Université de Montréal

Uplift modelling and performance criteria assessment

Par

Anas Kobbi

Département des sciences de la décision

HEC Montréal

Mémoire présenté à HEC Montréal
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en Intelligence d'affaires

Avril 2018

RÉSUMÉ

La modélisation incrémentale est une méthode élégante permettant de cibler des individus en fonction de leur probabilité de répondre positivement à un traitement (ou offre promotionnelle), sachant qu'ils sont exposés au dit-traitement. Cette méthode implique de séparer la population en deux groupes, traitement et contrôle, permettant d'isoler efficacement l'effet du traitement. Cependant cette méthodologie implique aussi l'exclusion de mesures d'évaluation de la performance du modèle telles que le taux de bonnes classifications ou la courbe ROC, ne pouvant pas être utilisées simplement parce qu'un individu ne peut être assigné aux deux groupes en même temps. Bien que plusieurs critères d'évaluation du modèle aient été suggérés, leur performance réelle dans l'évaluation d'un modèle demeure inconnue. Bien que l'efficacité de la méthode incrémentale est davantage illustrée dans cette étude, les résultats proposés démontrent explicitement que Qini et q0, une mesure dérivée de Qini, demeurent robustes dans diverses conditions de simulation des données. Nos résultats lèvent aussi le voile sur d'autres critères moins connus, à savoir la mesure de répétabilité R^2 et le critère de Tau, tous deux montrant une bonne fiabilité mais spécialement dans des conditions de simulation favorable où la complexité des jeux de données est diminuée.

Mots-clés: uplift, qini, valeur incrémentale, true lift

ABSTRACT

Uplift modeling is an elegant method used to target individuals based on their likelihood to respond positively to a treatment (i.e. promotional offer), given they received the treatment. However, as this methodology involves splitting the population into treatment and control groups, traditional metrics such as misclassification rate or ROC curve cannot be used to assess the model's predictive efficiency as an individual can only be assigned to one of the groups. While several model assessment criteria have been suggested, their performance has been poorly discussed. While the uplift method is further illustrated in this study, the proposed results demonstrate that Qini's and q0, a Qini-related metric, are robust and reliable under various simulated conditions. Our results additionally investigate more novel criteria, namely R^2 and Tau, showing great results under specific favorable settings.

Keywords: uplift modelling, qini, incremental model, true lift

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

AKNOWLEDGEMENT

First, I would like to express my heartfelt gratitude to my advisor Prof. Denis Larocque for his continuous support of my M.Sc. studies and research. His guidance was exceptionally valuable and allowed me to tackle this challenging research subject with sharpness and efficiency. Denis shared with me his world-class expertise and rigorous methodology all while being a great listener, values I will keep close for the upcoming years. I could simply not imagine conducting this research without Denis' precious support and insights.

I would like to thank my family who have had my back since the very beginning, displaying their unconditional support both through sleepless nights and times of success. Special thanks to my father Riadh for his patience, motivation, encouragements and especially colossal knowledge of the data science industry.

I also thank Prof. François Bellavance for enlightening me with the first sparkles of passion towards the uplift modelling, eventually building up to this research.

Finally, I would also like to thank the members of the jury for their time.

INTRODUCTION

The advent of e-commerce and internet purchasing has grown so popular in the last decades that it has turned the retail industry into an even more competitive landscape than ever before (Heinemann & Schwarzl, 2010). Often described as a seismic shift, it has turned merchants' budgeting into an even more difficult discipline – all while growing customer approach’s complexity (Hughes & al, 2009). It is now clear that direct marketing is under overwhelming pressure to deliver excellent results while minimizing vendor expenses and losses (Gregory, 2015).

The promotion of new products through email is nowadays a common marketing action for most retailers (Hughes & al, 2009). Such promotional campaigns usually aim to maximize the gain from this marketing action. While sending promotional e-mails is usually a low-cost marketing tool, it is in most cases proscribed to reach out to the entire customer’s base for obvious reasons, as it often results into spending resources on uninterested or churned customers (Hughes & al, 2009). A wiser and more modern approach is to profile the customer base beforehand in order to break the targeting down to individuals who are more likely to respond positively to this marketing action. Following the application of business rules, the marketing action (i.e. sending a personalized email in this case) is carried out on a selected number of customers. A certain increment or benefit from this action is subsequently measured. As Rukstales & Hansotia (2002) pointed out, database marketing traditionally targets the optimization of said response rate. Such models should allow the retailer to exclusively target customers who are likely to respond positively to the promotion. Yet, this is almost never the case.

CONVENTIONAL METHODOLOGY DRAWBACK

It has been demonstrated in the literature that the application of the conventional methodology involves several flaws. Further continuing with the promotional e-mail example, a trendy retailer sends a personalized email to a potential customer interested in a stylish pair of boots. The customer clicks on the web link attached to the email, shows genuine interest in the product and considers the discounted price at display. However, the customer does not buy the product immediately from the website but rather goes in store a couple days later. Does this purchase count as revenue measured by email, given that the purchase was not completed online? Another example would be promos in grocery stores. A mother of three does her grocery regularly at the same grocer twice a week, for convenience and proximity purposes. Every visit implies a 2-digits bill and half a dozen articles, among which the usual bread, milk and a can of tuna. The grocer targets this client through a promo email announcing a sale on the tuna. Of course, as every other visit, this customer grabs the tuna and throws it in her basket. Is this purchase attributable to the email sent to the customer, even if the customer would have bought the article regardless of the offer?

These two examples illustrate some of the shortcomings of the traditional method. Specifically, the major drawback of models optimizing the response rate is that they include in practice clients that are not necessarily interested in the promotional campaign. These models fail to make the crucial distinction between customers who:

1. Make a purchase with the promotion.
2. Make a purchase only because of the promotion.

The first category of clients has no real value to the campaign. As mentioned, there are customers who would buy regardless of whether there is a promotion or not. Hansotia & Rusktales (2002) made a valid point by demonstrating that this group of customers is included in most campaigns, regardless of their lack of incremental profit. More precisely, all potential customers targeted by a marketing action can be split as follows:

		Response if treated	
		Yes	No
Response if not treated	Yes	**Spontaneous response**	**To avoid**
	No	**Target**	**No effect**

Table 1. Customers distribution based on treatment and response variables (Kondareddy et al., 2016).

Even if they embody optimal clients, individuals who respond regardless if they are treated or not have no incremental value, and thus logically do not require targeting. On the other hand, individuals who do not respond to targeting are of even less interest, as they decrease the lift on the increment. Finally, individuals who respond positively when untreated but negatively when treated, although few in numbers, are extremely important in retention activities (Guelman et al., 2015). Several studies show cases where incremental value modeling has transformed conventional retention failing campaigns into blatant success (Kane et al., 2014; Siegel, 2013). It is therefore clear that the target category regroups customers who respond only when reached out. By avoiding other potential client categories through better targeting, it is both possible to improve incremental revenue while minimizing campaign costs. This is the main purpose of uplift modeling.

GENERAL UPLIFT METHODOLOGY

For more than half a century, this methodology has been used in clinical trials and drug development studies (Jackson et al, 2011). In the context of health sciences, this technique historically made possible the measurement of actual benefit from the treatment by eliminating most forms of methodological bias, such as the common placebo effect (Weisberg & Pontes, 2015). The first publications describing the application of uplift processes in a business context appeared in late 90’s. The first method suggested by Radcliffe & Surry (1999) was to determine the clients who are most likely to respond positively according to their characteristics. The major innovation of the proposed methodology was splitting the target sample into two categories: treatment and control. The treatment group thus gathered targeted individuals who will receive the marketing action (i.e. e-mail, phone call, etc.). On the other hand, the control group will not receive the treatment but will nevertheless be monitored for analytical purposes. For example, the retailer might collect data on the individuals by measuring their generated income, the overall customer behavior and so on.

	Treatment	Control	Increment
Model	A	B	A-B
Random	C	D	C-D
Model - Random	A-C	B-D	(A-B)-(C-D)

Table 2. Incremental value measurement with treatment and control groups (Lo, 2002)

The uplift method considers that a baseline incremental value can be obtained even by randomly targeting individuals from the database. In Table 2, A, B, C and D cells are aggregate values that may represent sales, response rates or other performance metrics (Lo, 2002). If the value of A > value of C, we presume that the model performs well in targeting the "good" clients as it provides better results than random targeting. It would be considered a good model according to conventional standards. However as mentioned earlier, this type of approach does not eliminate customers who would buy regardless of whether they receive the treatment or not – as it focuses on the broader segment of individuals who are simply more likely to buy. As Lo explains, this key information is obtained only when the aggregate values for the control group are considered. Consequently, if A tends to be very close to B, it becomes clear that even if the predictive model performs well (A> C) there is no real value related to the campaign. A ≈ B simply means that the group receiving the treatment brought the same value as the untreated group (Lo, 2002). The fact that the model performs better than random targeting is therefore not sufficient, as it’s necessary to consider the difference between the treatment group and the control group. Therefore, the proposed gain for uplift modeling is (A-B) - (C-D), given that all these values are statistically significant (Lo, 2002). Mathematically, the modeled value is as follows:

$$P^T(Y = 1|x_i) - P^C(Y = 1|x_i)$$

where Y is the response variable. Y = 1 represents a response (purchase) and Y = 0 an absence of response (no purchase), X_i is a vector of independent variables while T refers to the treatment group and C to the control group. This measure is therefore an estimate of the net gain of the treatment (Soltys et al, 2014). Since real-life marketing campaigns usually have a limited budget, the customer database is divided into deciles, where only the highest performing deciles (thus with the highest incremental values) are targeted for the future campaign. This will maximize the likelihood of targeting clients who are most likely to buy only because of the promotion, while leaving spontaneous responders or those who respond negatively out of the bag.

THE CURRENT PRACTICAL METHODS

Two-models approach

The simplest method proposed in several studies involves the development of two models. Categorized as indirect, this approach involves fitting one model for the control group and another for the treatment group, independently of each other and each separating the responders from the non-responders. The key step of this approach is the subtraction of the control score from the treatment score. This difference is then interpreted as the incremental response rate at the individual level (Hansotia & Rukstales, 2002). It is reported that this approach performs efficiently in practice when limited to the best deciles involving the highest uplift (Zhao, 2012). However, the basic hypothesis of the model implies that even if the individuals in the control and treatment group are different, the individuals grouped in the best deciles of the two groups remain similar and thus, comparable. Based on this hypothesis, it becomes therefore possible to subtract the individuals scores (Hansotia & Rukstales, 2002).

True uplift modeling approach

On the other hand, the direct method proposed by Victor Lo (Lo, 2002) involves combining the two treatment and control samples while creating a dummy variable T for treatment. This binary variable has a value of 1 if the individual has been part of the treatment group and zero if the individual was in the control group, independently of other variables. Lo suggests adding interaction terms involving T and the independent variables. By multiplying variable T with the other independent variables, we thus obtain two types of interest variables: the independent variables for the control group and the independent variables in interaction with the treatment variable. It is consequently possible to develop only one model, both on the control and treatment group jointly. Opposing the previous approach, this method avoided errors induced by fitting two separate models and calculating the scores on the difference between the two.

Lai's approach

A different uplift modeling approach was to reorganize the four client groups from Table 1 into two sets. Lai's method concretely allocates customers into two categories distributed as good vs bad segments (Lai, 2006). Customers who will not buy because of the treatment and customers who do not buy whether they receive a treatment or not are both referred to as "bad" customers (or negative uplift). On the other hand, customers who buy with or without treatment, or only with treatment, are referred to as "good customers" (or positive uplift). It is then a matter of modeling the likelihood of a client to be a good client (Lai, 2006).

Latest suggested approaches

A myriad of other methods was suggested in the last five years. An example would be the z-transform model, which is based on logistic regression or a bayesian classifier to construct a single model (Jaskowski & Jaroszewicz, 2012). A variant of Lai's method has also been proposed, generalized with added weights for standardization of predictions (Kane et al, 2014). Furthermore, the most recent methods combine modern techniques such as random forests (Guelman et al, 2015). The most recent approach to-date is based on creative reverse thinking, modelling the probability for an individual to be part of the treatment group knowing that there is a response (or non-response). The authors of this proposed model, named reflective uplift, claim that combining this approach to Lai's positive/negative concepts gives models more robust to face noise and sources of bias (Shaar et al, 2016).

CURRENT MODEL ASSESSMENT MEASURES

1. Known observed value

A first possible measure of model performance is to use data for which the outcome is known. Thus, the uplift predicted value by the model is then compared with the uplift known value, at the individual scale. A certain measure of precision for the model is consequently obtained. Studies such as Shaar et al (2016) and Lo (2002) put this method in use and evaluated the robustness of their novel uplift models on simulated data sets. By controlling the components of the dataset and knowing the actual uplift value, several assessment criteria can be obtained: the correlation coefficient between the predicted mean and the actual uplift, the AUC difference between the predicted value curve and the actual value (Shaar et al, 2016). Nevertheless, because this assessment method requires knowing the actual uplift value, this method is not usable in practical cases.

2. Uplift curve

Uplift curves, also called incremental gain curves, have been used to visualize a model's performance (Soltys et al, 2014). This curve requires the sample to be divided into training and validation. After modeling the uplift value on the training sample, the model is used to score the validation sample, both for the control and treatment group. We then sort the uplift scores estimates from highest to lowest. The observations are then separated into groups, often in deciles, once again in descending order. Consequently, the first decile regroups the top 10% of the estimated uplift scores. The observed uplift value for each group is then calculated:

$$(1)\ Observed\ uplift = \frac{Nb\ Treatment\ responses}{Nb\ Treatment} - \frac{Nb\ Control\ responses}{Nb\ Control}$$

$$(2)\ Subtracted\ uplift = Nb\ Treatment\ responses - Nb\ Control\ responses$$

$$(3)\ Real\ uplift = {}^{Observed\ uplift}\!/_{Decile}$$

The first calculation method (1) is used by Shaar et al (2016) and compares the success rate in the treated group vs control group. It can be isolated to specific deciles or applied to the whole model results, as an overall performance metric. The second method (2) instead focuses on absolute numbers in each group. Because it directly provides a difference in individuals and is more simple, this metric is generally more popular through literature (Lo, 2002; Radcliffe, 2007). These results are subsequently plotted on a graph of observed gain for each decile. The net gain predicted by the model is thus visually obtained for each portion of the population, ordered from the most to the least interesting deciles. Figure 1 below is an example of the uplift curve for 2 different models, based on subtracted uplift method.

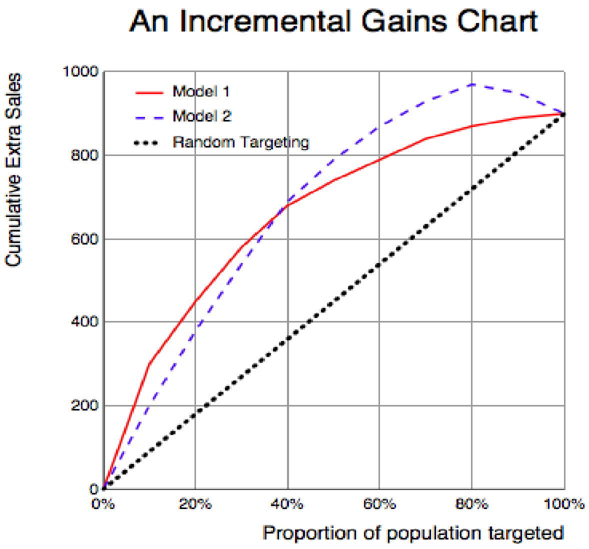


Figure 1. Incremental gains chart including random targeting, as seen in Radcliffe & Surry, 2011.

In practice, the incremental gain curve is extremely valuable in comparing models. Unlike conventional gain charts which focus on the response rate, the uplift curve is often represented on a graph of predicted incremental sales based on the number (or proportion) of individuals treated (Radcliffe 2007). The graph also often includes for comparative purposes a diagonal, representing the effect of random targeting on the population. This diagonal is drawn between the points (0,0) and (N, n), where N is the number (or proportion) of targeted from the population and n the gains obtained if the whole population is treated (Radcliffe, 2007). In some cases where costs, benefits and other ROI metrics need to be integrated, it becomes possible to include and measure the gain achieved by appropriate targeting. This provides an extremely useful graph of the profitability of the campaign for each developed uplift model.

3. Area under the uplift curve

The area under the uplift curve (AUUC) simplifies the performance of the model to a simple estimate (Jaskowski and Jaroszewicz, 2012). It provides a measure of the success rate of treatment according to the model. If the AUUC is measured between 0-100% inclusively on the x-axis, it can be interpreted as a measure of the campaign's success when 100% of the population is targeted. This measure is often subtracted from a random targeting curve, as described in detail with the next criterion.

4. GINI

4.1 Gini Coefficient

The Gini coefficient is well known throughout literature. Adapted initially for classification tree models with algorithms such as CHAID or CART, the Gini coefficient is known as an impurity index of the model (Kane et al., 2014). However, Gini coefficient applied to uplift models is an estimate of the overall fit of the model. Several Gini calculations have been suggested over time (Radcliffe, 2007; Radcliffe, 2011; Kane et al., 2014). An estimate of Gini can be obtained from a conventional gain curve graph or ROC chart by measuring the AUC between the developed model and random curve.

$$Gini\ Difference = Model\ AUC - Random\ AUC$$

On the other hand, Gini can also be a ratio. Optimal targeting is defined as assigning the highest uplift scores to all the "good" clients (responds to treatment) when compared to "bad" clients (not responding). To illustrate this concept and without taking negative effects into account, an optimal model would hence allow us for example to have 100 individuals who respond positively to the treatment when we treat 100 individuals. Thus, the Gini coefficient can also be measured as follows:

$$Gini\ Ratio = \frac{Model\ AUC - Random\ AUC}{Optimal\ AUC - Random\ AUC}$$

This measure therefore varies between -1 and 1, where a model correctly ranks all the responders at 1 or erroneously targets all the non-responders at -1 with a lower performance than random targeting. The closer this estimate is to 1, the better the model (Radcliffe, 2007).

4.2 Gini Top %

In real-life retail contexts, there is always a risk attributable to reaching clients who will be annoyed by the treatment, or simply will not make a purchase regardless of the circumstances. Targeting this group of individuals causes a negative effect, induced by a loss (i.e. related to the cost of the offer). Consequently, it is clear that the greater the number of individuals targeted is the greater the chances of inducing negative effect. Given that a campaign budget is often narrow, marketing promotions often have to be optimized to maximize their benefit which translates to better selecting the best potential responders. As a result, Gini Top % was proposed to provide an estimate of Gini only for a certain portion of the population, often limited to the first 2 or 3 deciles when sorted by predicted uplift value (Radcliffe, 2007). Based on this criterion it is thus possible to compare several models of uplift based on their efficiency to capture individuals with the highest potential within the best deciles.

4.3 Gini repeatability metric

Finally, under certain circumstances an uplift model might only be effective on a certain sample and entirely lose its predictive value on another. It has been pointed with several cases through literature that a model could be so sensitive to sampling noise or extreme values that the best deciles become inverted (Kane et al., 2014). Therefore, a variant of the Gini metric has been proposed. By performing a linear regression with lift as a dependent variable and the deciles as independent (Figures 2-3), we obtain an estimate of the classification quality of the deciles.

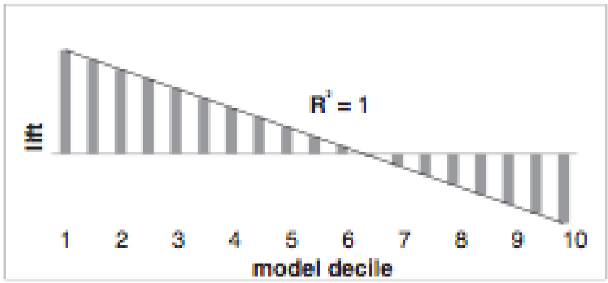


Figure 2. Gini repeatability metric on a lift chart, as seen in Kane et al (2014).

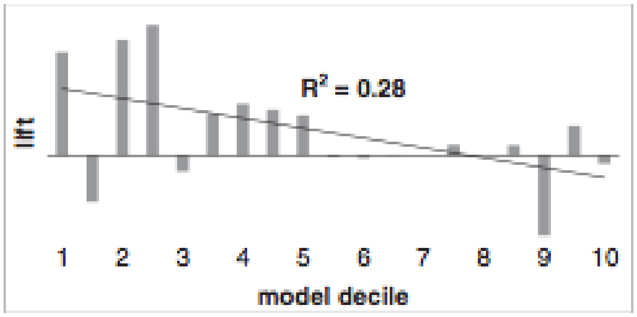


Figure 3. Example of Gini repeatability metric demonstrating a low performance, as seen in Kane et al (2014).

This calculation method has been proposed as the Gini measure of repeatability (Kane et al, 2014). Furthermore, the authors interpret it as a coefficient of determination (R^2) measured on a lift chart and based on the predictions on the validation sample. This measure essentially ensures that the best deciles suggested by the model maintain their robustness. In general, the closer R^2 is to 1 the better the model. However, it is suggested in practice that a model is considered "good" when its R^2 is between 0.3 and 0.5 (Kane et al., 2014).

5. QINI

5.1 Qini Coefficient

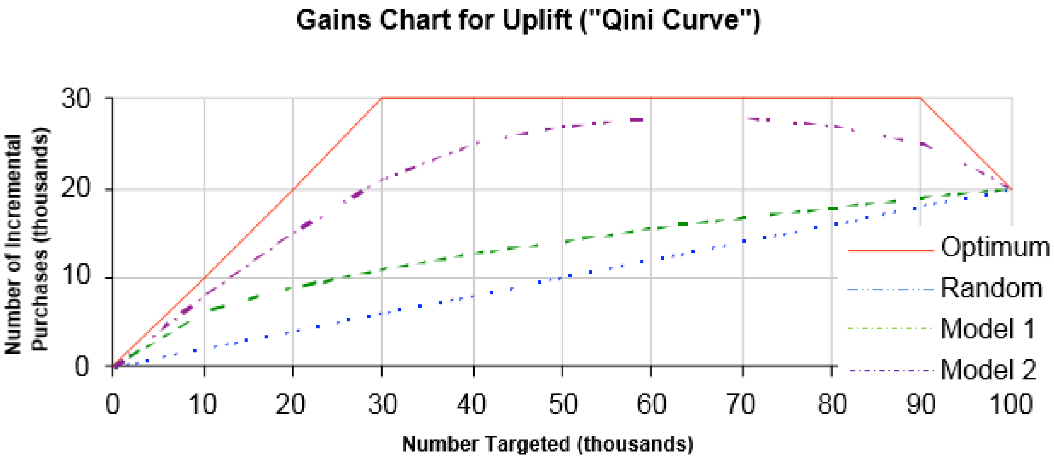


Figure 4. Qini Curve demonstrating an optimum curve versus random, as seen in Radcliffe (2007).

The general measure of Qini is a variant of the Gini coefficient applied specifically to uplift models. While an estimate of Gini is obtained on a graph of a conventional gain curve (Y = number of responses), the Qini coefficient is measured on the uplift curve (Y = incremental gain). Qini remains simply a variation of Gini, the main difference being Qini is a specialized measure of the AUUC (area under uplift curve) while Gini is a broader measurement of AUC. Qini measures are therefore calculated very similarly to the latter:

$$(1)\ Qini\ Difference = Model\ AUUC - Random\ AUUC$$

$$(2)\ Qini\ Ratio = \frac{Model\ AUUC - Random\ AUUC}{Optimal\ AUUC - Random\ AUUC}$$

Q can therefore be a ratio (2) or a difference (1). As they are chiefly based on the predictive model’s AUC and thus set on a similar baseline, both methods are expected to provide similar results. Both methods however seem to be accepted and substituted through literature – as a result we will keep both methods to calculate this criterion. Similarly, for the Gini coefficient, the Q ratio varies between -1 and 1 and its interpretation is carried out in the same fashion. The Qini thus provides information on the performance of an uplift model summed up in an estimate (Guelman et al, 2015).

5.2 Qini Continuous

The Qini coefficient can also be adapted to a continuous target variable, as a variant Qc has been proposed for this purpose (Radcliffe, 2007). This estimate ultimately makes possible the measurement of a certain gain amount "per head" (Radcliffe, 2007). This is relevant information in a context where not all clients provide the same amounts of money. However, this criterion will not be used in the context of this study.

5.3 q_0 coefficient

As highlighted with the Gini coefficient, it is important to consider the possibility of negative effects when more people are treated. The q_0 coefficient has been suggested regarding this problem. It is calculated as follows:

$$q_0 = \frac{Model\ AUC - Random\ AUC}{Zero\ downlift\ AUC}$$

This criterion uncovers the zero-downlift curve metric, which is defined as an optimal curve where negative effects are not considered (Radcliffe, 2007). To note that this concept is different from the previously introduced optimal curve, which is obtained with a model that would only capture positive responders in the best deciles (as illustrated in figure 4). On the other hand, the zero-downlift curve revolves around a scenario where we assume the absence in the population of individuals who would not respond because they are treated. As a result, and because it uses this theoretical scenario as baseline, q_0 penalizes for a large proportion of targeted people and thus limits in scale (Radcliffe & Surry, 2011). Usually varying between -1 and 1, it has been noted that this measure may under certain circumstances exceed 100% (Radcliffe, 2007; Radcliffe & Surry, 2011). The q0 coefficient is a key indicator of the maximum of the population that can be reached without inducing negative effects.

6. Tau criterion

Tau is a relatively novel criterion yet elegant through its simplicity. As opposed to previously mentioned criteria which were mainly built on incremental gains and Qini curves, Tau is measured directly on the predictive estimate of uplift (Gutierrez and Gérardy, 2016). In the presence of a sample with the outcome Y, treatment W and covariates X, it was previously established that if the outcome is binary (as in our case), then the treatment effect or uplift at x is defined as:

$$\tau(x) = P(Y = 1|W = 1, X = x) - P(Y = 1|W = 0, X = x)$$

Tau however further involves the concept of propensity score. The propensity score is the probability that a subject receives the treatment, defined as:

$$p(x) = P(W = 1|\, X = x).$$

Define the transformed response (Athey and Imbens, 2015)

$$Y^* = Y[\frac{W}{p(x)} - \frac{(1-W)}{(1-p(x))}].$$

Under the conditional independence assumption, which means that the treatment assignment is independent of the potential outcomes, conditional on the covariates, we have

$$E[Y^*|X = x] = \tau(x).$$

This means that the treatment effect can be estimated by modeling Y^*. However, in some applications, the propensity score is known (e.g. if the treatment/control assignments was done at random and then $p(x)$ =1/2), and in others it is not. If the propensity score is unknown, it can be estimated to obtain the estimated transformed response

$$Y^{**} = Y(\frac{W}{\hat{p}(x)} - \frac{(1-W)}{\left(1-\hat{p}(x)\right)}).$$

Gutierrez and Gérardy (2016) proposed using

$$MSE({Y_i}^{**}, \hat{\tau}) = \frac{1}{n}\sum_{i}^{n}({Y_i}^{**} \; - \; \hat{\tau}_i)^2$$

where $\hat{\tau}_i$ is the estimation of the lift, as the performance measure. Tau is distinctive by its minimalism and the fact that it's aimed directly at evaluating the lift predictions, as opposed to indirect methods such as AUC measurement.

7. Other performance criteria

In general, we would tend to choose the uplift model with the highest Qini, R^2 coefficient and Qini Top %. This is what Radcliffe & Surry (2011) refers to respectively as Validated Qini, Monotonicity and Maximum Impact. It is also important, once again in the context of a limited budget for a campaign, to consider the maximum gain at the threshold imposed by the budget, which is often referred to as Impact at cutoff (Radcliffe & Surry, 2011). Another criterion would be the Range of Predictions, which measures variations in model predictions between different deciles (Radcliffe & Surry, 2011).

Finally, it has been suggested to measure the Spearman correlation coefficient between real uplift measures of the sample and the predicted values of the model (Shaar et al, 2016). The underlying idea would be to measure the model precision. Another metric suggested by Shaar et al. was to sum all of the positive uplift curve points above the random curve. The authors propose this metric as an indicator for the effectiveness of the model.

The current problem is that despite the diversity of uplift models developed through literature, there is to our knowledge a lack of studies comparing the criteria and evaluating their assessment performance. Namely, when many uplift models are in competition, is it possible to select the best one with these criteria? We will investigate the performance of these criteria, as model selection tools, with a simulation study.

PERFORMANCE CRITERIA LIMITATION

As explained above, due to the uplift scheme, the causal inference problem makes it impossible to find the same individual both in the treatment and control group. To measure the gain difference between the treated and control groups, uplift models postulate that individuals in the deciles with the highest % in the treatment group remain comparable with those in the control group. This is a basic assumption for an uplift model, shared by all the uplift schemes. As it might seem like a leap of faith, this hypothesis is nonetheless reported to work well in practice (Rzepakowski and Jaroszewicz, 2011). However, the sample size is therefore a crucial aspect for a correct assessment of the models. Since several performance measures are based on the AUC difference between the treatment and control groups, N must be large enough per decile for the assumption to be valid, making the top individuals in these groups comparable. When modeling a binary variable (eg purchase = 1), it is reported in the literature that the number of individuals in each group must be higher than the product of the multiplication of the global uplift by total population size (Radcliffe & Surry, 2011):

$$Min\,N = Uplift\ \times\ N.$$

In addition, adding weights is recommended when the groups have very different sizes (Rzepakowski and Jaroszewicz, 2011).

METHODOLOGY

GENERATING THE DATA SET

To compare the performance criteria, we settled for simulated data models. Since the true lift are unknown in real life contexts, performance analysis through simulation remains the most viable and accurate way to compare the performance of model criteria.

Two data generating processes (DGP) have been used in this study. Each has 5 covariates X_i, a binary treatment variable (W) and a binary response variable (Y). Precisely, the data are in the form (Y, W, X), where:

- Y is the outcome; Y = 1 indicates the person responded, i.e. a success and Y = 0 indicates the person did not respond, i.e. failure. See below got the generating process.
- W is a treatment indicator (W = 1 indicates that the person received the treatment and W = 0 indicates it did not).
- X is a vector of covariates for each individual. It has the multivariate normal distribution with 0 mean vector, unit variances and a correlation of 0.3 between each pair of variables.

Two DGPs are considered for Y. The first one involves main effects only while the second one includes transformations of the covariates and interactions between them. In all cases, the treatment assignment W follows a Bernoulli distribution with a success probability of 0.5. For each observation generated by each DGP, the probability of response $p = P(Y = 1)$ is given by $p = 1/(1 + \exp(g(X, W))$. The DGPS are:

DGP 1: $$g(X, W) = -0{,}3 \times \begin{pmatrix} -4 + x1 + x2 + x3 + x4 + x5 + 0{,}5 \times W + \\ 3 \times (-1.5 \times W \times x1 + W \times x2 + W \times x3 + \\ W \times x4 + W \times x5) \end{pmatrix}$$

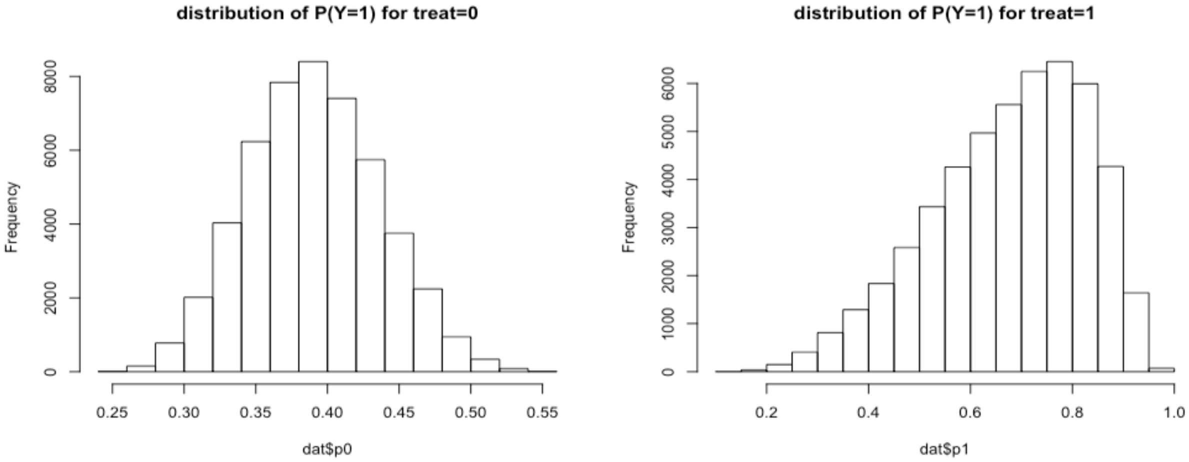


Figure 5. Distribution of the assigned probability of purchase when individuals are treated (right) or not treated (left) in a typical dataset containing 50K observations generated by the first DGP.

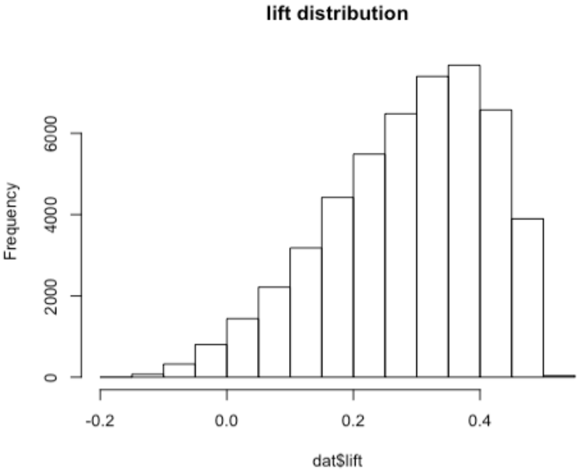


Figure 6. Distribution of lift value across a dataset containing 50K observations generated by the first DGP.

Figures 5 and 6 show key variables distribution in a sample of 50000 observations, based on this first DGP. Individuals who are not treated (treat=0) have a lower probability of making the action (P(Y=1)) versus individuals who are treated (treat=1). This can be visually seen, as a higher proportion of treated individuals are associated with a performed action.

DGP 2: $g(X, W) = -0{,}5 \times \begin{pmatrix} -2 + x1 + x2 + x3 + x4 + x5 + x1^2 + x2 \times x3 + \\ 4 \times W + 4 \times W \times x1 + 3 \times W \times x2 \times x3 \end{pmatrix}$

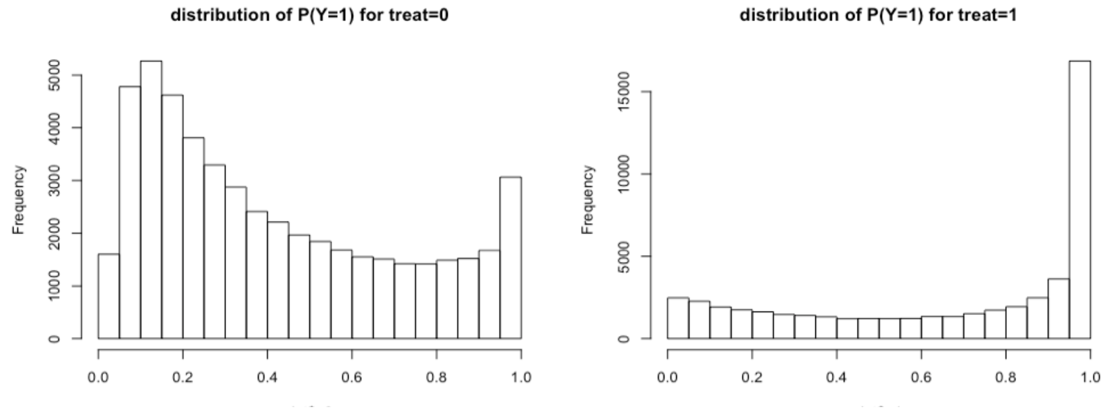


Figure 7. Distribution of probability of purchase when individuals are treated (right) or not treated (left) in a typical dataset containing 50K observations generated by the 2nd DGP.

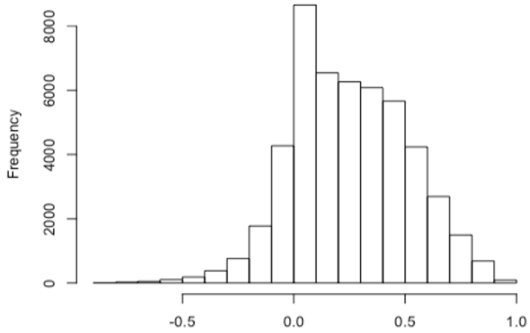


Figure 8. Distribution of lift value across a dataset containing 50K observations generated by the 2nd DGP.

While the lift was previously more centered around 0,4, the second DGP further distributes it around zero, attributing a low uplift effect to the majority of the individuals in the dataset. Similarly, to a real-life context, a minority of observations have negative uplift which are traditionally classified as “individuals not to bother”. This DGP includes a higher proportion of individuals with negative lift, penalizing more for a bad model that would include them in the targeting. Finally, there is also a greater proportion of individuals attributed Y=1 when treat =1, as compared to the first DGP.

Two probabilities have been also created in each dataset based on variable *p*:

$$p_0 = P(Y = 1)\ when\ w = 0$$
$$p_1 = P(Y = 1)\ when\ w = 1$$

where p_0 is the probability of a purchase if the individual is in the control group while p_1 is the probability of purchase if the individual is treated. These probabilities highlight the advantage of simulating data as it becomes possible to assign the same observation in both the control and treatment group. This allows each observation to be associated with an estimate of the lift, calculated as:

$$lift = p_1 - p_0.$$

This is qualified as "real" uplift value, which is once again not available in a real-world framework. Consequently, the closer the predicted uplift estimate is to this value the better the model. Further breakdown of a typical dataset can be found below in Figures 5-8, through the distributions of p_0, p_1 and the lift value. As one would expect, treated individuals (treat=1) have a higher proportions of responders (Y=1), while non-treated individuals (treat=0) have a higher proportion of non-responders.

To test the criteria performance, we decided to use each DGP to generate 6 different training sample sizes: 500, 1000, 2000, 5000, 10000 and 50000 observations. For each sample size, 100 data sets are generated and used to build the different models. When a dataset is generated, it is used to build and validate each of the models. Each criterion is computed for each model, leading to as many different values per criterion. Through a same run, we then highlight the best model based on the criterion (i.e. higher or lower value depending on the criterion) and assess the criterion pick based on MSE/MAE metrics. This data is then compiled for each run, further extended to several sample sizes to assure the robustness of our study.

Additionally, a single test data set is generated containing 10,000 observations and all the variables mentioned above, including true lift.

MODELING & PERFORMANCE ASSESSMENT

Four models have been selected for testing in the context of this study:

- LAI (uplift ccif algorithm based on Lai's method)
- RF (uplift random forest algorithm)
- LOG (uplift logistic regression algorithm, with interactions based on Lo's method)
- 2-MODELS (two random forests, one for treated and one for control observations)

The Random Forest (RF) model required using of the R Uplift package (Guelman, 2014), which allows us to estimate the uplift metric by specifically identifying independent variables and a treatment flag. The uplift RF functions are then used to build the model and group the sample observations in deciles based on their uplift predictions. Each random forest is based on 100 trees with a maximum depth of 3 to assure trees variety (limiting the splitting to 8 terminal nodes) as well as maximize computing power through the iteration, using the KL split method.

The Logistic Regression Model (LOG) uses the GLM function in R to develop generalized linear models, specifically binomial models in our case. Given that we are relying here on Lo's method deemed to improve predictive accuracy, an interaction term between each independent variable and the treatment flag is added. The performance function of the uplift package is then used to group predictions into deciles.

The LAI model is based on the randomForest package (Liaw & Wiener, 2002) to build simple random forests. The Lai method is applied by creating a new variable titled 'Class'. Class takes a value of 1 when treat = 1 and y = 1, or treat = 0 and y = 0, otherwise 0. As previously explained, this method seeks to group individuals based on their cost-effectiveness in order to assign a probability to be part of one of the two classes. Each random forest is based on 100 trees which predictions are grouped in deciles with the performance function of the uplift package.

The 2-model method relies on the randomForest package to build simple random forests. By dividing the sample according to the treatment variable, the original sample is divided into further treatment vs. control samples. By using the randomForest function of the package, we develop a model separately for treated and untreated. The difference between the two models is then calculated, which translates into an estimate of uplift for each observation. Similarly, to the previous models, the performance function of the uplift package is finally used to group the deciles.

Each of these models has been developed by splitting in half the sample into a training data set on which the model is built while the remaining second half is used to validate the model. It is in fact out of this validation process that the described model performance metrics are extracted and compiled. This method is thus representative of a real-life situation where we are exposed to a sample on which to build our model, a situation where we are forced to rely on approximate metrics and criteria. Following the validation process, a real-life context would imply picking the most effective model based on a specific performance criterion – to build the campaign or further targeting. As part of this study, this decision is pushed further – we will examine every model each criterion recommends and evaluate their accuracy on a test sample (10,000 observations).

As described in "generating the data-set" section, the test sample contains a true lift variable assigned to every observation, which is in fact the true uplift value. Thus, the closest the estimated lift will be to the true lift value, the better the model. We will use the validation set to compute the performance criteria metrics – e.g. at training sample size of 2000 that we split in two parts of 1000, the first part will be used to fit the model and measure the $\hat{\tau}_i$ for the observations in the second part.

On the other side, the accuracy of the lift estimation by the models will be measured with both the MAE and MSE criteria as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|P_i - \hat{P}_i| \qquad\qquad MSE = \frac{1}{n}\sum_{i=1}^{n}(P_i - \hat{P}_i)^2$$

Furthermore, and as explained above, each model is built on each of the datasets for a total of 600 runs per model, divided into 6 sample sizes. For each run, each of the 4 models is therefore attributed 6 performance criteria and 2 precision measuring criteria. The performance criteria would then be taken individually, and each would allow us to select the best model for each run. To measure the predictive power of a potentially selected model on the test sample, MAE and MSE criteria are used, both based on the difference between the predicted uplift and the actual uplift in the test sample. In addition, since different sample sizes are used, the results will be compared across different setups to identify whether there is a correlation between the prediction robustness of the criteria. This method has the advantage of making the models comparable within the same run.

In summary, all of the following criteria will be used in this study in order to assess their performance in choosing the best models, in different sample sizes and conditions.

Criteria	**Calculation method**	**Evaluated on**
Qini (Difference)	• AUC trapezoidal method measured on R • Model AUC – Random AUC	Validation sample
Qini (Ratio)	• AUC trapezoidal method measured on R • Measured in terms of proportion vs random AUC	
Qini Top 20 (Difference)	• Derived from Qini Difference • Applied to first 2 deciles	
q0	• Qini difference weighted with optimal curve	
Repeatability metric (*R2*)	• Coefficient of determination applied to uplift curve	
Tau	• Lift estimation based on treatment and response effects	
MAE	Mean absolute error of estimated uplift vs. true lift	Test sample

MSE	Mean squared error of estimated uplift vs. true lift

Table 3. Model performance assessment criteria and their calculation method used in this work.

RESULTS

As the objective of this study was to analyze the model performance criteria, we will be exclusively assessing the different criteria and not the models themselves.

SIMPLE DATASET (DGP 1)

As previously described the first data generating process (DGP) produces a relatively simple dataset, without any interaction between the independent variables. While simulating a minimal interaction between the treatment and independent variables, the samples produced therefore have very little noise which makes it easier to model the uplift value.

SAMPLE SIZE N=2000 WITH SIMPLE FRAMEWORK

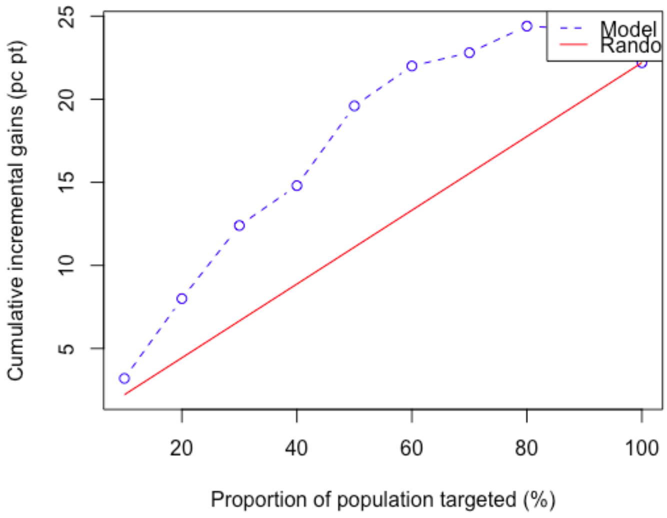


Figure 9. Example of a random forest uplift model performance versus random at N=2000.

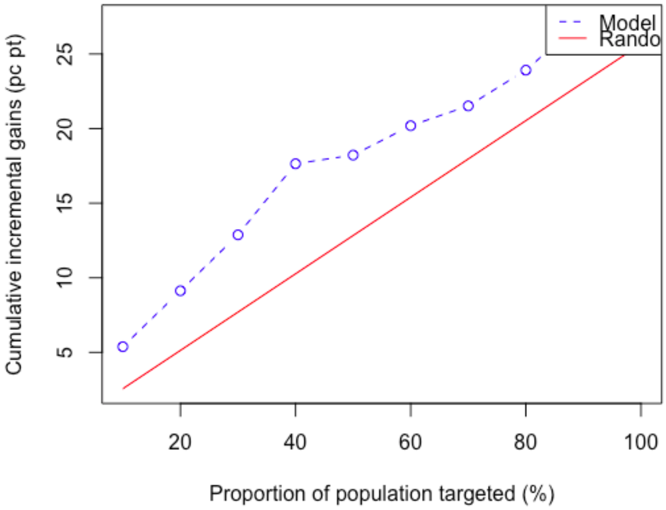


Figure 10. Example of a logistic regression uplift model performance versus random at N=2000.

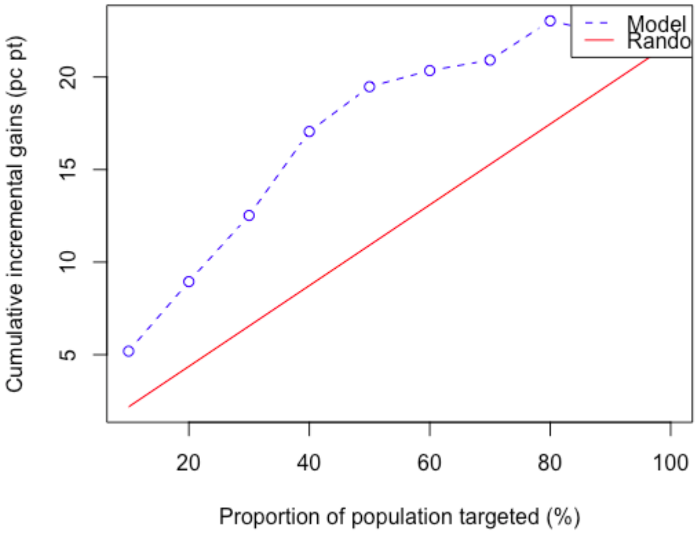


Figure 11. Example of a LAI uplift model performance versus random at N=2000.

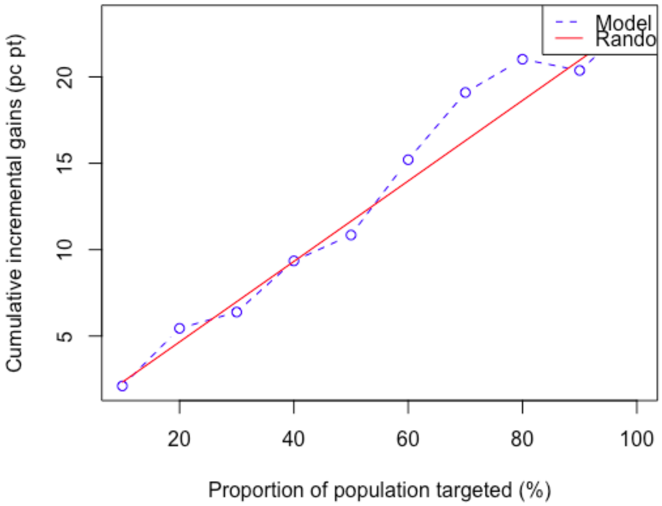
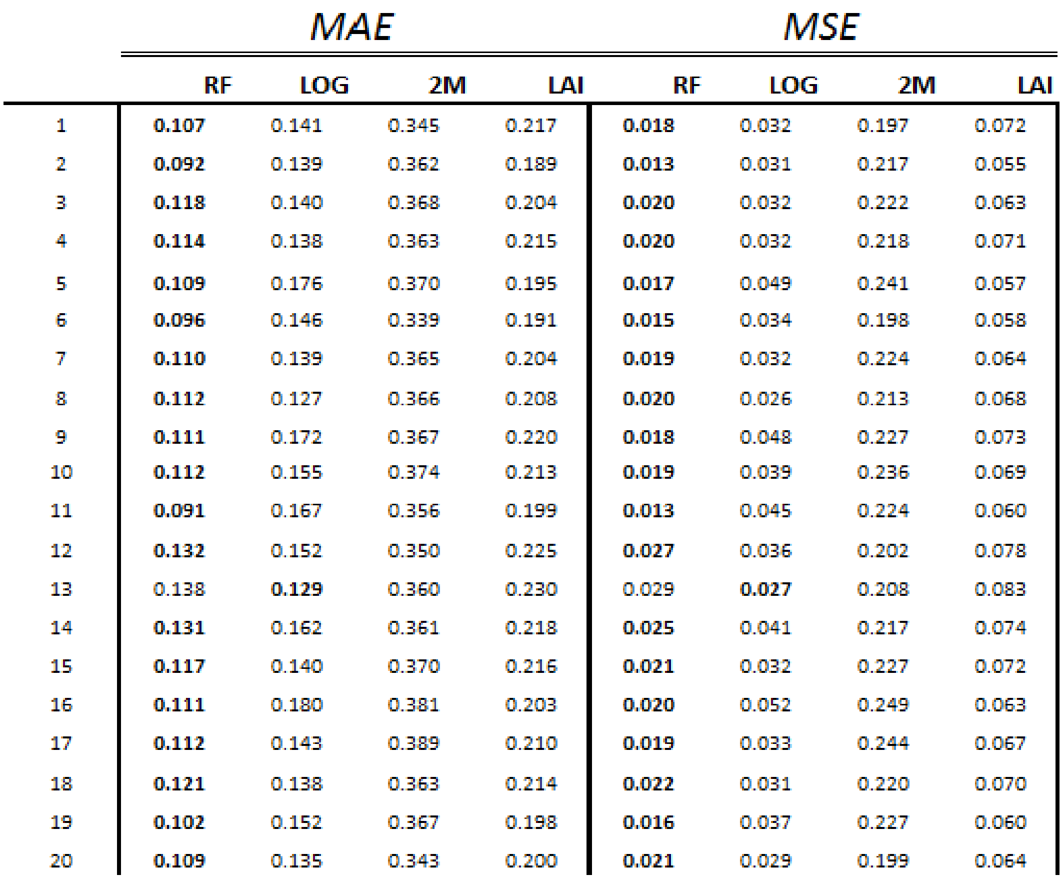


Figure 12. Example of a 2-models method performance versus random at N=2000.

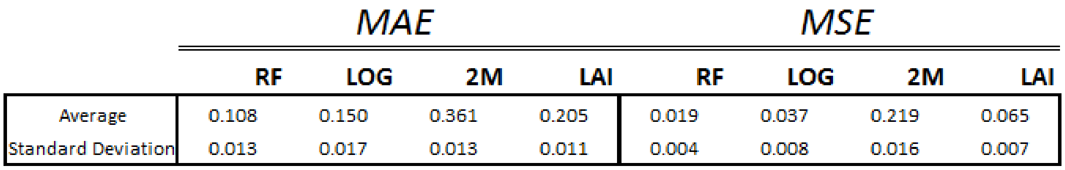
The models remained consistent across the runs in terms of performance, whether according to the MAE or MSE. Figures 9 to 12 above show each model’s individual performance in a single run, plotting the cumulative incremental gains based on the proportion of population targeted. Each graph additionally illustrates the random targeting curve shown as the diagonal line. In this particular run all models except the two-models method seem to perform well, i.e. visually have a bigger AUC compared to the random curve. However, as we are operating on simulated data and therefore have the real uplift value for each individual, we can further rank the models based on their accuracy, using MAE and MSE metrics.



	MAE				MSE			
	RF	LOG	2M	LAI	RF	LOG	2M	LAI
1	**0.107**	0.141	0.345	0.217	**0.018**	0.032	0.197	0.072
2	**0.092**	0.139	0.362	0.189	**0.013**	0.031	0.217	0.055
3	**0.118**	0.140	0.368	0.204	**0.020**	0.032	0.222	0.063
4	**0.114**	0.138	0.363	0.215	**0.020**	0.032	0.218	0.071
5	**0.109**	0.176	0.370	0.195	**0.017**	0.049	0.241	0.057
6	**0.096**	0.146	0.339	0.191	**0.015**	0.034	0.198	0.058
7	**0.110**	0.139	0.365	0.204	**0.019**	0.032	0.224	0.064
8	**0.112**	0.127	0.366	0.208	**0.020**	0.026	0.213	0.068
9	**0.111**	0.172	0.367	0.220	**0.018**	0.048	0.227	0.073
10	**0.112**	0.155	0.374	0.213	**0.019**	0.039	0.236	0.069
11	**0.091**	0.167	0.356	0.199	**0.013**	0.045	0.224	0.060
12	**0.132**	0.152	0.350	0.225	**0.027**	0.036	0.202	0.078
13	0.138	**0.129**	0.360	0.230	0.029	**0.027**	0.208	0.083
14	**0.131**	0.162	0.361	0.218	**0.025**	0.041	0.217	0.074
15	**0.117**	0.140	0.370	0.216	**0.021**	0.032	0.227	0.072
16	**0.111**	0.180	0.381	0.203	**0.020**	0.052	0.249	0.063
17	**0.112**	0.143	0.389	0.210	**0.019**	0.033	0.244	0.067
18	**0.121**	0.138	0.363	0.214	**0.022**	0.031	0.220	0.070
19	**0.102**	0.152	0.367	0.198	**0.016**	0.037	0.227	0.060
20	**0.109**	0.135	0.343	0.200	**0.021**	0.029	0.199	0.064

Table 4. MAE and MSE metrics results for each model at N=2000, for the first 20 runs.

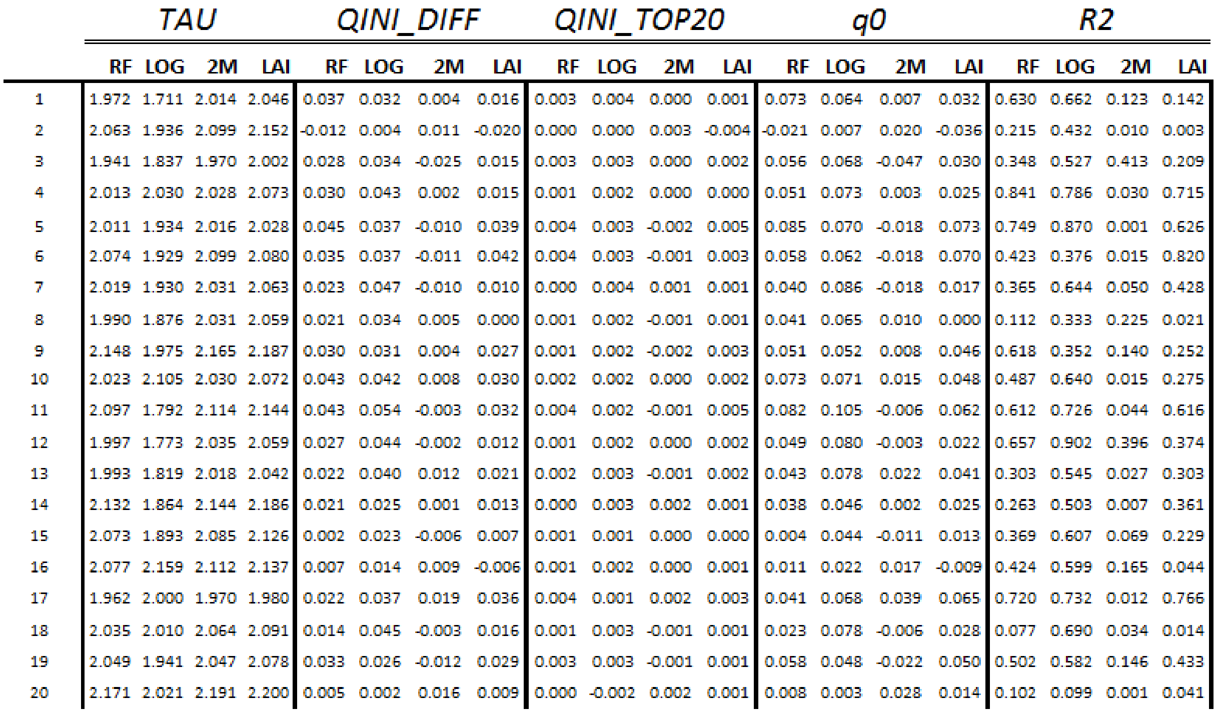
Table 4 demonstrates a quantified performance assessment for each model, based on MAE and MSE metrics results for the first 20 runs at N=2000. As each run's best model (i.e. lowest error metric) is highlighted in bold, we observe that random forests model (RF) seems to be the most effective model within these first runs closely followed by the logistic regression (LOG), inverting the order only in the 13th run.



	MAE				MSE			
	RF	LOG	2M	LAI	RF	LOG	2M	LAI
Average	0.108	0.150	0.361	0.205	0.019	0.037	0.219	0.065
Standard Deviation	0.013	0.017	0.013	0.011	0.004	0.008	0.016	0.007

Table 5. MAE and MSE metrics overall results for each model at N=2000 after 100 runs.

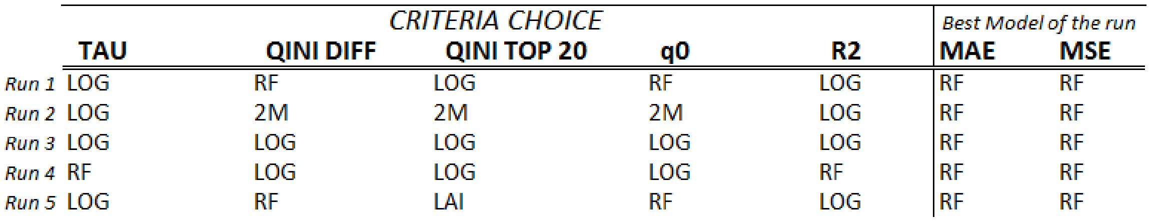
Overall results in table 5 demonstrate that RF remains the best model in terms of accuracy (i.e. lowest average) followed by LOG while LAI ranks third. The two-models method (2M) remains the least performing model, confirming the lack of effectiveness suggested by Figure 12. Since we can classify the models in order of precision, it is possible to quantify criterion performance among each other, by comparing each criterion's selected model.



	TAU				*QINI_DIFF*				*QINI_TOP20*				*q0*				*R2*			
	RF	LOG	2M	LAI	RF	LOG	2M	LAI	RF	LOG	2M	LAI	RF	LOG	2M	LAI	RF	LOG	2M	LAI
1	1.972	1.711	2.014	2.046	0.037	0.032	0.004	0.016	0.003	0.004	0.000	0.001	0.073	0.064	0.007	0.032	0.630	0.662	0.123	0.142
2	2.063	1.936	2.099	2.152	-0.012	0.004	0.011	-0.020	0.000	0.000	0.003	-0.004	-0.021	0.007	0.020	-0.036	0.215	0.432	0.010	0.003
3	1.941	1.837	1.970	2.002	0.028	0.034	-0.025	0.015	0.003	0.003	0.000	0.002	0.056	0.068	-0.047	0.030	0.348	0.527	0.413	0.209
4	2.013	2.030	2.028	2.073	0.030	0.043	0.002	0.015	0.001	0.002	0.000	0.000	0.051	0.073	0.003	0.025	0.841	0.786	0.030	0.715
5	2.011	1.934	2.016	2.028	0.045	0.037	-0.010	0.039	0.004	0.003	-0.002	0.005	0.085	0.070	-0.018	0.073	0.749	0.870	0.001	0.626
6	2.074	1.929	2.099	2.080	0.035	0.037	-0.011	0.042	0.004	0.003	-0.001	0.003	0.058	0.062	-0.018	0.070	0.423	0.376	0.015	0.820
7	2.019	1.930	2.031	2.063	0.023	0.047	-0.010	0.010	0.000	0.004	0.001	0.001	0.040	0.086	-0.018	0.017	0.365	0.644	0.050	0.428
8	1.990	1.876	2.031	2.059	0.021	0.034	0.005	0.000	0.001	0.002	-0.001	0.001	0.041	0.065	0.010	0.000	0.112	0.333	0.225	0.021
9	2.148	1.975	2.165	2.187	0.030	0.031	0.004	0.027	0.001	0.002	-0.002	0.003	0.051	0.052	0.008	0.046	0.618	0.352	0.140	0.252
10	2.023	2.105	2.030	2.072	0.043	0.042	0.008	0.030	0.002	0.002	0.000	0.002	0.073	0.071	0.015	0.048	0.487	0.640	0.015	0.275
11	2.097	1.792	2.114	2.144	0.043	0.054	-0.003	0.032	0.004	0.002	-0.001	0.005	0.082	0.105	-0.006	0.062	0.612	0.726	0.044	0.616
12	1.997	1.773	2.035	2.059	0.027	0.044	-0.002	0.012	0.001	0.002	0.000	0.002	0.049	0.080	-0.003	0.022	0.657	0.902	0.396	0.374
13	1.993	1.819	2.018	2.042	0.022	0.040	0.012	0.021	0.002	0.003	-0.001	0.002	0.043	0.078	0.022	0.041	0.303	0.545	0.027	0.303
14	2.132	1.864	2.144	2.186	0.021	0.025	0.001	0.013	0.000	0.003	0.002	0.001	0.038	0.046	0.002	0.025	0.263	0.503	0.007	0.361
15	2.073	1.893	2.085	2.126	0.002	0.023	-0.006	0.007	0.001	0.001	0.000	0.000	0.004	0.044	-0.011	0.013	0.369	0.607	0.069	0.229
16	2.077	2.159	2.112	2.137	0.007	0.014	0.009	-0.006	0.001	0.002	0.000	0.001	0.011	0.022	0.017	-0.009	0.424	0.599	0.165	0.044
17	1.962	2.000	1.970	1.980	0.022	0.037	0.019	0.036	0.004	0.001	0.002	0.003	0.041	0.068	0.039	0.065	0.720	0.732	0.012	0.766
18	2.035	2.010	2.064	2.091	0.014	0.045	-0.003	0.016	0.001	0.003	-0.001	0.001	0.023	0.078	-0.006	0.028	0.077	0.690	0.034	0.014
19	2.049	1.941	2.047	2.078	0.033	0.026	-0.012	0.029	0.003	0.003	-0.001	0.001	0.058	0.048	-0.022	0.050	0.502	0.582	0.146	0.433
20	2.171	2.021	2.191	2.200	0.005	0.002	0.016	0.009	0.000	-0.002	0.002	0.001	0.008	0.003	0.028	0.014	0.102	0.099	0.001	0.041

Table 6. Classification results per performance criteria for each of the 4 models at N=2000, for the first 20 runs.

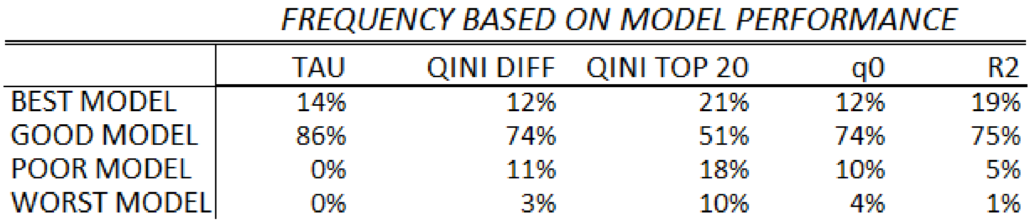
Above table 6 shows the first 20 runs using the first DGP at sample size N=2000. As previously described each run is associated with a different sample but produced by the same DGP (DGP1). Each model is ran on the same sample then classified by each of the five criteria, making it possible to classify each model's performance among a sample criterion. We observe that the models show great reliability at N = 2000, as shown below.



	CRITERIA CHOICE					Best Model of the run	
	TAU	QINI DIFF	QINI TOP 20	q0	R2	MAE	MSE
Run 1	LOG	RF	LOG	RF	LOG	RF	RF
Run 2	LOG	2M	2M	2M	LOG	RF	RF
Run 3	LOG	LOG	LOG	LOG	LOG	RF	RF
Run 4	RF	LOG	LOG	LOG	RF	RF	RF
Run 5	LOG	RF	LAI	RF	LOG	RF	RF

Table 7. Extract of the first five runs and the chosen model per criteria with at N=2000.

The results are very interesting when evaluated run by run. For example, for the first run at N=2000, the highest values for Qini difference and q0 are converging towards the random forests (RF) model while Qini TOP 20 tilts towards logistic regression (LOG). Solely Qini Diff and q0 have effectively chosen the best model in the first run as RF shows to be the best model based of the run, having the lowest value both for MSE and MAE. It is interesting to note that Qini Top 20 can take negative values despite positive Qini values as shown in the 13th run (table 6). This observation stresses the fact that while some models may have a positive Qini curve, they may perform poorly within the first deciles – sometimes worse than random. While Qini Top 20 did not choose the best model in any of the first 5 runs shown in Table 7, it is clear that this criterion should be taken into account as it provides complementary information that none of the other criteria deliver.



FREQUENCY BASED ON MODEL PERFORMANCE

	TAU	QINI DIFF	QINI TOP 20	q0	R2
BEST MODEL	14%	12%	21%	12%	19%
GOOD MODEL	86%	74%	51%	74%	75%
POOR MODEL	0%	11%	18%	10%	5%
WORST MODEL	0%	3%	10%	4%	1%

Table 8. Criteria performance after model classification following 100 runs at N=2000.

Table 8 demonstrates the frequency for each criterion to choose each model. A model is considered “Best model” when it has the highest ranking of the run (i.e. the lowest error metric), “Good Model” when it is ranked second best, etc. Overall, the best results at this small sample size are attributed to Tau criterion which achieves a successful 100% chance of picking the best or second-best model. R^2 is the second best criterion from this perspective, with a 94% chance of choosing a good model or better. While R^2 has been reported to measure the robustness of a model’s deciles ranking coherence (Kane et al., 2014), it is surprising to find it among the best criteria as it’s not directly based on the predictive estimate unlike Qini or Tau. Qini as well as q0, which is another Qini-based metric, both show 86% of likelihood to choose the best or second best model. While not presented in these results, it should be noted here that the type of calculation (difference vs ratio) doesn’t seem to affect the predictive accuracy, as the two criteria have reportedly the same performance. On the other hand, Qini TOP 20 shows however a relatively unstable performance. This criterion has the highest chance of choosing both the best model (21%) and the worst (10%) with an overall 72% chance of picking a good model or better.

It is obvious that a model can be recommended by multiple criteria within a same run, which could simplify the decision-making process or on the contrary, further increase its complexity. An example can be found in the 4th run shown in Table 7, where three of the five criteria tilt towards the logistic regression – This convergence would most probably lead us into picking this model for further targeting on the test sample without necessarily knowing that it’s not the best model of the run. While it is simple in this run as most of the criteria point almost-unanimously towards the same model, it’s possible to match weights with each criterion to facilitate the decision-making process. As a matter of fact, if a campaign budget is extremely tight it would be wise to give the Q TOP 20 more weight than the Qini criterion which takes the whole sample into account. These weights clearly depend on the limits and objectives of the real-life context, which are not taken into account in our simulated data.

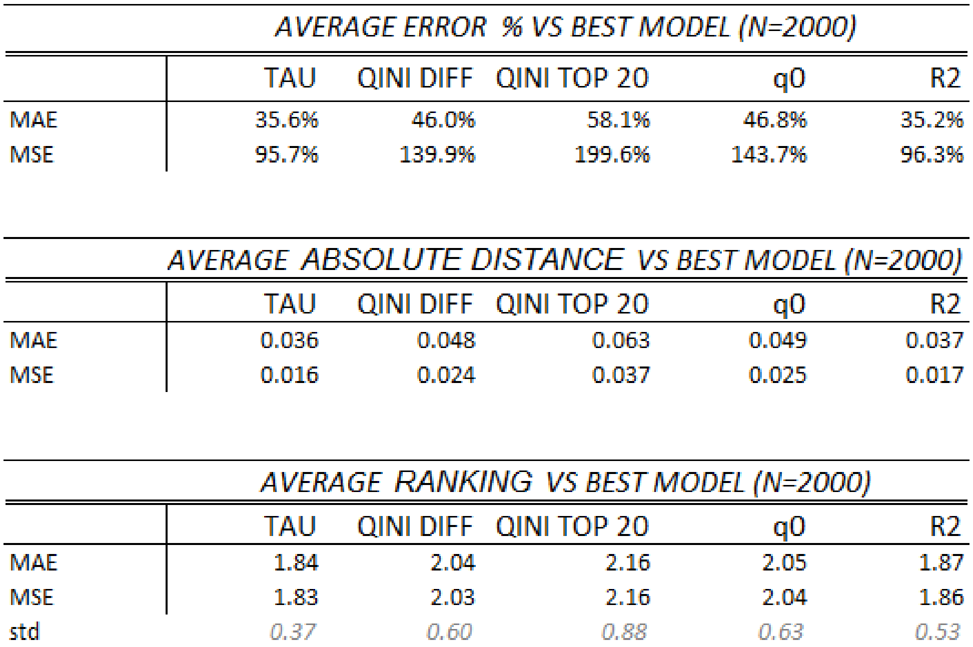
However, a criterion choosing a poorly ranked model in a run doesn’t necessarily indicate the criterion’s bad performance – the difference between 2 models ranked differently, sometimes even the best and worst models, can be extremely minimal. In order to provide the full picture here we would need to quantify the distance between each rank. Since each model is associated with an MSE and MAE per run, it is possible to classify the models by precision. For a given run, models are classified from 1-4, where 1 is the most efficient model of the run and 4 the least. Since each criterion chooses one model out of 4 per run, it is therefore possible to not only rank the models but also the criteria in terms of difference compared to the run’s best model.
Three metrics can be elaborated here:

$$Criteria\ \% \ = Chosen\ Model's\ MSE^{\star}\ \div Lowest\ MSE^{\star}\ of\ the\ run$$

$$Criteria\ Metric\ Distance = Chosen\ model's\ MSE^{\star} - Best\ model's\ MSE^{\star}$$

$$Criteria\ Ranking = Chosen\ model's\ ranking - Best\ model's\ ranking$$

The further a criterion ends from the run’s most precise model (in terms of MSE or MAE), the higher the criteria’s rank will be. Equally, a criterion that chooses the best model of the run will simply have a difference of 0%. This classification is demonstrated in Table 9, where each criterion is associated with a degree of difference compared to the best model of the run.



AVERAGE ERROR % VS BEST MODEL (N=2000)

	TAU	QINI DIFF	QINI TOP 20	q0	R2
MAE	35.6%	46.0%	58.1%	46.8%	35.2%
MSE	95.7%	139.9%	199.6%	143.7%	96.3%

AVERAGE ABSOLUTE DISTANCE VS BEST MODEL (N=2000)

	TAU	QINI DIFF	QINI TOP 20	q0	R2
MAE	0.036	0.048	0.063	0.049	0.037
MSE	0.016	0.024	0.037	0.025	0.017

AVERAGE RANKING VS BEST MODEL (N=2000)

	TAU	QINI DIFF	QINI TOP 20	q0	R2
MAE	1.84	2.04	2.16	2.05	1.87
MSE	1.83	2.03	2.16	2.04	1.86
std	*0.37*	*0.60*	*0.88*	*0.63*	*0.53*

Table 9. Performance assessment metrics values per criterion after 100 runs at N=2000.

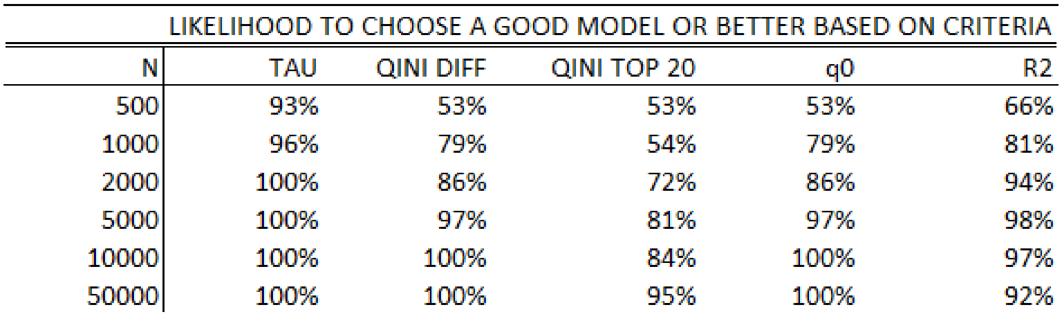
Once again at N=2000, i.e. small sample size, we observe that the best criteria seem to be both Tau and the coefficient of determination R^2, with the lowest error rate compared to the run’s best model respectively 36% and 35% (MAE). In other words, Tau leads us to choose a model on average 36% less accurate in terms of MAE or 96% in terms of MSE. Qini Diff and q0 rank 3rd with an average error % ranging between 46-47% (MAE).

This enhanced perspective slightly changes the ranking of the best criterion. While R^2 previously ranked second in terms of chances to choose a good or best model of the run as opposed to Tau, there doesn't seem to be a significant difference between both criteria. We hypothesize R^2 was previously ranked lower due to its overall 5% chance to choose a poor model (vs 1% for Tau). Quantifying the distance between the chosen models now proves that the difference is in fact minimal and both Tau and R^2 have the best accuracy. Coherently, Tau and R^2 seem to also be the best criteria when we consider absolute difference between the chosen model and the run's model, both in terms of MSE (0,02) and MAE (0,04). Again the same conclusion can be drawn regarding the ranking of the chosen model, where both Tau and R^2 seem to converge towards the best model of the run with the lowest model average ranking (1,9). Qini Diff and q0 are relatively efficient in choosing a good model (2,0-2,1), choosing on average a model closer to the second best model of the run.

As Tau is a direct measure of accuracy comparing the model's predictive estimate to the actual uplift value, it's not a surprise to find it amongst the best criteria. However as previously mentioned, this is not the case for R^2 as it's the only criterion that is not directly based on an approximation of the lift or on the area under the curve. This criterion does better than Qini and its derivatives on all fronts while it's merely based on deciles predictive coherence, at N=2000.

VARIOUS SAMPLE SIZES WITH A SIMPLE FRAMEWORK

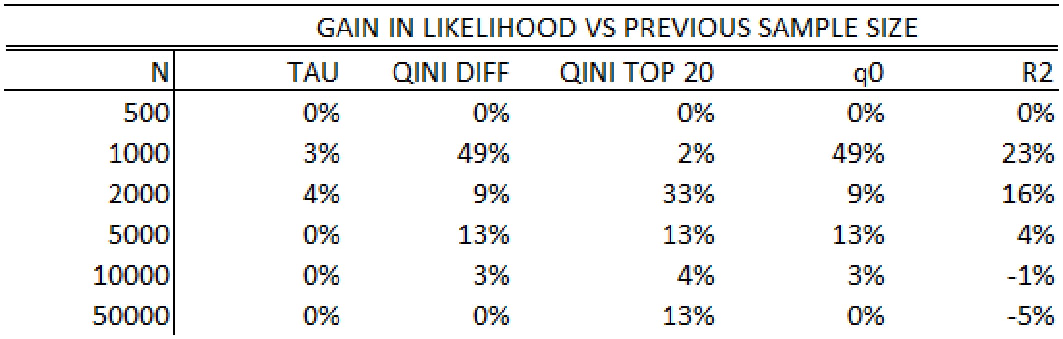
We have exclusively based our previous results on 2000 observations-based setups, considered relatively small sample sizes. Both Tau criteria and the coefficient of determination R^2 seem to be so far the best performance criterion at this level of the analysis, in terms of the MSE/MAE difference and likelihood of choosing the best model. The next step is logically to validate the consistency of these results on different sample sizes.



LIKELIHOOD TO CHOOSE A GOOD MODEL OR BETTER BASED ON CRITERIA					
N	TAU	QINI DIFF	QINI TOP 20	q0	R2
500	93%	53%	53%	53%	66%
1000	96%	79%	54%	79%	81%
2000	100%	86%	72%	86%	94%
5000	100%	97%	81%	97%	98%
10000	100%	100%	84%	100%	97%
50000	100%	100%	95%	100%	92%

Table 10. Likelihood distribution for each criterion to choose the best or second-best model of a run based on sample size.

As one would naturally expect, the larger the sample size the better the accuracy tends to be. The difference is even more spectacular for Qini criterion and its derivatives. Qini Diff jumps from a 53% chance to choose at least a good model (ranked second or better) at the smallest sample size to 100% at the highest sample size. We previously observed at N=2000 that Tau seemed to have the best likelihood at 100%. Tau remains in fact the best criterion at sample sizes lower than 2000 observations with a likelihood varying between 93 and 100%. The trend however changes with increased sample size where all the criteria (except Qini Top 20) have at least 97% chance to choose a good model or better. Additionally, while the likelihood seems to have reached a peak at N=50000, each additional increment in the sample size brought an important gain for most criteria.



GAIN IN LIKELIHOOD VS PREVIOUS SAMPLE SIZE					
N	TAU	QINI DIFF	QINI TOP 20	q0	R2
500	0%	0%	0%	0%	0%
1000	3%	49%	2%	49%	23%
2000	4%	9%	33%	9%	16%
5000	0%	13%	13%	13%	4%
10000	0%	3%	4%	3%	-1%
50000	0%	0%	13%	0%	-5%

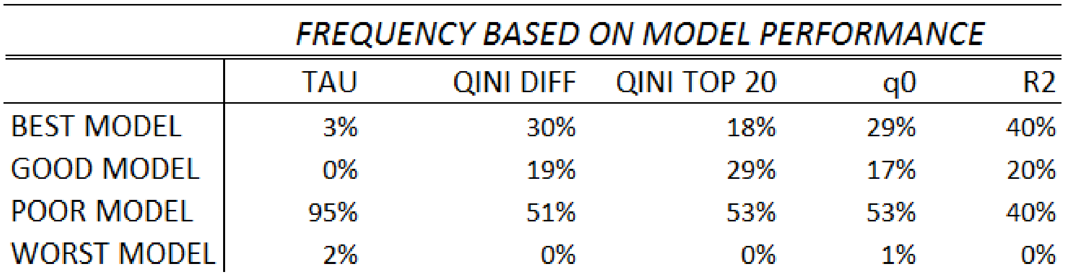
Table 11. Gain in likelihood for each sample size N as compared to previous size for each criterion.

Table 11 further emphasizes the gain in likelihood to choose a good model as opposed to the likelihood to the lower sample size. Even though most models reach a perfect score at N = 50000, it is clear that this accuracy peaks at a certain maximum sample size between 10000 and 50000. We however observe the greatest gain in performance when comparing from 500 to 1000. Doubling the sample size increased the likelihood for Qini Diff and q0 by almost half. Moving from 5000 to 10000 observations does not however provide such a spectacular gain, as it's a 13% increase or less for all criteria. While this demonstrates the critical importance of maximizing the sample size, 5000 observations seem to be necessary for the proper functioning of the named criteria, while demonstrating that it is not necessary to exceed this number. Finally, it is interesting to note that R^2 accuracy decreases with the sample size at 10000 observations and higher.

MORE COMPLEX DATASETS (DGP 2)

As all previous results were based on a simpler data-generating process (DGP), the next step is to test out the same criteria on more complex datasets. With this second DGP, we are looking at samples with interaction between independent variables and the treatment variable as well as interaction between the independent variables themselves.

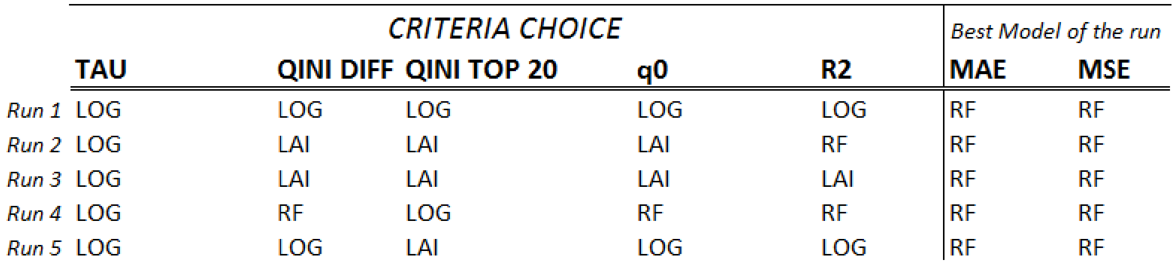
SAMPLE SIZE N=2000 WITH A COMPLEX FRAMEWORK



	FREQUENCY BASED ON MODEL PERFORMANCE				
	TAU	QINI DIFF	QINI TOP 20	q0	R2
BEST MODEL	3%	30%	18%	29%	40%
GOOD MODEL	0%	19%	29%	17%	20%
POOR MODEL	95%	51%	53%	53%	40%
WORST MODEL	2%	0%	0%	1%	0%

Table 12. Criteria performance after model classification following 100 runs on more complex datasets at N=2000.

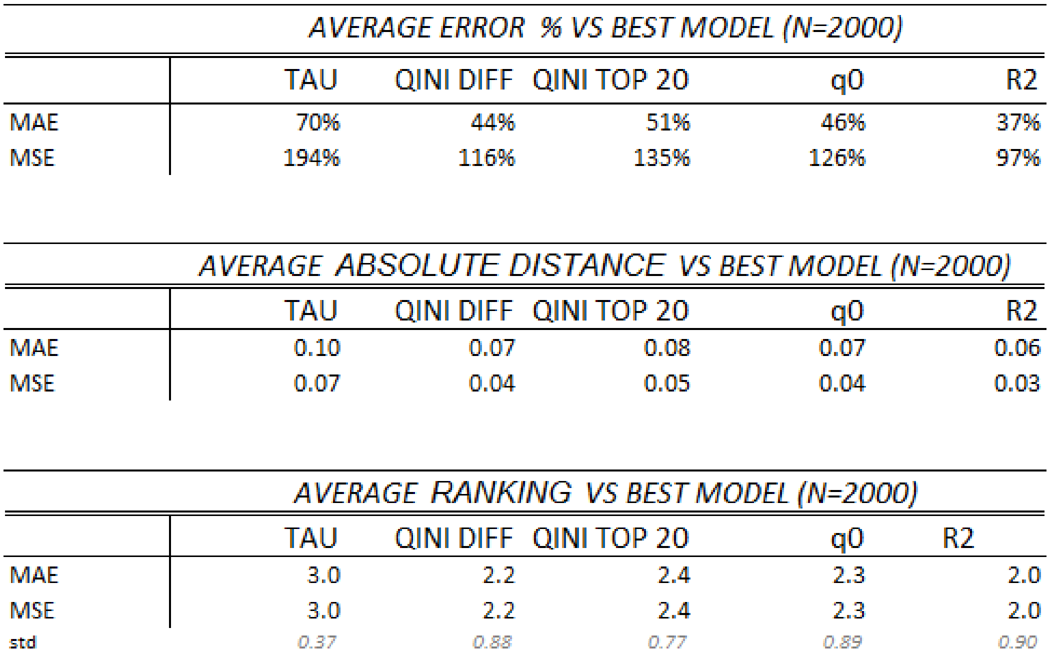
As the datasets increased in complexity, we expect the criteria to be less effective and have a higher error margin. It is to note that while Tau showed the best results with DGP 1, it is in fact the least effective criterion when applied to DGP 2 with only a 3% chance to choose at least a good model. R^2 on the other hand remains effective with a 60% chance to choose a second-ranked model or better, however importantly lower than the previous 94% with the first DGP. Interestingly, Qini Difference and q0 have an increased chance of choosing the best model of the run, as opposed to previously estimated at 12%. This increase in accuracy towards the best model is however offset by an alarmingly increased chance of choosing a poor model (3rd ranked out of 4) of more than 50% for all Qini-derivated criteria.



	CRITERIA CHOICE					Best Model of the run	
	TAU	QINI DIFF	QINI TOP 20	q0	R2	MAE	MSE
Run 1	LOG	LOG	LOG	LOG	LOG	RF	RF
Run 2	LOG	LAI	LAI	LAI	RF	RF	RF
Run 3	LOG	LAI	LAI	LAI	LAI	RF	RF
Run 4	LOG	RF	LOG	RF	RF	RF	RF
Run 5	LOG	LOG	LAI	LOG	LOG	RF	RF

Table 13. Extract of the first five runs and the chosen model per criteria with a complex dataset at N=2000.

It's also interesting to note that Tau seems to choose more often the logistic regression than any other criteria, as demonstrated in Table 13.



AVERAGE ERROR % VS BEST MODEL (N=2000)

	TAU	QINI DIFF	QINI TOP 20	q0	R2
MAE	70%	44%	51%	46%	37%
MSE	194%	116%	135%	126%	97%

AVERAGE ABSOLUTE DISTANCE VS BEST MODEL (N=2000)

	TAU	QINI DIFF	QINI TOP 20	q0	R2
MAE	0.10	0.07	0.08	0.07	0.06
MSE	0.07	0.04	0.05	0.04	0.03

AVERAGE RANKING VS BEST MODEL (N=2000)

	TAU	QINI DIFF	QINI TOP 20	q0	R2
MAE	3.0	2.2	2.4	2.3	2.0
MSE	3.0	2.2	2.4	2.3	2.0
std	*0.37*	*0.88*	*0.77*	*0.89*	*0.90*

Table 14. Performance assessment metrics values per criterion after 100 runs with more complex datasets at N=2000.

Further quantifying each criterion accuracy confirms the coefficient of determination R^2 efficiency over the other criteria. With the lowest average error at this sample size, R^2 leads us to choose a model on average 37% less accurate in terms of MAE than the best model of the run, almost twice better than Tau. While the average absolute distance compared to the run's best model is minimal between R^2 (0,06) and Qini-related criteria (0,07), Tau shows to be further behind with 0,10 absolute MAE difference with the best model. While LOG performed well with the first DGP, it is always ranked in the 3rd position in this case behind RF and LAI. Tau seems therefore to have a bias towards the logistic regression which it chooses almost systematically, leading to its bad performance to choose a good or best model of the run.

VARIOUS SAMPLE SIZES WITH A COMPLEX FRAMEWORK

The coefficient of determination R^2 seems to be the best performance criterion at N=2000, both in terms of the MSE/MAE difference and likelihood of choosing the best model, closely followed by Qini and q0. The next step is once again to validate the coherence of these observations on different sample sizes.

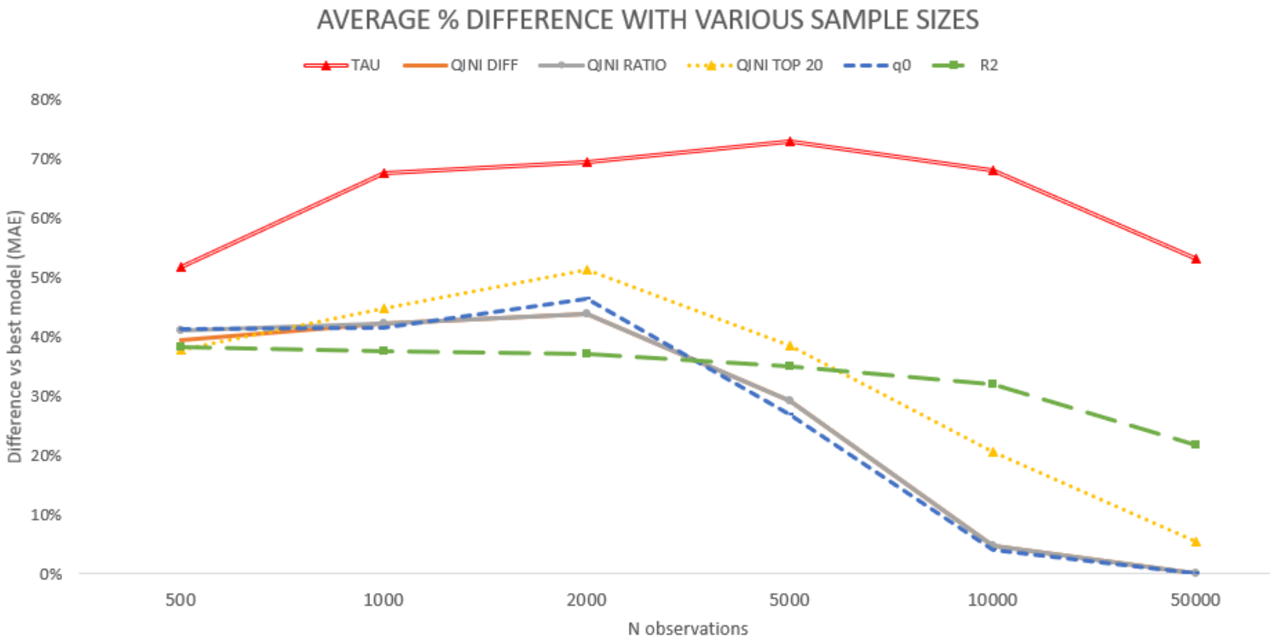
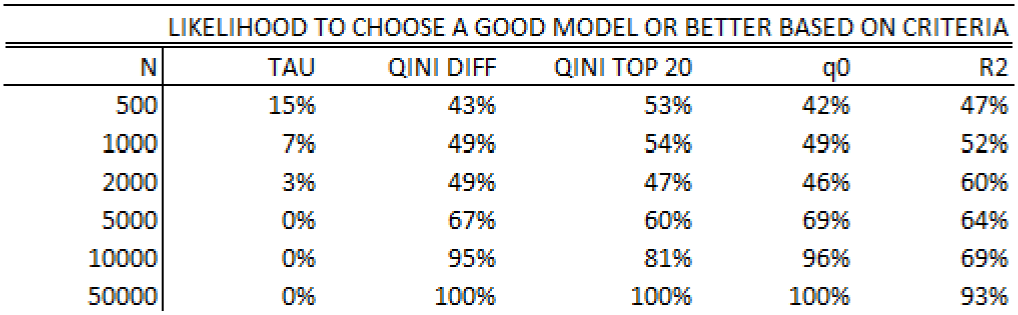


Figure 13. Sample sizes effect on criteria accuracy with more complex datasets.

Figure 13 makes it possible to visually apprehend the gain provided by each additional size increment. While the results at 50000 are as spectacular as with the simpler DGP, we observe here for most criteria that the gain between N = 10000 and N = 50000 is not visually as important as between N = 5000 and N = 10000. If we take q0 as an example, while the curve seems to stabilize at N=2000 as compared to N=1000, there is a significant decrease in % error at N=5000 and higher. It is also interesting to note that while most criteria see their error peak at N=2000 to further decrease, this is not the case for Tau. Tau criterion in fact curiously never sees its error % decrease lower than the level reached at the lowest sample size. As previously mentioned, while Qini ratio calculation will not be shown in the results it is also interesting to note that once again among the two shown methods to calculate Qini, whether by measuring a difference or by ratio, criterion accuracy remains identical. q0 is also highly correlated with the Qini criterion, which is expected as its relatively similar calculation method is also based on the area under the uplift curve. Thus when N is greater than 5000 observations it is the criterion q0 that displays the best results followed closely by Qini.



LIKELIHOOD TO CHOOSE A GOOD MODEL OR BETTER BASED ON CRITERIA

N	TAU	QINI DIFF	QINI TOP 20	q0	R2
500	15%	43%	53%	42%	47%
1000	7%	49%	54%	49%	52%
2000	3%	49%	47%	46%	60%
5000	0%	67%	60%	69%	64%
10000	0%	95%	81%	96%	69%
50000	0%	100%	100%	100%	93%

Table 15. Likelihood distribution for each criterion to choose the best or second-best model of a run based on sample size in a more complex framework.

While Tau was the best criterion with more simple datasets at sample sizes lower than 5K, we now observe the opposite. Tau is the least effective criterion even at N=50K, with an alarming 0% chance to choose a good or best model of the run. As mentioned earlier, this criterion seems to be correlated with the logistic regression's less effective performance with more complex datasets. Tau seems to almost systematically tilt towards the LOG model which tends to be ranked 3rd out of 4 most of the time. This bias explains the extremely poor performance of this criterion.

Finally, we observe an overall decrease in accuracy for all the criteria with increased complexity of the datasets. For example, at N=5000, q0 shows the be the best criterion with 69% chance to choose a good or best model, however previously set at an average of 97% with the simpler datasets (Table 9). For this criterion alone and at an identical sample size, we observe a performance loss of roughly 29% due to added noise and correlation in the samples. It is interesting to note that from N=10000 and beyond, most criteria almost regain the levels of performance previously achieved with simpler datasets (Table 9).

DISCUSSION AND CONCLUSION

Uplift modeling is an elegant and simple method for targeting individuals who are most likely to respond positively to an offer or treatment. This technique has proven to be extremely valuable in various fields: health sciences, marketing and even politics (Eric Siegel, 2013). Up to this date, Uplift modelling is still adapting to the growing technology/algorithms and new variations are regularly proposed in the literature. Conversely as previously described, the main difficulty associated with this technique resides in correctly evaluating the performance of the uplift model before deploying it. Since an individual cannot be both in the treated and control group, it is simply currently not easy to assess the future performance on an uplift model based on a training sample.

Uplift models, when properly executed, perform in most cases better than random. Models that performed poorly compared to random based on Qini (AUC model vs AUC random) remain possible but rare. It is clear that there is thus no doubt regarding the effectiveness of the uplift model, as this point has been made through literature in numerous occasions. Under our test conditions and with the described DGP, random forest models performed best, followed by Lai's method, logistic regression based on Lo’s interaction technique, and then the two-model method. Based on our findings, the following conclusions can be drawn:

1. SAMPLE SIZE EFFECT
   a. Overall, the performance of the criteria greatly varies according to the sample size. In a simple setup (DGP 1) where there were no interactions between variables, sample sizes as small as 10,000 observations provided excellent results for R^2, Qini, Tau and q0, allowing a significant gain in accuracy as opposed to sample sizes of 5,000 observations or less. Conversely, a sample size below 1,000 observations drastically reduced the predictive ability of the criteria to unreliable levels.

b. The same conclusion can be drawn with more complex datasets (DGP 2) which include interactions between variables and a higher % of "individuals not to bother", further penalizing models with poor targeting. With the exception of Tau which performs poorly under these conditions, increased data complexity increases the sample size threshold required to observe effective performance from the criteria when compared to simpler setups. As a result, most criteria were effective only starting from sample sizes of 10,000 observations and higher.

2. CRITERIA RANKING

a. In datasets where there are no interactions between variables, Tau and R^2 proved to be effective criteria to underline the best (i.e. lowest error metric) or second-best model of the run at smaller sample sizes (2,000 observations and lower). At higher sample sizes, Qini and q0 excel in choosing the best or second best model with nearly a 100% average accuracy across all 100 runs. Regardless if it is measured with the difference between AUC model and AUC random or a ratio between both, Qini shows to be a consistent criterion that remained coherent when the sample size was sufficiently large. Among the five evaluated criteria, this criterion ensures choosing on average the closest model to the most precise model of a run, the precision being expressed in terms of MSE or MAE metrics. Following a similar pattern, q0 shows to be a very useful criterion in providing results overall coherent with Qini. While both metrics derive from AUC measuring, q0 assesses model performance compared to the best scenario (optimal curve), where there is no down lift effect. A recommendation therefore would be possibly to use a combination of both criteria.

b. In datasets with increased complexity, all the criteria saw their performance greatly decline with sample sizes smaller than 10,000 observations. R^2 is no longer the best criterion, replaced instead by Qini and its derivates. In fact, R^2 fails to reach a 100% average accuracy with more complex datasets, which leads us to think that this criterion is mostly effective in simple samples but completely loses its interest in complex and more realistic samples. Qini-related metrics remain reliable at sample sizes greater than 10,000 observations, regardless of the complexity of the datasets.

c. Tau criterion gave the best results in simple setups, achieving a 99% likelihood to choose a good model or better from a sample size as small as 2,000 observations. However, it is extremely difficult here to assess the reliability of this criterion as it has completely lost its effectiveness in more complex conditions. Due to its bias towards the logistic regression which performs poorly in the presence of interaction between variables and noise, our findings remain non-conclusive towards this particular criterion.

3. While Qini TOP 20 performs better than the determination coefficient (R^2) or Tau, it remains below Qini or q0 levels when N is larger than 2,000. However, Qini TOP 20 remains a particular criterion due to its calculation derived from Qini but applied exclusively to the first 2 deciles of the model, i.e. isolating the AUC gain if we only targeted the best individuals. As observed previously, while a model's overall Qini might be positive, its Qini TOP 20 might be negative. This finding stresses therefore the importance of the first 2 deciles in the deployment of a treatment-based campaign, as they are found to be crucial in evaluating model performance.

Although these learnings were based on two different DGP assessing simple versus more complex datasets, they bring questions to a facet of uplift modeling little discussed in the literature. It would be interesting to undertake the same study while significantly extending the number of runs in order to add further robustness to the results. Creating different datasets, playing with noise, covariance or the number of variables for example would be extremely relevant in this context.

Finally, it might also be interesting to consider new performance assessment methods for uplift models. Although most of the performance criteria known to date to our knowledge are included in this study, a new solution might include cloning individuals according to their profile and further dividing the samples into more accurate test and control samples.

REFERENCES

[1| Athey Susan and Imbens Guido, Recursive Partitioning for Heterogeneous Causal Effects (2015), Proceedings of the National Academy of Sciences of the United States of America 113, p.7353-7360.

[2] Gregory Jonathan, The internet of things: revolutionizing the retail industry (2015), Accenture.

[3] Guelman Leo, Guillén Montserrat and Pérez-Marin Ana M., Uplift random forests (2015), Cybernetics and systems : an international journal, Vol. 64, p.230-248.

[4] Gutierrez Pierre and Gérardy Jean-Yves, Causal inference and uplift modelling A review of litterature (2016), JMLR: Workshop and Conference Proceedings 67, p.1-13.

[5] Hansotia Behram J. and Rukstales Bradley, Direct marketing for multichannel retailers : issues, challenges and solutions (2002), Journal of database marketing, vol.9 (3), 259-266.

[6] Hansotia Behram J. and Rukstales Bradley, Incremental value modeling (2002), Journal of interactive marketing, vol. 16 (3), p. 35-46.

[7] Heinemann Gerrit and Schwarzl Christoph, New Online Retailing (2010), Gabler, 1st edition, p.211-215

[8] Hughes Arthur Middleton and Sweetser Arthur, Successful e-mail marketing strategies: from hunting to farming (2009), Racom Communications.

[9] Jaskowski Maciej and Jaroszewicz Szymon, Uplift modeling for clinical trial data (2012), Institute of Computer Science, Polish Academy of Sciences, Poland.

[10] Jaroszewicz Szymon and Rzepakowski Piotr, Uplift modeling with survival data (2014), Institute of Computer Science, Polish Academy of Sciences, Poland.

[11] Kane Kathleen, Lo Victor S.Y and Zheng Jane, Mining for the truly responsive customers and prospects using true-lift modeling: comparison of new and existing methods (2014), Journal of marketing analytics, vol. 2 (4), p.218-238.

[12] Kondareddy Sankara Prasad, Agrawal Shruti and Shekhar Shishir, Incremental response modeling based on segmentation approach using uplift decision trees (2016), Springer, p.54-63.

[13] Lai Lily Yi-Tong, Influential marketing: a new direct marketing strategy addressing the existence of voluntary buyers (2006), Simon Fraser university, Canada.

[14] Lo Victor S.Y., The true lift model – a novel data mining approach to response modeling in database marketing (2002), SIGKDD explorations, vol. 4 (2) p.78-86.

[15] Radcliffe Nicholas J., Using control groups to target on predicted lift : building and assessing uplift models (2007), Direct Marketing Analytics Journal, UK, p.14-21.

[16] Radcliffe Nicholas J. and Surry Patrick D., Real-world uplift modelling with significance-based uplift trees (2011), Stochastic solutions, p.1-31.

[17] Rzepakowski Piotr and Jaroszewicz Szymon, Decision trees for uplift modeling with single and multiple treatments (2011), Knowledge and Information Systems, Springer, vol. 32 (2), p.303-327.

[18] Shaar Atef, Abdessalem Talel and Segard Olivier, Pessimistic uplift modeling (2016), Cornell University Library, p.1-9.

[19] Siegel Eric, Predictive analytics: the power to predict who will click, buy, lie, or die (2013), Wiley.

[20] Siegel Eric, The real story behind Obama's election victory (2013) The Fiscal Times, URL: http://www.thefiscaltimes.com/Articles/2013/01/21/The-Real-Story-Behind-Obamas-Election-Victory

[21] Soltys Michal, Jaroszewicz Szymon and Rzepakowski Piotr, Ensemble methods for uplift modeling (2014), Data Mining and Knowledge Discovery, Springer, vol.29 (6), p.1531-1559.

[22] Weisberg Herbert I. and Pontes P. Victor, Post hoc subgroups in clinical trials: anathema or analytics (2015), Clinical Trials, vol. 12 (4), p.357-364.

[23] Zhao Ryan, improve marketing campaign ROI using uplift modelling (2012), Analytics Resourcing Centre Limited, p.1-21.