Quality measures for uplift models

Patrick D. Surry
Pitney Bowes Business Insight*
Patrick.Surry@pb.com

Nicholas J. Radcliffe
Stochastic Solutions Limited† &
University of Edinburgh‡
Nicholas.Radcliffe@StochasticSolutions.com

ABSTRACT

Uplift models, which predict an individual's *change* in behavior due to a specific intervention such as a marketing campaign, have become increasingly popular in both business and scientific fields because they allow direct comparison between the cost and potential return of a discretionary action. Because the *change* in behavior due to an action can not be observed for any isolated individual (since they cannot be both subject to the action and not), traditional approaches both for fitting models and evaluating model quality do not apply directly. Although a variety of methods for fitting uplift models continue to be proposed, surprisingly little work has been done on metrics for assessing and comparing model quality.

This paper reviews potential metrics by analogy with traditional models, including nominal, ranked and parametric measures. In particular, we recommend the Qini coefficient (Radcliffe, 2004) and establish a strong mathematical correspondence with the traditional Gini coefficient which inspired it. The Qini is shown to measure the strength of (anti-) correlation between uplift rate and targeting depth as ordered by model score. Several equivalent geometric interpretations of Gini and Qini are demonstrated, including a generalization of the traditional Lorenz curve. This yields both a more effective practical approach for calculating the Qini coefficient, as well as suggesting an extension of the Kolmogorov-Smirnov statistic to uplift models.

Categories and Subject Descriptors

KDD [**Process**]: Evaluation metrics and methods; KDD [**Algorithms/Models**]: Ranking; Other statistical models; KDD [**Application Area**]: Web and ecommerce; Business; KDD [**Theory**]: Foundations; Statistical foundations; Information theory

Keywords

Uplift modeling, Incremental modeling, Differential response modeling, Lift, Net lift, True lift, Gini, Qini

1. INTRODUCTION

Uplift modeling is rapidly gaining mindshare in the analysis of large-scale business-to-consumer marketing programs (Radcliffe & Surry, 1999; Hansotia & Rukstales, 2001, 2002a; Lo, 2002; Courtheoux, 2003; Lai, 2004; Manahan, 2005), as well in other domains such as personalized medicine (Cai *et al.,* 2009). Alternatively known as incremental modeling, differential response modeling, true lift or net lift modeling, uplift modeling[1] allows the marketer to address a more relevant business problem than is possible with traditional modeling techniques. Typical models for churn, response or credit risk predict the likelihood of an event, or more generally, the level of a continuous outcome such as spend or customer lifetime value, often restricted to those individuals targeted by some marketing activity. In contrast, an uplift model uses both treated and control data to predict the likely *change* in the outcome that will result from a particular marketing action.

The development of uplift modeling techniques has suffered from a lack of accepted quality metrics for models, making it difficult to choose between models for a given problem instance, and to assess the relative efficacy of models across problems. Practitioners tend to resort to *ad hoc* comparisons of model "power", most commonly based a chart such as figure 1 showing actual uplift by decile of model score. For example, several authors suggest comparing the uplift within the top k deciles to the overall uplift (typically for $k <= 4$, e.g. Lo, 2002; Hansotia & Rukstales, 2002b; Lai, 2004; Larsen, 2010). In practice we have also often observed qualitative criteria applied, such as seeking a monotone decreasing pattern of uplift across deciles.

While comparison of uplift above a fixed cutoff is a simple and convenient metric, it is not a altogether satisfactory. Changing the cutoff can reverse the results of a comparison,

*125 Summer Street, Boston, MA, USA.
†37 Queen Street, Edinburgh, EH2 1JX, UK.
‡Department of Mathematics, King's Buildings, Edinburgh, EH9 3JZ, UK.

[1]We prefer the term uplift when comparing treated and control outcomes to avoid confusion with the common usage of lift in comparing (traditional) model targeteing to random selection (e.g. Piatetsky-Shapiro & Steingold, 2000; Piatetsky-Shapiro & Masand, 1999).

and unless confidence intervals are calculated and presented (a practice we strongly encourage), can also misleadingly suggest one model is "better" than another when in fact they are statistically indistinguishable. In numerous practical examples we have seen confidence intervals on uplift deciles that do not separate *any* deciles from the overall uplift, and we must resort to pentiles or even terciles to find meaningful differences.

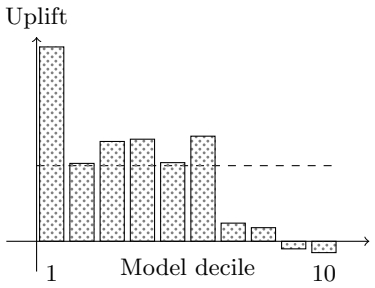

Figure 1: A typical presentation of results from an uplift model, showing uplift for each decile of the model. Model quality might be measured by the ratio of uplift between the top and bottom decile, or between the top decile and the overall uplift.

This paper surveys potential metrics by analogy with traditional models, including nominal, ranked and parametric measures. In particular, the Qini coefficient (Radcliffe, 2004), inspired by the Gini coefficient (Gini, 1912), is recommended. A mathematical correspondence between the two is established, in which the Gini measures the strength of (anti-) correlation between outcome rate and targeting depth (as ordered by model score), and Qini measures the strength of (anti-) correlation between uplift rate and targeting depth (again ordered by model score). Several equivalent geometric interpretations of Gini and Qini are demonstrated. This yields both a more effective practical approach for calculating the Qini coefficient, as well as a generalization of the Kolmogorov-Smirnov statistic for uplift models.

2. QUALITY MEASURES

A wide variety of quality metrics are employed for traditional modeling applications, ranging from nominal nonparameteric measures on $2 \times k$ contingency tables (confusion matrix, chi-square, information gain), through ordinal (ranked) metrics on contigency tables or sorted outcomes (Gini coefficient, Kolmogorov-Smirnov statistic), to parametric statistics such as R^2, divergence and maximum likelihood measures. For example see Malthouse (2001).

In the uplift case, point-wise parametric measures can not be applied since the true point-wise uplift outcome is unknown (an individual can not be both treated and not treated to measure the change in behavior).

Nominal measures on 2×2 contigency tables are useful to evaluate behavior at a predetermined cutoff, for example difference in lift at the k-th decile as noted above. They have also been employed in split criteria for uplift decision trees: indeed our original method (Radcliffe & Surry, 1999, 2011) measures the signficance of the interaction term in a linear model of the contingency table. Hansotia & Rukstales (2002a) evaluate simple differences in uplift across a cutoff, but it seems certain that relative population sizes must also be considered.

Measures on $2 \times k$ contingency tables are useful both to evaluate multi-way splits and as a variable selection technique for nominal variables. Rzepakowski & Jaroszewicz (2010) and Larsen (2010) present alternative generalizations of the Kullback-Leibler divergence to the uplift case. Rzepakowski & Jaroszewicz employ it as split criterion to build an uplift decision tree, while Larsen presents his definition as a variable selection technique, couched in the terminology of credit scoring as a Net Information Value (extending the weight of evidence and information value used in traditional binary models). It is unclear whether either form preserves a straightforward interpretation as the rate of information gain supporting a postitive versus negative outcome. Rzepakowski & Jaroszewicz (2010) also suggest the squared Euclidean distance between contigency table probabilities

Ranked methods are promising as a bridge towards parameteric methods (Conover & Iman, 1981), both for model performance evaluation (using the ranking induced by the model), and for ordinal and continuous variable selection. The Kolmogorov-Smirnov statistic is used with traditional binary-outcome models and measures performance at the "best" cutoff (that which minimizes the sum of positive and negative misclassification rates). However it is not as useful in summarizing model performance across a range of cutoffs, where the Gini coefficient is preferred. For example, the ranked list of binary outcomes 1111010101010000 and 0101111100000101 have equal K-S values (50%) but very different Ginis (75% vs. 25%), since the latter ranking performs poorly for very low or high cutoffs.

In the remainder of this paper, we explore the generalization of the Gini coefficient to uplift models, and suggest an similar generalization of the K-S statistic.

3. THE GINI COEFFICIENT

Consider a predictive model $\mathcal{M}$ that induces a rank ordering on a target population, allowing us to observe the point-wise outcome rate $f_{\mathcal{M}}(x) : [0,1] \rightarrow \mathbb{R}$ where x is the targeting depth.[2] Figure 2 shows a typical plot of outcome rate $f(x)$ (dropping the subscript $\mathcal{M}$ to simplify notation) versus targeting depth, x, with the average outcome rate $\bar{f} = \int_0^1 f(x)dx$.

Further define the cumulative outcome rate, F, as a function of x, namely $F(x) = \int_0^x f(y)dy$, resulting in the standard "gains chart" shown in figure 3. Note that $F(1) \equiv \bar{f}$, and $f(x) \equiv \frac{dF}{dx}$.

The Gini coefficient is traditionally used to measure the de-

[2]Note that the point-wise outcome rate is presented here as a (piecewise) continuous function of x for generality and simplicity. This embraces the typical discrete case of making predictions for a finite set of individuals, by spreading the probability density associated with any discrete score (including groups of tied scores) equally across an equivalent interval of continuous ranks. In short, this results in $f_{\mathcal{M}}(x)$ finite and piecewise continuous on $[0,1]$, and thus integrable.

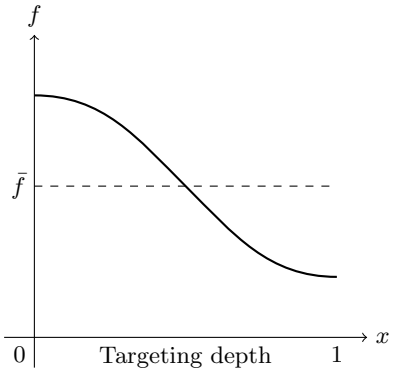

Figure 2: A typical plot of observed outcome rate versus targeting depth as ranked by model score. Because individuals with the best predicted outcomes are targeted first ($x = 0$), a good model tends to induce an observed outcome curve that decreases to the right, but in all cases f is finite and piecewise continuous (see footnote). For a binary outcome, $f \in [0, 1]$, and if $\mathcal{M}$ produces no ties, $f \in \{0, 1\}$.

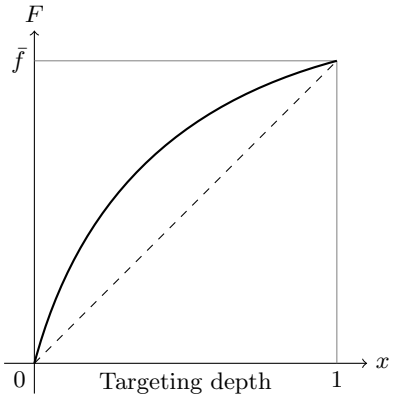

Figure 3: A typical "gains chart" plots the cumulative outcome rate, $F = \int_0^x f(y)dy$, as a function of targeting depth. A random model produces no gain over the diagonal, while a good model tends to produce the bow-shaped curve sketched here because the model ranks higher outcomes to the left.

gree of income inequality within a population, and is defined using a Lorenz curve (Lorenz, 1905) that plots the cumulative proportion of income as a function of the cumulative population size (sorted from lowest to highest income). The Lorenz curve is simply a (reversed) gains chart in which the model is equal to the outcome, i.e. the population is ranked by observed rather than predicted outcome. This is equally applicable to any non-negative outcome, and as shown later, generalizes to arbitrary real-valued outcomes.

In traditional credit scoring and direct marketing usage, where a binary outcome is the most common target for prediction (e.g. response indicator, credit-default flag or stay/go churn outcome), the Gini coefficient, $G_{\mathcal{M}}$, for a predictive model, $\mathcal{M}$, is normally measured using a "receiver operating characteristic" (ROC) curve (Green & Swets, 1966). The ROC curve plots the cumulative proportion of negative versus positive outcomes, with the population sorted by model score, as illustrated in figure 4. $G_{\mathcal{M}}$ is defined as the ratio of the (signed) shaded area under the curve to the area of the upper triangle (representing the ordering induced by the best possible model, $\mathcal{M}^*$, which ranks all positive outcomes before any negative ones).

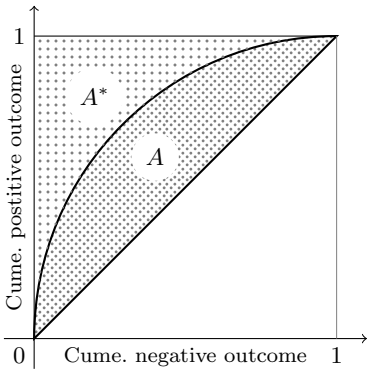

Figure 4: For a binary outcome, the Gini coefficient $G_{\mathcal{M}}$ is defined for a model $\mathcal{M}$ using the "Receiver Operator Characteristic". The population is sorted by model score (most likely to be positive first), and the cumulative rate of positive outcomes is plotted against the cumulative rate of negative outcomes. $G_{\mathcal{M}}$ is equal to the ratio $A/$ (where A^* includes A). The ROC curve also provides a graphical interpretation of the K-S statistic: the intersection of the tangent drawn parallel to the diagonal $y = x$ gives the cutoff that minimizes the sum of the positive and negative misclassification rates, with the tangent's intercept equal to the K-S value.

LEMMA 1. *For a binary outcome, the value of $G_{\mathcal{M}}$ defined using the ROC curve is equal to the value derived from the gains chart of figure 3. That is, $G_{\mathcal{M}}$ can be equivalently interpreted as the the ratio of the average excess cumulative outcome induced by $\mathcal{M}$ to the same quantity induced by the "best case" model ordering, $\mathcal{M}^*$, as illustrated in figure 5. This means we need only consider the gains chart formulation for both binary and continuous outcomes.*

PROOF. In the binary case, figure 5 can be derived by a

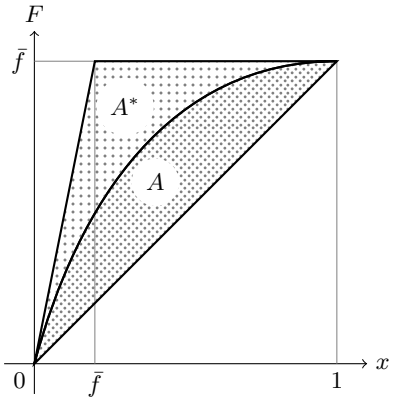

Figure 5: The Gini for income inequality was originally defined using the Lorenz curve, by plotting cumulative proportion of income against cumulative proportion of population and calculating $G = A/A^*$. (In the original formulation, income was sorted from smallest to largest, which simply rotates the area A below the diagonal.) This definition is more general than that based on the ROC curve (see figure 4) since it supports continuous as well as binary outcomes. In the binary case, the perfect model results in the area $A^* = \bar{f}(1-\bar{f})/2$ as shown, since the outcome can at most accumulate at a rate of one unit per individual. For a continuous model, the area A^* could be defined by sorting the actual outcome from smallest to largest (though the authors are unaware of this being done in practice), or by using a vertical line from the origin to $F = \bar{f}$, resulting in area $A^* = \bar{f}/2$.

simple affine transformation of figure 4, namely:

$$(x, y) \mapsto \left(\frac{x-y}{1-\bar{f}}, \frac{y}{\bar{f}}\right) \tag{1}$$

Because an affine transformation scales area by a constant factor ($\bar{f}(1-\bar{f})$ in this case), it is clear that the ratio of areas A/A^* is preserved. (This is easy to see intuitively by imagining integration of both areas along the vertical axis in figure 5: then G takes the form $\int(x_\Delta - x_A)dF / \int(x_\Delta - x_A^*)dF$ where both terms in each integrand are shifted by the same function of F and scaled by a fixed constant.) Thus G is equivalently defined in either diagram. □

THEOREM 1. *The Gini coefficient, $G_\mathcal{M}$, is simply the normalized (anti-) covariance between the outcome rate (induced by the model $\mathcal{M}$) and targeting depth.*[3] *Specifically:*

$$G_\mathcal{M} \equiv \frac{\text{cov}(f, x)}{\text{cov}(f^*, x)} \tag{2}$$

where $f^*(x)$ is the point-wise outcome of the best achievable model as a function of rank.

PROOF.

$$\begin{aligned}
\text{cov}(f, x) &\equiv \int_0^1 (f - \bar{f})(x - \tfrac{1}{2})dx && (3)\\
&= \int_0^1 (f \cdot x)dx - \bar{f}\int_0^1 x dx \\
&\quad - \tfrac{1}{2}\int_0^1 f(x)dx + \tfrac{1}{2}\bar{f}\int_0^1 dx && (4)\\
&= \int_0^1 (f \cdot x)dx - \tfrac{1}{2}\bar{f} && (5)
\end{aligned}$$

Now, $\frac{d(xF)}{dx} \equiv 1 \cdot F + x\frac{dF}{dx}$ by the product rule, and $\frac{dF}{dx} \equiv f$, so that $f \cdot x = \frac{d(xF)}{dx} - F$ and

$$\begin{aligned}
\text{cov}(f, x) &= xF\Big|_0^1 - \int_0^1 F(x)dx - \tfrac{1}{2}\bar{f} && (6)\\
&= -\left(\int_0^1 F(x)dx - \tfrac{1}{2}\bar{f}\right) && (7)
\end{aligned}$$

But the quantity in parentheses is exactly the area labeled A in figure 5: the first term representing the total area between the x axis and the curve, and the second term removing the area of the triangle under the diagonal from $(0, 0)$ to $(1, \bar{f})$.

If we now calculate the covariance for the best case ordering $f^*(x)$ induced by $\mathcal{M}^*$, we similarly derive the area labeled A^* in figure 5, so that:

$$G_\mathcal{M} \equiv \frac{\text{cov}(f, x)}{\text{cov}(f^*, x)} \tag{8}$$

For example in the binary case,

$$G_\mathcal{M} = \frac{2\int_0^1 F(x)dx - \bar{f}}{\bar{f}(1-\bar{f})} \tag{9}$$

□

[3] Lerman & Yitzhaki (1984) independently establish a similar result for the traditional Gini coefficient of income inequality.

4. THE QINI COEFFICIENT

For the uplift case, we are interested in the strength of relationship between the uplift rate induced by an uplift model $\mathcal{M}$ and the targeting depth. Similarly to above, we define the point-wise outcome rates induced by $\mathcal{M}$, for both the treated and control populations, as $f_T(x)$ and $f_C(x)$ respectively. This is sketched in 6, with the uplift defined as

$$u(x) \triangleq f_T(x) - f_C(x) \tag{10}$$

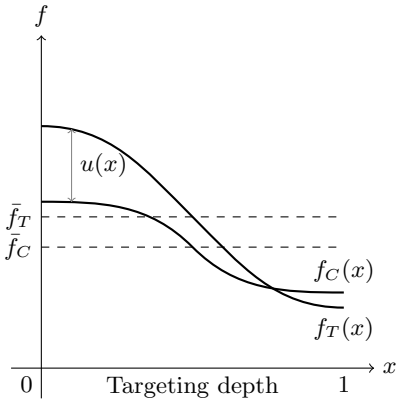

Figure 6: An uplift model induces a ranking on both the treated and control populations independently, resulting in the outcome rates f_T and f_C as a function of targeting depth (c.f. figure 2). We define uplift point-wise as the difference between the two rates.

As before, we further define the cumulative outcome and uplift rates shown in figure 7:

$$F_T(x) = \int_0^x f_T(y)dy \tag{11}$$

$$F_C(x) = \int_0^x f_C(y)dy \tag{12}$$

$$U(x) \triangleq F_T(x) - F_C(x) \tag{13}$$

Note also that $\bar{u} \equiv U(1)$ and $\frac{dU}{dx} \equiv u$.

THEOREM 2. *Radcliffe's Qini coefficients (Q, q_0) are simply normalized (anti-) covariances between uplift and treatment rate, specifically:*

$$Q \equiv \frac{\text{cov}(u, x)}{\text{cov}(u^*, x)} \tag{14}$$

$$q_0 \equiv \frac{\text{cov}(u, x)}{\text{cov}(u_0^*, x)} \tag{15}$$

where $u^*(x)$ is the uplift rate from the best potential model ranking in the binary case, and $u_0^*(x)$ is the uplift rate resulting from the best potential model ranking with zero downlift (see figures 10 and 12 respectively). In contrast to the Gini coefficient, these best potential ranking might be unachievable by any actual model, even with perfect information.

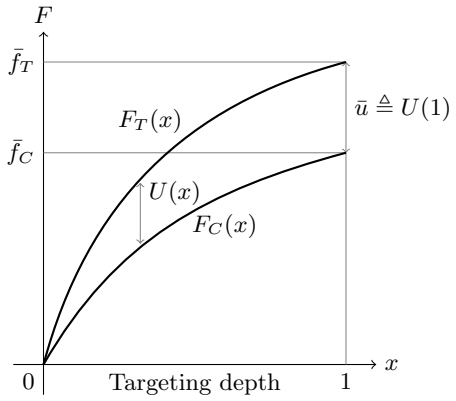

Figure 7: Analagously to figure 3, we define cumulative outcome rates F_T and F_C for the treated and control populations respectively, with the cumulative uplift $U(x)$ equal to their difference.

PROOF.

$$\begin{aligned}
\text{cov}(u, x) &= \int_0^1 (u - \bar{u})(x - \tfrac{1}{2})dx && (16)\\
&= \int_0^1 \left((f_T - f_C) - (\bar{f_T} - \bar{f_C})\right)(x - \tfrac{1}{2})dx && (17)\\
&= \int_0^1 (f_T - \bar{f_T})(x - \tfrac{1}{2})dx \\
&\quad - \int_0^1 (f_C - \bar{f_C})(x - \tfrac{1}{2})dx && (18)\\
&= \left(\int_0^1 F_T(x)dx - \tfrac{1}{2}\bar{f_T}\right) && (19)\\
&\quad - \left(\int_0^1 F_C(x)dx - \tfrac{1}{2}\bar{f_C}\right) \quad \text{(by (7))} && (20)\\
&= \int_0^1 U(x)dx - \tfrac{1}{2}\bar{u} && (21)
\end{aligned}$$

Now (21) corresponds exactly to Radcliffe's definition of Q, q_0 based on the area under the cumulative uplift curve (see figure 8), less the area under the diagonal: i.e. the average excess cumulative uplift.

It simply remains to normalize by the covariance for the "best case" model ordering. Radcliffe argues two different scenarios, for Q in the binary case and q_0 in the general case.

For Q, we define $u^*(x)$ by assuming the best possible assortment of the binary outcomes, where all positive treated outcomes occur first, and all positive control outcomes occur last. This is shown in figures 9 & 10. Of course, it might be that no model could achieve such an assortment even with perfect information: for example, no model can separate duplicate individuals whose outcome is independent of treatment. This is in constrast to the traditional binary outcome model, in which a perfect model is at least theoretically achievable.

There are various scenarios to consider for the ultimate shape of $u^*(x)$ (see Radcliffe, 2007 for details), but to illustrate us-

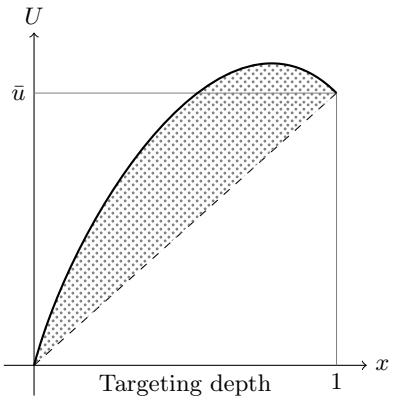

Figure 8: Radcliffe's cumulative uplift curve defining the Qini coefficients simply plots $U(x)$ as a function of targeting depth. This can be viewed as a Lorenz curve for the arbitrary real-valued outcome $u(x)$. The unscaled Qini value is defined as the average excess cumulative uplift (the shaded area).

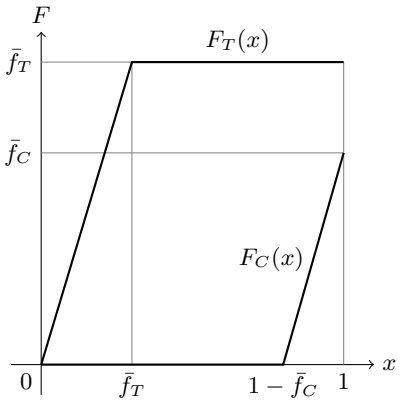

Figure 9: In the binary outcome case, the best model would rank treated individuals such that all positive outcomes appeared before any negative ones, and vice versa for control individuals. Unlike the traditional model, it might be that no uplift model could achieve such an ordering even with perfect information.

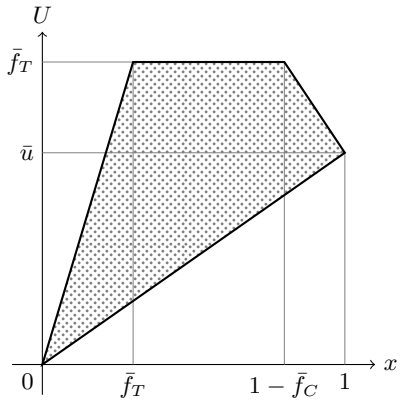

Figure 10: The best possible cumulative uplift curve in the binary case, $U^*(x) = \int_0^x u^*(y)dy$, is derived directly from figure 9. Note that there are several scenarios defining the extremes of the curve based on the relative sizes of f_T and f_C, with the version shown perhaps the most common.

ing the example sketched in figure 10 (where $\bar{f_T} < 1 - \bar{f_C}$ and $f_T > f_C$) we have

$$-\operatorname{cov}(u^*, x) = \bar{f_T}(1 - \bar{f_T})/2 + \bar{f_C}(1 - \bar{f_C})/2 \qquad (22)$$

Although Q as originally defined does not apply to continuous models, it appears possible to generalize by defining u^* based on a a Lorenz curve for the observed treated outcomes (sorted from best to worst) the reverse for the observed control outcomes (from worst to best). However this still seems likely to result in a very weak upper bound.

Because a value of $Q = 1$ is potentially not even theoretically achievable, and to cope with continuous outcomes, Radcliffe also defines q_0, based on a "best case" model with no negative uplift (no "downlift"). Here we define $u_0^*(x)$ using the sketches in figures 11 & 12. As for the Gini coefficient, we distinguish the binary and continuous outcome cases. In the binary example, we effectively push back the positive control outcomes as far as possible without creating negative uplift. There are a family of such scenarios in which $f_T(x) = 1$, $f_C(x) = 0$ for $x <= \bar{u} \triangleq \bar{f_T} - \bar{f_C}$ and $f_T(x) - f_C(x) = \bar{u}$ for $x > \bar{u}$ (see figure 11), all leading to the same $u_0^*(x)$ enclosing area

$$-\operatorname{cov}(u_0^*, x) = \bar{u}(1 - \bar{u})/2 \qquad (23)$$

In the continuous outcome case we allow all uplift to occur at rank 0, so that the relevant area is simply $\bar{u}/2$. □

The equivalence of (20) and (21) suggests that we can avoid the computational difficulties[4] of working directly with $U(x)$,

[4]Unlike traditional outcome rates, direct estimates of uplift rates are formed by subtracting quotients with different denominators. This means that uplift values are not intuitively additive. Even with normal random variations in treatment rate, different estimates can arise from (say) forward and backward accumulation. Indeed, if systematic

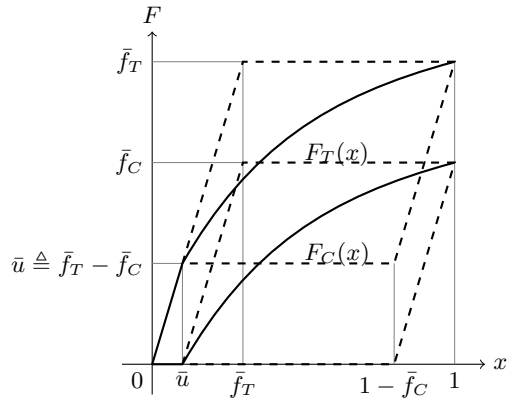

Figure 11: Both to cope with continuous outcomes, and because the optimum model assumed by Q typically hugely over-estimates what is practically achievable, the alternative q_0 is defined based on the best potential uplift ranking with no negative uplift. In the binary case sketched here, this is achieved by first ranking enough positive treated outcomes and negative treated outcomes to account for all uplift, and then balancing treated and control outcomes equally for the remainder of the curve (in one of a variety of possible ways). For the continuous case, we typically allow all uplift to occur at $x = 0$ with the outcomes matched thereafter. (Similary to figure 9, an alternative might be to allow all uplift to occur as rapidly as supported by the actual outcomes, sorting treated from best to worst and control from worst to best.)

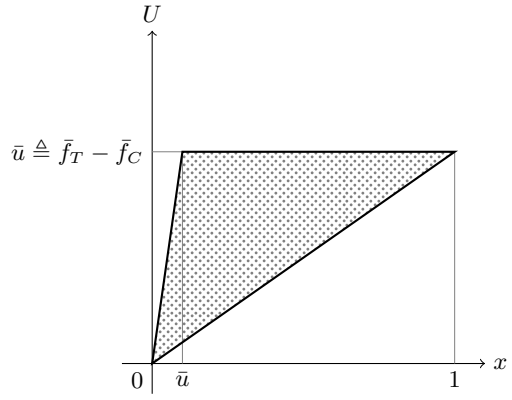

Figure 12: The best potential cumulative uplift curve with no downlift, $U_0^*(x) = \int_0^x u_0^*(y)dy$, as derived from figure 11. The binary case is illustrated, in which uplift can accumulate at a rate of at most one unit per individual.

the cumulative uplift (Radcliffe, 2007; TwoCitepRadcliffeSurry2011), and instead calculate the Gini-like coefficients for f_T and f_C independently (still ordered by uplift score). These can then be combined to achieve the desired Q or q_0 value. For example in the binary case with non-zero uplift, we can derive:

$$q_0 = \frac{\bar{f}_T(1-\bar{f}_T)G_T - \bar{f}_C(1-\bar{f}_C)G_C}{\bar{u}(1-\bar{u})} \tag{24}$$

Figure 7 also suggests a variant of the Kolmogorov-Smirnov statistic for uplift models. In traditional binary models, the K-S value is defined as largest difference in cumulative positive and negative outcome rates $\max_x F_+(x) - F_-(x)$. The corresponding targeting depth x^* identifies the cutoff that minimizes the sum of misclassification errors, $1 - F_+(x^*) + F_-(x^*)$.

Analogously, in the uplift case, we can calculate

$$\max_x F_T(x) - F_C(x) \equiv \max_x U(x) \tag{25}$$

which identifies the point of maximum excess cumulative uplift, the cutoff which minimizes the sum of positive control outcomes above the cutoff and negative treated outcomes below it, $1 - F_T(x^*) + F_C(x^*)$.

5. REFERENCES

T. Cai, L. Tian, P. H. Wong, and L. J. Wei, 2009. Analysis of randomized comparative clinical trial data for personalized treatment selections. Technical report, Harvard University Biostatistics Working Paper Series. Paper 97.

W. J. Conover and R. L. Iman, 1981. Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, 35(3):124–129.

R. J. Courtheoux, 2003. Modeling multi-channel response behavior. *Interactive Marketing*, 5(1).

C. Gini, 1912. Variabilità e mutabilità. In T. Pizetti E, Salvemini, editor, *Reprinted in Memorie di Metodologica Statistica*. Libreria Eredi Virgilio Veschi (Rome).

D. M. Green and J. M. Swets, 1966. *Signal detection theory and psychophysics*. John Wiley (New York).

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

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

B. Hansotia and B. Rukstales, 2002b. Incremental value modeling. *Journal of Interactive Marketing*, 16(3):35–46.

L. Y. Lai, 2004. Influential marketing: A new direct marketing strategy addressing the existence of voluntary buyers. Master's thesis, University of British Columbia.

K. Larsen, 2010. Net lift models. Presented at Predictive Analytics World (Washington, D.C.).

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

treatment bias is present, Simpson's Paradox can lead to arbitrarily large differences in estimates between seemingly equivalent calculation approaches.

M. O. Lorenz, 1905. Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, 9(70):209–219.

E. C. Malthouse, 2001. Assessing the performance of direct marketing scoring models. *Journal of Interactive Marketing*, 15(1):49–62.

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

G. Piatetsky-Shapiro and B. M. Masand, 1999. Estimating campaign benefits and modeling lift. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 185–193. ACM.

G. Piatetsky-Shapiro and S. Steingold, 2000. Measuring lift quality in database marketing. *ACM SIGKDD Explorations Newsletter*, 2(2):76–80.

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 and P. D. Surry, 2011. Uplift modelling: Theory and practice. Technical Report In Preparation, Stochastic Solutions & Portrait Software.

N. J. Radcliffe, 2004. Uplift modelling: Next generation targeting. Presented at OR46 Conference (York, UK).

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

P. Rzepakowski and S. Jaroszewicz, 2010. Decision tress for uplift modeling. In *Proceedings of the 2010 International Conference on Data Mining*, pages 441–450. IEEE Computer Society.

R. I. Yerman and S. Yitzhaki, 1984. A note on the calculation and interpretation of the Gini index. *Economics Letters*, 15(3-4):363–368.