



Stochastic Solutions

Hillstrom's MineThatData Email Analytics Challenge:

An Approach Using Uplift Modelling

Abstract

Kevin Hillstrom, through his MineThatData blog, made available a dataset describing two email campaigns and a control group and issued a challenge to analyse that dataset and answer various questions. This paper uses Uplift Modelling to take up Hillstrom's challenge. In the paper, we look at three different formulations of the problem and use Uplift Models to analyse each of the campaigns using those formulations. Broadly, our conclusions are that both campaigns had a positive impact overall and that both can be modelled successfully, allowing us to identify subpopulations particularly suitable and unsuitable for these campaigns. We also identified some segments for which the Women's mailing, in particular, appeared to reduce rather than increase spending; such effects were less prominent with the Men's campaign. We also observed that while average spend among purchasers increased significantly with the Women's campaign, it declined slightly for recipients of the Men's campaign. Furthermore, an extremely small number of people (of the order of 50) were responsible for over half of the incremental spend, making modelling challenging. Indeed, the problem in tackling this entire analysis was the difficulty of estimating most statistics reliably given the relatively small samples available and the low purchase rate. Despite these obstacles, by using fairly simple models and a variety of methods for controlling noise, we believe that the insights we present are fairly robust.

1. Introduction

This paper is a response to the MineThatData¹ E-Mail Analytics And Data Mining Challenge (see Appendix) in which Kevin Hillstrom has made available 64,000 records each describing a customer. One third of these customers were randomly chosen to receive an email referred to as the Men's email, a second random third received a different email (the Women's email) and the remaining customers served as a control, receiving neither email. Hillstrom has asked a number of questions, which we address in the remainder of the paper.

¹ <http://minethatdata.blogspot.com>, March 20th, 2008.

2. The Data

Hillstrom's dataset contains 64,000 records, almost perfectly equally divided between two mailings ("Men's" and "Women's") and an untreated control group.

Table 1: Data Volumes

Mailing	Men	Women	None	Total
Count	21,307	21,387	21,306	64,000
%	33.29%	33.42%	33.29%	100.00%

Analysis of the proportions broken down by other variables provided strongly supports Hillstrom's claim that the individuals were allocated to the three groups entirely randomly, greatly simplifying analysis.

Hillstrom provided three outcome (dependent) variables indicating whether people visited the site during a two-week outcome period, whether they purchased at the site ("conversion") during that period, and how much customers spent during the outcome period (zero, naturally, for those who didn't).

3. Q1: Which Campaign Performed Better?

In order to answer Hillstrom's first question, in principle we could consider three classes of success measure—those based on the success of the campaign in driving incremental spend (amount), those based on its success in driving incremental spend occasions (purchase frequency) or those based on the campaign's success in driving visits. In the absence of any other steer, it is natural to favour the first, as incremental spend achieved is most closely related to campaign profitability. In fact, however, the overall numbers appear to tell a very simple and consistent story, with the Men's campaign outperforming the Women's on all three measures, as shown in Table 2.

Table 2: Incremental Impact of the Two Mailings

Mailing	Visit Rate			Purchase Rate			Spend / head		
	Treated	Control	Uplift	Treated	Control	Uplift	Treated	Control	Uplift
Men's	18.28%	10.62%	7.66%	1.25%	0.57%	0.68%	\$1.41	\$0.65	\$0.77
Women's	15.14%	10.62%	4.52%	0.88%	0.57%	0.31%	\$1.07	\$0.65	\$0.42

As we can see, overall the men’s campaign produced a higher incremental impact (“uplift”) on all measures—a spend increase of 77¢ per head against 42¢ per head for the women’s campaign, an increase in purchase rate of 0.68 percentage points (pp) against 0.31pp and an increase in visit rate of 7.66 percentage points against 4.52pp for the women’s campaign.

(This table also addresses Hillstrom’s second question, “How much incremental sales per customer did the two versions of the campaign drive?”)

Note that while these dollar amounts per head sound modest, these figures show that the Men’s mailing more than doubled spend, and the Women’s campaign increased it by some 65%!

4. Robustness of Results

Before moving on, we might should pause to consider the robustness of these results. At first blush, they look very strong. After all, the classes of measures of incremental impact we have looked at all appear to tell a consistent story: the uplift in visit rate is quite large; the increases in spends, while modest, and not large enough to justify the cost of a paper mailing, are quite good for a low-cost email campaign; and even the impact on the proportion of customers purchasing, while modest, don’t require too many decimal places to register.

The traditional way to assess robustness is to use statistical tests to calculate *p* values and thus the level of statistical significance. However, even leaving aside important questions of whether any particular data meets the assumptions of the test (which is often not the case), we can gain much more insight simply by looking at different random samples of the data. Tables 3 and 4 compare the same three metrics as used previously for five different random 50% samples of the data and the full sample. (The control customers are the same for the corresponding 50% samples in the men’s and women’s campaigns.)

Table 3: Stability of the Impact of the Men’s Mailing (50% samples)

Men's Sample	Visits			Purchase			Spend		
	Treated	Control	Uplift	Treated	Control	Uplift	Treated	Control	Uplift
50% 1	18.11%	10.64%	7.47%	1.30%	0.51%	0.79%	\$1.49	\$0.47	\$1.01
50% 2	18.51%	10.65%	7.86%	1.23%	0.55%	0.68%	\$1.40	\$0.54	\$0.86
50% 3	18.03%	10.89%	7.13%	1.25%	0.47%	0.78%	\$1.28	\$0.45	\$0.84
50% 4	18.35%	10.46%	7.89%	1.38%	0.55%	0.83%	\$1.48	\$0.67	\$0.80
50% 5	18.56%	10.36%	8.20%	1.31%	0.61%	0.70%	\$1.44	\$0.71	\$0.73
100%	18.28%	10.62%	7.66%	1.25%	0.57%	0.68%	\$1.41	\$0.65	\$0.77

Table 4: Stability of the Impact of the Women’s Mailing (50% samples)

Women's Sample	Visits			Purchase			Spend		
	Treated	Control	Uplift	Treated	Control	Uplift	Treated	Control	Uplift
50% 1	15.16%	10.64%	4.52%	0.92%	0.51%	0.40%	\$1.23	\$0.47	\$0.76
50% 2	15.15%	10.65%	4.50%	0.85%	0.55%	0.30%	\$1.09	\$0.54	\$0.55
50% 3	15.43%	10.89%	4.53%	0.87%	0.47%	0.40%	\$1.09	\$0.45	\$0.64
50% 4	15.26%	10.46%	4.80%	0.90%	0.55%	0.35%	\$1.06	\$0.67	\$0.39
50% 5	15.33%	10.36%	4.97%	0.93%	0.61%	0.32%	\$1.23	\$0.71	\$0.52
100%	15.14%	10.62%	4.52%	0.88%	0.57%	0.31%	\$1.07	\$0.65	\$0.42

The good news is that these figures strongly confirm that across all the samples the uplifts in visit rate and purchase rate are broadly similar and in all cases the Men's Campaign outperforms the Women's. On spend, the variance is rather larger, with uplifts varying from \$0.73 to \$1.01 for the Men's Campaign and from \$0.39 to \$0.76 for the Women's Campaign. However, while these ranges overlap slightly, it remains clear that the Men's Campaign performed significantly better than the Women's (and indeed, comparing corresponding samples, the Men's Campaign achieved higher sales uplift than the Women's in each of the 50% samples.)

The size of the variances on 50% samples is, nevertheless, enough to give us significant pause before tackling Hillstrom's next questions, which concern identifying the best and worst 10,000 customers from the campaign. 10,000 is rather less than 20% of the 64,000 in the population, so if we were going to use some kind of validation methodology (e.g. a 50% test/training split) we would be looking at subpopulations less than 10% of the whole. To get a feel for the variances involved in that, we randomly assigned each customer to one of ten samples and again measured the same quantities, both for recipients of the Men's and the Women's mailings (Tables 5 and 6).

Table 5: Stability of the Impact of the Men's Mailing (10% samples)

Men's Sample	Visits			Purchase			Spend		
	Treated	Control	Uplift	Treated	Control	Uplift	Treated	Control	Uplift
1	16.38%	10.25%	6.13%	0.79%	0.41%	0.38%	\$0.87	\$0.36	\$0.51
2	19.15%	10.95%	8.20%	1.34%	0.63%	0.71%	\$1.71	\$0.82	\$0.89
3	17.67%	10.47%	7.20%	1.11%	0.56%	0.54%	\$0.95	\$0.60	\$0.35
4	17.50%	10.18%	7.32%	0.74%	0.48%	0.26%	\$0.63	\$0.70	-\$0.07
5	19.82%	11.15%	8.67%	1.48%	0.66%	0.82%	\$1.91	\$0.51	\$1.40
6	18.94%	10.64%	8.31%	1.28%	0.57%	0.71%	\$1.00	\$0.46	\$0.54
7	18.25%	10.60%	7.65%	1.42%	0.58%	0.84%	\$1.31	\$0.82	\$0.49
8	18.19%	10.98%	7.21%	1.60%	0.74%	0.86%	\$1.99	\$1.10	\$0.89
9	19.08%	11.02%	8.06%	1.59%	0.58%	1.01%	\$2.12	\$0.57	\$1.54
10	17.78%	9.90%	7.88%	1.19%	0.52%	0.67%	\$1.67	\$0.58	\$1.09

Table 6: Stability of the Impact of the Women's Mailing (10% samples)

Women's Sample	Visits			Purchase			Spend		
	Treated	Control	Uplift	Treated	Control	Uplift	Treated	Control	Uplift
1	15.33%	10.25%	5.08%	1.16%	0.41%	0.75%	\$0.94	\$0.36	\$0.58
2	14.86%	10.95%	3.91%	0.79%	0.63%	0.16%	\$0.95	\$0.82	\$0.13
3	15.65%	10.47%	5.18%	1.09%	0.56%	0.53%	\$1.48	\$0.60	\$0.88
4	16.43%	10.18%	6.25%	0.94%	0.48%	0.45%	\$1.18	\$0.70	\$0.49
5	15.57%	11.15%	4.42%	1.15%	0.66%	0.50%	\$1.16	\$0.51	\$0.64
6	15.49%	10.64%	4.85%	0.73%	0.57%	0.16%	\$1.05	\$0.46	\$0.59
7	14.84%	10.60%	4.23%	0.93%	0.58%	0.35%	\$1.01	\$0.82	\$0.20
8	14.00%	10.98%	3.02%	0.71%	0.74%	-0.03%	\$0.97	\$1.10	-\$0.13
9	14.37%	11.02%	3.35%	0.81%	0.58%	0.23%	\$1.22	\$0.57	\$0.65
10	14.84%	9.90%	4.94%	0.51%	0.52%	-0.01%	\$0.75	\$0.58	\$0.17

As can be seen clearly from these tables, when we go down to 10% of the population, the variances in both of the measures we are most interested in (incremental impact on purchase rate and spend) become enormous. For the Men's campaign, our estimate of uplift in spend varies from a high of +\$1.54 to a low of -\$0.07, while the estimate of the uplift in purchase rate varies from +1.01 percentage points to +0.26pp. Similarly, for the Women's campaign, the estimates of impact on incremental spend vary from +\$0.88 to -\$0.13, and on incremental purchase rate from +0.74pp to -0.03pp.

This is clearly going to make it extremely hard to have confidence in any targeting of the campaign, certainly at cell sizes around 10,000.

However, taking courage in both hands we press on.

5. Q3 & Q4: Identify the Best and Worst 10,000 Targets

5.1. Broad Approach: Uplift Modelling

Uplift modelling²⁻⁵ is a relatively new class of modelling techniques that are, in principle, ideally suited to answering these two questions. Uplift models seek to predict, for each individual in some population, the incremental impact of a specific activity on some outcome of interest.

In the case of a binary outcome, such as purchase, we define the uplift U as

$$U = P(\text{purchase} \mid \text{treatment}) - P(\text{purchase} \mid \text{no treatment}) \quad (1)$$

where $P(A \mid B)$ denotes the probability of A given B . In this case, obviously, the treatments of interest are the Men's and Women's mailings.

In the case of a continuous outcome, such as spend, we instead have

$$U = E(\text{spend} \mid \text{treatment}) - E(\text{spend} \mid \text{no treatment}) \quad (2)$$

where $E(A \mid B)$ is the expectation value of A given B .

Given a vector of predictors (independent variables) \mathbf{x} , a binary uplift model for purchase predicts

$$P(\text{purchase} \mid \mathbf{x}; \text{treatment}) - P(\text{purchase} \mid \mathbf{x}; \text{no treatment}) \quad (3)$$

while a continuous uplift model for spend predicts

$$E(\text{spend} \mid \mathbf{x}; \text{treatment}) - E(\text{spend} \mid \mathbf{x}; \text{no treatment}). \quad (4)$$

Note that this is very different from a traditional response model, which predicts merely $P(\text{purchase} \mid \mathbf{x}; \text{treatment})$ or $E(\text{spend} \mid \mathbf{x}; \text{treatment})$, and also from a traditional penetration model, which predicts something like $P(\text{purchase} \mid \mathbf{x}; \text{existing marketing mix})$ or $E(\text{spend} \mid \mathbf{x}; \text{existing marketing mix})$.

² Radcliffe N. J. & Surry, P. D. (1999). "Differential response analysis: Modelling true response by isolating the effect of a single action." Proceedings of Credit Scoring and Credit Control VI. Credit Research Centre, University of Edinburgh Management School.

³ Lo, V. S. Y.. (2002). "The true lift model". ACM SIGKDD Explorations Newsletter. Vol. 4 No. 2, 78-86. 1

⁴ Radcliffe, N. J. (2007). "Generating Incremental Sales: Maximizing the incremental impact of cross-selling, up-selling and deep-selling through uplift modelling", Stochastic Solutions White Paper, 2007. Available from <http://StochasticSolutions.com/papers.html>

⁵ Radcliffe, N. J. (2007). "Uplift Modelling FAQ", The Scientific Marketer (blog). <http://ScientificMarketer.com/2007/09/uplift-modelling-faq.html>

If we are able to build a reasonably granular uplift model that accurately predicts for each individual, the impact on their purchasing behaviour of the two treatments, then we can simply take the 10,000 people with the highest predicted positive uplift and the 10,000 people with the lowest (or most negative) predicted uplift and they will be ideal candidates for inclusions and exclusion respectively.

Various approaches to uplift modelling have been discussed in the literature, with good pointers being available in the references.²⁻⁵

On the basis of the evidence discussed earlier, we might choose to concentrate only on the Men's mailing, since overall that was close to twice as effective as the Women's mailing, but we might also consider the possibility that the Women's mailing, while being less successful overall, may have been very successful for a subsegment, and indeed might perform better than the Men's for some people.

5.2. Formulation

Before actually attempting to build models, we first need to make some decisions about problem formulation. There are different ways in which the emails could generate their impact. Broadly, the base possibilities are as follows:

- the emails could cause more people to visit the site (i.e. drive visiting);
- the emails could cause more of the people who visit the site to purchase;
- the emails could cause people who purchase to spend more.

Obviously, these possibilities are not mutually exclusive, and all three effects may be present. We do need to take a view on this, however, because our approach to modelling will potentially be very different in the three cases.

A natural way to decompose the purchase behaviour is as follows:

$$E(\text{spend}) = E(\text{spend} \mid \text{purchase}) \times P(\text{purchase} \mid \text{visit}) \times P(\text{visit}). \quad (5)$$

We can estimate these for the two mailings from the data as in Table 7.

Table 7: Estimating the components of Equation 5

	Average Spend (purchasers)	Purchase Rate (visitors)	Visit Rate	Mean Spend
Men's Mailing	\$112.91	6.86%	18.28%	\$1.41
Women's Mailing	\$121.30	5.84%	15.14%	\$1.07
Control	\$113.41	5.39%	10.62%	\$0.65
Men's Mailing Uplift +	-\$0.50	1.46%	7.66%	\$0.77
Women's Mailing Uplift +	\$7.89	0.44%	4.52%	\$0.42
Men's Mailing Uplift x	99.56%	127.13%	172.14%	217.88%
Women's Mailing Uplift x	106.95%	108.22%	142.61%	165.06%

The results are fascinating. The first three rows show estimates of the three components of the expression in equation 5, for three segments (Men's Mailing, Women's Mailing, No Mailing). Specifically, 'Average Spend (purchasers)' estimates the expected spend given purchase, 'Purchase Rate (visitors)' estimates the probability of purchase given a visit, and 'Visit Rate' is the proportion of customers who visit the site. Their product, of course, is guaranteed to equal the mean spend.

The next three rows show the (additive) uplifts for these three quantities, i.e. the difference between these quantities for the two mailed groups versus the control. Notice the radically different patterns. Among purchasers, the average spend is actually slightly lower for those who received the Men's mailing than for those who received nothing, while for those who received the Women's it is nearly \$8.00 higher. For those who visited, among the recipients of the Men's mailing, the purchase rate increased by a modest-sounding 1.46 percentage points, while for recipients of the Women's mailing, the purchase rate increased only one third as much—by 0.44pp. Finally, as we have already seen, the Men's mailing drove the visit rate by over 7 percentage points, whereas the effect of the Women's mailing was an increase of less than 5pp.

While we normally advocate modelling uplift as an additive quantity,⁶ it is also interesting to look at the multiplicative impact of the mailings. These are shown in the last two rows. The difference between the two mailings is startling. Whereas the Men's Mailing had hardly any effect on spend among purchasers (less than 0.5%, and negative), the Women's Mailing increased spend among purchasers by nearly 7%. But the impact of the Men's Mailing on both purchase rate (conversion) and visit rate was high, generating increases of 27% and 72% respectively. The Women's mailing was very different, generating corresponding respective increases of only 8% and 43%.

⁶ This is largely because money is additive. Other things being equal, we would rather have a 10% increase in spend from a \$1,000 per visit customer than a 100% increase in spend from a \$50 per visit customer.

It would obviously be useful to know more about the two mailings. The nature of the difference in responses might suggest that the Men's Mailing featured discounts and the Women's did not (or that the discounts were larger for the Men's Mailing). This would appear to be consistent with the higher uplifts in visit frequency and purchase frequency but marginal decline in spend for recipients of the Men's Mailing, and smaller effects but an increase in average purchase size for recipients of the Women's Mailing. (If true, it would also have implications for the relative profitabilities of the mailings, and thus possibly on which is considered more successful; but that goes beyond any data that we have.)

The strongest implication is that if we are going to model one thing, impact on visit frequency, particularly for the Men's Mailing, is probably the one to focus on. For the Men's mailing, it appears that conversion rate is next most significant factor.

For the Women's Mailing, modelling incremental impact on Visit Frequency also makes sense, as this still dominates the overall increase in spend per head; however, the choice of a second component to model is less clear here.

The fact that this analysis suggests that impact on visit frequency is the most important driver is also good news from a statistical perspective, since the uplift in visit rate is significantly higher than the uplift in purchase rate. Furthermore, as can be seen from Tables 5 and 6, the variations in our uplift estimates for visit rate over 10% samples, while not small, are much more manageable than are those for the other uplifts.

5.3. Uplift Models for Visit Rate

We used The Quadstone System from Portrait Software,⁷ and in particular the Uplift Optimizer⁸ package to build uplift models. Uplift Optimizer supports all stages of the modelling process, from data preparation through variable selection, *ad hoc* data exploration, data transformation, direct construction of uplift models and assessment of model performance, reporting and scoring. A variety of modelling methods are available; for this study, we used bagged tree-based models and simpler indicator models.

⁷ <http://portraitsoftware.com>

⁸ http://portraitsoftware.com/Products/portrait_uplift_optimizer

Based on observations in the previous section, we initially constructed binary uplift models to try to identify customers who were strongly driven to the web site by the email campaigns. Slightly surprisingly, despite the fact that the Men’s campaign had a larger overall impact than did the Women’s mailing, the patterns in the Women’s mailing appeared much more stable and predictable.

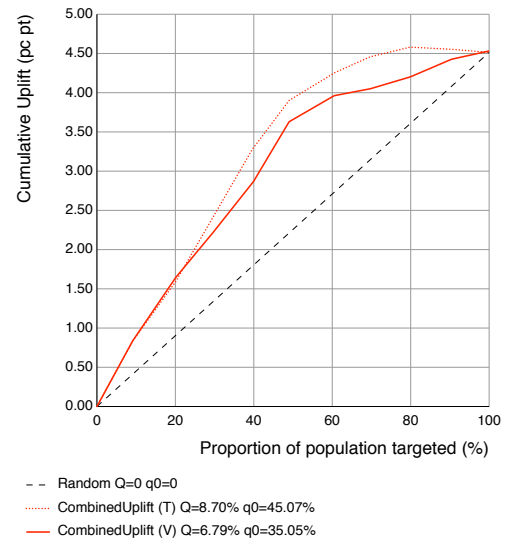
Table 8, below, shows a heat map of the top five most predictive variables identified for the Women’s campaign by the variable selection procedure in Uplift Optimizer, in descending order of predictiveness. Each field has been automatically binned, and the three numbers show uplifts over three different random partitions of the data, with a heat map colour scheme helping to highlight instabilities.

Table 8: Analysis of uplift as a function of variables, including stability analysis (Women’s Email)

Uplift (visit)	1.06% and under			4.52%			7.98% and over
recency	1 to 2	3 to 5	6 to 9	10 to 12			
	4.28%	3.43%	5.36%	4.39%			
	4.56%	3.82%	5.11%	4.17%			
	5.12%	3.48%	5.67%	4.48%			
history	[29.99 - 51]	[51 - 116.5]	[116.5 - 211.5]	[211.5 - 382]	[382 - 3345.93]		
	4.77%	5.30%	3.54%	2.76%	5.36%		
	3.34%	5.31%	4.22%	3.81%	5.49%		
	4.74%	5.85%	3.56%	3.81%	5.63%		
mens	0	1					
	7.13%	2.12%					
	7.08%	2.32%					
womens	0	1					
	1.10%	7.04%					
	1.06%	7.26%					
newbie	0	1					
	4.10%	4.68%					
	3.74%	5.16%					
	4.20%	5.29%					

The result of using these variables to construct an uplift model for visit rate, based on the Women’s Mailing, is shown on the Qini Graph on the right. A Qini graph⁹ is a generalization of a Gains Chart to the case of uplift models. Thus, as with a conventional Gains Chart, the population is sorted, from left to right, by a score, from those to be targeted first (the best prospects) to those to be targeted last (the worst prospects). The vertical axis then shows the cumulative increase in visits, expressed in percentage points, when targeting a given proportion of the population. So the graph starts at (0,0) because obviously we achieve no additional visits when we target no one. Similarly, it ends at (100%, 4.52 pc pt) because, as we saw in Table 2, the overall impact of the Women’s mailing is to increase visits to the website by 4.52 percentage points. The quality of our score determines the shape of the Qini curve. If we target a random x% of the population, we expect to achieve x% the incremental impact of targeting everyone, so the diagonal line represents a random targeting strategy or (equivalently) a random score. In this case, our score is significantly better than random, because when for example, we target 20% of the population, we increase visits by over 1.5 pc pt, which is more than one third of the effect of targeting everyone. Thus curves that are bowed above the diagonal represent useful models. The two measures, Q and q₀, quantify (on different scales) the improvement over random targeting, with larger numbers representing better performance.⁹ The dashed line shows the performance on the training data (a random 50% chosen for building the model) and the solid line shows the performance on the validation data (not used to build the model). We will examine how this can be used to optimize campaign targeting in the next section.

Figure 1: An Uplift Model for Visit Rate (Women’s Mailing)



⁹ Radcliffe, N. J. (2007). “Using Control Groups to Target on Predicted Lift: Building and Assessing Uplift Models”, Direct Marketing Analytics Journal, Direct Marketing Association, 2007.

We also built an uplift model for visit rate for the Men’s Mailing. Table 9 shows the top five five variables, as binned, chosen by the variable selection procedure, showing stability over three random partitions of the data.

Table 9: Analysis of uplift as a function of variables, including stability analysis (Men’s Email)

Uplift (visit)	6.38% and under		7.66%		9.33% and over
recency	1 to 3	4 to 8	9 to 12		
	8.56%	7.25%	7.39%		
	8.45%	6.79%	7.45%		
history_segment	1) \$0 - \$100	2) \$100 - \$200	3) \$200 - \$350	4) - 7) >\$350	
	6.70%	7.31%	8.28%	9.28%	
	6.71%	6.38%	8.39%	9.33%	
history	[29.99 to 91)	[91 to 254)	[254 to 3345.93]		
	6.62%	7.73%	8.86%		
	6.69%	6.99%	8.98%		
mens	0	1			
	6.98%	8.38%			
	7.11%	7.94%			
womens	0	1			
	7.21%	8.16%			
	6.52%	8.40%			
	7.01%	8.21%			

Again we have constructed an uplift model, shown in the Qini Graph in Figure 2. Although Qini values cannot be directly numerically compared, it is clear that this model is less impressive than that for the women's mailing less bowed and showing a somewhat weaker validation. Nevertheless, it will certainly allow us to target better than randomly if our goal is to increase traffic to the site.

We now need to examine how these models work for optimizing the campaign.

5.4. Using Visit Uplift Models to Optimize the Campaign

Our analysis in section 5.2 suggested that the strongest effect of each campaign was to drive visitors to the site. Now that we have uplift models allowing us to assess for which customers the campaigns were more effective in this regard, we can see whether this allows us to optimize the campaign. Figure 3 shows the uplift in visits by decile for the Men's (red) and Women's (black) uplift models (best to worst); while we might like better separation across deciles 2 and 5, these are not bad. Unfortunately, however, Figure 4 shows the spend uplift by decile, which is largely useless.

Figure 2: An Uplift Model for Visit Rate (Men's mailing)

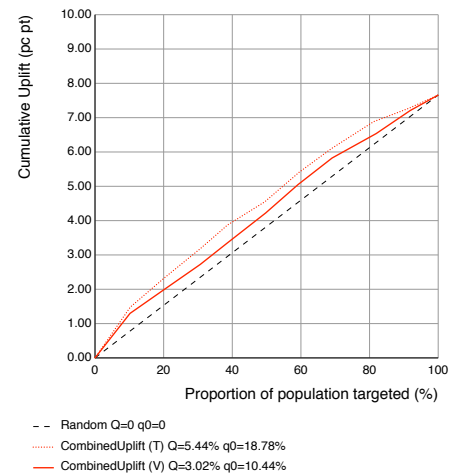


Figure 3: Increase in Visit Rate by (Visit) Uplift Decile

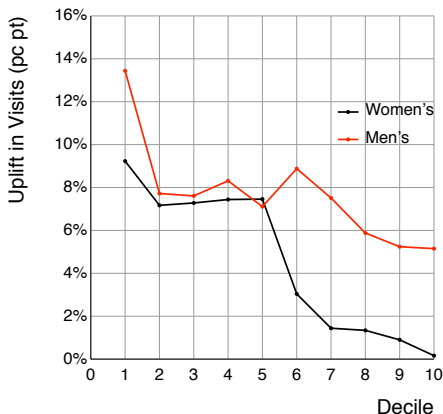
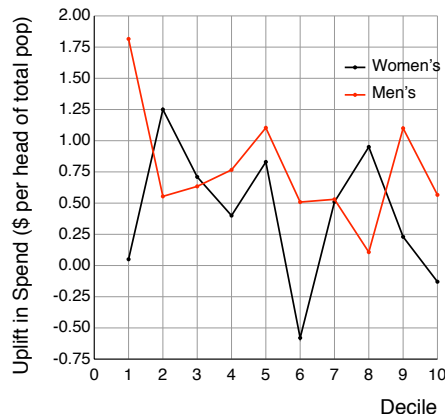


Figure 4: Increase in Spend by (Visit) Uplift Decile

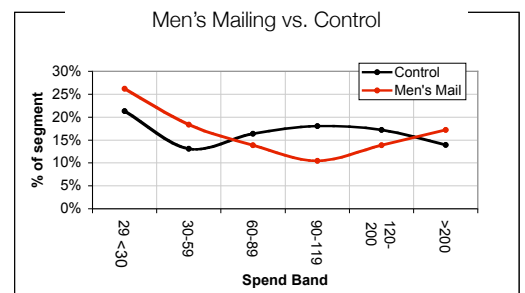


effect on visit rates was so much larger than other impacts this was the key quantity to model needs to be re-examined. We can certainly first try weighting the models by estimates of the other components of Equation 5, but this does not, in fact, help substantially. The fundamental problem, as we shall now see, is that there are strong interactions between the impact of the campaigns on visit rate and on spend, and these significantly compromise our initial line of attack.

5.5. The Spend Distributions for Mailed and Non-Mailed

Our problem becomes clearer if we look at the spend distributions for the mailed and non-mailed customers. Figure 5 show these for the Men's mailed population in comparison with the control.

Figure 5: Distribution of Sales Value: Men's Mailing vs. Control



We noted in Section 5.2 that the spend level among purchasers was marginally lower for those in receipt of the Men's mailing than for the control, but assumed this was a small effect, perhaps down to noise. In fact, Figure 5 suggests otherwise: we can see that spend a lot more of the spend is at low levels in the mailed population and much less of the spend is in the range \$60-\$200. At first, it may seem surprising that this graph is consistent with the Mailed spend being only \$0.50 lower than the control. It all becomes clearer, when we look at the raw data, in Table 10.

Table 10: Analysis of the Difference in Spend among Purchasers (Men's Mailing)

spend band	Frequency (purchasers)				Average Spend		Total Segment Spend		Segment as % of total		Approx Uplift	Seg %
	Control	Men's	Control	Men's	Control	Men's	Control	Men's	Control	Men's		
<30	26	70	21.31%	26.22%	\$29.00	\$29.00	\$754	\$2,030	5.45%	6.73%	\$1,276	7.82%
30-59	16	49	13.11%	18.35%	\$41.69	\$42.31	\$667	\$2,073	4.82%	6.88%	\$1,406	8.62%
60-89	20	37	16.39%	13.86%	\$73.25	\$71.27	\$1,465	\$2,637	10.59%	8.75%	\$1,172	7.19%
90-119	22	28	18.03%	10.49%	\$101.36	\$100.79	\$2,230	\$2,822	16.12%	9.36%	\$592	3.63%
120-200	21	37	17.21%	13.86%	\$153.10	\$159.41	\$3,215	\$5,898	23.24%	19.56%	\$2,683	16.45%
>200	17	46	13.93%	17.23%	\$323.82	\$319.28	\$5,505	\$14,687	39.79%	48.72%	\$9,182	56.29%
Total	122	267	100.00%	100.00%	\$113.41	\$112.91	\$13,836	\$30,147	100.00%	100.00%	\$16,311	100.00%

While the first block of the table shows the data graphed in Figure 5, and the second blocks shows segment average and total spends, it is the last four columns that really shed light. The first two of these, (Segment as % of total) show how much of the total spend each spend band accounts for, both for the control and for the group receiving the Men's mailing. Notice that for the mailed group, very nearly 50% of the total spend comes from people spending over \$200. And notice also that this is just 46 people.¹⁰ In terms of uplift, things are even starker. If we make a rough approximation of the amount of spend uplift contributed by each spend band, (last two columns), we find that over 50% of the increase comes from people spending over \$200. It is this that allows the average spend among purchasers for the Mailed Men to be only \$0.50 lower than the control group.

A similar analysis of the Women's mailing reveals some differences, but perhaps larger similarities (Figure 6 and Table 11).

¹⁰ In fact, 12 people in the data spent \$499. 6 received the Men's mail, 4 received the Women's mail, and 2 were controls. A single \$500 purchase increases the per-head spend in a 20k segment by 2.5¢.

Figure 6: Distribution of Sales Value: Women's Mailing vs. Control

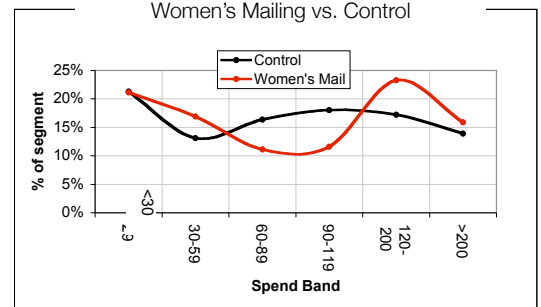


Table 11: Analysis of the Difference in Spend among Purchasers (Women's Mailing)

spend band	Frequency (purchasers)				Average Spend		Segment Spend		Segment as % of total		Approx Uplift	Seg %
	Control	Women's	Control	Women's	Control	Women's	Control	Women's	Control	Women's		
<30	26	40	21.31%	21.16%	\$29.00	\$29.00	\$754	\$1,160	5.45%	5.06%	\$406	4.47%
30-59	16	32	13.11%	16.93%	\$41.69	\$42.31	\$667	\$1,354	4.82%	5.91%	\$687	7.56%
60-89	20	21	16.39%	11.11%	\$73.25	\$76.81	\$1,465	\$1,613	10.59%	7.04%	\$148	1.63%
90-119	22	22	18.03%	11.64%	\$101.36	\$105.05	\$2,230	\$2,311	16.12%	10.08%	\$81	0.89%
120-200	21	44	17.21%	23.28%	\$153.10	\$157.14	\$3,215	\$6,914	23.24%	30.16%	\$3,699	40.70%
>200	17	30	13.93%	15.87%	\$323.82	\$319.10	\$5,505	\$9,573	39.79%	41.76%	\$4,068	44.76%
Total	122	189	100.00%	100.00%	\$113.41	\$121.30	\$13,836	\$22,925	100.00%	100.00%	\$9,089	100.00%

Again, among the lower part of the distribution, there is some depression of spend, though it is less marked than in the case of the Men's Mailing, but this is strongly compensated for by a marked increase in the proportion of high-spenders, not just at the \$200+ level but also in the \$120–\$200 range.

5.6. Modelling Spend Uplift Directly

Contrary to our earlier approach, it therefore looks as if we need to model the uplift in spend directly. However, we know this will be very hard for several reasons. First, the total number of spenders is small. Secondly, in the case of the Men's Mailing, just 45 people (out of a mailing of 21,000), account for more than half of the incremental spend, while for the Women's mailing 85% of the incremental spend comes from just 84 people. Thirdly, we are attempting a second-order modelling problem, predicting the difference in behaviour between two populations (the mailed and the non-mailed).

We have tried two different approaches.

The first is a direct, continuous approach: we simply build an uplift model with the spend variable as the outcome. This is problematical, because around 99% of the population has zero spend, so the distribution is very skewed, but in principle is possible and directly models what we want. We are also helped by the fact that the decision trees we use are less affected by population skews than are regression models.

The second approach, which takes into account the fact that well over 50% of the uplift in spend comes from individuals who spend large amounts, is to reformulate the problem as a binary problem, creating a binary outcome variable that is 'spend over x', for some amount x. We set x to \$60, giving us 148 recipients of the Men's Mailing who meet the criterion and 117 recipients of the Women's mailing.

We found that the continuous approach worked better for the Women's mailing, while the binary approach worked more reliably for the Men's mailing.

5.7. An Uplift Model (M) for the Men's Mailing

We found that the large variances for the Men's Mailing meant that we needed to use one of the simpler uplift modelling techniques, namely an indicator model, which we will denote M. This model simply assigned 1 point for each of the following:

- Historic Spend being over \$350
- Historic Spend being over \$160
- The customer being a multi-channel customer.

Thus each customer can score 0 to 3, resulting in a 4-way segmentation, with the best being multi-channel customers with historic spend over \$350.

As a result of the extreme simplicity of this model (Figure 7, Table 12), stability is quite good. To emphasize this, we made ten random 50% selections on the data, which resulted in uplifts for the top segment as shown in Table 13.

Table 12: Spend uplift by score (Men's Mailing)

Score (M)	0	1	2	3
Count	21,548	9,764	7,803	3,498
Spend Uplift	\$0.55	\$0.67	\$1.13	\$1.54

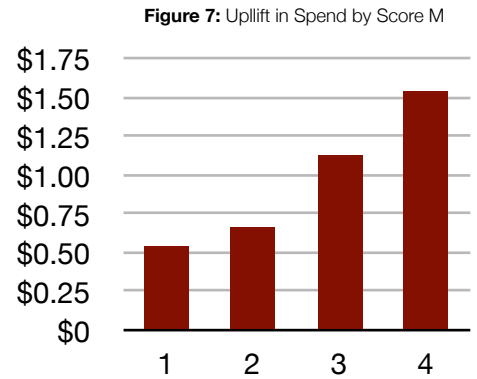


Table 13: Spend uplift for score 3 over ten different random 50% samples

\$1.39	\$1.60
\$1.69	\$1.47
\$1.59	\$1.61
\$1.50	\$1.71
\$1.49	\$1.38

5.8. An Uplift Model (W) for the Women's Mailing

The Women's Mailing produced a much stronger (and more complicated) Uplift Model, W, directly predicting the incremental spend that would result. The Qini graph for this is shown in Figure 8. While there is certainly a degree of overfitting (with the training data showing a stronger Qini than the validation data), overall this is a strong result. Several points to note particularly are:

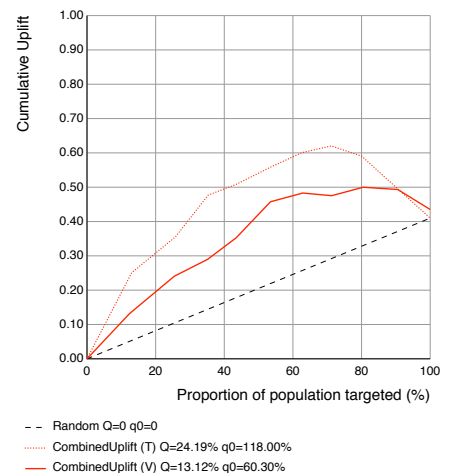
1. The model identifies (even on validation) 20% of the population that delivers around half the total incremental spend (cumulative uplift \$0.20 per head of total population at 20% against \$0.42 targeting the whole population).
2. Only around 50% of the population needs to be targeted to get the same effect as targeting 100%.
3. There is reasonably strong evidence of some negative effects for the last 10–20% of the population, again, even on validation. This suggests that the Women's email is actually detrimental to sales for a portion of the population.

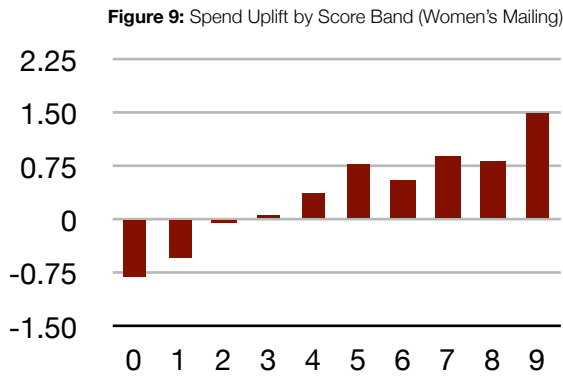
This model is harder to describe, being a bagged tree model, but the summary in Table 12, which shows the average score assigned to each bin of each variable gives a good insight, emphasizing that the model favours particularly customers with high historical spend over \$500, but also to some extent those with low historical spend, multi-channel customers and newbies. Figure 9 shows uplift in spend as a function of score band for W.

Table 12: Average score by bin for a Spend Uplift Model for the Women's Mailing

Uplift Score (W)	-0.3521 and under			0.4225				1.2962 and over
history_segment	1) \$0 - \$100 0.6874	2) \$100 - \$200 0.6313	3) \$200 - \$350 -0.3431	4) \$350 - \$500 0.1213	5-7) >\$500 1.2962			
channel	Multichannel 0.9325	Phone 0.2102	Web 0.6582					
history	[29.99 to 51) 0.8674	[51 to 116.5) 0.4203	[116.5 to 211.5) 0.6382	[211.5 to 382) -0.3521	[382 to 3345.93] 0.8913			
newbie	0 0.2282	1 0.7548						

Figure 8: Qini Curve (Incremental Gains Chart) for Uplift Model W for Women's Mailing





5.9. Q3: Selecting the Best 10,000 Prospects

To choose the best 10,000 to receive the mailing, we obviously start with people to whom the Men's Uplift model assigned a score of M=3. This provides 5,249 people, and estimating their uplift in spend from the 3,498 of them who did not receive the Women's mailing, we get an expected uplift of \$1.54 per head.

If we then consider those who scored M=2, this adds in another 11,751 people with an expected uplift of \$1.13. However, this gives us significantly more people than we want (17,000 in total).

If we look at the figures, Table 13, it's clear that the Score 2's who whose channel is web are the best, with an uplift of \$1.61. Therefore we would add in those, giving us a total of 9,933 customers with an expected uplift of \$1.58. If we then really needed to find another 67 customers, we would choose a random 67 of the people with M=2 and channel as phone.

Table 13: Performance of different M=2 segments

M=2 Customers	Count	Spend Uplift
Phone	4,554	0.97
Web	4,684	1.61
history < \$350	2,513	0.49

5.10. Q4: Selecting the Worst 10,000 Prospects

We now have 2 models and need to select the best and the worst 10,000 candidates for a mailing.

Models are somewhat but by no means completely correlated.

Given that the Men's Mailing performed better, we need to select the 10,000 worst prospects for that. Bottom score (M=0) identifies 32,237 people with an average uplift in response to the Men's Mailing of 55¢. However, if we then attempt to select from those with M=0 the 10,000 with the lowest W score (a predicted spend uplift of under 44¢, corresponding roughly to scorebands 0 to 4) we get 12,952 (because of ties). We can assess the expected spend uplift from mailing those with the Men's mailing by removing those who actually received the Women's mailing and measuring the spend uplift on the remainder. The result is an uplift from the Men's Mailing of just 5¢. So if we were to exclude 10,000 people, we would exclude a random 10,000 out of those with M=0 and W < 5.

These would be people who

- Have a historical spend under \$160
- Are not multichannel.

They will also tend to be people who (from Table 12) who

- were not newbies
- whose channel was Phone.

[Indeed, if we take exactly that population, we get 8,023 people, and assessing their uplift on the 5307 of them who did not receive the Women’s mailing, we find that the impact of them Men’s Mailing on them was to depress their spend by an average of just over 25¢.]

6. Q5–8: Other Issues

We have largely addressed the remaining questions (see Appendix) with the possible exception of characterising the differences between the impacts of the mailings. Obviously the strongest differences between the mailings concerned performance: whereas the Men’s mailing generated 77¢ per head, the Women’s generated only 42¢ of incremental spend; moreover, there is some serious evidence of negative effects in some segments from the Women’s mailing, whereas this is less apparent for the Men’s (notwithstanding the final point in the previous section).

To understand the relationship between the scores, we have examined the joint distributions. Table 14 shows the (additive) difference between the actual proportion in each cell and that which would be expected if the two scores were independent (uncorrelated), with red cells showing significantly higher densities than expected and blue cells showing areas in which the proportion is noticeably less than expected.

Table 14: Joint Score Distribution against Expected if Independent

		M			
		0	1	2	3
W	0	-3.89%	5.45%	-0.79%	-0.77%
	1	-1.04%	2.01%	-0.16%	-0.82%
	2	0.45%	-2.00%	2.29%	-0.75%
	3	-3.83%	1.34%	2.66%	-0.16%
	4	0.61%	1.76%	-1.64%	-0.73%
	5	1.04%	-2.31%	1.95%	-0.67%
	6	1.81%	-0.76%	-2.48%	1.43%
	7	4.57%	-2.12%	-1.69%	-0.76%
	8	4.41%	-2.05%	-1.63%	-0.73%
	9	-4.13%	-1.32%	1.50%	3.95%

Obviously, if the scores were positively correlated, we would see a strong red pattern down the leading diagonal and blue towards the top right and bottom

left. In fact, apart from the highest scores, where normality returns, the bottom nine deciles for score from the Women's mailing and the bottom three score bands for the model from the Men's Mailing show marked anti-correlations. In fact, the correlation coefficient overall between the scores is essentially zero (0.000934). and if the top score bands are removed the correlation coefficient is -0.46443 .

One particularly marked difference is people with historical spend in the range \$200 to \$400. For these people, the Men's Mailing increases spend by \$1.35 per head, whereas the Women's Mailing it reduces spend by \$0.37 per head.

Conversely, people with historical spend under \$100 reacted more positively than average to the Women's mailing (with an uplift of 62¢ per head against an overall uplift of 42¢) whereas for them Men's Mailing this group reacted less positively than average, increasing spend by only 53¢ against an overall increase of 77¢ per head.

7. Conclusion

Analysis of Hillstrom's dataset and the two campaigns it describes has proved challenging (appropriately) and interesting. As usual, when attempting to analyse the incremental impact of campaigns, data volumes were a real problem, as was illustrated in section 4, where we pointed to the difficulty even of forming reliable estimates of overall impact using 10% samples.

Despite these challenges, we believe we have fairly strong conclusions that hold up reasonably well when tested against different subsamples of the data. In terms of the fundamental questions, it is extremely clear that the Men's Mailing was more effective than the Women's, overall, in terms of generating increased sales, increased purchase rate and increased site visits. The best segments to target that we have found are multi-channel customers with higher historical spends (over \$160, and especially over \$350). We also identified a group of customers who appeared to be negatively affected by the Women's email, and combining the Indicator Uplift model that we built to identify people who were significantly more likely to spend over \$60 when receiving the Men's Mailing with the Incremental Spend Model from the Women's Mailing allowed us to identify a very low-performing group for whom neither email was effective, and indeed some segments for whom one or both appeared to reduce the level of spend.

Appendix: Hillstrom's Challenge

From <http://minethatdata.blogspot.com/2008/03/minethatdata-e-mail-analytics-and-data.html>

March 20, 2008

The MineThatData E-Mail Analytics And Data Mining Challenge

It is time to find a few smart individuals in the world of e-mail analytics and data mining! And honestly, what follows is a dataset that you can manipulate using Excel pivot tables, so you don't have to be a data mining wizard, just be clever!

Here is a link to the [MineThatData E-Mail Analytics And Data Mining Challenge dataset](#): The dataset is in .csv format, and is about the size of a typical mp3 file. I recommend saving the file to disk, then open the file (read only) in the software tool of your choice.

This dataset contains 64,000 customers who last purchased within twelve months. The customers were involved in an e-mail test.

- 1/3 were randomly chosen to receive an e-mail campaign featuring Mens merchandise.
- 1/3 were randomly chosen to receive an e-mail campaign featuring Womens merchandise.
- 1/3 were randomly chosen to not receive an e-mail campaign.

During a period of two weeks following the e-mail campaign, results were tracked. Your job is to tell the world if the Mens or Womens e-mail campaign was successful.

Historical customer attributes at your disposal include:

- **Recency:** Months since last purchase.
- **History_Segment:** Categorization of dollars spent in the past year.
- **History:** Actual dollar value spent in the past year.
- **Mens:** 1/0 indicator, 1 = customer purchased Mens merchandise in the past year.
- **Womens:** 1/0 indicator, 1 = customer purchased Womens merchandise in the past year.
- **Zip_Code:** Classifies zip code as Urban, Suburban, or Rural.
- **Newbie:** 1/0 indicator, 1 = New customer in the past twelve months.
- **Channel:** Describes the channels the customer purchased from in the past year.

Another variable describes the e-mail campaign the customer received:

- **Segment**
 - Mens E-Mail
 - Womens E-Mail
 - No E-Mail

Finally, we have a series of variables describing activity in the two weeks following delivery of the e-mail campaign:

- **Visit:** 1/0 indicator, 1 = Customer visited website in the following two weeks.
- **Conversion:** 1/0 indicator, 1 = Customer purchased merchandise in the following two weeks.
- **Spend:** Actual dollars spent in the following two weeks.

Ok, that represents the basics.

By April 30, you are encouraged to write a paper that answers the following questions. Winning submissions will receive a copy of my book, **Hillstrom's Multichannel Forensics**, currently available at **ForBetterBooks** and **Amazon.com**. There's nothing wrong with winning a book valued at \$95, is there??

I will give away at least one book, and as many as three books, depending upon entries within the following categories:

- **The E-Mail Blogosphere:** If we get enough entries, I will give away one book to the e-mail blogger who provides the most insightful answer.
- **The Direct Marketing Industry:** The best answer among direct marketing and e-mail marketing professionals and e-mail marketing vendors will receive a book. ***In addition, I'll publish well-written and insightful answers received from any qualified e-mail marketing vendor. In other words, you'll earn an opportunity to advertise for free to the MineThatData community, a community of more than 1,200 subscribers and daily visitors.***
- **The Data Mining Community:** Data Mining professionals and University students are encouraged to send in entries, with the best-written and most insightful response receiving a free book.

Here are the questions you are encouraged to answer.

- Which e-mail campaign performed the best, the Mens version, or the Womens version?
- How much incremental sales per customer did the Mens version of the e-mail campaign drive? How much incremental sales per customer did the Womens version of the e-mail campaign drive?
- If you could only send an e-mail campaign to the best 10,000 customers, which customers would receive the e-mail campaign? Why?
- If you had to eliminate 10,000 customers from receiving an e-mail campaign, which customers would you suppress from the campaign? Why?
- Did the Mens version of the e-mail campaign perform different than the Womens version of the e-mail campaign, across various customer segments?
- Did the campaigns perform different when measured across different metrics, like Visitors, Conversion, and Total Spend?
- Did you observe any anomalies, or odd findings?
- Which audience would you target the Mens version to, and the Womens version to, given the results of the test? What data do you have to support your recommendation?

E-mail your responses to me by 11:59pm on Wednesday, April 30, 2008. Good luck, and have fun analyzing the information! Dazzle our readers with your insights --- feel free to share your findings in the comments section of this post.

Acknowledgements

Stochastic Solutions and the author would like to thank Portrait Software for making its Quadstone System and in particular its Uplift Optimizer software available for this project, and for providing feedback on this paper.

We would also like to thank Kevin Hillstrom for issuing the MineThisData E-Mail Analytics And Data Mining Challenge, and for making an interesting, clean dataset available for analysis. (And given the rarity of this in practice, we would especially like to acknowledge the apparently completely unbiased selection of the three mail cells: if only 'twere always thus!)

Author

Nicholas J. Radcliffe

Nicholas Radcliffe is founder and director of Stochastic Solutions Limited. He was previously a founder and Technical Director of Quadstone Limited, where he led the development of their approach to uplift modelling and numerous client engagements.

Nicholas is also a Visiting Professor in the Department of Mathematics and Statistics at University of Edinburgh, where he works in the Operations Research group.

Nicholas.Radcliffe@StochasticSolutions.com

Stochastic Solutions

Stochastic Solutions is a consultancy that focuses on customer targeting and on combinatorial and numerical optimization problems. The company has special expertise in uplift modelling, customer strategy and stochastic optimization.