Discovery and Data Mining. 92-98.

Kiefer, J. 1959. Optimal experimental designs. J. R. Stat. Soc., series B, vol. 21, pp. 272-- 304.

Lewis D. and Gale W. 1994. Training text classifiers by uncertainty sampling. In Proc. of Int'l ACM Conf. on Research and Development in Information Retrieval, pages 3--12, 1994.

Makadok, Richard, & Barney, Jay 2001. Strategic factor market intelligence: An application of information economics to strategy formulation and competitor intelligence, Management Science, 47 (12): 1621-1638.

Mitchel, T. M. 1997. Machine Learning. WCB/McGraw Hill, 1997.

Moe, W. and Fader, P. S. 2004. Dynamic Conversion Behavior at e-Commerce Sites. Management Science, 50 (3), 326-335.

Mookerjee V. and Mannino M. 1997. Redesigning case retrieval to reduce Information acquisition costs, Information Systems Research 8, 1, pp. 51-68.

Pednault, E., N. Abe, B. Zadrozny, H. Wang, W. Fan and C. Apte 2002. Sequential Cost Sensitive Decision Making with Reinforcement Learning. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

New York Times 2003a. Digging for nuggets of wisdom. The New York Time, October 16, 2003.

Perlich, C., Provost F., and Simonoff, J. S. 2003. Tree induction vs. logistic regression: A learning-curve analysis. Journal of Machine Learning Research 4 (2003) 211-255.

Provost, F. and Domingos P. 2003. Tree induction for probability-based ranking, Machine Learning Research, 52:3, 199-215, 2003.

Quinlan, J. R. 1993. C4.5: Programs for machine learning. Morgan Kaufman, San Mateo, California.

Resnik, P. and Varian. H. R. 1997. Recommender systems. Communications of the Association for Computing Machinery, 40(3):56-- 59.

Robins H. 1952. Some aspects of the sequential design of experiments. Bulletin American Mathematical Society, 55:527-535, 1952.

Rosset, S., Neumann, E., Eick, U., Vatnik, N. (2003). Lifetime Value Models for Decision Support. Data Mining and Knowledge Discovery Journal, Vol. 7, 321-339.

Roy, N. and McCallum, A. 2001. Toward optimal active learning through sampling estimation of error reduction. In Proceedings of the International Conference on Machine Learning, 2001.

Saar-Tsechansky M., and Provost F. 2004. Active sampling for class probability estimation and ranking. Machine Learning 54:2 2004, 153-178.

Seung H. S., Opper, and H. Smopolinsky. 1992. Query by committee. In Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory, 287-294, 1992.

Simon H. A. and Lea, G. (1974). Problem solving and rule induction: A unified view. In L.W. Gregg (ed.), Knowledge and Cognition. Chap. 5. Potomac, MD: Erlbaum, 1974.

Wall Street Journal. 1997. "Secret recipes" let Harrah's target its low-rollers at the individual level.

New York Times, 2003b. Making Intelligence A Bit Less Artificial, The New York Times, May 1 2003.

West P. M., Brockett, P. L., and Golden L. L. 1997. A Comparative Analysis of Neural Networks and Statistical Methods for Predicting Consumer Choice. Marketing Science,16 (4).

Winston, P. H. 1975. Learning structural descriptions from examples. In "The Psychology of Computer Vision", P. H. Winston (ed.), McGraw-Hill, New York, 1975.

Zadrozny B. and Elkan C. 2001. Learning and making decisions when costs and probabilities are both unknown. In Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining. 204-212.