Real-World Uplift Modelling with Significance-Based Uplift Trees

Nicholas J. Radcliffe[a,b]
& Patrick D. Surry[c]
Nicholas.Radcliffe@StochasticSolutions.com
Patrick.Surry@pb.com

[a]Stochastic Solutions Limited, 37 Queen Street, Edinburgh, EH2 1JX, UK.
[b]Department of Mathematics, University of Edinburgh, King's Buildings, EH9 3JZ.
[c]Pitney Bowes Business Insight, 125 Summer Street, 16th Floor, Boston, MA 02110, USA.

Abstract

This paper seeks to document the current state of the art in 'uplift modelling'—the practice of modelling the *change* in behaviour that results directly from a specified treatment such as a marketing intervention. We include details of the Significance-Based Uplift Trees that have formed the core of the only packaged uplift modelling software currently available. The paper includes a summary of some of the results that have been delivered using uplift modelling in practice, with examples drawn from demand-stimulation and customer-retention applications. It also surveys and discusses approaches to each of the major stages involved in uplift modelling—variable selection, model construction, quality measures and post-campaign evaluation—all of which require different approaches from traditional response modelling.

1 Organization

We begin by motivating and defining uplift modelling in section 2, then review the history and literature of uplift modelling (section 3), including a review of results. Next, we devote sections, in turn, to four key areas involved in building and using uplift models. We start with the definition of quality measures and success criteria (section 4), since these are a conceptual prerequisite for all of the other areas. We then move on to the central issue of model construction, first discussing a number of possible approaches (section 5), and then detailing the core tree-based algorithm that we have used successfully over a number of years, which we call the Significance-Based Uplift Tree (section 6). Next, we address variable selection (section 7). This is important because the best variables for conventional models are not necessarily the best ones for predicting uplift (and, in practice, are often not). We close the main body with some final remarks (section 8), mostly concerning when, in practice, an uplift modelling approach is likely to deliver worthwhile extra value.

2 Introduction

2.1 Predictive Modelling in Customer Management

Statistical modelling has been applied to problems in customer management since the introduction of statistical credit scoring in the early 1950s, when the consultancy that became the Fair Isaac Corporation was formed (Thomas, 2000).[1] This was followed by progressively more sophisticated use of predictive modelling for customer targeting, particularly in the areas of demand stimulation and customer retention.

As a broad progression over time, we have seen:

1. *penetration* (or *lookalike*) *models,* which seek to characterize the customers who have already bought a product. Their use is based on the assumption that people with similar characteristics to those who have already bought will be good targets, an assumption that tends to have greatest validity in markets that are far from saturation;

2. *purchase models,* which seek to characterize the customers who have bought in a recent historical period. These are similar to penetration models, but restrict attention to the more recent past. As a result, they can be more sensitive to changes in customer characteristics across the product purchase cycle from early adopters through the mainstream majority to laggards (Moore, 1991);

3. *'response' models,* which seek to characterize the customers who have purchased in apparent 'response' to some (direct) marketing activity such as a piece of direct mail. Sometimes, the identification of 'responders' involves a coupon, or a response code ('direct attribution'), while in other cases it is simply based on a combination of the customer's having received the communication and purchasing in some constrained time window afterwards[2] ('indirect attribution'). 'Response' models are normally considered to be more sophisticated than both penetration models and purchase models, in that they at least attempt to connect the purchase outcome with the marketing activity designed to stimulate that activity.

All of these kinds of modelling come under the general umbrella of 'propensity modelling'.

In the context of customer retention, there has been a similar progression, starting with targeted acquisition programmes, followed by models to predict which customers are most likely to leave, particularly around contract renewal time. Such 'churn' or 'attrition' models are now commonly combined with value estimates allowing companies to focus more accurately on retaining value rather than mere customer numbers.

[1] Thomas reports that David Durand of the US National Bureau of Economic Research was the first to suggest the idea of applying statistical modelling to predicting credit risk (Durand, 1941).

[2] More complex inferred response rules are sometime used to attribute particular sales to given marketing treatments, but these appear to us to be rather hard to justify in most cases.

2.2 Measuring Success in Direct Marketing: the Control Group

The primary goal of most direct marketing is to effect some specific change in customer behaviour. A common example of this is the stimulation of extra purchasing by a group of customers or prospects. While there may be subsidiary goals, such as brand awareness and the generation of customer goodwill, most marketing campaigns are primarily evaluated on the basis of some kind of return-on-investment (ROI) calculation.

If we focus, initially, on the specific goal of generating incremental revenue, it is clear that measurement of success is non-trivial, because of the difficulty of knowing what level of sales would have been achieved had the marketing activity in question not been undertaken. The key, as is well known, is the use of a control group, and it is a well-established and widely recognized best practice to measure the incremental impact of direct marketing activity by comparing the performance of the treated group with that of a valid control group chosen uniformly at random[3] from the target population.

2.3 The Uplift Critique of Conventional Propensity Modelling

While there is a broad consensus that accurate *measurement* of the impact of direct marketing activity requires a focus on incrementality through the systematic and careful use of control groups, there has been much less widespread recognition of the need to focus on incrementality when *selecting* a target population. None of the approaches to propensity modelling discussed in section 2.1 is designed to model incremental impact. Thus, perversely,

> *most targeted marketing activity today, even when measured on the basis of* incremental *impact, is targeted on the basis of* non-incremental *models.*

It is widely recognized that neither penetration models nor purchase models even attempt to model changes in customer behaviour, but less widely recognized that so-called 'response' models are also *not* designed to model incremental impact. The reason they do not is that the outcome variable[4] is necessarily set on the basis of a test such as "purchased within a 6-week period after the mail was sent" or the use of some kind of a coupon, code or custom link. Such approaches attempt to tie the purchase to the campaign activity, either temporally or through a code. But while these provide some evidence that a customer has been influenced by (or was at least aware of) the marketing activity, they by no means guarantee that we limit ourselves to incremental purchasers. These approaches can also fail to record genuinely incremental purchases from customers who *have* been influenced but for whatever reason do not use the relevant coupon or code.

For the same reasons that we reject as flawed *measurement* of the incrementality of a marketing action through counting response codes or counting all purchases within

[3] Strictly, the control group does not have to be chosen in this way. It can certainly be stratified, and can even be from a biased distribution if that distribution is known, but this is rarely done as it complicates the analysis considerably. Although we have sometimes used more complicated test designs, for clarity of exposition, we assume uniform random sampling throughout this paper.

[4] the dependent variable, or target variable

a time window, we must reject as flawed *modelling* based on outcomes that are not incremental if our goal is to model the change in behaviour that results from a given marketing intervention (as it surely should be if our success metric is incremental).

A common case worth singling out arises when a response code is associated with a discount or other incentive. If a customer who has already decided to buy a given item receives a coupon offering a discount on that item, it seems likely that in many cases the customer will choose to use the coupon. (Indeed, it is not uncommon for helpful sales staff to point out coupons and offers to customers.) Manifestly, in these cases, the sales are not incremental[5] whatever the code on coupon may appear to indicate. Indeed, in this case, not only were marketing costs increased by including the customer, but incremental revenue was reduced—from some perspectives, almost the worst possible outcome.

2.4 The Unfortunately Named 'Response' Model

We suspect that the very term 'response modelling' is a significant impediment to the wider appreciation of the fact that so-called 'response models' in marketing are not reliably incremental. The term 'response' is (deliberately) loaded and carries the unmistakable connotation of causality. At the risk of labouring the point, the Oxford English Dictionary's first definition (Onions, 1973, p. 1810) of response is:

> **Response. 1.** An answer, a reply. **b.** *transf.* and *fig.* An action or feeling which answers to some stimulus or influence.

While it is unrealistic for us to expect to change the historic and accepted nomenclature, we encourage the term 'response' model to be used with care and qualification. As noted before, our preferred term for models that genuinely model the incremental impact of an action is an 'uplift model', though as we shall see, other terms are also used.

2.5 Conventional Models and Uplift Models

Assume we partition our candidate population randomly[6] into two subpopulations, T and C. We then apply a given treatment to the members of T and not to C. Considering first the binary case, we denote the outcome $O \in \{0, 1\}$, and here assume that 1 is the desirable outcome (say purchase).

A conventional 'response' model predicts the probability that O is 1 for customers in T. Thus a conventional model fits

$$P(O = 1 \mid \boldsymbol{x}; T), \qquad \textit{(conventional binary 'response' model)} \tag{1}$$

[5] We are assuming here that the coupon was issued by the manufacturer, who is indifferent as to the channel through which the item is purchased. A coupon from a particular shop could cause the customer to switch *to that shop,* but again the coupon alone does not establish this, as the customer could and might have bought from that shop anyway.

[6] When we say randomly, we more precisely mean *uniformly, at random*, i.e. each member of the population is assigned to T or C independently, at random, with some fixed, common probability $p \in (0, 1)$

where $P(O = 1 \mid \boldsymbol{x}; T)$ denotes "the probability that $O = 1$ given that the customer, described by a vector of variables $\boldsymbol{x}$, is in subpopulation T". Note that the control group C does not play any part of this definition. In contrast, an uplift model fits

$$P(O = 1 \mid \boldsymbol{x}; T) - P(O = 1 \mid \boldsymbol{x}; C). \qquad \textit{(binary uplift model)} \tag{2}$$

Thus, where a conventional 'response' model attempts to estimate the probability that customers will purchase *if* we treat them, an uplift model attempts to estimate the *increase* in their purchase probability if we treat them over the corresponding probability if we do not. The explicit goal is now to model the *difference* in purchasing behaviour between T and C.

Henceforth, we will not explicitly list the $\boldsymbol{x}$ dependence in equations such as these, but it should be assumed.

We can make an equivalent distinction for non-binary outcomes. For example, if the outcome of interest is some measure of the size of purchase, e.g, revenue, R, the conventional model fits

$$E(R \mid T) \qquad \textit{(conventional continuous 'response' model)} \tag{3}$$

whereas the uplift model estimates

$$E(R \mid T) - E(R \mid C) \qquad \textit{(continuous uplift model)} \tag{4}$$

3 History & Literature Review

The authors' interest in predicting incremental response began around 1996 while consulting and building commercial software for analytical marketing.[7] At that time, the most widely used modelling methods for targeting were various forms of regression and trees.[8] The more common regression methods included linear regression, logistic regression and generalised additive models (Hastie & Tinshirani, 1990), usually in the form of scorecards. Favoured tree-based methods included classification and regression trees (CART; Breiman *et al.,* 1984) and, to a lesser extent, ID3 (Quinlan, 1986), C4.5 (Quinlan, 1993), and AID/CHAID (Hawkins & Kass, 1982; Kass, 1980). These were used to build propensity models, as described in the introduction. It quickly became clear to us that these did not lead to an optimal allocation of direct marketing resources for reasons described already, with the consequence that they did not allow us to target accurately the people who were most positively influenced by a marketing treatment.

We developed a series of tree-based algorithms for tackling uplift modelling, all of which were based on the general framework common to most binary tree-based methods (e.g. CART), but using modified split criteria and quality measures. Tree methods

[7] The Decisionhouse software was produced by Quadstone Limited, which is now part of Pitney Bowes.

[8] Both of these classes of methods remain widely used, though they have been augmented by recommendation systems that focus more on product set rather than other customer attributes, using a variety of methods including collaborative filtering (Resnick *et al.,* 1994) and association rules (Piatetsky-Shapiro 1991). Bayesian approaches, particularly Naïve Bayes models (Hand & Yu, 2001), have also grown in popularity over this period.

usually begin with a growth phase employing a greedy algorithm (Cormen *et al.,* 1990). Such greedy algorithms start with the whole population at the root of the tree and then evaluate a large number of candidate splits, using an appropriate quality measure. The standard approach considers a number of splits for each (potential) predictor.[9] The best split is then chosen and the process is repeated recursively (and independently) for each subpopulation until some termination criterion is met—usually once the tree is large. In many variants, there is then a pruning phase during which some of the lower splits are discarded in the interest of avoiding overfitting. The present authors outlined our approach in a 1999 paper (Radcliffe & Surry, 1999), when we used the term *Differential Response Modelling* to describe what we now call *Uplift Modelling.*[10] At that point, we did not publish our (then) split criterion, but we now give details of our current, improved criterion in section 6. Other researchers have developed alternative methods for the same problem independently, unfortunately using different terminology in almost every case.

Various results from the approach described in this paper have been published elsewhere, including:

- US Bank found that a 'response' model was spectacularly unsuccessful for targeting a (physical) mailshot promoting a high-value product to its existing customers. When the whole base was targeted, this was profitable (on the basis of the value of incremental sales measured against a control group), but when the top 30% identified by a conventional 'response' model was targeted, the result was almost exactly zero incremental sales (and a resulting negative ROI). This was because the 'response' model succeeded only in targeting people who would have bought anyway. An uplift model managed to identify a different 30% which, when targeted, generated 90% of the incremental sales achieved when targeting the whole population, and correspondingly turned a severely loss-making marketing activity into a highly successful (and profitable) one (Grundhoefer, 2009).[11]

- A mobile churn reduction initiative actually increased churn from 9% to 10% prior to uplift modelling. The uplift model allowed a 30% subsegment to be identified. Targeting only that subsegment reduced overall churn from 9% to under 8%, while reducing spend by 70% (Radcliffe & Simpson, 2008). The estimated value of this to the provider was $8m per year per million customers in the subscriber base.

- A different mobile churn reduction initiative (at a different operator) was successful in reducing churn by about 5 percentage points (pp), but an uplift model was able to identify 25% of the population where the activity was marginal or

[9] independent variable

[10] Various versions of the algorithm were implemented in the Decisionhouse software over the years, and used commercially, with increasing success.

[11] US Bank also developed an in-house approach to uplift prediction called a matrix model. This was based on the idea of comparing predictions from a response model on the treated population with a natural buy rate model built on a mixed population. Prediction from both models were binned and segments showing high uplift were targeted. This produced a somewhat useful model, but one that was less than half as powerful as a direct, significance tree-based uplift model.

counterproductive. By targeting only the identified 75% overall retention was increased by from 5 pp to 6 pp, (i.e. 20% more customers were saved) at reduced cost (Radcliffe & Simpson, 2008). This was also valued at roughly $8m per year per million customers in the subscriber base.

We also published an electronic retail analysis based on a challenge set by Kevin Hillstrom (Radcliffe, 2008) of MineThatData (Hillstrom, 2008).

Maxwell *et al.* (2000), at Microsoft Research, describe their approach to targeting mail to try to sell a service such as MSN. Like us, they base their approach on decision trees but they simply build a standard tree on the whole population (treated and control) and then force a split on the treatment variable at each leaf node. The primary limitation with this approach is that splits in the tree are not chosen to fit uplift; it is simply the estimation at the end that is adapted. The authors do not compare to a non-uplift algorithm, but report benefits over a mail-to-all strategy in the range $0.05 to $0.20 per head.

Hansotia & Rukstales (2001, 2002) describe their approach to what they call *Incremental Value Modelling,* which involves using the difference in raw uplifts in the two subpopulations as a split criterion. This, indeed, is a natural approach but has the obvious disadvantage that population size is not taken into account, leading to an overemphasis on small populations with high observed uplift in the training population.

Lo (2002) has maintained a long-term interest in what he calls *True Lift Modelling* while working in direct marketing for Fidelity Investments. He developed an approach which is based on adding explicit interaction terms between each predictor and the treatment. Having added these terms he performs a standard regression. To use the model, he computes the prediction with the treatment variable set to one (indicating treatment) and subtracts the prediction from the model with the treatment variable set to zero. Lo has used this approach to underpin direct marketing at Fidelity for a number of years (Lo, 2005) with good success.

Manahan (2005) tackles the problem from the perspective of a cellular phone company (Cingular) trying to target customers for retention activity around contract renewal time. As Manahan notes, an extra reason for paying attention in this case is that there is clear evidence that retention activity backfires for some customers and has the net effect of driving them away. Manahan calls his method a *Proportional Hazards* approach, and the paper is couched in terms of survival analysis (hence the 'hazards' language), but on close reading it appears that the core method for predicting uplift is, like Hansotia & Rukstales (2001), the 'two-model' approach, i.e. direct subtraction of models for the treated and untreated populations. Manahan uses both logistic regression and neural network models, and finds that in his case the neural approach is more successful. (Manahan creates rolling predictions of customer defection rates from his uplift models and compares these with known survival curves both as a form of validation and an input to model selection.)

As well as these published approaches, we have seen many organizations try the natural approach of modelling the two populations (treated and control) separately and subtracting the predictions. This has the advantages of simplicity and manifest correctness. Unfortunately, as we discuss in section 5, in our experience, in all but the simplest cases it tends to fail rather badly. (We refer to this as the 'two-model' approach to uplift

modelling.)

More recently, Larsen (2010) has reported work at Charles Schwab using what he calls *Net Lift Modeling.* His approach is much closer to ours in that it fundamentally changes the quantity being optimized in the fitting process to be uplift (net lift). He does this using a modification of the *weight of evidence* transformation (Thomas, 2000) to produce the *net weight of evidence,* which is then used as the basis of fitting using either a K-Nearest Neighbours approach (Hand, 1981) or a Naïve Bayes approach (Hand & Yu, 2001). Larsen also proposes using a net version of 'information value' (the *net information value*) for variable selection.

Finally, Rzepakowski & Jaroszewicz (2010) have proposed a tree-based method for uplift modelling that is based on generalizing classical tree-building split criteria and pruning methods. Their approach is fundamentally based on the idea of comparing the distributions of outcomes in the treated and control populations using a divergence statistic and they consider two, one based on the Kullback-Leibler divergence and another based on a Euclidean metric. Although we have not performed an experimental comparison yet, we note that their approach is designed partly around a postulate stating that if the control group is empty the split criterion should reduce to a classical splitting criterion. This does not seem natural to us; a more appropriate requirement might be that when the *response rate* in the control population is zero the split criterion should reduce to a classical case. We are also concerned that their proposed split conditions are independent of overall population size, whereas our experience is that this is critical in noisy, real-world situations. Finally, it is troubling that the standard definition of uplift (as the difference between treated and control outcome rates) cannot be used in their split criterion because of an implicit requirement that the measure of distribution divergence be convex.

4 Quality Measures and Success Criteria

Given a valid control group, computing the uplift achieved in a campaign is straightforward, though subject to relatively large measurement error. Assessing the performance of an (uplift) model is more complex.

We have found the uplift equivalent of a gains curve, as shown in Figure 1, to be a useful starting point when assessing model quality (Radcliffe, 2007; Surry & Radcliffe, 2011). Such incremental gains curves are similar to conventional gains curves except that they show an estimate of the cumulative incremental impact on the vertical axis where the conventional gains curve shows the cumulative raw outcome.

If we have predetermined a cut-off (e.g. 20%), we can use the uplift directly as a measure of model quality: in this case, Model 1 is superior[12] at 20% target volume because it delivers an estimated 450 incremental sales against the estimated 380 incremental sales delivered by Model 2. At target volumes above 40%, the situation reverses.

[12]For simplicity, we are not specifying, here, whether this is training or validation data, nor are we specifying error bars, though we would do so in practice, giving more weight to validation performance and taking into account estimated errors.

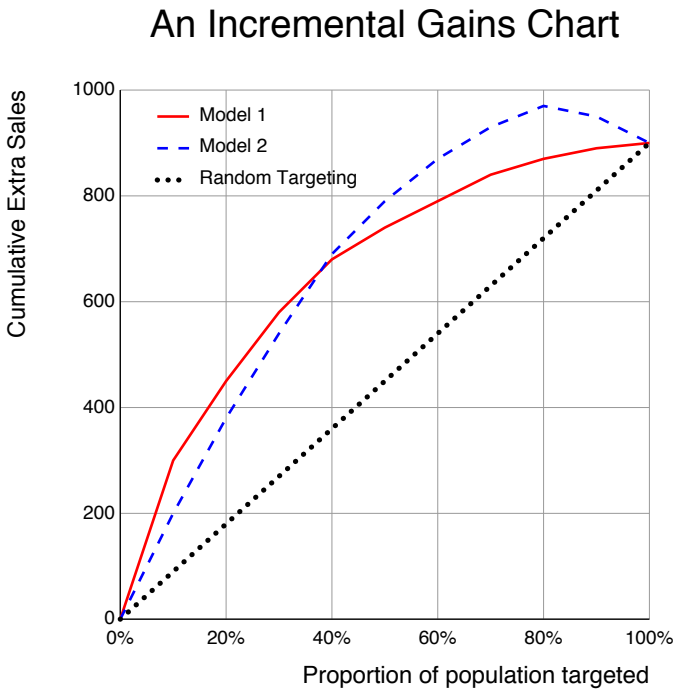

Figure 1: This *incremental gains curve* shows the effect of targeting different proportions of the population with two different models. In each case, people are chosen in descending order of quality as ranked by the model in question. The vertical axis shows an estimate, in this case, of the number of *incremental* sales achieved. This estimate is produced by comparing the cumulative purchase rate, targeting by model score, in the treated and control populations (section 4.1). The vertical axis can alternatively be labelled 'uplift', and measured in percentage points. The diagonal shows the effect of random targeting. Note that using Model 2, more incremental sales are achieved by targeting 80% of the population than by targeting the whole; this is because of negative effects in the last two deciles. In cases where the focus is on revenue or value, rather than conversion, the vertical axis is modified to show an estimate of the cumulative incremental sales value, rather than the volume.

Given cost and value information, we can determine the optimal cutoff for each model and choose the one that leads to the highest predicted campaign profit. Figure 2 is derived directly from the incremental gains curve by applying cost and value information, illustrated here with the cost of treating each 1% of the population set to $1,000 and the profit contribution from each incremental sale set to $150. Using these figures, we can go further and say that Model 2 is better in the sense that it allows us to deliver a higher (estimated) overall campaign profit (c. $70,000 at 60%, against a maximum of slightly over $60,000 at 40% for Model 1), if that is the goal.[13]

Since Model 1 performs better than Model 2 at small volumes while Model 2 performs better than Model 1 (by a larger margin) at higher target volumes, we might borrow the notion of *dominance* from multi-objective optimization (Louis & Rawlins, 1993), and say that neither model *dominates* the other (i.e. neither is better at all cutoffs).

Notwithstanding the observation that different models may outperform each other at different target volumes, it is useful to have access to measures that summarize performance across all possible target volumes. Qini measures (Radcliffe, 2007) do this, and we will outline them below after a few introductory points.

4.1 Segment-based vs. Pointwise Uplift Estimates: Non-Additivity

The core complication with uplift modelling lies in the fact that we cannot measure the uplift for an individual because we cannot simultaneously treat and not treat a single person. For this reason, developing any useful quality measure based on comparing actual and observed outcomes at the level of the individual seems doomed to failure.

Given a valid treatment-control structure, we can, however, estimate uplift for different segments, provided that we take equivalent subpopulations in each of the treated and control and that those subpopulations are large enough to be meaningful. This includes the case of a population segmented by model score. Thus it is legitimate for us to estimate the uplift for customers with scores in the range (say) 100–200 by comparing the purchase rates of customers with scores in this range from the treated and control populations.

In going down this route, however, we need to be aware that uplift estimates are not, in general, additive (see Table 1). This is because of unavoidable variation in the precise proportions of treated and control customers in arbitrary subsegments.[14]

4.2 Qini Measures

Qini measures are based on the area under the incremental gains curve (e.g. Figure 1). This is a natural generalization of the gini coefficient, which though more commonly defined with reference to the area under a receiver-operator characteristic (ROC) curve, can equivalently be defined with reference to the conventional gains curve. Because the incremental gains curve is so intimately linked to the qini measure, we tend to refer to

[13] This statement assumes that the uplift estimates are accurate. In general, uplift cannot be estimated as accurately as a purchase rate can be.

[14] a phenomenon related to Simpson's 'Paradox' (Simpson, 1951)

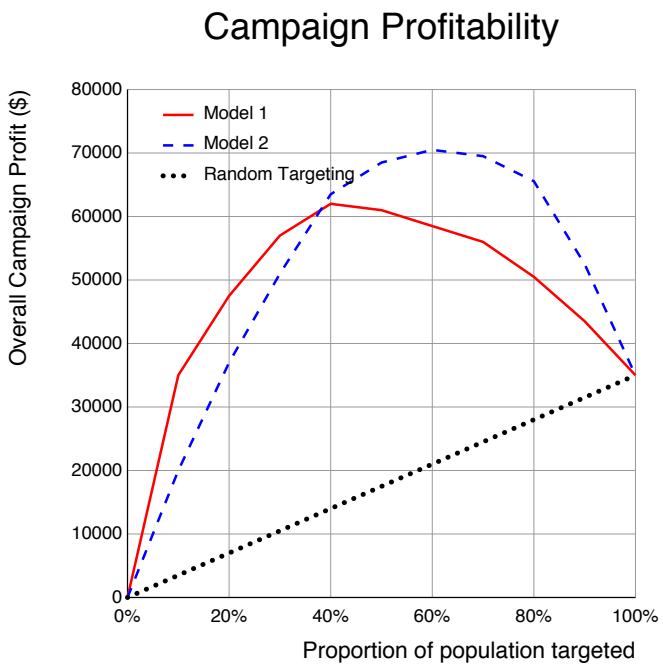

Figure 2: This graph shows the profit achieved using the models from Figure 1 at different targeting volumes on the assumption that an incremental sale delivers net profit contribution of $150 and that targeting each 1% of the population has a cost of $1,000. For Model 1, profit is maximized by targeting 40% of the population, generating 680 incremental sales and an overall campaign profit of $62,000, whereas for Model 2, the optimum target volume is 60%, generating 870 incremental sales and a campaign profit of $70,500.

Table 1: An illustration of the non-additivity of uplift. In this table, the columns are population size (n), the number of sales (# sales) and the proportion of the population that purchased (% rate). The weighted average (which in this case, is the same as the unweighted average) of the uplift estimates from the two segments, 0.497pp (percentage points), is *not* equal to the direct uplift estimate from the overall population (0.500pp). This is *not* the result of rounding error.

	Below average score			**Above or average score**			**Overall**		
	n	# sales	% rate	n	# sales	% rate	n	# sales	% rate
Treated	19,950	384	1.925%	20,050	464	2.314%	40,000	848	2.120%
Control	10,050	122	1.214%	9,950	202	2.030%	20,000	324	1.620%
Total	30,000			30,000			60,000		
Uplift			0.711pp			0.284pp			0.500pp

the curve on an incremental gains graph as a *qini curve.* Qini curves and qini measures are discussed in detail in Radcliffe (2007) but key features include:

1. *Gini.* The gini coefficient is defined as the ratio of two areas. The numerator is the area between the actual gains curve and the diagonal corresponding to random targeting. The denominator is the same area but now for the optimal gains curve. This optimal gains curve is achieved by a model that assigns higher scores to all the responders than to any of the non-responders and leads to a triangular gains curve with slope 1 at the start[15] and 0 after all the purchasers have been 'used up'. Thus gini coefficients range from $+1$, for a model that correctly ranks all purchasers ahead of all non-purchasers, to 0, for a model that performs as well (overall) as random targeting, to -1 for the worst possible model, which ranks each non-purchaser ahead of all purchasers.

2. q_0 *— a direct analogue of gini.* Because of the possibility of negative effects,[16] even for a binary response variable, the optimum incremental gains curve is less obvious, though can be calculated straightforwardly (details in Radcliffe, 2007). However, we do not generally use this for scaling the qini measure, partly because this theoretical optimum is often an order of magnitude or more larger than anything achievable, and partly because it is not well defined for non-binary outcomes. Instead, we often scale with respect to the so-called *zero downlift optimum,* which is the optimal gains curve if it is assumed that there are no negative effects. This version of the qini coefficient is denoted by q_0 and is defined as the ratio of the area of the actual incremental gains curve above the diagonal to the zero-downlift incremental gains curve. It should be noted, however, that if the overall uplift is zero, the area of the zero-downlift qini curve will also be zero, leading to an infinite result for q_0.

3. Q *— the more general qini measure.* Although the q_0 measure is a useful direct analogue of gini for binary outcomes, the more generally useful measure is the unscaled qini measure Q. This is defined simply as the area between the actual incremental gains curve in question and the diagonal corresponding to random targeting. This is scaled only to remove dependence on the population size N, dividing by N^2, where necessary.

4. *Calculational Issues and Non-Additivity.* Uplift estimates are not strictly additive. For this reason, some ways of calculating the qini coefficient are more subject to statistical variation than others (see Surry & Radcliffe, 2011).

4.3 Success Criteria and Goals

The qini measure is the most concrete and direct measure of overall model performance we currently have, but we need to discuss further what it means for an uplift model to be 'good'.

[15] assuming both axes use the same scaling

[16] Negative effects arise when, for some or all segments of the population, the overall impact of the treatment is to reduce sales.

With a conventional model, we can directly compare the predictions from the model to the actual outcomes, point-by-point, both on the data used during model building and on validation data. There is no equivalent process for an uplift model.

We can directly compare the predictions of the model over different subpopulations. The qini measure does this, with the subpopulations being partially defined by the predictions themselves, i.e. we work from the highest scores to the lowest. There is necessarily a degree of arbitrariness in terms of exactly how we define the segments, but this is not a large problem.

The qini measure, like gini, is a measure only of the rank ordering performed by the model. For many purposes this is sufficient, particularly in cases where there is a relatively small, fixed volume to be treated and the model is to be used simply to pick the best candidates.

For some purposes, however, calibration matters, especially when picking cut-off points, i.e. sometimes the actual correspondence between the predicted uplift for a segment and the actual uplift is important.

Even perfect accuracy of predictions at segment level is no guarantee of utility of the model, because uplift predictions can be arbitrarily weak. For example, we could define a set of random segments, and predict the population average uplift for each of them. We would expect excellent correspondence between our useless predictions and reality[17] but the model would be of no help to use because it makes no interesting predictions. Thus we need not only *accurate* predictions at segment level, but a range of *different* predictions in order for a model to have any utility.

In general, we consider all of the following when evaluating uplift models:

- *Validated qini* (i.e. comparing two models, the one with a higher qini on validation data will generally be preferred);[18]
- *Monotonicity of incremental gains:* for reasonably large segments, we like each segment to have lower uplift than the previous (working from high predicted uplift to lower predicted uplift);
- *Maximum impact:* where negative effects are present, we have regard to the highest predicted uplift we can achieve, especially when the incremental gains have a stable, monotonic pattern;
- *Impact at cutoff:* Sometimes, the cutoff is predetermined or determined by a profit calculation based on the predictions: in these cases, we are of course concerned with the performance at the cutoff.
- *Tight validation:* as with conventional models, we are more confident when the pattern seen in validation data is very similar to that in the data used to build the model (though the inherently larger errors associated with measuring uplift mean that validation is rarely as tight as with conventional models).

[17] compromised only by statistical variation

[18] qini values for the same dataset and outcome are comparable, but comparing qini values across different datasets and outcomes can be less meaningful.

- *Range of Predictions.* As noted, other things being equal, the more different the predictions of the model for different segments, the more useful is the model.

When, in later sections, we describe approaches as 'successful' or 'unsuccessful' without specifying a criterion, we are generally referring to a combination of most or all of these performance considerations.

5 Approaches to Building Uplift Models

We first discuss the obvious 'two-model' approach to uplift modelling (section 5.1) and why it tends not to work well (sections 5.2 and 5.3), and some approaches based on additive models (section 5.4). We then discuss the shared character of more fruitful approaches to building uplift models (section 6) before discussing the tree-based approach in detail (section 6.1).

5.1 The Two Model Approach

As noted in section 3, the most straightforward approach to modelling uplift is to build two separate models, one, M_T, for the treated population and another, M_C, for the control population, and then to subtract their predictions to predict uplift ($M_U = M_T - M_C$). The main advantages of this approach are (1) simplicity (2) manifest correctness *in principle* and (3) lack of requirement for new methods, techniques and software. Unfortunately, while the approach works well in simple cases, such as are often constructed as tests, our experience and that of others (e.g. Lo, 2002) is that this approach rarely works well for real-world problems. We will now attempt to explain why this should be the case, accepting that our explanation will be partial because there is much that is still not understood about uplift modelling.

The authors believe that multiple factors contribute to the practical failure of the two-model approach in real-world situations:

1. *Objective.* The two-model approach builds two independent models, usually on the same basis, using the same model form and candidate predictors. Each is fitted on the basis of a fitting objective that makes no reference to the other population. Thus, viewed from an optimization perspective, when we build two separate models, nothing in the fitting process is (explicitly) attempting to fit the difference in behaviour between the two populations.

   In effect, this approach is based upon the (correct) observation that *if* the two models were perfect, in the sense of making correct predictions for every customer under both treated and control scenarios, *then* the uplift prediction formed by subtracting them would also be perfect. It does not, however, follow that subtracting two 'good' independent models will necessarily lead to a 'good' uplift model. (Before we could even assign meaning to such a statement, we would first need to define our metrics for model quality for both the non-uplift and the uplift models.)

   While attempting to fit a given objective is no guarantee of success, other things being equal, the authors believe that approaches in which the optimizer

has knowledge of the actual goal of the exercise have a material and exploitable advantage over approaches in which this is not the case.

2. *Relative Signal Strength.* Carrying on from the previous point, for typical applications of uplift modelling, it is common for the size of the uplift to be small compared with the size of the main effect. For example, in marketing response applications, it would not be atypical for the purchase rate in the control group to be around 1% while the rate in the treated group was around 1.1%, representing an uplift of 0.1pp. This leads the modelling process to concentrate its efforts (both in selecting variables and fitting) on the main effect. To the extent, therefore, that there is any conflict or disagreement, priority will tend to be given to the main effect.

3. *Variable Selection/Ranking/Weighting.* The complete model-building process typically starts from a set of candidate predictors and then reduces these to a subset that actually appear in the model. For example, step-wise methods are common in regression, and tree-based methods, by their nature, prioritize and use only a subset of the available variables in the general case.

   It is both theoretically possible and observable in practice that the most predictive variables for modelling the non-uplift outcome can be different from those most predictive for uplift. In some cases, this leads the two model approach to discard key predictors for uplift entirely, thus limiting their ability to capture the uplift pattern.

4. *Noise, Signal-to-Noise and Fitting Errors.* We suspect that by fitting the two models independently, we leave ourselves more open to the possibility that the fitting errors in the two models might combine in unfortunate ways, thus degrading the quality of fit of the difference model in a way that can perhaps be avoided by specifically attempting to control that error as part of the direct modelling approach. We are not, however, able to make this idea mathematically precise at this point.

5.2 The Two Model Approach Failure: Illustration

We illustrate the points above with the simplest example we have found that demonstrates the effects discussed.

We created an entirely synthetic dataset with 64,000 records. We created predictor variables, x and y, each of which consisted of uniform-random integers from 0–99. We also randomly assigned each member of the population to be either the treated or control segment (in this case, in roughly equal numbers). We then created an outcome variable, o. For the controls, this was drawn from $U[0, x)$, i.e. a uniform random deviate over the half-open interval from 0 to x. For treated members of the population, o was drawn from

$$U[0, x) + U[0, y)/10 + 3, \tag{5}$$

where the two random deviates are independent. This models a situation in which there is a strong background dependency between the outcome and the predictor x, a small

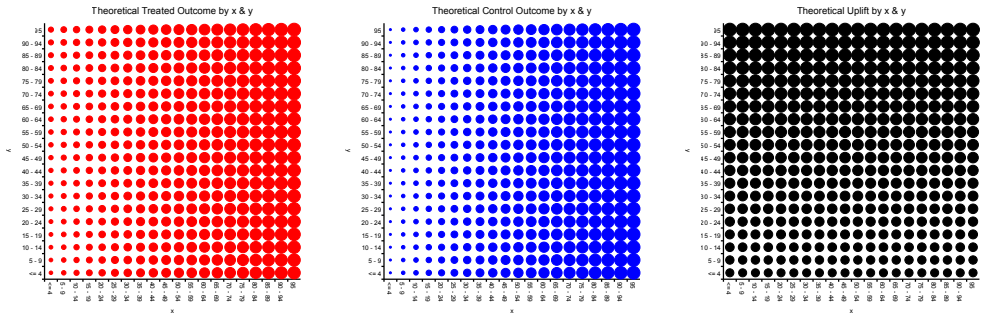

Figure 3: The left-hand plot shows the theoretical mean outcome as a function of x and y for the treated population, using a bin-width of 5, and making the area of the disc proportional to the outcome. The middle plot is the same illustration for the control population. The right-hand plot shows the theoretical uplift in a similar manner. (Note that the plots are scaled independently to allow the variation of uplift with x (horizontal axis) and y (vertical axis) more clearly. If the scales were the same, the areas of the circles in the right-hand plot would be around a factor of ten smaller.)

overall uplift from the treatment (the +3 term) and a weaker systematic uplift effect for increasing values of y.

The underlying relationship that the data samples is therefore as shown in Figure 3.

Figure 4 shows the same plots for the sampled outcomes. Naturally, there is sampling error, with the result that the observed patterns are different from the theoretical ones.

We then used the two-model approach, using simple regression trees (Breiman *et al.,* 1984). The two trees built are shown in Figure 5. They happen to have identical structure (reflecting the fact that the main effect controlling the outcome is similar in the two cases) but different estimates associated with the nodes.

In contrast, Figure 6 shows a comparison of two uplift trees. The first was built using the methods described later in this paper, and directly models uplift. The second tree is formed by taking the difference of the two trees built separately, from Figure 5, which is capable of being represented very simply because the tree structure produced for the treated and control populations was identical in this case. We shall refer to this as a 'difference tree'.

5.3 The Two Model Approach Failure: Discussion

It should be clear that, in this case, the two-model approach has produced a much less useful model than the direct modelling approach. The two-model approach does not even refer to the key predictor, y, has only a small variation in uplift values at the and its q_0 measure is 8.44% against 25.72% for the (direct) uplift model. These models were built without using cross-validation, but the comparison would be even more

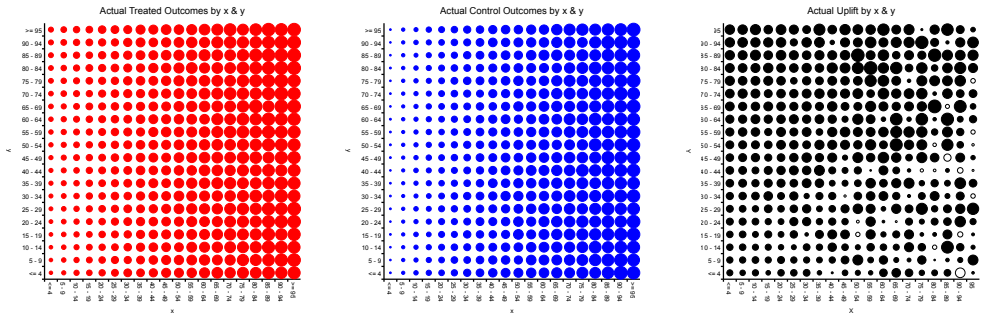

Figure 4: These plots are equivalent to those in the previous figure except that they now show the *observed* mean outcomes from the sampled data. In the case of the third plot, showing uplift, a small number of the observed uplifts are negative; these are shown as empty circles. Again note that the plots are scaled independently to allow the variation of uplift with x and y more clearly.

dramatic if it were employed since the pattern produced by the uplift tree is essentially correct, whereas the pattern produced by the difference tree has in fact succeeded only in modelling noise.

This example focuses primarily on a failure mode of the two model approach that arises when the modelling concentrates on the wrong variables. It might be felt by some readers that regression would be less 'fooled' in this case, and there is some truth in that. The pattern in the dataset chosen is well suited to fitting by linear regression, and if two simple linear regression models are built, and coefficients for x and y are deliberately retained (rather than omitting the much less predictive y effect, as many step-wise approaches would do), the coefficients for x largely cancel, leaving those for y. In this case, the q_0 qini of the resulting model is 19.45%—still lower than the direct tree-based uplift model, but far better than that for the difference tree. But it should be further noted that this illustration was designed specifically to illustrate one failure mode, and happens to be particularly well suited to solution by linear regression. If a more general pattern is used, or a more sophisticated modelling method is employed (e.g. a generalized additive model), the failure reasserts itself. It should also be noted that whereas leaving two variables in a regression is reasonable, variable reduction is typically an important part of the modelling process, and if omitted will itself degrade that process.

5.4 Additive Models for Uplift Modelling

A widely used variation of regression is the generalized additive model (Hastie & Tinshirani, 1990), and in particular scorecard models, in which each predictor variable x is replaced with a function (or *transformation*) of a set of binary indicators ('dummy variables'). The indicators partition the range of the predictor x and for any value of x, exactly one of the indicators is 1 while the others are zero. When the outcome is binary,

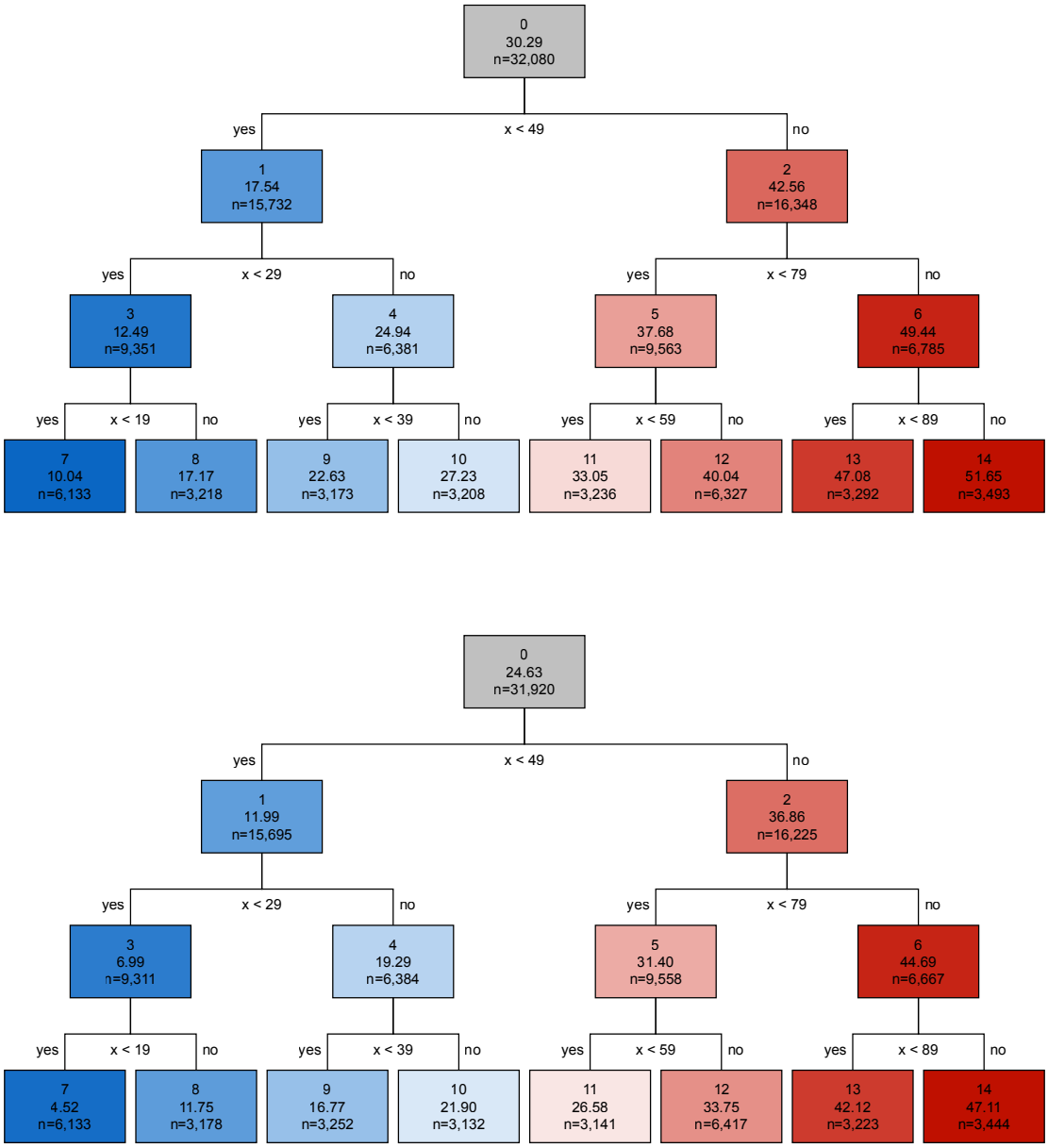

Figure 5: These two trees are simple regression trees built using the CART regression tree splitting algorithm, considering splits near multiples of ten. (No pruning was employed for this simple example, nor was any validation performed, these not being relevant to the illustration.) The upper tree is for the treated population and the lower tree is for the control population.

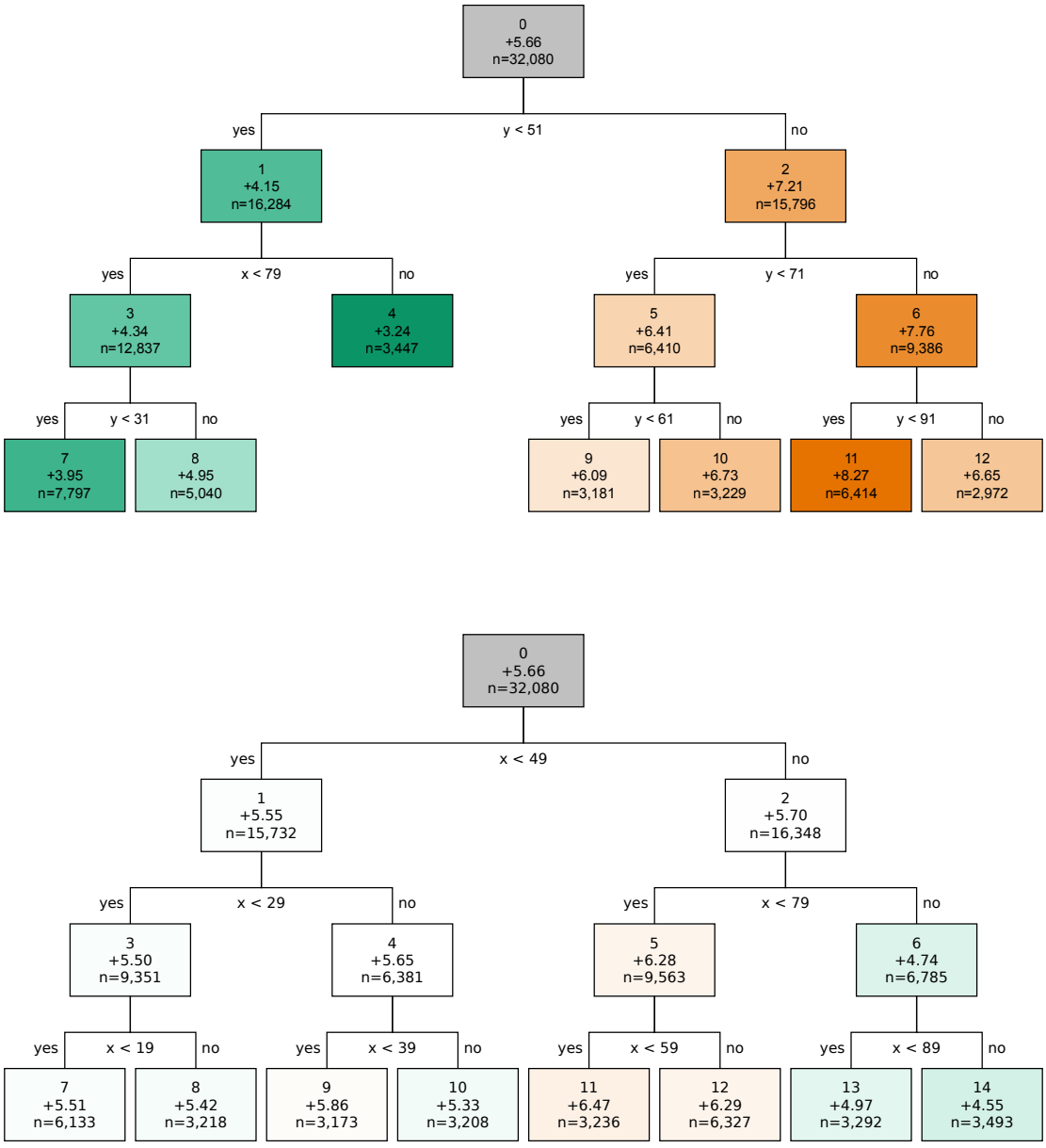

Figure 6: The upper tree is an uplift tree built directly using the tree-build method described in section 6.1. The lower tree is formed by subtracting the node values from the non-uplift regression trees shown in the previous figure built over the treated and control populations. Note the much greater variation in the uplifts at the leaf nodes of the upper tree ($8.27 - 3.95 = 4.32$ against $6.29 - 4.97 = 1.32$), particularly when compared with the root node (5.66).

a common choice for the transformation is the so-called *weight of evidence,* defined as the log of the odds ratio. So the weight of evidence for the ith bin is

$$w_i = \ln\left(\frac{p_i}{q_i}\right) \tag{6}$$

where $p_i = P(Y = 1|X = i)$ and $q_i = 1 - p_i$. There is also the *adjusted weight of evidence,* which simply subtracts off the overall log odds, i.e.

$$w_i^* = \ln\left(\frac{p_i}{q_i}\right) - \ln\left(\frac{p}{q}\right) \tag{7}$$

where p is the overall outcome rate $P(Y = 1)$ and $q = 1 - p$.

The transformation can be viewed as a two-level model, in which rather than modelling the outcome directly in terms of the predictor variables, we first build a simple model by averaging the outcome over various bins and then fit the linear model in terms of the output from that first model. Larsen (2010) proposes a method of uplift modelling based on adapting this approach to use a 'net' version of the adjusted weight of evidence:[19]

$$\Delta w_i^* = w_{i,T}^* - w_{i,C}^*. \tag{8}$$

The use of a net weight of evidence is helpful, but still leaves the question of how to perform the regression itself. Larsen's approach is similar to the method of Lo (2002), in that he performs a logistic regression (using the net adjusted weight of evidence transformation) with separate parameters for the treated and control cases.

Although we have not tried this method ourselves, it seems plausible that the combination of Lo's basic method with the added benefit of the net adjusted weight of evidence transformation could work well.

6 The Significance-Based Uplift Tree

When modelling uplift directly (with a single model) the core problem is that outcomes cannot be measured at the level of the individual. This has implications for the fitting process itself and for measuring model quality. Given the absence of individual outcomes, we use estimated outcomes for segments (subpopulations).

A natural class of models to consider is tree-based models, since these are intrinsically based on segments; these are introduced in section 6.1 and extended to continuous outcomes in 6.3. (Regression methods based on binned variable transformations are another promising candidate, since the transformation can group values from a segment, as discussed earlier in section 5.4).

We now describe the approach to uplift modelling currently favoured by the authors, which forms the core of the uplift modelling available in the *Portrait Uplift* product from Pitney Bowes. The key features of the current approach are:

- the signficance-based splitting criterion (section 6.2).

[19] Larsen refers simply to *the weight of evidence,* but his formulae make it clear that it is the adjusted version that he uses.

- variance-based pruning (section 6.5);
- use of bagging (section 6.6);
- pessimistic qini-based variable selection (section 7.4).

The significance-based pruning criterion was introduced to the software (but no algorithmic details were published) in 2002; the other features were added between 2002 and 2007.

6.1 Tree-based Uplift Modelling

The criteria used for choosing each split during the growth phase of standard divisive binary tree-based methods (notably CART and Quinlan's various methods) trade off two desirable properties:

- maximization of the difference in *outcome* between the two subpopulations;
- minimization of the difference in *size* between them.

These tend to be inherently in conflict, as it is usually easy to find small populations that exhibit extreme outcome rates—for example, splitting off one purchaser (a segment with a 100% purchase rate) from the main population.

We approach split conditions for uplift trees from the same perspective, the difference being that the outcome is now the uplift in each subpopulation, rather than a simple purchase rate or similar.

The method of Hansotia & Rukstales (2001) is simply the result of ignoring the trade-off and directly using the difference in uplifts (which they call "$\Delta\Delta p$") as the split criterion. Notwithstanding their reported success, we have not had good results with this approach.

We have also tried using qini directly as the split quality measure. Qini does take into account both population size and uplift on each side of the split, but although we find qini useful for assessing the overall performance of uplift models, we have had only limited success in using it as a split criterion. The fact that qini only measures rank ordering is probably a factor here.

It is also possible to take an *ad hoc* approach, whereby the difference in population sizes is treated as some sort of penalty term to adjust the raw difference in uplifts. If the (absolute) difference in uplifts is Δ and the two subpopulation sizes are N_L and N_R, a natural candidate penalized form might be

$$\Delta \; / \left(\frac{N_L + N_R}{2 \min(N_L, N_R)} \right)^k , \tag{9}$$

for some k. This penalty is 1 when the populations are even and increases as they become more uneven. Another obvious alternative penalized form might be

$$\Delta \left(1 - \left| \frac{N_L - N_R}{N_L + N_R} \right|^k \right) \tag{10}$$

for some k. Here, the penalty is zero for equal-sized populations and increases to 1 as their sizes diverge. (In both cases, k, is a parameter to be set heuristically.) Our earliest method, described in Radcliffe & Surry (1999), was of this general form. However, we failed to find any *ad hoc* penalty scheme that worked well across any useful range of real-world problems.

6.2 The Significance-Based Splitting Criterion

Our current significance-based splitting criterion fits a linear model to each candidate split and uses the significance of the interaction term as a measure of the split quality. Considering first the case of a binary outcome we model the response probability as

$$p_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \tag{11}$$

where

- p is the response probability;
- i is T for a treated customer and C for a control;
- j indicates the side of the split (L or R);
- μ is a constant related to the mean outcome rate;
- α quantifies the overall effect of treatment;
- β quantifies the overall effect of the split;
- the interaction term, γ, captures the relationship between the treatment and the split.

Without loss of generality, we can set

$$\alpha_C = \beta_L = \gamma_{CL} = \gamma_{CR} = \gamma_{TL} = 0. \tag{12}$$

Then, γ_{TR} is the difference in uplift between the two subpopulations, which is exactly the quantity in which we are interested.

For a given dataset with known (binary) outcomes and values for the candidate split, we can determine γ_{TR} and its significance by fitting the linear model defined by equations 11 and 12 using standard weighted (multiple) linear regression. We then use the significance of γ_{TR} as the quality measure for the split. The significance of parameters in a multiple linear regression is given by a t-statistic (Jennings, 2004). Note that the relevant t-statistic tests the significance of the estimator for γ_{TR} *given that the other variables are already in the model,* thus isolating the effect of the split on uplift, which is what we are interested in.

In practice, since we do not care about which side of the split is higher, it suffices to maximize the square of the statistic

$$t^2\{\gamma_{TR}\} = \frac{\gamma_{TR}^2}{s^2\{\gamma_{TR}\}} \tag{13}$$

where our parameter γ_{TR} is estimated using the observed difference in uplift between the left and right populations, $U_R - U_L \equiv (p_{TR} - p_{CR}) - (p_{TL} - p_{CL})$, and

$$s^2\{\gamma_{TR}\} = MSE \times C_{44}, \tag{14}$$

where MSE is the mean squared error from the regression and C_{44} is the (4,4)-element of $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$. Here, $\mathbf{X}$ is the design matrix for the regression

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & X_{13} \\ 1 & X_{21} & X_{22} & X_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & X_{n3} \end{pmatrix}, \tag{15}$$

in which X_{i1} indicates whether the ith record was treated, X_{i2} indicates whether the ith record was on the right of the split, and $X_{i3} = X_{i1}X_{i2}$ is the interaction indicator, set to 1 when the record is treated and on the right of the split. We have four degrees of freedom in the regression (corresponding to μ, α_T, β_R and γ_{TR}) so the $MSE = SSE/(n-4)$ where n is the number of observations.

Simplifying $\mathbf{C}$, we find that

$$C_{44} = \frac{1}{N_{TR}} + \frac{1}{N_{TL}} + \frac{1}{N_{CR}} + \frac{1}{N_{CL}}, \tag{16}$$

the reciprocal harmonic mean of the cell sizes, N_{ij}. In practice, therefore, the expression we actually use for split quality is

$$t^2\{\gamma_{TR}\} = \frac{(n-4)(U_R - U_L)^2}{C_{44} \times SSE}, \tag{17}$$

with

$$SSE = \sum_{i \in \{T,C\}} \sum_{j \in \{L,R\}} N_{ij} p_{ij}(1 - p_{ij}). \tag{18}$$

6.3 Continuous or Ordered Outcomes

Up to this point, we have focused on cases in which the outcome that we wish to model is binary, but there are also common cases in which the outcome is continuous.[20] Perhaps the most common case is a campaign where the goal is to increase customer spend, so that the uplift in question is the incremental spend attributable to the treatment (equation 4).

The significance-based split criterion defined by equations 11 and 13 is based on fitting a linear model, so can be extended to the continuous case without modification, i.e. the p_{ij} in equation 11 can now be replaced with a general outcome O_{ij}. The calculation of the SSE in equation 18 becomes slightly more complicated, but the formulae are otherwise unchanged.

NOTE: In classical response modelling, if the "response" rate is low, it is common

[20] or more generally, an ordered outcome, which can include a discrete ordinal.

practice to adopt a two-stage approach to estimating outcomes by expressing the expected revenue R as

$$E(R) = P(R > 0) \cdot E(R\,|\,R > 0). \tag{19}$$

This would then be modelled using a binary model for the first term and weighted by a continuous model built only on people in the training data with non-zero spend. This approach is used because a large number of zeros can often limit the ability of a continuous model to fit the non-zero values accurately.

Similarly, modelling uplift in spend directly with a single uplift model is best suited to situations in which the outcome for most people is a non-zero spend.

If most people will not, in fact, spend, it may well be that a two stage approach, with a binary uplift model and a conventional continuous model over the spenders will be a more effective approach. The continuous model, in this case, should be built on the purchasers in the treated population, rather than the whole population.

6.4 Multiplicative Uplift and Modelling

The authors have a strong bias towards modelling uplift as an additive phenomenon, i.e. towards measuring uplift as an absolute difference between mean outcomes in treated and control populations. An alternative approach would be to measure uplift multiplicatively, i.e. to use the ratio of the outcomes in the treated and control populations. We have developed, but never use, a version of the significance-based uplift tree for such multiplicative uplift, and we have chosen not to document it here.

The reason we favour measuring uplift as a difference is we have been unable to find any plausible application in which it would be rational to act based on the ratio of outcomes rather than the difference.

For example, consider a marketing application in which there is a choice of targeting segment A, in which our treatment increases response rate from 1% to 1.5%, or segment B, in which the response rate is raised from 5% to 6%. Other things being equal (including the segment sizes), the value of targeting segment B is twice that of targeting segment A, even though the uplift ratio in segment A is 50% compared with only 20% in segment B. Money is additive.

From a utilitarian perspective, the same is true with medical interventions: even if a drug completely cures all patients of type A, with an untreated mortality rate of 1%, and "only" reduces mortality from 90% to 30% is segment B, it is clear that targeting patients in segment B reduces mortality 60 times more than does targeting segment A, notwithstanding the infinite uplift ratio in segment A.

This seems to be the invariable pattern, at least when the true outcome is being measured.

It might, in principle still be preferable to model uplift as a ratio if the underlying impact of the treatment were believed to be multiplicative, because in such cases it seems likely that a more accurate model would result. In practice, however, to understand the (additive) impact of action, it would still be necessary to model the outcome without treatment and then to compute the uplift as a difference, leading to similar problems to those identified with the two model approach earlier. We do not, therefore, recommend using a definition of uplift as a ratio.

6.5 Variance-Based Pruning and Limiting Split Choices

Most classical decision tree algorithms involve first building a deep tree by recursive splitting and then pruning the resulting bushy tree by removing unhelpful splits. For example, both CART and Quinlan's algorithms use this approach. The pruning phase is intended to reduce overfitting and commonly uses some form of cross-validation.

The main reason for adopting such a two-stage approach is that trees are highly non-linear models in which predictions can depend strongly on the interaction between variables. Where there are strong non-linear patterns in the data, it may be that an enabling split higher up a tree is only marginally useful by itself, but becomes more useful lower down when combined with a further split, so true value of a split cannot always be determined without further splitting.

In the most extreme form of the growth algorithm, splitting continues until all nodes are pure (i.e. have identical outcomes for all records) or no further useful splits can be found that increase purity. It is common practice, however, when working with large datasets to limit the smallest population size allowable, and possibly also the maximum depth of the tree. It is also common to consider only a subset of possible splits, rather than literally every possible univariate split, as is recommended in CART (Breiman *et al.,* 1984).

All of these points apply if anything more strongly in the case of typical uplift models. The challenges with uplift modelling include (1) overall uplift is often small compared to the background effect (2) the control population is commonly smaller than the treated population, often by a factor of ten or more (3) uplift is a second-order phenomenon, with consequently large errors associated with the estimates. For all of these reasons, the single biggest challenge with uplift modelling tends to be producing stable models in the face of these difficulties.

The approach to pruning we have found most successful for significance-based uplift trees involves resampling the training population (with replacement) k times; we usually take k=8. Numbering the resulting resampled populations 1–k, we train on population 1 and then evaluate the stability of the tree with reference to populations 2–k. We measure the uplift in population 1 at each node, and then estimate its standard deviation across the remaining seven populations. Splits (and their descendents) are removed if the uplift at either child node exhibits a standard deviation greater than some pre-determined threshold. The exact details are not important except to note that it is the deviation from the mean in population 1 that we measure, because that is the uplift estimate that the tree will use. It is hard to generalize, because it varies so much from dataset to dataset, but for typical marketing problems, where background rates might be 1–3% and uplift might be 0.1 to 2 percentage points, we most often use pruning thresholds in the range 0.5% to 3%.

It should be emphasized that resampling and pruning all happens within a training population; any population held back for validation of the final model is not involved.

6.6 Bagging for Stability

In practice, once a sensible approach to modelling and variable selection has been identified, the main practical difficulty encountered when building uplift models is achiev-

ing model stability. The difficulty arises from the combination of trying to model a second-order phenomenon and the typically low strength of the interaction relative to the background effect. In plainer words, uplift modelling is particularly hard in practice because it is often applied to situations in which the overall impact being modelled is modest.

In addition to using pessimistic qini estimates in the variable selection, we tend to employ a number of other mechanisms for increasing robustness, including, most importantly, bagging (Breiman, 1997). As with variable selection, we most commonly form $n = 8$ populations by resampling the training set (with replacement). We then build a model on one of the populations and validate it on the other partitions, rejecting the model if it fails to validate appropriately. We build a number of such models (typically $b = 10$ or 20), each using different resamplings, and average their predictions. Employing this approach, we often succeed on problems that we cannot model effectively using a single tree. This approach is related to, though different from, the random forests approach suggested by Breiman (2001).

7 Variable Selection

Variable selection is recognized as an important part of any model-building process. Our experience is that it is actually of greater importance in uplift modelling than in conventional modelling.

7.1 Conventional Motivations for Variable Selection

In the context of conventional modelling, there are a number of different motivations for reducing the set of variables prior to model building, depending on the type of model being built, the build method and the purpose to which the model is to be put. Some of the more important motivations include:

1. *Reducing the dimensionality of the model and the likelihood of overfitting.* For modelling approaches in which all available variables are used (such as traditional multiple regression), there is a clear need to reduce the number of variables (if large) to control the number of degrees of freedom. Performing a regression with a large number of independent parameters will tend to result in severe overfitting.[21]

2. *Avoiding correlation.* There are a number of difficulties with using strongly correlated variables in modelling, particularly if the model is supposed to be causal. If a pair of strongly correlated variables exists, there is generally freedom to increase the weight on one and decrease the weight on the other, leading, at a minimum, to interpretational challenges, and in practice often to unstable results and numerical errors.

[21] Transformations, as discussed in section 5.4 are also, among other things, a way of reducing the number of degrees of freedom.

3. *Improving model quality and stability.* For some build methods, removing variables can actually improve model quality even on the training data. As a simple example of the principle, standard greedy tree-building methods do not, in general, produce optimal trees, and it can be the case that removing a variable that will be used for splitting at one level will actually result in a better tree when more levels are built.

   Validation considerations increase this motivation. In almost all models, it is better to remove a variable that will cause instability before the model build proper than to leave it in. If (ultimately) left in, it will normally degrade the model. If taken out at some later stage, its damage will often have been done, either by leaving other model parameters set suboptimally, or by having effectively removed the chance for better variables to be included in the model. For example, in the standard tree-building approach the tree is first built (greedily) on a training population, and then pruned using a validation population. There is no "try re-building with alternative variables" phase, so splits removed during pruning have prevented other splits, that might not have been pruned, from appearing in the tree.

4. *Improving model interpretability.* When models are to be interpreted, different variables may lead to different interpretations. Moreover, given a pair of correlated variables, x_1 and x_2, both useful as predictors of the outcome, o, it may be, for example, that x_1 actually drives x_2 and o; in this case, there is a strong advantage in using x_1 rather than x_2 as the predictor, particularly if other factors drive x_2, but not o, so that in future x_1 might remain a better predictor if x_1 and x_2 diverge.

7.2 Motivations for Variable Selection in Uplift Modelling

In the context of uplift modelling, while all of these considerations remain, our experience suggests that motivation 3 comes to the fore. This is because uplift models are *second-order* models in the sense that they model the difference between two outcomes, rather than a direct outcome. Because of this, and also because in many practical cases the uplift is small relative to the direct outcomes, the risk of overfitting increases markedly.

Indeed, we could state this more strongly. Before we developed suitable variable selection procedures for uplift modelling, it was common for our uplift models to fail to validate to any useful degree (or to be pruned right back to the root during the pruning phase). In this sense, we have found that for many practical problems, good variable selection is an absolute prerequisite for uplift modelling.

7.3 Desirable Properties of Variable Selection Methods

Given the above, there are three main attributes we might seek from the predictors chosen by a variable-selection procedure for uplift modelling:

1. predictiveness;

2. stability (robustness);

3. independence.

The first is fundamental: if we don't have variables that are predictive, by definition we have little hope of building a useful model.

We believe that the second, as discussed above, is even more important for uplift models than for conventional models, because the risk of instability (in the sense of failure to validate) is that much higher for second-order models.

The third, though desirable, is probably less important, particularly when using tree-based methods. Other things being equal, it is certainly better to identify a set of candidate predictors that capture different aspects of the relationship and thus have low correlation. To some extent, step-wise methods for regression seek to achieve this. In practice, however, we find that if we are effective in removing variables that lead to instability, and in including primarily variables with good correlation, our trees are usually quite effective at combining them to produce useful uplift models.

7.4 Pessimistic Qini Estimates

The simplest approach to variable selection is to use a quality measure to rank the candidate variables and then to take the best ones identified, usually either choosing a fixed number, or all those above some threshold, perhaps again subject to a threshold quality.

We have had success using this approach with quality measures based on qini. However, in light of the need for stability/robustness, as highlighted above, we have found it useful to modify our qini estimates to be more pessimistic, thereby further reducing the likelihood of choosing variables that will lead to unstable models.

There are various ways if making such a pessimistic, or 'low' qini estimate (LQE). Our approaches are mostly based on subtracting some kind of estimate of the spread of the qini distribution. To do this, we most often resample the training population in a manner similar to that used with bagging (Breiman, 1997). Thus, in our currently favoured approach, we produce n different resampled variants of the population (typically $n = 8$), all of which have the same size as the original and which are formed by resampling the training population with replacement. We then calculate the qini[22] in each population to form a set of estimates $\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_n$. We then adjust the qini estimate by subtracting off some multiple of s, the sample standard deviation. While the most natural multiple to take off is probably something like $1/\sqrt{n}$ ($\simeq 0.35$ for $n = 8$), we tend to use a larger multiple than this, often in the range 0.5 to 1.0, and to some extent use this factor as a control parameter for the variable selection. Most often, we start with it at 0.5, then increase it if we find we are having problems with model stability and reduce it if we find stability is acceptable and we are seeking to increase the predictive power of the model.

Many variations on this basic theme are possible.

[22] usually using the unscaled qini, Q, to avoid problems should the overall uplift be zero

7.5 Net Information Value

As noted above, Larsen (2010) has proposed the *net information value* as a method for variable selection for uplift modelling with binary outcomes.

The (ordinary) information value (I) is formed from the adjusted weight of evidence w_i^* by forming a weighted sum of the weights of evidence over the bins, weighting each weight of evidence by the difference in outcome rates (1 and 0) in the bin, i.e.

$$I = \sum_i w_i^* \left(P(O = 1|X = i) - P(O = 1|X = i)\right) \tag{20}$$

This is a common metric used for variable selection in conventional modelling. Larsen proposes a modification of this using his net weight of evidence and the four likelihoods associated with each bin (outcomes 1 and 0 for treated and control).

Larsen defines net information value I_N as essentially

$$\sum_i \Delta w_i^* \Big(P(X = i|O = 1)_T P(X = i|O = 0)_C - P(X = i|O = 0)_T P(X = i|O = 1)_C \Big). \tag{21}$$

We have not experimented with this approach.

8 Closing Remarks

The authors have been using uplift modelling commercially in various forms for some twelve years now, and have been surprised that use of the approach has grown as slowly as it has. However, the use of uplift modelling, in various forms, is now on the rise. We hope this more detailed paper discussing the techniques will encourage that continued growth, since, from the authors' perspective, it is now clearly demonstrated not only that uplift modelling is the correct formulation of the problem that many marketers are trying to solve, but also that the techniques presented here and elsewhere can, in many cases, produce measurable results that are variously superior in terms of profitability, cost and targeting volume, while reducing negative effects. We will close by discussing what we have learned, heuristically, about when uplift modelling is more and less effective relative to the alternatives.

8.1 When is Uplift Modelling Worthwhile?

While we would argue that, in principle, uplift modelling is *almost always,* from a theoretical perspective, the correct way to formulate marketing response modelling, it is not the case that in practice, it will always produce superior results, nor that even when it will, the improvement always justifies the extra complexity. We suggest the following checklist when deciding whether or not to take an uplift modelling approach.

- *Existence of a valid control group.* Straightforwardly, if no valid control group exists in the historical data, uplift modelling cannot normally be employed. (But one can be created for the next iteration of the campaign.)

- *Volume.* Not only must a control group exist, but it must be large enough to support uplift modelling. One rule of thumb we use is that the control group[23] needs to be at least ten times larger for modelling than it does for simple measurement of incremental response. A second rule of thumb is that, for modelling binary outcomes, the product of the overall uplift and the size of each population should be at least 500. So, if the overall uplift is 0.1%, this means that both the treated and the control group need to be at least 500,000.

- *Negative Effects.* The likely (or known) presence of negative effects is a strong reason to consider uplift modelling seriously. Conventional modelling is inherently unable to handle negative effects, and negative effects are doubly harmful, in that they both carry a cost of action and reduce the overall impact of the activity. It is not uncommon, particularly in the area of retention, for uplift modelling to deliver more value by identifying populations where negative effects are prevalent than from ranking the parts of the population where the impact is neutral or positive. Negative effects tend to occur most when uplift is small relative to the background outcome rate, where inaction on the part of the customer leads to a positive outcome for the supplier, where interventions are intrusive, and where customers targeted are already unhappy with the supplier, which is why uplift modelling has been perhaps most widely used and most effective in customer retention applications.

- *Complex Customer Influences.* Where the customer is subject to many influences (advertising, multiple communications, in-branch activity etc.), the risk of attributing sales incorrectly to a particular piece of marketing activity is larger than where there are fewer interventions, so the difference between uplift models and conventional models tends to be more marked.

- *Anticorrelated Outcomes.* A situation we commonly see in retail environments is that direct marketing activity appears to have the most positive effect on high-spending customers. Where this is the case, uplift and sales are positively correlated, and therefore a conventional model and an uplift model are more likely to rank the population in similar ways. While we first conceived uplift modelling in a retail sales context, and have had some positive results in this area, overall we have found the benefits of uplift modelling are often smaller in retail environments, where we often see these positive correlations, than in some other areas.

  Conversely, almost by definition, when the (non-incremental) outcome is anticorrelated with the incremental impact of marketing activity, a conventional model is likely to perform particularly badly, and the benefit of uplift modelling may be larger. This is the case when, for example, a marketing offer succeeds in driving up purchase rate most among a group of customers who normally do purchase little, but where that group remains a relatively low purchasing group. The two cases are illustrated in figure 7.

[23] Technically, the smaller of the control group and the treated group, but usually this is the control group

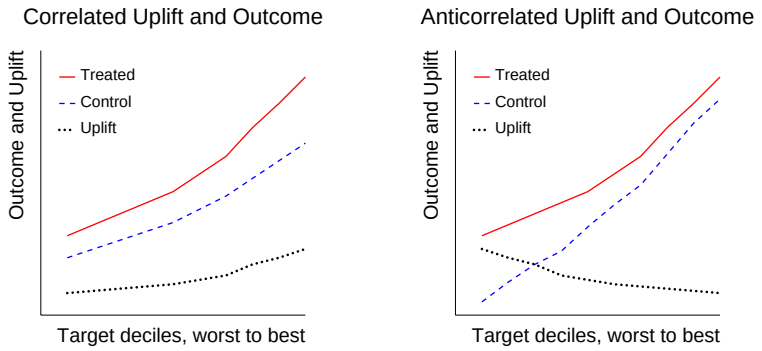

Figure 7: The graph on the left illustrates positive correlation between outcome (top two lines) and uplift (bottom line). A conventional model built on either population should rank this population well for uplift. The graph on the right illustrates negative correlation between outcome (ascending lines) and uplift (descending line). A conventional model built on either population will tend to rank the worst group by uplift as the best targets.

- *Background Rate and Brand Recognition.* Another situation in which conventional analysis may well be misled is when the background outcome rate is high. As an extreme example, sending an incentive to an existing regular weekly customer and then counting a purchase the following week as a 'response' is probably unwise. On the other hand, a little known brand that targets people who are unlikely to have heard of the company or product might reasonably be more confident that apparent responses really are incremental even without a control group, and in such cases, an uplift model may offer little or no benefit, or may actually perform worse than a conventional 'response' model.

8.2 The Trade Off

Ultimately, with uplift modelling, there is a trade-off between modelling the right thing (which uplift modelling, applied appropriately, does) against the added complexity of modelling and increased data requirements that are inherent in the second-order problem formulation that uplift modelling involves. We have tried to give guidance above as to when that trade-off genuinely favours the uplift approach and when it may fail to deliver value. We close with the wise and germane words of Tukey (1962):

> *Far better an approximate answer to the right question . . . than the exact answer to the wrong question.*[24]

[24]The full quote is *"Far better an approximate answer to the right question, which is often vague, than the exact answer to the wrong question, which can always be made precise."* In our case, there is nothing vague about the right question, but it is the case that, as second-order models, uplift models usually have larger error bars than conventional models, and are in this sense "more approximate".

Acknowledgements

Many people contributed to the development of uplift modelling over a period that spanned more than a decade. We would like to recognize and thank, in particular, David Signorini and Tim Harding, as well as Ian Bradbury, Ann Gould, Elaine Farrow, Bob Fletcher, Sandy Nicholson, Rob Simpson, Brian Gibb, Ian Flockhart and others too numerous to list, all of whom contributed to the ideas, implementation, testing and other aspects of the research programme that resulted in uplift modelling as it is today.

References

L. Breiman, J. Freidman, R.A.Olshen, and C. Stone, 1984. *Classification and Regression Trees*. Wadsworth.

L. Breiman, 1997. Arcing classifiers. Technical report, Statistics Department, University of California at Berkley.

L. Breiman, 2001. Random forests. *Machine Learning*.

D. M. Chickering and D. Heckerman, 2000. A decision-theoretic approach to targeted advertising. In *Sixteenth Annual Conference on Uncertainty in Artificial Intelligence, Stanford, CA*.

T. H. Cormen, C. E. Leiserson, and R. L. Rivest, 1990. Chapter 16, greedy algorithms. In *Introduction to Algorithms*, page 329. MIT Press and McGraw-Hill.

D. Durand, 1941. Risk elements in consumer instalment financing. Technical report, National Bureau of Economic Research, New York.

M. D. Grundhoefer, 2009. Raising the bar in cross-sell marketing with uplift modeling. In E. Siegel, editor, *Predictive Analytics World Conference, Washington, D.C.* Prediction Impact Inc.

D. J. Hand and K. Yu, 2001. Idiot's Bayes — not so stupid after all? *International Statistical Review*, 69:385–399.

D. J. Hand, 1981. *Discrimination and Classification*. John Wiley (Chichester).

B. Hansotia and B. Rukstales, 2001. Incremental value modeling. In *DMA Research Council Journal*, pages 1–11.

B. Hansotia and B. Rukstales, 2002. Direct marketing for multichannel retailers: Issues, challenges and solutions. *Journal of Database Marketing*, 9(3):259–266.

T. J. Hastie and R. J. Tibshirani, 1990. *Generalized Additive Models*. Chapman & Hall.

D. M. Hawkins and G. Kass, 1982. Automatic interaction detection. In D. Hawkins, editor, *Topics in Applied Multivariate Analysis*, pages 269–302. Cambridge University Press (Cambridge).

K. Hillstrom, 2008. The minethatdata e-mail analytics and data mining challenge. http://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html.

K. Jennings, 2004. Statistics 512: Applied linear models, topic 3, chapter 5. Technical report, Perdue University.

G. V. Kass, 1980. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29(2):119–127.

K. Larsen, 2010. Net lift models.

V. S. Y. Lo, 2002. The true lift model. *ACM SIGKDD Explorations Newsletter*, 4(2):78–86.

V. S. Y. Lo., 2005. Marketing data mining – new opportunities. In J. Wang, editor, *Encyclopedia of Data Warehousing and Mining*. Idea Reference Group.

S. J. Louis and G. J. E. Rawlins, 1993. Pareto optimality, GA-easiness and deception. In S. Forrest, editor, *Proceedings of the Fifth International Conference on Genetic Algorithms*. Morgan Kaufmann (San Mateo, CA).

C. Manahan, 2005. A proportional hazards approach to campaign list selection. In *SAS User Group International (SUGI) 30 Proceedings*.

G. Moore, 1991. *Crossing the Chasm: Marketing and Selling Technology Products to Mainstream Customers*. Harper Business Essentials.

C. T. Onions, editor, 1973. *The Shorter Oxford English Dictionary*. Clarendon Press (Oxford).

G. Piatetsky-Shapiro, 1991. Discovery, analysis, and presentation of strong rules. In G. Piatetsky-Shapiro and W. J. Frawley, editors, *Knowledge Discovery in Databases*. AAAI/MIT Press (Cambridge, MA).

J. R. Quinlan, 1986. Induction of decision trees. *Machine Learning*, 1(1):81–106.

J. R. Quinlan, 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann (San Mateo, CA).

N. J. Radcliffe and R. Simpson, 2008. Identifying who can be saved and who will be driven away by retention activity. *Journal of Telecommunications Management. Henry Stewart Publications*, (to appear).

N. J. Radcliffe and P. D. Surry, 1999. Differential response analysis: Modeling true response by isolating the effect of a single action. In *Proceedings of Credit Scoring and Credit Control VI*. Credit Research Centre, University of Edinburgh Management School.

N. J. Radcliffe, 2007. Using control groups to target on predicted lift: Building and assessing uplift model. *Direct Marketing Analytics Journal, An Annual Publication from the Direct Marketing Association Analytics Council*, pages 14–21.

N. J. Radcliffe, 2008. Hillstrom's MineThatData email analytics challenge: An approach using uplift modelling. Technical report, Stochastic Solutions Limited. Available from http://stochasticsolutions.com/pdf/HillstromChallenge.pdf.

P. Resnick, N. Iacovou, , M. Suchak, P. Bergstrom, and J. Riedl, 1994. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of ACM Conference on Computer Supported Cooperative Work*, pages 175–186. Chapel Hill.

P. Rzepakowski and S. Jaroszewicz, 2010. Decision trees for uplift modeling. *IEEE Conference on Data Mining*, pages 441–450.

E. H. Simpson, 1951. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, 13:238–241.

P. D. Surry and N. J. Radcliffe, 2011. Quality measures for uplift models. *submitted to KDD2011*.

L. C. Thomas, 2000. A survey of credit and behavioural scoring: forecasting nancial risk of lending to consumers. *International Journal of Forecasting*, 16(2):149–172.

J. W. Tukey, 1962. The future of data analysis. *Annals of Mathematical Statistics*, 33(1):1–67.