***Keywords*** Uplift modeling · model evaluation · supervised learning · causal inference

# UPLIFT MODEL EVALUATION FOR RANDOMIZED CONTROL TRIALS

A PREPRINT

**Robert S. Yi**
Wayfair
4 Copley Place, Floor 7
Boston, MA 02116
robert@ryi.me

**William T. Frost**
Wayfair
4 Copley Place, Floor 7
Boston, MA 02116
frost.williamfrost@gmail.com

July 6, 2020

**ABSTRACT**

We provide an overview of methods to evaluate the effectiveness of models in finding heterogeneous treatment effects in randomized control trials ("uplift models"), and we introduce two novel evaluation curves: the adjusted Qini curve and the efficiency curve.

## 1 Introduction

Uplift models seek to estimate individual treatment effects, $\tau(\mathbf{x})$, defined as:

$$\tau(\mathbf{x}) = E[Y|\mathbf{X} = \mathbf{x}, W = 1] - E[Y|\mathbf{X} = \mathbf{x}, W = 0] \tag{1}$$

where $\mathbf{X}$ represents a feature vector, $\mathbf{x}$ represents a particular observation of values for this feature vector, $W$ represents a binary treatment vector $\in \{1, 0\}$, and $\tau$ the treatment effect (Holland, 1986; Rubin, 2005). An abundance of work has been done on the subject in the last decade (Radcliffe and Surry, 1999; Lo, 2002; Radcliffe and Surry, 2011; Surry and Radcliffe, 2011; Jaskowski and Jaroszewicz, 2012; Kane et al., 2014; Athey and Imbens, 2015; Guelman, 2015; Gutierrez and Gérardy, 2017; Künzel et al., 2019), but substantially less work has exhaustively covered techniques for proper evaluation of such models (Surry and Radcliffe, 2011). While we will give a brief introduction to the subject, our focus here will be on the latter. Typically, machine learning models are evaluated by measuring the error between the quantity modeled and the true, observed value. But in the case of uplift models, we cannot observe the true individual treatment effect $\tau_i$ (the fundamental problem of causal inference (Holland, 1986)) and therefore lack a source of truth to which to compare our predictions.

While we will discuss a clever object that circumvents this problem (the Transformed Outcome object), it is actually not always in the best interest of practitioners to prioritize reduction of predictive error. In practice, the problem to be solved is typically not "what is the individual treatment effect", but rather "who should we target with our treatment", and the machine leraning problem quickly becomes one of identification rather than measurement.

For example, an uplift model might be used during a direct mail campaign to find a subset of individuals to target that are particularly incremental to direct mail flyers. In a case like this, the model's ability to identify the highly incremental individuals is more important than that model's ability to correctly predict individual treatment effects. On the other hand, if this ad campaign were through online display ad vendors where a precise valuation is required to set an auction bid, an accurate estimate is paramount. Although these two concerns should theoretically be one and the same, in practice, some care is required to ensure that both are addressed (Athey and Imbens, 2016). And for each aim, different evaluation schemes are warranted.

In what follows, we will discuss common ways of evaluating a model's ability to rank customers according to $\tau$, briefly discuss estimation of $\tau$, and introduce some novel methods for special use cases.

## 2 Measuring error using the transformed outcome

The Transformed Outcome, introduced by Athey and Imbens (2015), provides a clever solution to the fundamental problem of causal inference by simply transforming outcomes according to a treatment label so that the transformed quantity, in expectation, represents lift. The transformed outcome is defined as

$$Y_i^* = Y_i \frac{W_i - p}{p(1-p)}, \tag{2}$$

where $p$ is the probability of receiving a treatment, $Y_i$ is the outcome label, and $W_i$ is a binary treatment vector $\in \{1, 0\}$. It is simple to show that $E[Y_i^*] = \tau_i$, where $\tau_i$ is the true heterogeneous treatment effect:

$$E[Y_i^*|\mathbf{x}] = \frac{Y_i}{p} P(Y = 1|\mathbf{x}, W = 1) P(W = 1|\mathbf{x}) \tag{3}$$

$$+ \frac{-Y_i}{1-p} P(Y = 1|\mathbf{x}, W = 0) P(W = 0|\mathbf{x}) \tag{4}$$

$$= Y_i(P(Y = 1|\mathbf{x}, W = 1) - P(Y = 1|\mathbf{x}, W = 0)) \tag{5}$$

Moreover, $Y_i^* = \tau_i + \nu_i$, where $\nu_i$ is some error, which, in the case of a randomized trial, should be orthogonal to any features $X_i$. $E[\nu_i|X_i]$ should therefore always be 0. The mean squared error between a predicted quantity and the transformed outcome can then be shown to be equivalent to the mean squared error between the prediction and the true individual treatment effect (following Hitsch and Misra (2018)):

$$E[(Y_i^* - \hat{\tau}_i)^2|X_i] = E[(\tau_i + \nu_i - \hat{\tau}_i)^2|X_i] \tag{6}$$

$$= E[(\tau_i - \hat{\tau})^2|X_i] + E[\nu_i|X_i]E[2(\tau_i - \hat{\tau}_i)] + E[\nu_i^2|X_i] \tag{7}$$

$$= E[(\tau_i - \hat{\tau})^2|X_i] + E[\nu_i^2] \tag{8}$$

The mean squared error between any prediction and the transformed outcome, therefore, can be used to represent the prediction error plus a residual term $E[\nu_i]^2$, which is independent of $X_i$. A convenient consequence of this equality is that a regression model can be trained towards the transformed outcome and, by minimizing MSE, result in predictions that scale correctly with the true treatment effect, $\tau$.

## 3 Evaluation curves

Often, a single evaluative number (like error against the transformed outcome) is not sufficient for decision-making. We generally care not about how accurately we predict the individual treatment effect, but rather how well our prediction separates positively incremental subjects ("persuadables") from unresponsive subjects ("sure things" or "lost causes"), or worse, negatively incremental subjects ("sleeping dogs"). It is often therefore instructive to obtain an estimate of the number of conversions, the revenue, or the efficiency a particular targeting policy would yield. Practitioners frequently turn to methods that order the individuals of an independent validation population according to the model's predictions, then calculate the treatment effects within groups binned by this ranking. We will describe these methods in the sections that follow. We first define here some common variables that we will use:

- $\phi$: the fractional position in a list of customers ranked by the model prediction; low $\phi$ indicates high predicted value. We also use $\phi$ as shorthand to represent the set of individuals between ranking 0 and $\phi$. $1 \setminus \phi$ therefore indicates the complement group of individuals from $\phi$ to 1.
- $n_{t,1}(\phi)$ and $n_{c,1}(\phi)$: the cumulative number of purchases in the treatment group and the control group in the set $\phi$, respectively.
- $n_t(\phi)$ and $n_c(\phi)$: the cumulative number of individuals (regardless of whether they purchase or not) in the treatment group and the control group up to $\phi$, respectively.
- $N_t$ and $N_c$: be the total counts of subjects in the treated and untreated groups, respectively.

### 3.1 Qini-style curves

A common approach to evaluating uplift models are Qini-style curves, which simply count the number of incremental conversions obtained as a function of individuals targeted (Radcliffe, 2007; Surry and Radcliffe, 2011). The word "Qini" was coined by (Radcliffe, 2007) and is derived from the concept of Gini coefficient, which employs a Lorenz

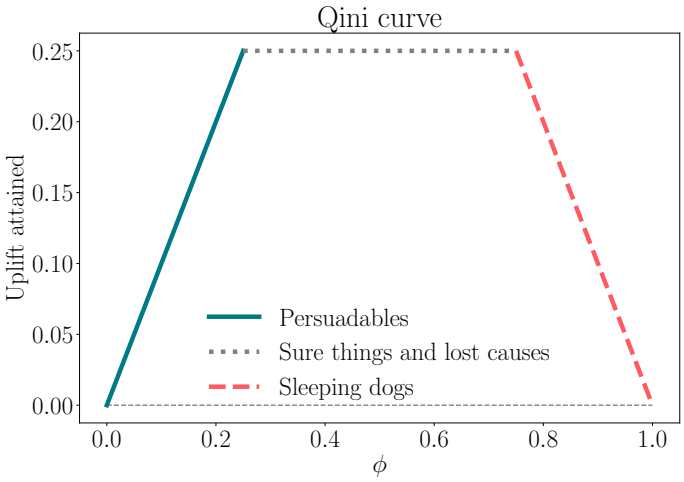


Figure 1: A Qini curve. 8 individuals are ranked above as follows: 2 persuadables, 2 sure things, 2 lost causes, and 2 sleeping dogs. Each pair contains one individual from the treatment group and one from the control group. In a scenario where treatment and control are evenly balanced, all evaluation curves should be identical.

curve to represent income inequality Lorenz (1905). We will briefly discuss three metrics below that offer different normalizations of this type of curve. All three curves are identical when $n_t(\phi) = n_c(\phi)$, and an example of such a curve is shown in Figure 1. But because this assumption breaks down when models find pockets of treatment imbalance, we suggest adopting one of the adjustments described in sections 3.1.2 or 3.1.3. Note that, while we discuss these curves under the assumption that the numerators represent counts, these numerators can be replaced with continuous values, such as revenue or margin.

Code for generating these curves in python can be found at `https://github.com/df-foundation/pylift`.

### 3.1.1 The Qini curve

The canonical Qini-style curve used to evaluate uplift models is known simply as the Qini curve Radcliffe (2007), measured as:

$$\text{Qini}(\phi) = n_{t,1}(\phi) - \frac{n_{c,1}(\phi)N_t}{N_c} \tag{9}$$

This is simply a measure of incremental purchases, where control group conversions are rebalanced according to the treatment/control split across the entire population being evaluated. This is often normalized to express a "net lift" rather than a direct count Guelman (2015) (although the two are visually equivalent):

$$\text{Qini'}(\phi) = \frac{n_{t,1}(\phi)}{N_t} - \frac{n_{c,1}(\phi)}{N_c} \tag{10}$$

While this curve is generally a reasonable evaluation tool, we note that if a model were to somehow unevenly differentiate those in the treatment group from the control group, these metrics could be deceptively inflated (more likely) or deflated.

### 3.1.2 The cumulative gain chart

To address uneven treatment/control splits in the ordering, the Qini curve can be normalized as follows (Gutierrez and Gérardy, 2017):

$$g(\phi) = \left(\frac{n_{t,1}(\phi)}{n_t(\phi)} - \frac{n_{c,1}(\phi)}{n_c(\phi)}\right)(n_t(\phi) + n_c(\phi))\,. \tag{11}$$

While we generally prefer this normalization, we note that in a situation where there is an uneven treatment/control split, one could observe an increase in the y-axis value of a Qini curve, even when no more treatment group successes

are being captured. When viewing the Qini curve's y-axis as the number of incremental conversions that have been obtained, this can be a bit odd. A more detailed discussion can be found in Appendix A.

#### 3.1.3 The adjusted Qini curve

To address this concern, we introduce another curve, normalized slightly differently, which we coin the adjusted Qini curve:

$$\text{Adjusted Qini}(\phi) = n_{t,1}(\phi) - \frac{n_{c,1}(\phi)n_t(\phi)}{n_c(\phi)}. \tag{12}$$

This curve is less susceptible to local fluctuations in $n_t(\phi)$ and $n_c(\phi)$ than the traditional Qini curve, avoids the odd quirk of the cumulative gain chart, while still converging to the Qini curve in the limit $n_t(\phi) \to n_c(\phi)$. It also only permits y-values to increase when there are successes in the treatment group. As with previous curves, this can be normalized to reflect net lift rather than net purchases as follows:

$$\text{Adjusted Qini'}(\phi) = \frac{n_{t,1}(\phi)}{N_t} - \frac{n_{c,1}(\phi)n_t(\phi)}{n_c(\phi)N_t}. \tag{13}$$

#### 3.1.4 The adjusted Qini vs. the cumulative gains curve

We concede that the difference between these curves (particularly between the adjusted Qini curve and the cumulative gains curve) is subtle. So to clarify the assumptions behind each, we consider a toy problem. Suppose we have ranked customers into two groups: 1 and 2. In Group 1, suppose the treated individuals have a 100% success rate, while untreated individuals have a 50% success rate. In Group 2, suppose the success rate is 50% for both the treatment and control populations.

Group 1 has a positive lift, while group 2 has zero lift. Now suppose group 1 has a treatment imbalance. For ease of calculation, suppose 75% of the group is in the treatment group, and only 25% is in the control group. Although there is an imbalance, if we had a model thac correctly ranked Group 1 first and Group 2 second, we'd expect any Qini-style curve to reach its maximum at the end of group 1.

In our common notation, we could, for illustration's sake, assign the following counts for Group 1:

$$n_{t,1}(\phi = 0.5) = 75$$
$$n_t(\phi = 0.5) = 75$$
$$n_{c,1}(\phi = 0.5) = 12.5$$
$$n_c(\phi = 0.5) = 25$$

And for both group 1 and group 2:

$$n_{t,1}(\phi = 1) = 75 + 25 = 100$$
$$n_t(\phi = 1) = 75 + 50 = 125$$
$$n_{c,1}(\phi = 1) = 12.5 + 25 = 37.5$$
$$n_c(\phi = 1) = 25 + 50 = 75$$

Because of the small control group, the Qini curve will appear artificially high for group 1. But on the other hand, the cumulative gains chart will be $0.25$ at the end of the first group, but at the end of the second, $0.3$. There appears to be lift, even though we are simply adding lost causes + sure things into the mix. This is shown in Figure 2.

The adjusted Qini curve counteracts this anomaly, by allowing increases to the y-axis to only come from the treatment group. The adjusted Qini curve assumes that the number of successes obtained in a bin are an accurate reflection of the number of successes we would obtain, regardless of the treatment/control split. Similarly, if there is a proportionally large increase in control group members in a bin, the adjusted Qini curve will not increase within this bin, while the cumulative gains curve might. In cases where the local maxima of the Qini-style curve is used to determine a cutoff for targeting, the adjusted Qini will generally give a more conservative cutoff, which may be preferable in the presence of budget constraints.

On the other hand, if you believe that the overall lift up to $\phi$ is a better representation of the population up to $\phi$, then the cumulative gains curve would be preferable. The curve increases because, although there may not be more conversions, we can guess that the new control group members would have converted at the average rate, had they been treated.

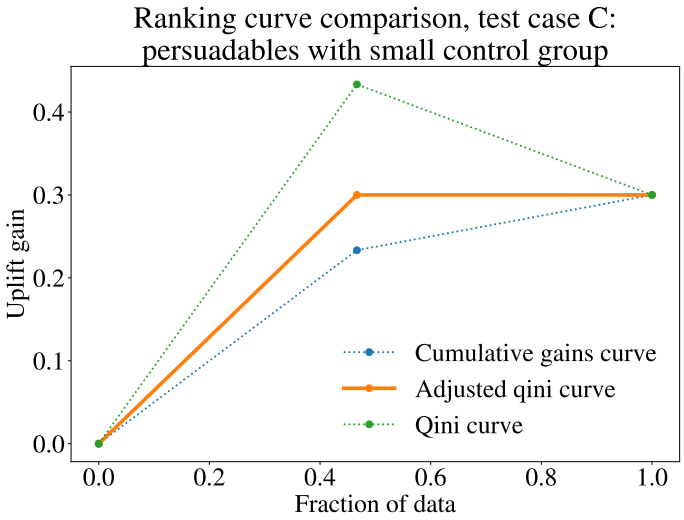


Figure 2: An illustration of how the adjusted Qini curve might produce the most accurate results. In this example, the first non-zero point at $x = 0.5$ contains only incremental individuals, while the second non-zero point at $x = 1$ contains only non-incremental individuals. Because of a treatment/control group size imbalance, the Qini curve overpredicts the lift, while the cumulative gains curve underpredicts it. The adjusted Qini curve produces an accurate result.

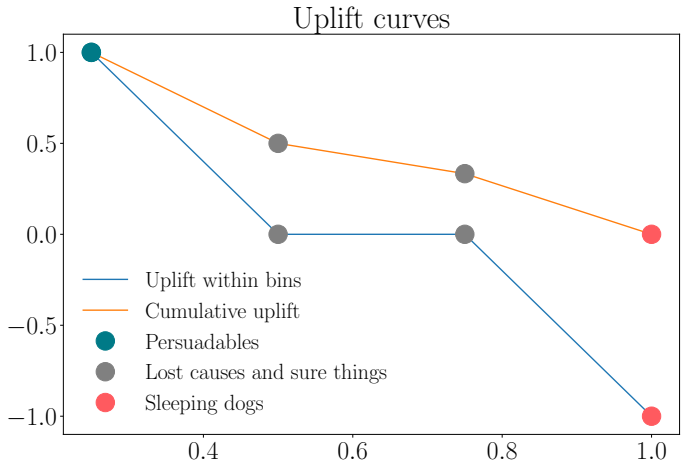


Figure 3: Uplift and cumulative uplift curves.

### 3.2 Uplift curve

It can sometimes be instructive to look at lift within each bin (rather than cumulatively) to measure the predicted incrementality of subjects by quantile. This curve can be calculated as:

$$\text{Uplift}(\phi) = \frac{\Delta n_{t,1}(\phi)}{\Delta n_t(\phi)} - \frac{\Delta n_{c,1}(\phi)}{\Delta n_c(\phi)}, \tag{14}$$

where the $\Delta$ indicates that we are no longer taking counts up to $\phi$, but rather, within a small slice of $\phi$. A curve of this type is shown in Figure 3.

### 3.3 Cumulative uplift curves

Similarly, the lift can also be measured up to each bin, but without scaling by the fraction of the population targeted to get a sense of, for example, how incremental a targeted population would be when you set your policy to target up to a

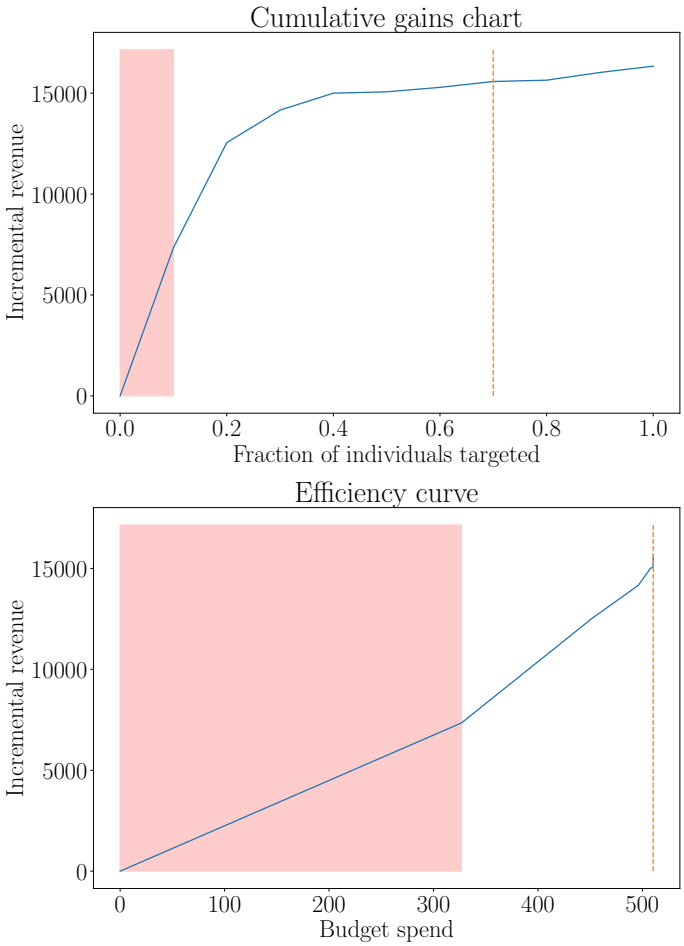


Figure 4: A (a) cumulative gains curve with the y-axis indicating incremental revenue, and (b) the corresponding efficiency curve, where the $x$-axis has been scaled by the model predictions. The orange vertical line indicates the same point with respect to $\phi$ (because the efficiency curve is scaled by the prediction, the individuals with $\hat{\tau} \leq 0$ take up no space on the x-axis). The shaded red region indicates a lower efficiency area – notice the slope of the efficiency curve is lower as a result of the model predictions being too high for these individuals.

threshold. This is defined as:

$$\text{Cumulative uplift}(\phi) = \frac{n_{t,1}(\phi)}{n_t(\phi)} - \frac{n_{c,1}(\phi)}{n_c(\phi)}. \tag{15}$$

An example of this curve can be seen in Figure 3.

### 3.4 Efficiency curve

We now introduce another curve that is primarily useful in scenarios where the uplift model being evaluated is used to generate a bid. This paradigm is common in advertising, where there are frequently second-price auctions that occur to determine who, of the parties interested in advertising, should win an ad spot. Because the optimal strategy in these auctions tends to be truthful bidding of one's estimated value, accurate estimation of the treatment effect can be as important as the ranking by incremental value of ad recipients. In this case, it can be useful to adjust the cumulative gains chart by scaling $\phi$ by the sum of predicted values $\hat{\tau}(\phi)$.

For convenience, we coin this curve the "efficiency curve". The $y$ axis of this curve still represents incremental value, but the $x$ axis now represents amount spent. In the scenario where individual treatment effect is precisely predicted, this curve should simply be a straight line with slope 1. In practice, the slope of this line can be used to indicate the expected efficiency $\varepsilon$ of the model targeting policy within this region. The ultimate targeting policy, then, can be adjusted accordingly. An example of this trade-off is shown in Figure 4.

### 3.5 Expected response curve ("Modified Uplift Curve")

The curves we have discussed so far place incremental responses within bins of ranked individuals on the y-axis, but we could alternatively place on the y-axis the total number of responses if we target up to $\phi$, following Zhao et al. (2017) (they term this the "Modified Uplift Curve"). That is, within each percentage $\phi$, we can calculate the expected response, defined as:

$$\max_{t=1,\cdot,K} E[Y|\text{target up to } \phi, T=t] \tag{16}$$

where $T$ indicates the treatment shown, $t$ is one of a number of possible $K$ treatments. This curve calculates the expected response if all individuals in a subset $\phi$ are targeted with the optimal treatment (determined by a model) and the remaining $1-\phi$ individuals are not targeted. The $y$-axis then reflects the total number of responses from both groups. In the simple case where $K=1$, there's an even, and there is no heterogeneity in the probability of treatment with $\phi$ (i.e. $P(W=1|\phi) \simeq P(W=1)$), the y-axis values can be written in our notation as follows:

$$\text{Expected Response} \sim n_{t,1}(\phi) + n_{c,1}(1 \setminus \phi) \tag{17}$$

A clear benefit of using this curve is that, when comparing different models, we can measure the value of our targeting policy in terms of the performance of a metric, overall, rather than simply the incremental value. Any bootstrapped error, then, should closely resemble the statistics one would expect from running an A/B test of two targeting policies against one another.

Equation 17 should be corrected to account for heterogeneities in $n_t(\phi)$ and $n_c(\phi)$, then normalized as follows;

$$\text{Expected Response} = \frac{n_{t,1}(\phi)}{n_t(\phi)}\phi + \frac{n_{c,1}(1 \setminus \phi)}{n_c(1 \setminus \phi)}(1-\phi) \tag{18}$$

This normalization assumes that the response rate in $\phi$ in the treatment group (and $1 \setminus \phi$ in the control group) reflects the response rates that would be attained in $\phi$ (and $1 \setminus \phi$) had the entire group been treated.

### 3.6 Adjusting for a heterogeneous treatment policy

While we have assumed a heterogeneous treatment policy, this is commonly not the case. To adjust for this, each individual response (before counting towards the Qini-style curves) should be scaled by the inverse of the individual treatment propensity $p_i$, such that our counting functions are redefined as follows:

$$\begin{aligned}
n_{t,1}(\phi) &= \sum_{i \in \phi} \frac{y_i}{2p_i} \\
n_{c,1}(\phi) &= \sum_{i \in \phi} \frac{y_i}{2(1-p_i)} \\
n_t(\phi) &= \sum_{i \in \phi} \frac{1}{2p_i} \\
n_c(\phi) &= \sum_{i \in \phi} \frac{1}{2(1-p_i)}.
\end{aligned}$$

Moreover, each individual's contribution to $\phi$ should also be scaled by $\frac{1}{2p_i}$ or $\frac{1}{2(1-p_i)}$ in treatment or control, respectively, such that $\phi$ faithfully represents the original population.

A similar adjustment can be applied to $N_t$ and $N_c$ to correct the original Qini curve, but because problems with this curve will be exacerbated with the adoption of a non-uniform treatment policy, we highly discourage its use in such cases.

### 3.7 Qini values and the ideal Qini curves

Optimal Qini-style curves provide both a visual metric of comparison as well as a means of normalizing AUC metrics for Qini-style curves ("Qini measures" (Radcliffe, 2007)). However, optimal curves can be calculated under different assumptions. In the following few sections, we discuss these different curves and their assumptions. Moreover, these vary depending on which Qini-style curve is being used. The second and third curves listed below are the same for all three types of curves listed above, while the first only only exists for the Qini curve.

To foster a more intuitive discussion, we adopt the following jargon in our discussion below:

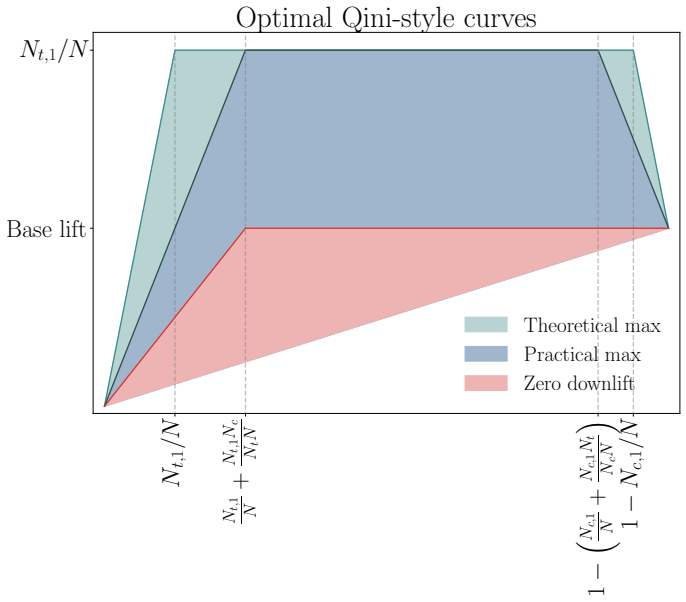


Figure 5: Optimal Qini-style curves.

- **Persuadable**: an individual who only responds when treated.
- **Sure thing**: an individual who responds, regardless of treatment.
- **Lost cause**: an individual who does not respond, regardless of treatment.
- **Sleeping dogs**: an individual who only responds when untreated.

#### 3.7.1 The theoretical maximum.

**Assumption:** *an individual is a "persuadable" if and only if* $(Y = 1, W = 1)$*, and an individual is a "sleeping dog" if and only if* $(Y = 1, W = 0)$.

The curve with the highest possible AUC can be obtained by assuming that the number of positive outcome cases in the treatment group are the only individuals who are "persuadables." The curve begins with a line that increases at a slope of 2, up to the point $(\frac{N_{t,1}}{N}, \frac{N_{t,1}}{N_t})$, where $N_{t,1}$ is the number of positive cases in the treatment group and $N$ is the total number of individuals. The curve then maintains a slope of 0 until the point $\phi = 1 - N_{c,1}/N$, at which point it decreases at a slope of $-2$ until $\phi = 1$ and the Qini value lowers to the base lift of the treatment, $\frac{N_{t,1}}{N_t} - \frac{N_{c,1}}{N_c}$.

Note that this curve is not well-defined for the adjusted Qini and cumulative gains curve, because of the $n_c\phi$ term in the denominator of a term in both equations. If we, however, were to assume that $n_c\phi$ was instead a small but negligible value, we would find that this ordering for the cumulative gains curve would produce a slope of 1, while for the adjusted qini curve, would have a slope of 2, taking on the same shape as the optimal Qini curve.

#### 3.7.2 The practical maximum

**Assumption:** *all* $(Y = 1, W = 1)$ *are "persuadables," all* $(Y = 1, W = 0)$ *are "sleeping dogs", and proportionate numbers of persuadables and sleeping dogs exist in* $(Y = 0, W = 0)$ *and* $(Y = 0, W = 1)$*, respectively (the practical maximum).*

This curve has the more restrictive assumption that each persuadable in the treatment group $(Y = 1, W = 1)$ must correspond to a persuadable in the control group $(Y = 0, W = 0)$. This curve is used in Radcliffe (2007). The same assumption can then be applied to the sleeping dogs, where $(Y = 0, W = 1)$ and $(Y = 1, W = 0)$. The curve therefore begins with a line that increases at a slope of 1, up to the point $(\frac{N_{t,1}}{N} + \frac{N_c}{N}\frac{N_{t,1}}{N_t}, \frac{N_{t,1}}{N_t})$. As with 3.7.1, the line then maintains a slope of 0 until the point $\phi = 1 - \frac{N_{c,1}}{N} - \frac{N_t}{N}\frac{N_{c,1}}{N_c}$, at which point it decreases with a slope of $-1$ until $\phi = 1$ and the Qini, once again, lowers to the base lift of the treatment.

#### 3.7.3 Zero-downlift curve

**Assumption:** *no negative effects (no "sleeping dogs").*

If we have no negative effects (or in our jargon, no "sleeping dogs"), the optimal Qini-style curve should never have a negative slope and is therefore capped by the base lift observed over the entire population. (Radcliffe, 2007) suggest using the area between this curve and the random targeting line to normalize $Q$, the area between the observed cumulative gains curve and the random targeting line, as it provides a more interpretable value than the theoretical or practical maxima, which can be orders of magnitude larger than the best possible Qini-style curve. They term this normalized AUC $q_0$.

While, in our estimation, this normalization produces the most intuitive metric, it behaves poorly if negative effects are expected and the base lift is, therefore, small.

## 4 Conclusion

We have provided an overview of methods that can be used to evaluate uplift models. As causal methods begin to garner wider adoption, uplift models are being increasingly used to leverage heterogeneities in experimental data. We hope our work here provides a step forward in making the evaluation of these models rigorous and explicit.

## References


Susan Athey and Guido Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113, 27 (2016), 7353–7360.

Susan Athey and Guido W Imbens. 2015. Machine learning methods for estimating heterogeneous causal effects. *stat* 1050, 5 (2015).

Leo Guelman. 2015. Optimal personalized treatment learning models with insurance applications. (2015).

Pierre Gutierrez and Jean-Yves Gérardy. 2017. Causal Inference and Uplift Modelling: A Review of the Literature. In *International Conference on Predictive Applications and APIs*. 1–13.

Günter J Hitsch and Sanjog Misra. 2018. Heterogeneous treatment effects and optimal targeting policy evaluation. *Available at SSRN 3111957* (2018).

Paul W Holland. 1986. Statistics and causal inference. *Journal of the American statistical Association* 81, 396 (1986), 945–960.

Maciej Jaskowski and Szymon Jaroszewicz. 2012. Uplift modeling for clinical trial data. In *ICML Workshop on Clinical Data Analysis*.

Kathleen Kane, Victor SY Lo, and Jane Zheng. 2014. Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods. *Journal of Marketing Analytics* 2, 4 (2014), 218–238.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences* 116, 10 (2019), 4156–4165.

Victor SY Lo. 2002. The true lift model: a novel data mining approach to response modeling in database marketing. *ACM SIGKDD Explorations Newsletter* 4, 2 (2002), 78–86.

Max O Lorenz. 1905. Methods of measuring the concentration of wealth. *Publications of the American statistical association* 9, 70 (1905), 209–219.

Nicholas Radcliffe. 2007. Using control groups to target on predicted lift: Building and assessing uplift model. *Direct Marketing Analytics Journal* (2007), 14–21.

Nicholas J Radcliffe and Patrick D Surry. 1999. Differential response analysis: Modeling true response by isolating the effect of a single action. *Credit Scoring and Credit Control VI. Edinburgh, Scotland* (1999).

Nicholas J Radcliffe and Patrick D Surry. 2011. Real-world uplift modelling with significance-based uplift trees. *White Paper TR-2011-1, Stochastic Solutions* (2011).

Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* 100, 469 (2005), 322–331.

Patrick D Surry and Nicholas J Radcliffe. 2011. Quality measures for uplift models. *submitted to KDD2011* (2011).

Yan Zhao, Xiao Fang, and David Simchi-Levi. 2017. A practically competitive and provably consistent algorithm for uplift modeling. In *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1171–1176.

# A A quick example demonstrating why the cumulative gains curve may increase when the uplift does not

We present a simple example to illustrate why the cumulative gains curve might be problematic when choosing a targeting threshold using a local maxima. Consider a model which ranks a number of treated customers first $\phi_0$ that purchase at some rate $p_t < 1$. Following them in the ranked list of predictions are some control group individuals that purchase at some other rate $p_c < 1$. Then for $\phi > \phi_0$, we can write

$$g(\phi) = (p_t\phi_0 - p_c(\phi - \phi_0))\phi. \tag{19}$$

This will be a concave down parabola with a local maxima at

$$\phi = \frac{(p_t + p_c)\phi_0}{2p_c}. \tag{20}$$

As long as $\frac{p_t+p_c}{2p_c} > 1$, or $p_t > p_c$, a local maxima will occur at $\phi > 0$, even though, by construction, the lift is not actually increasing. Thus, while this metric does diminish the false signal associated with an uneven treatment/control division with $\phi$ in the Qini curve, it may put the local maxima at a point where we are not gaining any more information about the treatment group. In plain terms, if we view the y-axis of the Qini-style curves as the total number of conversions captured by targeting a fraction $\phi$ of individuals, then the cumulative gains curve might erroneously suggest that we are still obtaining more conversions, as in the above example, even when we are not.