Profit-Optimal Model and Target Size Selection with Variable Marginal Costs

Alan Abrahams *[†], Firass Hathout*, Andreas Staubli*, and Balaji Padmanabhan*
** Department of Operations and Information Management, The Wharton School, University of Pennsylvania*
[†] University of Cambridge Computer Laboratory, Cambridge, United Kingdom
{asa28, balaji}@wharton.upenn.edu

Abstract

Various organizations, from catalogue order companies to credit card and insurance institutions, employ direct-mail-response as a core marketing strategy. As the demand of a given random selection of prospects is uncertain, many of these corporations use data mining techniques to characterize good prospects in their target audiences and improve the likelihood of response. Conventional approaches to model and target size selection in the field of data mining assume fixed marginal costs, and consequently static profit-to-cost ratios. In reality, however, marginal costs vary as a result of economies of scale or bulk discounts from suppliers. In this paper, we investigate the impact of variable marginal costs on profit-optimal model and target size selection. We prove analytically that profit-optimal target size selection depends on profit-to-cost ratios. Finally, we show that, to maximize profits, model and target-size selection should be cognizant of variable marginal costs.

1. Introduction

Direct-mail-response organizations frequently employ classification algorithms from the field of data mining for the purposes of predicting whether prospects are responders or non-responders [Rud2001]. A typical data mining exercise may employ one or more of a vast array of data mining techniques for classifying prospects: rule induction, decision trees, neural nets, logistic regression, naïve Bayes, or distance-based algorithms (e.g. k-nearest neighbour) to name a few. Each technique, in turn, may offer a variety of algorithms – for example: ID3, C4.5, C5.0, CHAID, and CART for decision trees; STAR, PRISM, CN2, and 1-R for rule induction [CHS98, Dun2003, Gro99, HK2001, HMS2001, SD94, TCC99, WF99]. Finally, a given algorithm may be seeded with different parameters.

The usage of various techniques, algorithms, and parameters results in the generation of multiple models, each with differing characteristics: recall and precision will vary across models and within models (for different target sizes). The challenge for direct-mail-response organizations is then to select the model and target size (mailing size) which maximizes profits. Previous model and target size selection approaches – see Sections 2 and 3 – have assumed fixed marginal costs, and consequently fixed profit-to-cost ratios. However, in real business environments, marginal costs will vary as target size (mailing numbers) and response numbers vary.

Variation in marginal costs may be due to economies of scale, bulk discounts provided by suppliers, and other factors encountered by companies using database marketing. For example, Royal Mail's direct mail service [RM2003] offers multi-tiered pricing, varying from about £3.50 per unit for 500 mailings, down to around £0.50 per unit for 20,000 mailings. Furthermore, discounts of up to 35% are offered for heavy users of parcel postage. This means that direct mail organizations can procure savings during both initial customer solicitation and final order fulfilment, through bulk discounts. In addition, companies may employ different means of promotion as target size changes: expensive personal calls may be made to a small number of highly likely buyers, whereas cheaper mailings may be sent to less likely prospects [LL98]. As another example of varying marginal costs, magazine publishers often encounter a financial penalty if their circulation is below a threshold agreed with their advertisers [Mal2002]. Per unit costs will therefore be higher for low circulation numbers, and the subscriptions department must delicately balance the costs of magazine promotion mailings against the severity of financial penalties on missed circulation targets.

In this paper, we propose a new mechanism for model and target size selection that is cognizant of variable marginal costs and profit-to-cost ratios. This paper is organized as follows. In Section 2, we provide a refresher on conventional model evaluation techniques like Confusion Matrices (§2.1), Lift Charts (§2.2), and Gain Charts (§2.3). Section 3 reviews previous work on model selection from the literature on data mining. In Section 4, we describe new analytic results that demonstrate that optimal target size depends on profit-to-cost ratios. We look, in turn, at various profit curve shapes for data mining models: increasing (§4.1), convex (§4.2), and decreasing (§4.3) under high, medium, and low profit-to-cost ratios respectively. For each curve, we provide the formulaic profit-to-cost ratios that produce curves of that shape. We then compare our approach to traditional techniques that assume fixed marginal costs, and we show that, in environments where costs vary with mailing and production volumes, our mechanism is able to discover higher profit models and target sizes (Section 5).

2. Background

Conventional approaches to model evaluation include confusion matrices, lift charts, and gain charts [TCC99]. We review each of these traditional techniques in the following sub-sections:

2.1. Confusion Matrices, Precision, and Recall

Table 1 below shows a sample confusion matrix for a given data mining model. The table can be interpreted as follows:

- *TN* (True Negatives): is the number of prospects predicted as being non-responders, which are actually non-responders.
- *FN* (False Negatives): is the number of prospects predicted as being non-responders, which are actually responders.
- *FP* (False Positives): is the number of prospects predicted as being responders, which are actually non-responders.
- *TP* (True Positives): is the number of prospects predicted as being responders, which are actually responders.

We can also derive the following figures:

- Using M to denote the total number of prospects to Mail, M is $FP + TP$.
- The *Recall* of the model is the percentage of respondents we obtained out of the total obtainable respondents: $Recall = TP / (FN+TP)$.
- Using RR_M to denote the Response Rate of the Model, RR_M is $TP/(FP+TP)$. RR_M is also known as the *Precision* of the model: precision is the percentage of predicted respondents that are actual respondents.
- Using RR_R to denote the Response Rate of mailing a Random Sample, RR_R is $(FN+TP)/(FN+TP+TN+FP)$. Note that RR_R is (theoretically) constant, irrespective of the size of the mailing (M).

		Predicted	
		No	**Yes**
Actual	**No**	*TN*	*FP*
	Yes	*FN*	*TP*

Table 1: Sample Confusion Matrix For a Given Model and Threshold

Typically, a Confusion Matrix is constructed for a given threshold. Only prospects who score above the given threshold are mailed, so the threshold determines the target mailing size. A reduction in threshold usually results in lower precision (RR_M), but higher recall (TP).

To illustrate the concept of a threshold, take the example of 14 scored prospects shown in Figure 1 below. Each prospect is denoted by a circle. The score assigned by the current model is shown within each circle. Actual responders are shown shaded, whereas actual non-responders are not shaded. For a classification threshold of 0.9, the recall is 37.5% (as 3 out of 8 actual responders are predicted) and the precision is 100% (as, of the 3 predicted responders, all are actual responders). Reducing the threshold to 0.6, the recall of the model is improved to 75% (as 6 out of 8 actual responders are predicted), but the precision of the model drops to 86% (as only 6 of the 7 predicted responders are actual responders).

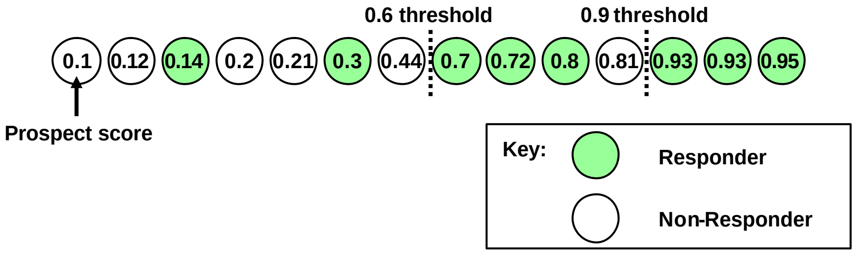

Figure 1: Scored Prospects

As we will show later (Sections 4 and 5), the profit optimal threshold (i.e. combination of precision and recall) depends on the profit-to-cost ratio for the business initiative that will be using the data mining model.

2.2. Lift Tables and Charts

To illustrate the value of a model above a random prospect selection, lift tables and charts can be constructed [LL98, Rud2001, WF99]. Lift tables and charts assume that prospects have been scored, and plot the results of mailing the top x% of prospects (as ranked by score). Lift is a measure of the improvement in response rate that the model provides over the random case. For example, in Table 2, mailing 10% of prospects (second data row of column A) nets 20% of respondents (second data row of column B) when using a hypothetical model. For a random mailing, sending to 10% of prospects would yield only 10% of respondents (second data row of column C). The lift of the model for the top 10% of prospects is therefore 2 (= B/C = 20% / 10%). Figure 2 is a graphical representation of the Lift Table, known as a Lift Chart. Notice that, for well-targeted mailings, models produce consistently more (cumulative) respondents than a random mailing.

A % of Prospects	B Model Recall (Cumulative Percent of Responses Correctly Predicted by Model)	C Cumulative Percent of Responses Correctly Predicted by Random Sample	D Cumulative Lift (= B/C)	E Model Response Rate
0%	0.0%	0%	n/a	n/a
10%	20.0%	10%	2.00	5.0%
20%	36.4%	20%	1.82	4.6%
30%	49.2%	30%	1.64	4.1%
40%	61.2%	40%	1.53	3.8%
50%	72.8%	50%	1.46	3.6%
60%	81.6%	60%	1.36	3.4%
70%	88.8%	70%	1.27	3.2%
80%	93.6%	80%	1.17	2.9%
90%	97.2%	90%	1.08	2.7%
100%	100.0%	100%	1.00	2.5%

Table 2: Lift Table

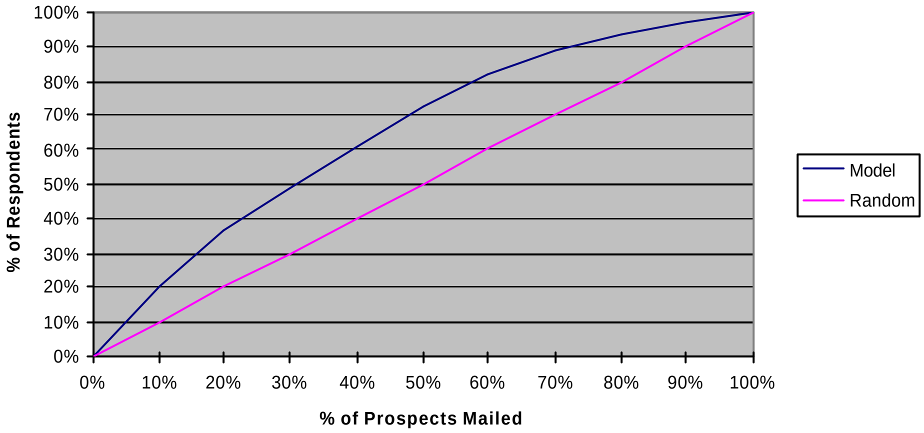

Figure 2: Lift Chart

2.3. Gain Tables and Charts

Lift Tables and Lift Charts do not give an illustration of the exact profitability of models. For this, Gain Tables and Charts have been employed [MP2003, Rud2001, US98]. The predicted profit of a targeted mailing[1] is given by:

$$\begin{aligned} \textit{Profit} &= \textit{Revenue} - \textit{Cost} \\ &= (\textit{Number of Respondents} \times (\textit{Revenue per Response} - \textit{Cost per Response}) \\ &\qquad - (\textit{Number Mailed} \times \textit{Cost per Mailing}) \\ &= (\textit{Number of Respondents} \times \textit{Profit per Response}) \\ &\qquad - (\textit{Number Mailed} \times \textit{Cost per Mailing}) \end{aligned}$$

$$\boxed{\textit{Profit} = (TP \times \textit{Profit per Response}) - ((FP+TP) \times \textit{Cost per Mailing}) \qquad \textbf{Formula 1}}$$

Assume a market size of 10,000 prospects, out of whom 250 are responders. Further, assume a Revenue per Response (i.e. Revenue per Sale) of $11, a Cost per Response (i.e. Cost per Sale) of $1,

[1] The *actual* profit will, of course, vary from this estimation as few models offer perfect predictions.

and a Cost per Mailing of $0.20. Using the response rates from the Lift Table above (Table 2), we obtain the Gain Table in Table 3 below, and the corresponding Gain Chart in Figure 3 below. It can be noticed that profit is maximized (at $840, highlighted), by mailing 60% of prospects (6,000 prospects) under these assumptions.

A	B	C
Number of Prospects	**Profit Using Model (Targeted Mailing)**	**Profit Using Random Mailing**
0	$0	$0
1,000	$300	$50
2,000	$510	$100
3,000	$630	$150
4,000	$730	$200
5,000	$820	$250
6,000	$840	$300
7,000	$820	$350
8,000	$740	$400
9,000	$630	$450
10,000	$500	$500

Table 3: Gain Table (Max profit highlighted)

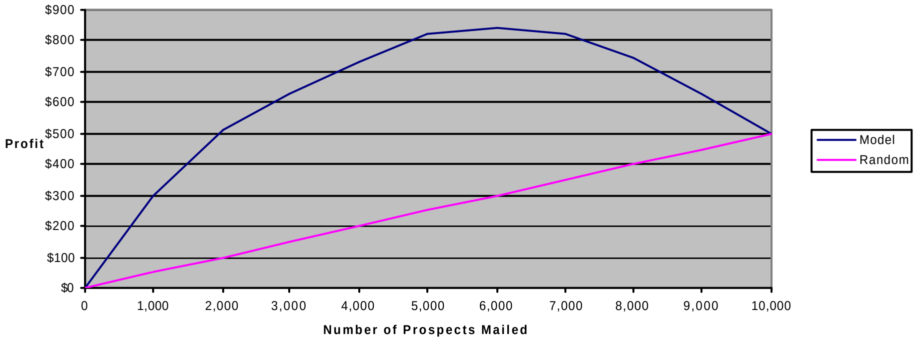

Figure 3: Convex Gain Chart

3. *Related Work*

Many traditional data mining course texts and software products [Dun2003, HK2001, HMS2001, SPSS2003] ignore business aspects when suggesting model quality metrics: profit-ignorant score and loss functions are used for model evaluation and selection. For example, Least Squares Error, Quadratic Loss and Information Loss provide metrics for the deviation between actual and predicted values (e.g. between actual class membership and predicted probability of being in a class). These measures of predictive accuracy (error rate) of a model are inappropriate for database marketing applications, as they presume that the costs of misclassification between classes are equal, which is highly unrealistic [AH99, Faw2001, LL98, DH2000, PF97, PFK98, PSS2000, Mal2002].

Lift and Gain Charts are, nevertheless, dealt with by some authors and software tools [Gro99, LZ99, MP2003, Rud2001, WF99, US98]. Often, though, authors and developers have assumed fixed marginal costs, and overlooked the common business case where variable marginal costs would impact upon model selection. For example, the Gain Chart shown in Figure 3 above is typical of current approaches, and is constructed using static marginal mailing and production costs irrespective of the number of prospects to mail or the number of responses.

4. Analytic Results

Substituting various different values for revenues and costs, we can notice that the shape of the Gain Chart varies. In the following sub-sections we investigate the effects of *high*, *medium*, and *low* profit-to-cost ratios on the shape of the Gain Chart curve. Further, we derive analytic results which show how target size selection depends on profit-to-cost ratios.

4.1. Increasing Profit Curve (High Profit-to-Cost Ratio = Seldom Exploitable Models)

Assuming a high profit-to-cost ratio (e.g. profit per response = $10; cost per mailing = $0.01; giving a profit-to-cost ratio of 1,000:1), we obtain the Gain Chart in Figure 4 below.

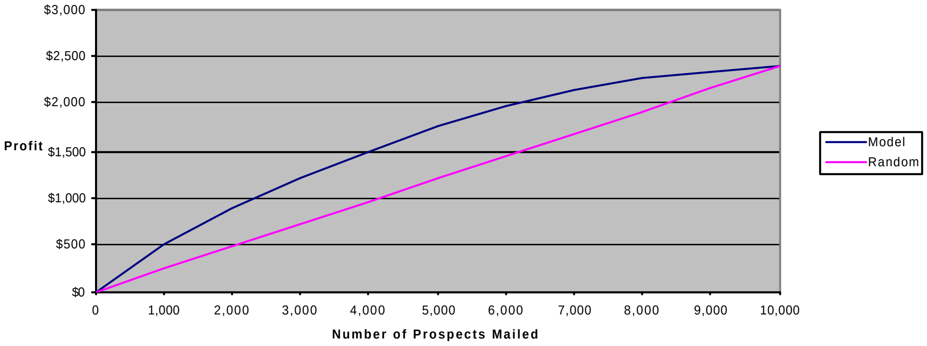

Figure 4: Increasing Gain Chart

With this profit and cost assumption, the profit from employing the model increases as the number of prospects mailed increases. Notice, however, that under this *high* profit-to-cost ratio presumption, the data mining model is not useful, as the maximum profit is obtained simply by mailing all prospects (i.e. using maximum target size). The data mining model is only helpful if some resource constraint prevents us from mailing all prospects, in which case the model will provide higher profits than the random case.

Analytically, the necessary and sufficient conditions for curves which give maximum profit when all prospects are mailed are as follows:

1. Profit from model is consistently greater than zero: ***Positive profit***
2. Profit from model is less than the profit from mailing all prospects: ***Non-maximal profit***

We will analyze each of these constraints in the following sub-sections:

4.1.1. Positive Profit

For the profit from a model to be consistently positive, we have:

Profit from Model > 0	***Constraint 1***

That is:

$$(TP \times ProfitPerResponse) - ((FP+TP) \times CostPerMail) > 0$$

… which gives …

$$(TP \times ProfitPerResponse) > ((FP+TP) \times CostPerMail)$$

In terms of profit-to-cost ratio we then have,

$$\frac{ProfitPerResponse}{CostPerMailing} > \frac{FP+TP}{TP}$$

… which simplifies to …

$$\frac{ProfitPerResponse}{CostPerMailing} > \frac{1}{Precision}$$ ***Formula 2 (Profitability Threshold)***

That is, for positive profits, the profit-to-cost ratio must exceed the inverse of the model's precision (response rate).

4.1.2. Non-Maximal Profit

For the profit from a model to be non-maximal, we have:

$$\textit{Profit from Model} = \textit{Profit by Mailing all Prospects}$$ ***Constraint 2***

That is:

$$\textit{Profit from Model} = \textit{Total Prospects} \times ((RR_R \times \textit{Profit per Response}) - \textit{Cost per Mailing})$$

Substituting **Formula 1** above for 'Profit from Model' we arrive at…

$$(TP \times \textit{Profit per Response}) - ((FP+TP) \times \textit{Cost per Mailing}) = \textit{Total Prospects} \times ((RR_R \times \textit{Profit per Response}) - \textit{Cost per Mailing})$$ ***Formula 3***

Now,

$$TP = \frac{FN+TP}{FN+TP+TN+FP} \times (FN+TP+TN+FP) \times \frac{TP}{FN+TP}$$

That is,

$$TP = RR_R \times \textit{Total Prospects} \times \textit{Recall}$$ ***Formula 4***

Also,

$$FP+TP = \frac{FN+TP}{FN+TP+TN+FP} \times (FN+TP+TN+FP) \times \frac{TP}{FN+TP} \times \frac{FP+TP}{TP}$$

Noticing that,

$$\frac{FP+TP}{TP} = \frac{1}{Precision}$$

… we see that …

$$FP+TP = \frac{RR_R \times TotalProspects \times Recall}{Precision}$$ ***Formula 5***

Substituting ***Formula 4*** and ***Formula 5*** into ***Formula 3*** , we see that, for the data mining model to be less profitable than mailing all prospects, we require that:

$$(RR_R \times TotalProspects \times Recall \times ProfitPerResponse) - (\frac{RR_R \times TotalProspects \times Recall}{Precision} \times CostPerMailing)$$
$$\leq TotalProspects \times ((RR_R \times ProfitPerResponse) - CostPerMailing)$$

Factoring *TotalProspects* from both sides, we get:

$$(RR_R \times Recall \times ProfitPerResponse) - (\frac{RR_R \times Recall}{Precision} \times CostPerMailing) \leq (RR_R \times ProfitPerResponse) - CostPerMailing$$

Thence:

$$((1 - \frac{RR_R \times Recall}{Precision}) \times CostPerMailing) \leq (RR_R \times (1 - Recall) \times ProfitPerResponse)$$

Thence:

$$\boxed{\frac{(1 - \frac{RR_R \times Recall}{Precision})}{(RR_R \times (1 - Recall))} \leq \frac{ProfitPerResponse}{CostPerMailing} \qquad \textbf{\textit{Formula 6 (Utility Bound)}}}$$

4.1.3. Summary

From §4.1.1 and §4.1.2, we see that, if our model profit curve is to have the general (increasing) shape shown in Figure 4 above then:

1. As seen in ***Formula 2*** (§4.1.1), the profit-to-cost ratio must be *higher* than the Profitabiilty Threshold (i.e. higher than the inverse of the response rate[2]), and
2. As seen in ***Formula 6*** (§4.1.2), the profit-to-cost ratio must be *higher* than the Utility Bound. If this holds, then the model's profits are lower than the profits obtained by mailing all prospects.

Table 4 below is the lift table for out hypothetical model of §2.2, extended with the Profitability Thresholds and Utility Thresholds computed from ***Formula 2*** and ***Formula 6***.

A % of Prospects	B Model Recall (Cumulative Percent of Responses Correctly Predicted by Model)	C Cumulative Percent of Responses Correctly Predicted by Random Sample	D Cumulative Lift (= B/C)	E Model Response Rate	F Profitability Threshold: For positive profit, profit-to-cost ratio must exceed inverse of response rate (1/E)	G Utility Bound: For model to be exploitable, profit-to-cost ratio must be less than
0%	0.0%	0%	n/a	n/a	n/a	n/a
10%	20.0%	10%	2.00	5.0%	**20**	45.00
20%	36.4%	20%	1.82	4.6%	22	50.31
30%	49.2%	30%	1.64	4.1%	24	55.12
40%	61.2%	40%	1.53	3.8%	26	61.86
50%	72.8%	50%	1.46	3.6%	27	73.53
60%	81.6%	60%	1.36	3.4%	29	86.96
70%	88.8%	70%	1.27	3.2%	32	107.14
80%	93.6%	80%	1.17	2.9%	34	125.00
90%	97.2%	90%	1.08	2.7%	37	**142.86**
100%	100.0%	100%	1.00	2.5%	40	n/a

Table 4: Profitability and Utility Table

From Table 4, it can be observed that:

- the model will be profitable if the profit-to-cost ratio exceeds the lowest Profitability Threshold figure (highlighted) in column F: 20:1[3].

[2] Response rate is synonymous with precision.

- the model will *not* be exploitable (except for the purposes of increasing market share) if the profit-to-cost ratio exceeds the highest Utility Bound figure highlighted in column G: 142.86:1. This is because, for profit-to-cost ratios higher than this Utility Bound, the company can maximize its profits simply by mailing all prospects.

Observe that, for Figure 4 above, the profit-to-cost ratio is higher than both the Profitability Threshold, and the Utility Bound, which ensures that the maximum profit from the model is always below the profit from mailing all prospects, even though the model profit is always increasing.

4.2. Convex Profit Curve (Medium Profit-to-Cost Ratio = Exploitable Models)

Assuming a medium profit-to-cost ratio (e.g. profit per response = $10; cost per mailing = $0.20; giving a profit-to-cost ratio of 50:1), we obtain the convex Gain Chart already shown in Figure 3 above. Observe that, here, the profit-to-cost ratio is between the Profitability Threshold and the Utility Bound.

Notice that, under this *medium* profit-to-cost ratio presumption, the data mining model allows us to maximize profits by choosing an intermediate target size: in our example (see Table 3 and Figure 3 above), mailing 60% of prospects will maximize profits.

Analytically, we can prove that the Gain Chart will be *convex*, with a higher maximum profit than mailing all prospects, under the following circumstances:

- *Profit from Model > 0*, and,
- *Profit from Model > Profit by Mailing all Prospects*

Using the same reasoning process as that used in §4.1 above, we find that the profit curve will be convex when the profit-to-cost ratio is *lower* than the Utility Bound, but *higher* than the Profitability Threshold.

4.3. Decreasing Profit Curve (Low Profit-to-Cost Ratio = Seldom Exploitable Models)

Assuming a low profit-to-cost ratio (e.g. profit per response = $10; cost per mailing = $0.75: giving a profit-to-cost ratio of 13.33:1 which is below the Profitability Threshold for our example model), we obtain the Gain Chart in Figure 5 below.

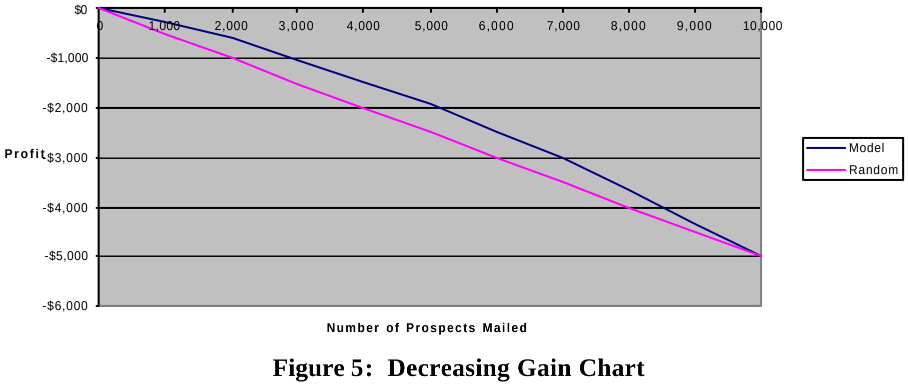

Figure 5: Decreasing Gain Chart

Notice that, under this *low* profit-to-cost ratio presumption, the data mining model is not useful, as we might as well mail no prospects (i.e. use minimum target size) in order to maximize our profits

[3] Because we have divided the data into deciles, these ratios are approximate. More accurate ratios can be obtained by dividing the data into percentiles or smaller segments when constructing the lift table. However, a downside to using smaller intervals is that the figures in the lift chart will capture the specific nuances of applying the model to the current test set, and may not generalize well to other test sets.

(minimize our losses). The data mining model is only helpful if we are seeking to increase market share by mailing at least some prospects, in which case the model will allow us to achieve our objective with lower losses than would be suffered with a random mailing.

Analytically, we can prove that the Gain Chart will be *decreasing* under the following circumstances:

$$\textit{Profit from Model} < 0$$

Using the same reasoning process as that used in §4.1 above, we find that the profit curve will be decreasing when the profit-to-cost ratio is *lower* than the Profitability Threshold.

5. Comparing Model and Target Size Selection with Fixed vs Variable Marginal Costs

It is clear from the analysis in Section 4 above that the usefulness of a data mining model is highly contingent upon the company's profit-to-cost ratio. Variable marginal costs will have an impact upon profit-to-cost ratios[4]: higher mailing and response sizes could reduce marginal costs through supplier discounts or economies of scale. Therefore, higher target sizes influence profit-to-cost ratio and hence the model selection decision, because different models produce different precision and recall for a selection of target sizes.

5.1. Model and Target Size Selection under *Fixed* Marginal Cost

Assume two new models – Model 1 and Model 2. Had we assumed static revenues per unit of $11 and **fixed marginal costs** (of $0.25 per unit for mailing, and $1 per unit for production) we could have obtained the profit curves shown in Figure 6 below[5]. For the case of fixed marginal costs, using Model 1 and mailing size of 4,000 provides maximum profit ($720).

5.2. Model and Target Size Selection under *Variable* Marginal Cost

Let us now assume static revenue of $11 per unit, and take into account **variable marginal costs** of mailing and production according to the following scheme:

Mailing Costs	
= 7,500 units mailed:	*$0.20 per unit*
> 7,500 units mailed:	*$0.18 per unit*

	Production Costs	**Production Profit**
= 150 units produced:	*$1 per unit*	*$10 per unit*
> 150 units produced:	*$0.50 per unit*	*$10.50 per unit*

Using these variable marginal costs, we obtain the profit curves shown in Figure 7 below[6]. Importantly, it becomes evident that, with variable marginal costs, Model 2 and mailing size 8,000 now become the profit-optimal solution ($965). Had we remained with Model 1 and mailing size 4,000 – the profit-optimal solution suggested by the fixed marginal cost scenario (§5.1 above) – we would have obtained a substantially lower profit ($796) in a world of variable marginal costs[7].

Various previous authors [AH99, DH2000, Faw2003, PF97, PFK98, PF2000] have offered an analytical exposition of the precise circumstances under which models of oscillating dominance emerge. These authors show that conventional approaches like Receiver Operating Characteristics (ROC) curves and Area Under Curve (AUC) analysis are unable to distinguish the optimal model in cases where model strength varies with target size. They describe advanced representation techniques which are able to show how optimal model choice is sensitive to different misclassification costs.

[4] Price uncertainty, which is outside the scope of this paper, is another factor that can result in varying profit-to-cost ratios.

[5] See Table 5 and Table 6 in the Appendix for the source data for Figure 6.

[6] See Table 5 and Table 7 in the Appendix for the source data for Figure 7.

[7] We have assumed throughout our discussion that threshold adjustment is the only mechanism for increasing the number of respondents. Note, however, that if the company is able to increase its customer database (e.g. through list purchase) it may be able to achieve economies of scale, and thence maximum profits, whilst still remaining with Model 1.

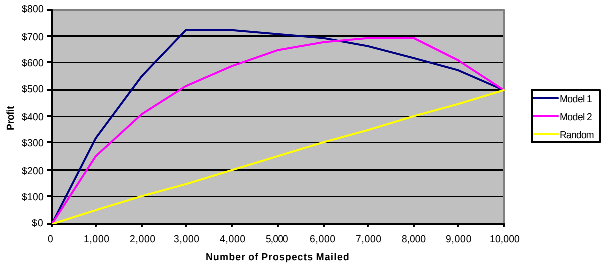

Figure 6: Gain Chart For Two Models Based on *Fixed* Marginal Costs

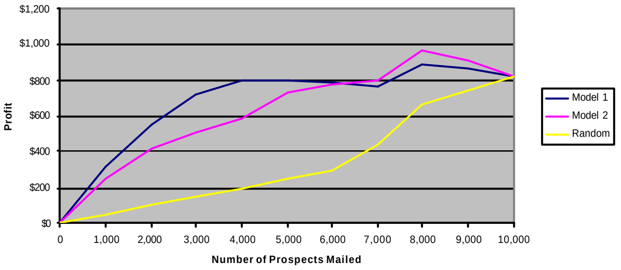

Figure 7: Gain Chart For Two Models Based on *Variable* Marginal Costs

6. Conclusion

In this paper, we have shown that the worth of a data mining model is largely determined by the profit-to-cost ratio of the particular business case being tackled. For very high profit-to-cost ratios (higher than both the Profitability Threshold and Utility Bound), models are seldom useful, as – in the absence of resource constraints – it is most profitable to solicit the entire population of prospects. For very low profit-to-cost ratios (lower than the Profitability Threshold), and assuming no market share growth imperative, losses can be minimized by cancelling the campaign altogether. It is for medium profit-to-cost ratios (between the Profitability Threshold and the Utility Bound) that models shine – helping to maximize profits by targeting a subset of likely respondents. Moreover, selection of a profit-optimal model is influenced by varying marginal costs. This is because varying marginal costs result in fluctuations in the profit-to-cost ratios, and hence would influence the solicitation strategy chosen.

For high and low profit-to-cost ratio scenarios, our results here are valuable because, based on early profit, cost, and model precision and response rate estimates[8], businesses can determine *in advance* whether it is worthwhile to spend money building data mining models. For medium profit-to-cost ratios, our results are valuable because they show that model selection is not driven solely by the model's own characteristics – rather, variable marginal costs have a meaningful effect on profit-optimal model and target size choice.

Acknowledgements

Alan Abrahams in funded by *Microsoft Research Cambridge*.

References

[AH99] Adams NM and Hand DJ. Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition*. 32. Elsevier Science Ltd. 1999. pp. 1139-1147.

[CHS98] Cabena P, Hadjinian P, Stadler R, Verhees J, and Zanasi A. *Discovering Data Mining: From Concept to Implementation*. Prentice Hall. 1998.

[DH2000] Drummond C and Holte RC. Explicitly Representing Expected Cost: An Alternative to ROC Representation. *Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining (KDD'2000)*. Boston, MA. 2000.

[Dun2003] Dunham MH. *Data Mining: Introductory and Advanced Topics*. Prentice Hall. 2003.

[Faw2001] Fawcett T. Using Rule Sets to Maximize ROC Performance. *2001 IEEE International Conference on Data Mining (ICDM'01)*. San Jose, CA. November 2001.

[Faw2003] Fawcett T. *ROC Graphs: Notes and Practical Considerations for Data Mining Researchers*. Hewlett Packard Laboratories Technical Report HPL-2003-4. 2003.

[Gro99] Groth R. *Data Mining: Building Competitive Advantage*. Prentice Hall. 1999.

[8] [PSM99] offers a domain-independent rule of thumb formula for estimating lift curves (precision and response rates) and resulting profits.

[HK2001] Han J and Kamber M. *Data Mining: Concepts and Techniques*. Morgan Kaufmann. 2001.

[HMS2001] Hand D, Mannila H, and Smyth P. *Principles of Data Mining*. MIT Press. 2001.

[LL98] Ling CX and Li C. Data Mining for Direct Marketing: Problems and Solutions. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98)*. New York, NY. 1998.

[LZ99] Levin N and Zahavi J. *Data Mining*. 1999. http://www.urbanscience.com/data_mining.pdf

[Mal2002] Malthouse EC. Performance-based Variable Selection for Scoring Models. *Journal of Interactive Marketing*. 16(4). Wiley Periodicals, Inc Autumn 2002. pp. 37-50.

[MP2003] MegaPuter Corporation. *PolyAnalyst 4.5*. 2003. http://www.megaputer.com/products/pa/index.php3

[PF97] Provost F and Fawcett T. Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions. *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD'97)*. Huntington Beach, CA. 1997.

[PFK98] Provost F, Fawcett T, and Kohavi R. The Case Against Accuracy Estimation for Comparing Induction Algorithms. *Proceedings of the Fifteenth International Conference on Machine Learning (ICML'98)*. Madison, WI. 1998.

[PF2000] Provost F, and Fawcett T. Robust Classification for Imprecise Environments. *Machine Learning*. 42(3). Kluwer Academic Publishers. pp. 203-231. 2000.

[PSM99] Piatetsky-Shapiro G, and Masand B. Estimating Campaign Benefits and Modeling Lift. *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD'99)*. San Diego, CA. 1999.

[PSS2000] Piatetsky-Shapiro G, and Steingold S. Measuring Lift Quality in Database Marketing. *SIGKDD Explorations*. 2(2). Association for Computing Machinery SIGKDD. December 2000. pp 76-80.

[RM2003] Royal Mail Group plc *Royal Mail Account Plus* (parcel service) and *DM Online* (Direct Mail Online). Available from http://www.royalmail.co.uk/ and http://www.dm-online.co.uk/

[Rud2001] Rud OP. *Data Mining Cookbook*. Wiley. 2001.

[SD94] Sestito S and Dillon T. *Automated Knowledge Acquisition*. Prentice Hall. 1994.

[SPSS2003] SPSS Inc *Clementine – Data Mining, Predictive Models*. 2003. http://www.spss.com/spssbi/clementine/index.htm

[TCC99] Two Crows Corporation. *Introduction to Data Mining (3rd Edition)*. 1999. http://www.twocrows.com/intro-dm.pdf

[US98] Urban Science. *GainSmarts*. 1998. http://www.urbanscience.com/

[vdP98] van der Putten, P. Data Mining in Direct Marketing Databases. In: Baets W (ed.) *Complexity and Management: A Collection of Essays*. World Scientific Publishers. Singapore 1999.

[WF99] Witten IH and Frank E. *Data Mining: Practical Machine Learning Techniques with Java Implementations*. Morgan Kaufmann. 1999.

Appendix

A	B	C	D	E	F	G	H
Number of Prospects	Number of Responses Correctly Predicted by Model 1 (Current Decile)	Number of Responses Correctly Predicted by Model 2 (Current Decile)	Number of Responses Correctly Predicted by Random Sample	Model 1		Model 2	
				Recall	Precision	Recall	Precision
0	0	0	0	0%	n/a	0%	n/a
1,000	52	45	25	21%	5.20%	18%	4.50%
2,000	43	36	25	38%	4.75%	32%	4.05%
3,000	37	30	25	53%	4.40%	44%	3.70%
4,000	20	28	25	61%	3.80%	56%	3.48%
5,000	19	26	25	68%	3.42%	66%	3.30%
6,000	18	23	25	76%	3.15%	75%	3.13%
7,000	17	21	25	82%	2.94%	84%	2.99%
8,000	16	20	25	89%	2.78%	92%	2.86%
9,000	15	12	25	95%	2.63%	96%	2.68%
10,000	13	9	25	100%	2.50%	100%	2.50%

Table 5: Recall and Response Rates for Two New Models, Model 1 and Model 2

A Number of Prospects	B Profit Using Model 1	C Profit Using Model 2	D Profit Using Random Mailing
0	$0	$0	$0
1,000	$320	$250	$50
2,000	$550	$410	$100
3,000	$720	$510	$150
4,000	$720	$590	$200
5,000	$710	$650	$250
6,000	$690	$680	$300
7,000	$660	$690	$350
8,000	$620	$690	$400
9,000	$570	$610	$450
10,000	$500	$500	$500

Table 6: Gain Table for Two Models, with *Fixed* Marginal Costs (Max Profit Highlighted)

A Number of Prospects	B Profit Using Model 1	C Profit Using Model 2	D Profit Using Random Mailing
0	$0	$0	$0
1,000	$320	$250	$50
2,000	$550	$410	$100
3,000	$720	$510	$150
4,000	$796	$590	$200
5,000	$796	$733	$250
6,000	$785	$774	$300
7,000	$763	$795	$438
8,000	$891	$965	$660
9,000	$869	$911	$743
10,000	$825	$825	$825

Table 7: Gain Table for Two Models, with *Variable* Marginal Costs (Max Profit Highlighted)