QINI-BASED UPLIFT REGRESSION

BY MOULOUD BELBAHRI,∗ ALEJANDRO MURUA,¶ OLIVIER GANDOUET‡ AND VAHID PARTOVI NIA§

Université de Montréal∗¶, TD Insurance‡ and École Polytechnique de Montréal§

Uplift models provide a solution to the problem of isolating the marketing effect of a campaign. For customer churn reduction, uplift models are used to identify the customers who are likely to respond positively to a retention activity only if targeted, and to avoid wasting resources on customers that are very likely to switch to another company. In practice, the uplift models performance is measured by the Qini coefficient. We introduce a Qini-based uplift regression model to analyze a large insurance company’s retention marketing campaign. Our approach is based on logistic regression models. We show that a Qini-optimized uplift model acts as a regularizing factor for uplift, much as a penalized likelihood model does for regression. This results in interpretable models with few relevant explanatory variables. Our results show that Qini-based parameter estimation significantly improves the Qini prediction performance of uplift models.

1. Introduction. This work proposes a methodology that identifies characteristics associated with a home insurance policy that can be used to infer the link between marketing intervention and policy renewal rate. Using the resulting statistical model, the goal is to predict which customers the company should focus on, in order to deploy future retention campaigns.

A subscription-based company loses its customers when they stop doing business with their service. Also known as customer attrition, customer churn can be a drag on the business growth. It is less expensive to retain existing customers than to acquire new customers, so businesses put effort into marketing strategies to reduce customer attrition. Customer loyalty, on the other hand, is usually more profitable because the company have already earned the trust and loyalty of existing customers. Businesses mostly have a defined strategy for mitigating customer churn. Organizations are able to determine their success rate in customer loyalty and identify improvement strategies using available data and learning about churn.

With the increasing amount of data available, a company tries to find the

∗¶Corresponding author: alejandro.murua@umontreal.ca

MSC 2010 subject classifications: casual inference, Kendall’s correlation, lasso, logistic regression, marketing, derivative-free optimization
causal effects of customer churn. The term causal, as in causal study, refers to a study that tries to discover a cause-effect relationship. The statement $A$ causes $B$ means that changing the value of $A$ will change the distribution of $B$. When $A$ causes $B$, $A$ and $B$ will be associated but the reverse is not, in general, true, since association does not necessarily imply causation. There exists two frameworks for discussing causation [Pearl, 2009]. We will consider the statistical framework for causal inference formally introduced by Rubin [1974], which uses the notation of counterfactual random variables. This framework is also associated with the potential outcome framework [Neyman, 1923], also known as the Rubin causal model [Holland, 1986]. Suppose a company decides to deploy a marketing campaign, and that customers are randomly divided into two groups. The first group is targeted with a marketing initiative (treatment group), and the second group serves as control (or baseline). A potential outcome is the theoretical response each customer would have manifested, had it been assigned to a particular group. Under randomization, association and causation coincide and these outcomes are independent of the assignment other customers receive. In practice, potential outcomes for an individual cannot be observed. Each customer is only assigned to either treatment or control, making direct observations in the other condition (called the counterfactual condition) and the observed individual treatment effects, impossible [Holland, 1986].

In marketing, it is common to compare the targeted customers’ mean response rate (treatment group) with the non-targeted customers’ mean response rate (control group). A campaign is considered successful if it succeeds in increasing the mean response rate of the treated group relative to the mean response rate of the control group. The difference in mean response rates is the increase due to the campaign. To further increase the returns of future direct marketing campaigns, a predictive response model can be developed. Response models [Smith and Swinyard, 1982, Hanssens et al., 2003, Coussement et al., 2015] of client behavior are used to predict the probability that a client responds to a marketing campaign (e.g. renews subscription). Marketing campaigns using response models concentrate on clients with high probability of positive response. However, this strategy does not necessarily cause the renewal. In other words, the customers could renew their subscription without marketing effort. Therefore, it is important to extract the cause of the renewal, and isolate the effect of marketing.

Data from one of the leader north-American insurers is at our disposal to evaluate the performance of the methodology introduced in this work. This company is interested in designing retention strategies to minimize its policyholders’ attrition rate. For that purpose, during three months, an ex-
Experimental loyalty campaign was implemented, from which policies coming up for renewal were randomly allocated into one of the following two groups: a treatment group, and a control group. Policyholders under the treatment group received an outbound courtesy call made by one of the company’s licensed insurance advisors, with the objective to reinforce the customers confidence in the company, to review their coverage and address any questions they might have about their policy. No retention efforts were applied to the control group. The goal of the study is develop models that will be used to identify which clients are likely to benefit from a call at renewal, that is, clients that are likely to renew their policy only if they are called by an advisor during their renewal period. Also, clients that should not be targeted could

- renew their policy on their own,
- cancel their policy whether they receive a call or not,
- cancel their policy only if they receive a call.

Table 1 shows the marketing campaign retention results. The observed difference in retention rates between the treated group and the control group is small, but there is some evidence of a slightly negative impact of the outbound call. Even if the difference is slightly negative, it may be the case that the campaign had positive retention effects on some subgroup of customers, but they were offset by negative effects on other subgroups. For example, one possible reason could be that some customers are already dissatisfied with their insurance policies and have already decided to change them before receiving the call.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Renewal rate by group for ( n = 20,997 ) home insurance policies.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control</td>
</tr>
<tr>
<td>Renewed policies</td>
<td>2,253</td>
</tr>
<tr>
<td>Cancelled policies</td>
<td>72</td>
</tr>
<tr>
<td>Renewal rate</td>
<td>96.90%</td>
</tr>
</tbody>
</table>

In a randomized experiment, researchers often focus on the estimation of average treatment effects. However, there might be a proportion of the customers that responds favorably to the marketing campaign, and another proportion that does not. A decision at the individual level based on average treatment effects would require adjustments because of the heterogeneity in responses that can be originated by many factors. Table 2 describes some of the \( p = 97 \) available explanatory variables in the dataset, in addition to the treatment (Called or Control) and outcome (Renewed or Cancelled the policy) variables.
Table 2
Descriptive statistics of some available variables for \( n = 20,997 \) home insurance policies.
Because of randomization, the treatment and control group means associated with each available predictor are not significantly different. For privacy concerns, we hide some values with \(*\).

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Control</th>
<th>Called</th>
<th>Diff Mean</th>
<th>Diff SD</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Credit Score</td>
<td>756.93</td>
<td>756.92</td>
<td>-0.00</td>
<td>1.46</td>
<td>( \mathbb{R}^+ )</td>
</tr>
<tr>
<td>Age (Years)</td>
<td>44.97</td>
<td>45.26</td>
<td>0.30</td>
<td>0.25</td>
<td>( \mathbb{R}^+ )</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>0.59</td>
<td>0.60</td>
<td>0.01</td>
<td>0.01</td>
<td>{0, 1}</td>
</tr>
<tr>
<td>Marital Status</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Divorced</td>
<td>0.02</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
<td>{0, 1}</td>
</tr>
<tr>
<td>Married</td>
<td>0.69</td>
<td>0.69</td>
<td>0.00</td>
<td>0.01</td>
<td>{0, 1}</td>
</tr>
<tr>
<td>Single</td>
<td>0.23</td>
<td>0.23</td>
<td>0.00</td>
<td>0.01</td>
<td>{0, 1}</td>
</tr>
<tr>
<td>Seniority (Years)</td>
<td>9.57</td>
<td>9.73</td>
<td>0.16</td>
<td>0.16</td>
<td>( \mathbb{R}^+ )</td>
</tr>
<tr>
<td>Policy Premium ($)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New Premium</td>
<td>*</td>
<td>*</td>
<td>-7.44</td>
<td>16.14</td>
<td>( \mathbb{R}^+ )</td>
</tr>
<tr>
<td>Old Premium</td>
<td>*</td>
<td>*</td>
<td>-5.18</td>
<td>15.12</td>
<td>( \mathbb{R}^+ )</td>
</tr>
<tr>
<td>Territory</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rural</td>
<td>0.06</td>
<td>0.06</td>
<td>0.00</td>
<td>0.01</td>
<td>{0, 1}</td>
</tr>
<tr>
<td>Products</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Auto and Home</td>
<td>0.86</td>
<td>0.85</td>
<td>-0.01</td>
<td>0.01</td>
<td>{0, 1}</td>
</tr>
<tr>
<td>Auto Policies Count</td>
<td>1.05</td>
<td>1.04</td>
<td>-0.01</td>
<td>0.01</td>
<td>( N )</td>
</tr>
<tr>
<td>Mortgage Count</td>
<td>0.66</td>
<td>0.67</td>
<td>0.01</td>
<td>0.01</td>
<td>( N )</td>
</tr>
<tr>
<td>Residences Count</td>
<td>1.07</td>
<td>1.08</td>
<td>0.01</td>
<td>0.01</td>
<td>( N )</td>
</tr>
<tr>
<td>Endorsement Count</td>
<td>1.98</td>
<td>2.00</td>
<td>0.02</td>
<td>0.03</td>
<td>( N )</td>
</tr>
<tr>
<td>Neighbourhood’s Retention</td>
<td>0.87</td>
<td>0.87</td>
<td>0.00</td>
<td>0.00</td>
<td>{0, 1}</td>
</tr>
<tr>
<td>Type of Dwelling</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Family House</td>
<td>0.69</td>
<td>0.70</td>
<td>0.00</td>
<td>0.01</td>
<td>{0, 1}</td>
</tr>
<tr>
<td>Duplex</td>
<td>0.03</td>
<td>0.03</td>
<td>-0.00</td>
<td>0.00</td>
<td>{0, 1}</td>
</tr>
<tr>
<td>Apartment</td>
<td>0.19</td>
<td>0.19</td>
<td>-0.00</td>
<td>0.01</td>
<td>{0, 1}</td>
</tr>
<tr>
<td>Year of Construction</td>
<td>1982.82</td>
<td>1983.52</td>
<td>0.71</td>
<td>0.58</td>
<td>( N )</td>
</tr>
<tr>
<td>Extra Options</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Option 1</td>
<td>0.20</td>
<td>0.20</td>
<td>0.00</td>
<td>0.01</td>
<td>{0, 1}</td>
</tr>
<tr>
<td>Option 2</td>
<td>0.73</td>
<td>0.73</td>
<td>0.00</td>
<td>0.01</td>
<td>{0, 1}</td>
</tr>
<tr>
<td>Option 3</td>
<td>0.71</td>
<td>0.72</td>
<td>0.01</td>
<td>0.01</td>
<td>{0, 1}</td>
</tr>
</tbody>
</table>

The so-called uplift model [Radcliffe and Surry, 1999, Hand and Yu, 2001, Lo, 2002] provides a solution to the problem of isolating the marketing effect. Instead of modeling the class probabilities, uplift attempts to model the difference between conditional class probabilities in the treatment (e.g., a marketing campaign) and control groups. Uplift modeling aims at identifying groups on which a predetermined action will have the most positive effect.

Assessing model performance is complex for uplift modeling, as the actual
value of the response, that is, the \textit{true} uplift, is unknown at the individual subject level. To overcome this limitation, one can assess model performance by comparing groups of observations. This is done through the Qini coefficient [Radcliffe, 2007], which plays a similar role as the Gini coefficient [Gini, 1997] in Economics. The Qini coefficient is a single statistics drawn form the Qini curve. This latter object is a generalization of the Lorenz curve [Lorenz, 1905] traditionally used in direct marketing for response models.

As in all regression-based modeling, an issue in uplift modeling is the ease of interpretation of the results. The model becomes harder to interpret when the number of potential explanatory variables, that is the \textit{dimension} of the explanatory variables increases. When the variable dimension is small, knowledge-based approaches to select the optimal set of variables can be effectively applied. When the number of potentially important variables is too large, it becomes too time-consuming to apply a manual variable selection process. In this case one may consider using automatic subset selection tools. Variable selection is an important step. It reduces the dimension of the model, avoids overfitting, and improves model stability and accuracy [Guyon and Elisseeff, 2003]. Well-known variable selection techniques such as forward, backward, stepwise [Montgomery et al., 2012], stagewise [Hastie et al., 2007], lasso [Tibshirani, 1996], and LARS [Efron et al., 2004b], among others, are not designed for uplift models. One might need to adapt them to perform variable selection in this context.

We propose a new way to perform model selection in uplift regression models. Our methodology is based on the maximization of a modified version of the Qini coefficient, \textit{the adjusted Qini}, that we introduced in Section 2.1. Because model selection corresponds to variable selection, the task is haunting and intractable if done in a straightforward manner when the number of variables to consider is large, e.g. $p \approx 100$, like in the case of the insurance data. To realistically search for a good model, we conceived a searching method based on an efficient exploration of the regression coefficients space combined with a lasso penalization of the log-likelihood. There is no explicit analytical expression for the adjusted Qini surface (nor for the Qini curve), so unveiling it is not easy. Our idea is to gradually uncover the adjusted Qini surface in a manner inspired by surface response designs. The goal is to find the global maximum or a reasonable local maximum of the adjusted Qini by exploring the surface near optimal values of the coefficients. These coefficient values are given by maximizing the lasso penalized log-likelihood. The exploration is done using Latin hypercube sampling structures [McKay et al., 2000] centered in a sequence of penalized estimates of the coefficients.

The rest of the paper is organized as follows. We first present the current
uplift models in Section 2 and Section 3 introduces the notation and details of Qini-based uplift regression. Sections 4 and 5 present the computational results of the proposed methodology on synthetic and real datasets. Final remarks and conclusion are given in Section 6.

2. Uplift modeling. Let $Y$ be the 0-1 binary response variable, $T$ the 0-1 treatment indicator variable and $X_1, \ldots, X_p$ the explanatory variables (predictors). The binary variable $T$ indicates if a unit is exposed to treatment ($T = 1$) or control ($T = 0$). Suppose that $n$ independent units are observed $\{(y_i, x_i, t_i)\}_{i=1}^n$, where $x_i = (x_{i1}, \ldots, x_{ip})$ are realisations of the predictors random variables. For $i = 1, \ldots, n$, an uplift model estimates

$$u(x_i) = \Pr(Y_i = 1 \mid x_i, T_i = 1) - \Pr(Y_i = 1 \mid x_i, T_i = 0),$$

where the notation $\Pr(Y_i = y_i \mid x_i, T_i = t_i)$ stands for the corresponding conditional probability. Uplift modeling was formally introduced in Radcliffe and Surry [1999] under the appellation of differential response modeling where a thorough motivation and several practical cases promoted uplift modeling in comparison with common regression or basic tree-based methods that were used to predict the probability of success for the treatment group. They showed that conventional models, which were referred to as response models, did not target the people who were the most positively influenced by the treatment. In [Radcliffe and Surry, 1999] and [Hansotia and Rukstales, 2002], the methods introduced are tree-based algorithms similar to CART [Breiman et al., 1984], but using modified split criteria that suited the uplift purpose. The method proposed by Hansotia and Rukstales [2002] uses the uplift’s absolute difference $\Delta = |u_l - u_r|$, where $u_l, u_r$ are the observed uplifts in the left and right child nodes, respectively. It is also possible to use the difference in node sizes as some sort of penalty term to adjust the differences in uplift [Radcliffe and Surry, 2011]. Other split criteria proposed in the literature are based on the $\chi^2$ statistic [Su et al., 2009, Radcliffe and Surry, 2011], which is usually a function of $\Delta^2$. All these splitting criteria rely on maximizing heterogeneity in treatment effects ($\Delta$).

2.1. Adjusted Qini. Evaluating uplift models requires the construction of the Qini curve and the computation of the Qini coefficient [Radcliffe, 2007]. The motivation to consider the Qini curve comes from the fact that a good model should be able to select individuals with highest uplift first. More explicitly, for a given model, let $\hat{u}_{(1)} \geq \hat{u}_{(2)} \geq \ldots \geq \hat{u}_{(n)}$ be the sorted predicted uplifts. Let $\phi \in [0, 1]$ be a given proportion and let $N_{\phi} = \{i : \hat{u}_i \geq \hat{u}_{(\lfloor \phi n \rfloor)}\} \subset \{1, \ldots, n\}$ be the subset of individuals with the $\phi n \times 100\%$
highest predicted uplifts $\hat{u}_i$ (here $\lceil s \rceil$ denotes the smallest integer larger or equal to $s \in \mathbb{R}$). Because $N_\phi$ is a function of the predicted uplifts, $N_\phi$ is a function of the fitted model. For a parametric model such as (9), $N_\phi$ is a function of the model’s parameters estimates, and should be denoted $N_\phi(\hat{\theta})$. To simplify the notation, we prefer to omit this specification.

As a function of the fraction of population targeted $\phi$, the relative incremental uplift is defined as

$$g(\phi) = \left( \sum_{i \in N_\phi} y_i t_i - \sum_{i \in N_\phi} y_i (1 - t_i) \left\{ \sum_{i \in N_\phi} t_i / \sum_{i \in N_\phi} (1 - t_i) \right\} \right) / \sum_{i = 1}^n t_i,$$

where $\sum_{i \in N_\phi} (1 - t_i) \neq 0$, with $g(0) = 0$. The incremental uplift has been normalized by the number of subjects treated in $N_\phi$. Note that $g(1)$ is the overall sample observed uplift, that is

$$g(1) = \left( \frac{\sum_{i = 1}^n y_i t_i}{\sum_{i = 1}^n t_i} \right) - \left( \frac{\sum_{i = 1}^n y_i (1 - t_i)}{\sum_{i = 1}^n (1 - t_i)} \right).$$

The Qini curve is constructed by plotting $g(\phi)$ as a function of $\phi \in [0, 1]$. This is illustrated in Figure 1. The curve can be interpreted as follows. The
$x$-axis represents the fraction of targeted individuals and the $y$-axis shows the incremental number of positive responses relative to the total number of targeted individuals. The straight line between the points $(0, 0)$ and $(1, g(1))$ in Figure 1 represents a benchmark to compare the performance of the model to a strategy that would randomly target subjects. In other words, when the strategy is to treat individuals randomly, if a proportion $\phi$ of the population is treated, we expect to observe an uplift equal to $\phi$ times the global uplift.

The Qini coefficient $q$ is a single index of model performance. It is defined as the area between the Qini curve and the straight line

$$q = \int_0^1 Q(\phi) \, d\phi = \int_0^1 \{g(\phi) - \phi g(1)\} \, d\phi,$$

where $Q(\phi) = g(\phi) - \phi g(1)$. This area can be numerically approximated using a Riemann method such as the trapezoid rule formula: the domain of $\phi \in [0, 1]$ is partitioned into $J$ panels, or $J + 1$ grid points $0 = \phi_1 < \phi_2 < \ldots < \phi_{J+1} = 1$, to approximate the Qini coefficient $q$ by its empirical estimation

$$\hat{q} = \frac{1}{2} \sum_{j=1}^J (\phi_{j+1} - \phi_j) \{Q(\phi_{j+1}) + Q(\phi_j)\} \times 100\%.$$  

In general, when comparing several models, the preferred model is the one with the maximum Qini coefficient [Radcliffe, 2007].

Another visualization associated with uplift model validation is based on the observed uplifts in each of the $J$ bins used to compute the Qini coefficient: a good model should induce a decreasing disposition of the observed uplifts in these bins. Figure 2 illustrates good and bad uplift models as barplots of observed uplifts associated with each of the $J$ bins. A decreasing disposition of the uplift values in the $J$ bins is an important property of an uplift model. To measure the degree to which a model does this correctly, we suggest the use of the Kendall rank correlation coefficient [Kendall, 1938]. The goal is to find a model that maximizes the correlation between the predicted uplift and the observed uplift. Let $B_k$ denote the $k$th bin $(\phi_k, \phi_{k+1}] \subseteq (0, 1]$, $k = 1, \ldots, J$. The Kendall’s uplift rank correlation is defined as

$$\rho = \frac{2}{J(J-1)} \sum_{i<j} \text{sign}(\bar{u}_i - \bar{u}_j) \text{sign}(\tilde{u}_i - \tilde{u}_j),$$

where $\bar{u}_k$ is the average predicted uplift in bin $B_k$, $k = 1, \ldots, J$, and $\tilde{u}_k$ is the observed uplift in the same bin $B_k$, that is,

$$\bar{u}_k = \frac{\sum_{i \in B_k} y_i t_i / \sum_{i \in B_k} t_i - \sum_{i \in B_k} y_i (1 - t_i) / \sum_{i \in B_k} (1 - t_i) }{ \sum_{i \in B_k} t_i / \sum_{i \in B_k} (1 - t_i) }.$$
Fig 2. Theoretical predicted uplift barplots with 5 panels corresponding to two different models. A good model should order the observed uplift from highest to lowest. The Kendall’s uplift rank correlation is $\rho = 1$ for the left barplot and $\rho = 0.6$ for the right barplot.

From a business point of view, this statistic and the associated barplot are easier to interpret than the Qini and the Qini curve. However, we do not advise the use of $\rho$ alone for model selection. If two models have the same $\hat{q}$, the preferred one should be the one with the highest $\rho$. But, when two models $\hat{q}$ differ, it is not clear that the preferred one should be the one with the highest $\rho$. In Figure 3, we show an example of two models where the one in the left has a perfect Kendall’s uplift rank correlation ($\rho = 1$) but with a Qini coefficient much smaller than the model on the right panel. In this scenario, the model with $\rho = 0.8$ may be preferred since its Qini coefficient is much higher.

We propose an appropriate combination of (4) and (5): the \textit{adjusted Qini coefficient} which is given by

\begin{equation}
\hat{q}_{\text{adj}} = \rho \max\{0, \hat{q}\}.
\end{equation}

Maximizing the adjusted Qini coefficient maximizes the Qini coefficient and simultaneously promotes grouping the individuals in decreasing uplift bins, which in turn result in concave Qini curves. In fact, the observed uplift $\bar{u}_k$ in the $k$th bin $B_k$ can be interpreted as the Qini curve’s slope between two points $\left(\phi_k, g(\phi_k)\right)$ and $\left(\phi_{k+1}, g(\phi_{k+1})\right)$, $k \in 1, \ldots, J$. This is shown in Section A of the appendix. For the moment, suppose that $\bar{u}_k \approx \nabla g_k \overset{\text{def}}{=} \{g(\phi_{k+1}) - g(\phi_k)\}/(\phi_{k+1} - \phi_k)$. That is, $\bar{u}_k$ approximates the derivative of the Qini curve at $\phi_k$. Maximizing the Kendall’s uplift rank correlation pushes the sequence $\{\bar{u}_k\}$ to be decreasing. In other words, it pushes $\text{sign}(\bar{u}_k - \bar{u}_{k+1})$ to be positive. Note that $\bar{u}_k - \bar{u}_{k+1} \approx \nabla g_k - \nabla g_{k+1}$, which can be seen as an
estimate of the second derivative of the Qini curve. Then, maximizing the Kendall’s uplift rank correlation is pushing for concavity in the Qini curve.

A note on the estimation of the Qini curve. The number of bins $J$ may be seen as a hyper-parameter. Its choice will certainly affect the computation of the adjusted Qini coefficient. In practice, the sample is divided into quantiles ($J = 5$) or deciles ($J = 10$). In order to have a hint on what adequate values for $J$ are, suppose that the relative incremental uplift function $g(\phi)$ is twice-differentiable, with bounded second derivative. Consider the trapezoid rule approximation to the integral $q$ based on $J$ bins. Let us assume that the bin sizes are proportional to $1/J$. It is well-known that under these assumptions the error of the approximation is order $O(1/J^2)$. Since $g(\cdot)$ is unknown, one need to estimate it with data. Let $\hat{g}_j$ be the estimate of $g(\phi_j)$, $j = 1, \ldots, J$. Suppose that $\hat{g}_j$ is obtained as a mean of $n/J$ random variables observed in the $j$-th bin. We suppose that these random variables are independent and identically distributed with mean $g(\phi_j)$ and a certain finite variance. The weak law of large numbers says that $\hat{g}_j$ converges to $g(\phi_j)$, and the error in this approximation is of order $O(J/\sqrt{n})$. It turns out that we need $J$ to minimize $\kappa_1/J^2 + \kappa_2J/\sqrt{n}$, where $\kappa_1, \kappa_2$ are constants. The solution is $J = O(n^{1/6})$. So, for example, if $n \approx 1000$, then the optimal $J \approx 3$. Hence, the usual values of $J = 5$ and $J = 10$ seem reasonable to estimate the Qini [Radcliffe, 2007].

2.2. Brief overview of previous work on uplift modeling. The intuitive approach to uplift modeling is to build two separated classification models. Hansotia and Rukstales [2001] used the two-model approach which consists in direct subtraction of models for the treated and untreated groups. The
QINI-BASED UPLIFT REGRESSION

asset of this technique is its simplicity. However, in many cases this approach performs poorly [Radcliffe and Surry, 2011]. Both models focus on predicting the class probabilities instead of making the best effort to predict the uplift, i.e., the difference between two probabilities. General discussions following differential response modeling and the two-model approach appeared in Hansotia and Rukstales [2002] where the technique known as incremental value modeling was introduced. This uses the difference in response rates in the two groups (treatment and control) as the split criterion of a regression tree. Also, Lo [2002] introduced the true lift modeling using a single standard logistic regression model which explicitly added interaction terms between each explanatory variable and the treatment indicator. The interaction terms measure the additional effect of each explanatory variable because of treatment. The model yields an indirect estimation of the causal effect by subtracting the corresponding prediction probabilities, which are obtained by respectively setting the treatment indicator variable to treated and control in the fitted model. The disadvantage with this solution is that it is not optimized with respect to the goodness-of-fit measures designed for uplift. Instead, the parameters are estimated with respect to the likelihood. Our results show that estimating the regression parameters by maximizing the adjusted Qini significantly improves the Qini prediction performance of uplift models.

Most current approaches that directly model the uplift causal effect are adaptations of classification and regression trees [Breiman et al., 1984]. Rzepakowski and Jaroszewicz [2010] propose a tree-based method based on generalizing classical tree-building split criteria and pruning methods. The approach is based on the idea of comparing the distributions of outcomes in treatment and control groups, using a divergence statistic, such as the Kullback-Leibler divergence or a modified Euclidean distance [Rzepakowski and Jaroszewicz, 2012, Guelman et al., 2012, Rzepakowski and Jaroszewicz, 2010]. Another non-parametric method is discussed in [Alemi et al., 2009, Su et al., 2012]. Therein the uplift is estimated from the nearest neighbors containing at least one treated and one control observation. This method quickly becomes computationally expensive when dealing with large datasets, because the entire dataset has to be stored in order to predict the uplift for new observations. For a more detailed overview of the uplift modeling literature, the reader is referred to the works of Kane et al. [2014], Gutierrez and Gérardy [2017] and Devriendt et al. [2018].

From a complexity point of view, parametric models are simpler than non-parametric ones such as regression trees, because for parametric models, the number of parameters is kept small and fixed. Although, for many
analysts prediction is the main target, from a business point of view, model interpretation is very important. Knowing which variables and how these variables discriminate between groups of clients is one of the main goal of uplift modeling for marketing. For these reasons, in this work we focus on parametric models. We develop our methodology for the logistic regression since interpretation of the odds ratios is well-known. However, our estimation procedure can be easily generalized to other parametric models.

3. Qini-based logistic regression for uplift. Logistic regression is a well-known parametric model for binary response variables. Given a \( p \)-dimensional predictor vector \( \mathbf{x}_i, i \in \{1, \ldots, n\} \), logistic intercept \( \theta_o \in \mathbb{R} \), and logistic regression coefficients \( \mathbf{\beta} \in \mathbb{R}^p \), the model is

\[
p_i = p_i(\theta_o, \mathbf{\beta}) = \Pr(Y_i = 1 \mid \mathbf{x}_i, \theta_o, \mathbf{\beta}) = \left(1 + \exp\{-\theta_o + \mathbf{x}_i^\top \mathbf{\beta}\}\right)^{-1}
\]

or, equivalently, \( \logit(p_i) = \theta_o + \mathbf{x}_i^\top \mathbf{\beta} \), where \( \logit(p_i) = \log\{p_i/(1-p_i)\} \).

Throughout the paper, the superscript \( ^\top \) stands for the transpose of a column vector or matrix. In the uplift context, one need to add explicit interaction terms between each explanatory variable and the treatment indicator. Let \( \gamma \) denote the treatment effect, \( \mathbf{\beta} \), the vector of main effects, \( \mathbf{\delta} \), the vector of interactions effects, and \( \theta_o \), the intercept. The model is

\[
p_i(\theta_o, \theta) = \Pr(Y_i = 1 \mid \mathbf{x}_i, t_i, \theta_o, \theta) = \left(1 + \exp\{-\theta_o + \gamma t_i + \mathbf{x}_i^\top [\mathbf{\beta} + t_i \mathbf{\delta}]\}\right)^{-1},
\]

where \( \theta = (\mathbf{\beta}, \gamma, \mathbf{\delta}) \), denotes all model parameters except for the intercept \( \theta_o \). The likelihood function associated with the uplift model is

\[
\mathcal{L}(\theta_o, \theta) = \prod_{i=1}^{n} p_i(\theta_o, \theta)^{y_i} \{1 - p_i(\theta_o, \theta)\}^{1-y_i},
\]

where \( \{y_i : i = 1, \ldots, n\} \) are the observed response variables. The maximum likelihood estimates of \( (\theta_o, \theta) \) will be denoted by \( (\hat{\theta}_o, \hat{\theta}) \), with \( \hat{\theta} = (\hat{\mathbf{\beta}}, \hat{\gamma}, \hat{\mathbf{\delta}}) \).

The predicted uplift associated with the covariates vector \( \mathbf{x}_{n+1} \) of a future individual is estimated by

\[
\hat{u}(\mathbf{x}_{n+1}) = \left(1 + \exp\{-(\hat{\theta}_o + \hat{\gamma} + \mathbf{x}_{n+1}^\top [\hat{\mathbf{\beta}} + \hat{\mathbf{\delta}}])\}\right)^{-1} - \left(1 + \exp\{-(\hat{\theta}_o + \mathbf{x}_{n+1}^\top \hat{\mathbf{\beta}})\}\right)^{-1}.
\]

We propose to select a regression model that maximizes the adjusted Qini coefficient. To realistically search for a good model, we conceived a searching method based on Latin hypercube sampling of the regression coefficients space combined with a lasso penalization of the log-likelihood. The procedure is explained in the following sections.
3.1. *Estimation of the Qini maximizer.* Because the adjusted Qini function is not straightforward to optimize with respect to the parameters, one needs to explore the parameters space in order to find the maximum of the adjusted Qini.

Latin hypercube sampling (LHS) is a statistical method for quasi-random sampling based on a multivariate probability law inspired by the Monte Carlo method [McKay et al., 2000]. The method performs the sampling by ensuring that each sample is positioned in a space $\Omega$ of dimension $p$ as the only sample in each hyperplane of dimension $p - 1$ aligned with the coordinates that define its position. Each sample is therefore positioned according to the position of previously positioned samples to ensure that they do not have any common coordinates in the $\Omega$ space. When sampling a function of $p$ variables, the range of each variable is divided into $M$ equally probable intervals. $M$ sample points are then placed to satisfy the Latin hypercube requirements; this forces the number of divisions, $M$, to be equal for each variable. Also this sampling scheme does not require more samples for more dimensions (variables); this independence is one of the main advantages of this sampling scheme. We use LHS to find the coefficient parameters that maximize the adjusted Qini. The procedure to search for the Qini maximizer is explained next. It is based on the lasso penalized likelihood and several LHS structures.

3.1.1. *Penalized log-likelihood.* In the context of linear regression, the effectiveness of penalization has been amply supported practically and theoretically in several studies. In order to decrease the mean squared error of least squares estimates, ridge regression [Hoerl and Kennard, 1970] has been proposed as a trade-off between bias and variance. This technique adds an $L_2$-norm penalization term to the least squares loss. The lasso (least absolute shrinkage and selection operator) penalization technique [Tibshirani, 1996] uses an $L_1$-norm penalization which sets some of the regression coefficients to zero (sparse selection) while shrinking the rest. The elastic net penalization technique [Zou and Hastie, 2005] linearly combines the $L_1$ and $L_2$-norms to provide better prediction in the presence of collinearity. Other penalization techniques such as scad [Fan and Li, 2001] and bridge regression [Frank and Friedman, 1993], offer interesting theoretical properties, including consistency.

Here, we focus on sparse estimation of the coefficients. That is, the selection of a small subset of features to predict the response. This is often achieved with a $L_1$-norm penalization. Given $\lambda \in \mathbb{R}^+$, in the context of linear regression, the lasso penalization [Tibshirani, 1996] finds the esti-
mate of the coefficients $\hat{\beta}(\lambda)$ that maximizes the penalized log-likelihood, say $\ell(\beta) + \lambda \sum_{j=1}^{p} |\beta_j|$. Setting the penalization constant $\lambda = 0$ returns the least squares estimates which performs no shrinking and no selection. For $\lambda > 0$, the regression coefficients $\hat{\beta}(\lambda)$ are shrunk towards zero, and some of them are set to zero (sparse selection). Friedman et al. [2007] proposed a fast pathwise coordinate descent method to find $\hat{\beta}(\lambda)$, using the current estimates as warm starts. In practice, the value of $\lambda$ is unknown. Cross-validation is often used to search for a good value of the penalization constant. The least angle regression (or LARS algorithm) efficiently computes a path of values of $\hat{\beta}(\lambda)$ over a sequence of values of $\lambda = \lambda_1 < \cdots < \lambda_j < \cdots < \lambda_{\min(n,p)}$, for which the parameter dimension changes [Efron et al., 2004a]. The entire sequence of steps in the LARS algorithm with $p < n$ variables requires $O(p^3 + np^2)$ computations, which is the cost of a single least squares fit on $p$ variables. Extensions to generalized linear models with nonlinear loss functions require some form of approximation. In particular, for the logistic regression case, which is our model of interest, Friedman et al. [2010] extend the pathwise coordinate descent algorithm [Friedman et al., 2007] by first, approximating the log-likelihood (quadratic Taylor expansion about current estimates), and then using coordinate descent to solve the penalized weighted least-squares problem. The algorithm computes the path of solutions for a decreasing sequence of values for $\lambda = \lambda_{\min(n,p)} > \cdots > \lambda_j > \cdots > \lambda_1$, starting at the smallest value for which the entire vector $\hat{\beta} = 0$. The algorithm works on large datasets, and is publicly available through the R package glmnet [Friedman et al., 2009], which we use in this work. In what follows, we will refer to the sequence of regularizing constant values given by glmnet as the logistic-lasso sequence.

3.1.2. Qini-optimized uplift regression. Recall the uplift model likelihood given in (9). The vector of parameters $\theta = (\beta, \gamma, \delta)$ is a $p' = (2p + 1)$-dimensional vector. Because of the considerations mentioned in the previous sections, in order to select an appropriate sparse model for uplift, we adapt the lasso algorithm to explore a relatively small set of reasonable models, so as to avoid an exhaustive model search. The penalized uplift model log-likelihood is given by

$$
\ell(\theta_o, \theta \mid \lambda) = \sum_{i=1}^{n} \left( y_i \log \left\{ p_i(\theta_o, \theta) \right\} + (1 - y_i) \log \left\{ 1 - p_i(\theta_o, \theta) \right\} \right) + \lambda \| \theta \|_1,
$$

where $p_i(\theta_o, \theta)$ is as in (8), and $\| \cdot \|_1$ stands for the $L_1$-norm. For any given $\lambda$, the parameters that maximize the penalized log-likelihood (10) are denoted
QINI-BASED UPLIFT REGRESSION

15

by

(\hat{\theta}_o(\lambda), \hat{\theta}(\lambda)) = \arg\max_{\hat{\theta}_o, \theta} \ell(\theta_o, \theta \mid \lambda).

Applying the pathwise coordinate descent algorithm to the uplift model, we get a sequence of critical penalization values $\lambda_1 < \cdots < \lambda_{\min\{n,p'\}}$ and corresponding model parameters $\{(\hat{\theta}_o(\lambda_j), \hat{\theta}(\lambda_j))\}_{j=1}^{\min\{n,p'\}}$ associated with different model dimensions $m \in \{1, \ldots, p'\}$.

3.1.3. The LHS search. For each $\lambda_j$, $j = 1, \ldots, \min\{n,p'\}$, we generate a LHS comprising $L$ points $\{\hat{\theta}(\lambda_j)\}_{l=1}^{L}$ in the neighborhood of $\hat{\theta}(\lambda_j)$, and evaluate the adjusted Qini on each of these points. The optimal coefficients are estimated as those coefficients among the $(\min\{n,p'\} \times L)$ LHS points that maximize the adjusted Qini. Figure 4 illustrates the procedure.

![Figure 4](image-url)

**Fig 4.** Example of a 2-dimensional Latin hypercube sampling for $\theta = (\theta_1, \theta_2)$. For the regularization constants $\lambda_1, \lambda_2$, there are two penalized estimates $\hat{\theta}(\lambda_j)$, $j = 1, 2$ (solid symbols in the left panel). The idea is to sample points (outlined symbols in the left panel) centered at $\hat{\theta}(\lambda_j)$, $j = 1, 2$ and then to compute $\hat{q}_{\text{adj}}$ for all these points (right panel).

3.1.4. A simpler estimate of the Qini-based uplift regression parameters. We also consider a simpler two-stage procedure to find a good uplift model. This one is based only on the penalized log-likelihood and does not require the posterior LHS-based search for the optimal coefficients. Let $\hat{q}_{\text{adj}}(\lambda)$ be the adjusted Qini coefficient associated with the model with parameters $(\hat{\theta}_o(\lambda), \hat{\theta}(\lambda))$. The first stage of the procedure solves

(12) \[ \hat{\lambda} = \arg\max \{ \hat{q}_{\text{adj}}(\lambda_j) : j = 1, \ldots, \min\{n,p'\} \}, \]
where as before, the sequence $\lambda_1 < \cdots < \lambda_{\min\{n,p'\}}$ is the logistic-lasso sequence. On the second-stage, a reduced model that only include those explanatory variables associated with non-zero entries of the estimated parameter $\hat{\theta}(\hat{\lambda})$ is fitted without penalization, that is, with $\lambda$ set to zero. The parameters are estimated with maximum likelihood. This yields the selected model. In our simulations, this model performs well. It also serves to show that the value of the penalization parameter $\hat{\lambda}$ that maximizes the Qini or adjusted Qini, is not necessarily the same as the one that maximizes the penalized log-likelihood.

4. Simulations. We conduct a simulation study to examine the performance of Qini-based uplift regression. More specifically, we compare the different proposed parameters estimation methods by varying both the complexity of the data, and the number of predictors in the model. In order to create realistic scenarios, we based our artificial data generation on the home insurance policy data described in the introduction. We take advantage of the opportunity to have real data in order to generate realistic scenarios. We proceed as follows. First, we fit a non-parametric model on a random sample $D$ of the home insurance policy data. Based on the resulting model, we can extract the probabilities 

$$p_1(x) = \Pr(Y = 1 \mid x, T = 1), \text{ and } p_0(x) = \Pr(Y = 1 \mid x, T = 0),$$

for any given value $x$. Then, we use these probabilities to generate synthetic data. We start by creating a bootstrap sample $S$ of size $n_S$ from $D$. For each observation $x_i \in S$, we generate a random vector $\tilde{y}_i = (\tilde{y}_{i0}, \tilde{y}_{i1})$, where $\tilde{y}_{i0}$ is the binary outcome of a Bernoulli trial with success probability $p_0(x_i)$, and $\tilde{y}_{i1}$ is the binary outcome of a Bernoulli trial with success probability $p_1(x_i)$, $i = 1, \ldots, n_S$. The augmented synthetic dataset $\{(x_i, t_i, \tilde{y}_i)\}_{i=1}^{n_S}$, which we are going to denote again by $S$, is the data of interest in the simulation. For each simulated dataset, we implement the following models:

(a) a multivariate logistic regression without penalization as in (9). This is the baseline model, and we will refer to it as Baseline.

(b) our Qini-based uplift regression model that uses several LHS structures to search for the optimal parameters (see Section 3.1.2). We denote this model by $Q+\text{LHS}$.

(c) our Qini-based uplift regression model that uses the simpler estimate of the regression parameters as explained in Section 3.1.4. We denote this model by $Q+\text{lasso}$.

Note that our $Q+\text{LHS}$ method is a derivative free optimization procedure. Another derivative free optimization method is the well-known Nelder-Mead
method [Nelder and Mead, 1965]. In order to obtain benchmarks for the LHS search, we implement the following Nelder-Mead Qini-based uplift regression models:

(d) \textit{Base+NM}, which initializes the Nelder-Mead algorithm with the maximum likelihood estimates (the \textit{Baseline} model solution) and which searches for coefficients that maximize the adjusted Qini coefficient.

(e) \textit{Q+NM}, which initializes the Nelder-Mead algorithm with coefficients from the lasso-sequence (the first-stage of \textit{Q+lasso}) and which searches for coefficients that maximize the adjusted Qini coefficient.

\textit{Data generation.} As discussed in Section 2.2, several tree-based methods have been suggested in the uplift literature. Here, we use the uplift random forest [Guelman et al., 2012] as the data generating process. We chose this method due to its simplicity, and because it is readily available in R through the package \texttt{uplift} [Guelman, 2014]. Algorithm 1 describes the associated methodology.

\begin{algorithm}
ange{Uplift Random Forest [Guelman et al., 2012]}
\begin{algorithmic}
\State $B \leftarrow \text{number of bootstrap samples}$
\For {$b = 1$ to $B$}
\State Draw a bootstrap sample of size $n_S$ with replacement from the data
\State Fit an uplift decision tree $T_b$ to the bootstrap data
\EndFor
\State Output the ensemble of uplift trees $T_b; b = \{1, 2, \ldots, B\}$ and the predicted probabilities $\Pr(Y = 1 \mid x, T = 1)$ and $\Pr(Y = 1 \mid x, T = 0)$ obtained by averaging the predictions of the individual trees in the ensemble.
\end{algorithmic}
\end{algorithm}

In our simulations, we vary two parameters: the depth of the trees used to fit the uplift random forests, and the number of variables $k$ considered when fitting the uplift logistic models. Algorithm 2 details the procedure.

We define 21 scenarios by varying two parameters: (i) the depth of the uplift random forest trees used to generate the synthetic data is either 1, 2 or 3, and (ii) the number of total covariates $k$ considered to build the forest model, $k \in \{10, 20, 30, 50, 75, 90, 97\}$. For scenarios 1-7, the depth is 1, and we vary $k$; for scenarios 8-14, the depth is 2; and for scenarios 15-21, the depth is 3. Each scenario was replicated 100 times.

The sample means of $\hat{q}(3)$, and $\hat{q}_{adj}(7)$ and their corresponding standard errors are reported in Tables 3, and 4, respectively. Since the conclusions are similar for the three tree depths, we report only the results associated with depth 3, that is, for the most complex model. For each comparison group, we also report the corresponding performance of an uplift random forest (RF) fitted to the synthetic data using all available predictors (i.e., $k = 97$).
Algorithm 2 Simulations
1: \( M \leftarrow \) number of simulations
2: for \( m = 1 \) to \( M \) do
3: Draw a stratified sample \( D_m \) of size \( n_S \) without replacement from the data
4: Fit an uplift random forest of a given tree depth to the sampled data \( D_m \)
5: Generate a complete (i.e., including the binary responses) synthetic data \( S_m \) with data \( D_m \)
6: Fit an uplift random forest of the same given tree depth to the synthetic data \( S_m \) using all \( p \) predictors
7: for each \( k \) do
8: Sample \( k \leq p \) random predictors for modeling
9: Fit the different uplift logistic models with \( k \) predictors on \( S_m \)
10: Output the average and standard errors of the metrics for each method

Table 3
Qini coefficient (\( \hat{q} \)) averaged over 100 simulations. Standard-errors are reported in parenthesis. The RF model (with \( k = 97 \) and depth=3) performance is 1.60 (0.039).
\( n = 5000 \) observations.

<table>
<thead>
<tr>
<th>( k )</th>
<th>Baseline</th>
<th>Q+lasso</th>
<th>Q+LHS</th>
<th>Base+NM</th>
<th>Q+NM</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.51 (0.023)</td>
<td>0.56 (0.020)</td>
<td>0.76 (0.021)</td>
<td>0.64 (0.019)</td>
<td>0.68 (0.021)</td>
</tr>
<tr>
<td>20</td>
<td>0.74 (0.020)</td>
<td>0.83 (0.022)</td>
<td>1.03 (0.024)</td>
<td>0.93 (0.025)</td>
<td>0.95 (0.020)</td>
</tr>
<tr>
<td>30</td>
<td>0.94 (0.023)</td>
<td>1.01 (0.023)</td>
<td>1.20 (0.023)</td>
<td>1.04 (0.024)</td>
<td>1.13 (0.026)</td>
</tr>
<tr>
<td>50</td>
<td>1.09 (0.039)</td>
<td>1.19 (0.025)</td>
<td>1.48 (0.026)</td>
<td>1.23 (0.031)</td>
<td>1.35 (0.029)</td>
</tr>
<tr>
<td>75</td>
<td>0.97 (0.073)</td>
<td>1.37 (0.036)</td>
<td>1.48 (0.033)</td>
<td>1.35 (0.059)</td>
<td>1.49 (0.056)</td>
</tr>
<tr>
<td>90</td>
<td>1.00 (0.084)</td>
<td>1.36 (0.045)</td>
<td>1.54 (0.036)</td>
<td>1.40 (0.071)</td>
<td>1.47 (0.064)</td>
</tr>
<tr>
<td>97</td>
<td>0.83 (0.083)</td>
<td>1.36 (0.055)</td>
<td>1.59 (0.037)</td>
<td>1.31 (0.088)</td>
<td>1.41 (0.079)</td>
</tr>
</tbody>
</table>

In Table 3, we compare the performance of the models according to the Qini coefficient \( \hat{q} \). We observe that performing variable selection driven by the adjusted Qini coefficient (\( Q+lasso \)) significantly improves the performance of the baseline model. As expected, the models using a LHS-driven optimization perform better. The performance of \( Q+lasso \) is similar to the Base+NM performance, and is slightly lower than \( Q+LHS \) and \( Q+NM \). Using the lasso-sequence in order to initialize the posterior searches, that is using \( Q+LHS \) or \( Q+NM \), improves the performance of the \( Q+lasso \) and Base+NM models. However, for the \( Q+NM \) solution, the standard error of the Qini coefficient increases with \( k \). It is almost twice the standard errors from \( Q+LHS \) for \( k \geq 75 \). Using all predictors (\( k = 97 \)) enables the \( Q+LHS \) models to achieve the same performance as the RF model.
In Table 4, we compare the main statistic of interest, that is, the adjusted Qini coefficient. These results corroborate the findings from the previous table. Guiding the variable selection by this statistic leads to significant improvements from the results of the baseline model. Similarly, estimating the parameters with a derivative-free maximization of the adjusted Qini coefficient improves the performance of the models in comparison to maximum likelihood estimation. The best results are obtained with models that make use of the lasso-sequence in order to explore the space of the parameters (Q+LHS and Q+NM). As in Table 3, the Q+LHS models give the best results. Moreover, when using all available predictors, the Q+LHS models outperform both the RF and the Q+NM models.

The difference in performance between Q+LHS and Q+lasso is significant. The left panel of Figure 5 display boxplots of the differences in performance between these two models in each simulation. It is clear that Q+LHS perform much better than Q+lasso most of the time. The relative performance of Q+lasso improves slightly when the number of predictors approaches the total number of predictors available ($p = 97$). The same pattern is observed in the difference of performance between Q+LHS and Base+NM (see right panel of Figure 5). This confirms the importance of the use of the lasso-sequence in order to estimate the model’s parameters.

Choosing an appropriate sparse model. Next, we compare the models selected by the Qini-based uplift regression and the classical lasso approach where the penalization constant is chosen by cross-validation on the log-likelihood. Consider the values of the logistic-lasso sequence $\lambda_1 < \cdots < \lambda_{\min\{n,p'\}}$ sorted according to the results of $Q+lasso$. That is, consider the permutation $(\pi_1, \pi_2, \ldots, \pi_{\min\{n,p'\}})$ of $(1, 2, \ldots, \min\{n, p'\})$ so that $\lambda_{\pi_{\min\{n,p'\}}} \leq \cdots \leq \lambda_{\pi_1}$, where the relation $\lambda_{\pi_i} \leq \lambda_{\pi_j}$ means that $\hat{q}_{\adj}(\lambda_{\pi_i}) \leq \hat{q}_{\adj}(\lambda_{\pi_j})$. We look at the value of $\tilde{\lambda} \in \{\lambda_1, \ldots, \lambda_{\min\{n,p'\}}\}$ that is chosen by cross-validation.
Fig 5. Comparison between Q+LHS and Q+lasso (left panel) and Q+LHS and Base+NM (right panel). Boxplots of the differences in terms of \( \hat{q}_{\text{adj}} \) as a function of the number of predictors used in the models over the 100 simulations with \( n = 5000 \) observations. The black lines represent the differences medians.

The penalty constant of the log-likelihood, and report its ranking based on the sorted Q+lasso sequence \( \lambda_{\pi_{\min(n,p')} } \leq \cdots \leq \lambda_{\pi_1} \). Comparing the two models is equivalent to check when lasso finds the “best” \( \lambda \), that is, when \( \lambda_{\pi_1} \) is equal to \( \tilde{\lambda} \). We repeated the simulation 100 times, each time using \( n = 5000 \) observations randomly selected from the full data set.

Fig 6. Barplot of the distribution of the Q+lasso rankings associated with \( \tilde{\lambda} \).

The barplot in Figure 6 shows that only 7% of the time \( \tilde{\lambda} \) also maximizes \( \hat{q}_{\text{adj}} \). Observe that 4% of the time \( \tilde{\lambda} \) is positioned 10 in the ranking, and 12% of the time, it is positioned 39. These results clearly show that choosing the penalization constant by cross-validation of the log-likelihood does not solve...
the problem of maximizing the adjusted Qini coefficient, and therefore, is not necessarily appropriate for uplift models.

5. Insurance data analysis. Recall the insurance data introduced in Section 1. The insurance company is interested in designing retention strategies to minimize its policyholders’ attrition rate. An experimental loyalty campaign was implemented, from which policies coming up for renewal were randomly allocated into one of the following two groups: treatment group, and control group. The goal of this section is to analyze the marketing campaign results so as to identify both the set of persuadable clients, and the set of clients that should not be disturbed.

Parameter estimation. We fit the Qini-based uplift regression $Q+LHS$ to the data using the methodology described in the previous sections. For comparison purposes, we also considered the model $Q+lasso$. Although, we are interested in interpretable parametric models, we also fit an uplift random forest ($RF$) as a benchmark for our comparison.

In order to choose the optimal value from the logistic-lasso sequence of penalization constant values $\{\lambda_1, \ldots, \lambda_{\min\{n,p\}}\}$, we use a 5-fold cross-validation on the adjusted Qini statistics. We compare the resulting models with the one yielded by applying the classical lasso approach, that is, with the model associated with the value of the penalization constant that maximizes the cross-validated log-likelihood. We will refer to this latter model as $MLE+lasso$. The two-stage approach was used in all the cases. The first stage estimates the best $\lambda$ in the logistic-lasso sequence by cross-validation. The second stage fits the non penalized logistic regression model with the subset of selected variables.

For the $Q+LHS$ model, for each $\lambda_j$, we perform a LHS search to directly maximize the adjusted Qini coefficient (see Appendix B for more details). In this case, applying the LHS search leads to the selection of the model associated with the penalization constant $\approx 3 \times 10^{-5}$, while for the $MLE+lasso$ logistic regression, it is $\approx 10^{-3}$. The number of selected variables are, respectively, 163 and 53 out of a total of 195 main and interaction effect terms.

If we use the simple search method described in Section 3.1.4, which was denoted by $Q+lasso$ in the previous section, the optimal value of the penalization constant is $\approx 4 \times 10^{-5}$. In this case, the number of selected variables is 156.

In order to have a fair comparison, we followed a process similar to the one applied to the $Q+LHS$ model to fit the uplift $RF$ model. The accuracy of a random forest can be sensitive to several training hyper-parameters: number of trees (from 10 to 200, with increments of 10 trees), maximum
depth on each tree (from 1 to 10), minimum number of observations per node (either 100, 200 or 500), and split criterion, either Euclidean distance or Kullback-Leibler divergence; see Guelman et al. [2012] for more details on the split criteria. The optimal RF hyper-parameters were those that maximized the adjusted Qini coefficient with a 5-fold cross-validation over the grid given by the possible values of the hyper-parameters. Hence, the chosen RF was composed of 100 trees of maximum depth 3, with a minimum of 200 observations per node, with trees splitted according to the Kullback-Leibler criterion. The final RF was fitted using all available data.

Figures 7 and 8 show the performance of the models in terms of the Qini curve and the uplift barplot (Kendall’s rank correlation), respectively.

![Graph showing models comparison](image)

**Fig 7.** Performance of the final models based on the Qini curves.

As expected, the Qini-based uplift regression models outperform the classic lasso approach, that is, the MLE+lasso model, both in terms of overall adjusted Qini coefficient (see Table 5) and in terms of sorting the individuals in decreasing order of uplift (see Figure 8). Surprisingly, the non-penalized model, that is, the model whose parameters are estimated by maximizing the non-penalized likelihood, yields \( \hat{q}_{\text{adj}} = 0.85 \) (with 195 non-zero coefficients) which is larger than \( \hat{q}_{\text{adj}} \) for MLE+lasso \( \hat{q}_{\text{adj}} = 0.39 \) (see Table 5). However, we see an increase in terms of \( \hat{q}_{\text{adj}} \) for Q+lasso \( \hat{q}_{\text{adj}} = 1.02 \). Therefore, for this dataset, eliminating \( \approx 20\% \) of the coefficients yielded a better up-
lift estimation in terms of adjusted Qini. Moreover, the performance of the Qini-based uplift regression is slightly lower, but comparable to the one of the RF model ($\hat{q}_{\text{adj}} = 1.07$), even though the RF is more complex. Indeed, with 100 trees of depth 3, the RF model can both model non-linearity and interactions between covariates, which makes interpretation of the final RF hard. However, the in-sample performance is similar to our models. This is interesting because it shows that it is possible to get powerful models without losing interpretation when estimating the parameters with the adjusted Qini function.

Based on the final models, we can identify both (i) the group of clients at the top 20% of predicted uplifts, that is, the clients to pursue in the marketing, and (ii) the group of clients at the bottom 20% of predicted uplifts, that is, the clients not to disturb with any marketing. The group of clients at the top 20% of predicted uplifts provides very strong return on investment cases when applied to retention activities. For example, by only targeting the persuadable customers in an outbound marketing campaign, the contact costs and hence the return per unit spend can be dramatically improved [Radcliffe and Surry, 2011]. We observe from Table 5 that the RF model finds the highest top 20% uplift group which presents an uplift of 6.41%, while the $Q+LHS$ model finds the lowest bottom 20% uplift group which shows an uplift of approximately −6%. Note that the overall uplift is approximately −0.5%. The performance of the $Q+\text{lasso}$ model is slightly lower than the one of the $Q+LHS$ model.

Model interpretation. Because the $Q+LHS$ model is a logistic model, we can interpret the results through its coefficients. The usual approach is that of the odds ratios. For a specific variable, the odds ratio is computed by

---

**Fig 8.** Performance of the $Q+LHS$ and MLE+lasso models based on uplift Kendall’s correlations. The left barplot corresponds to $Q+LHS$ ($\rho = 1$) and the right barplot to the MLE+lasso model ($\rho = 0.8$). The observed uplift was computed as in (6).
Table 5

Uplift results of the top and bottom 20% observed uplift groups for the models MLE, RF, Q+LHS, Q+lasso, and MLE+lasso. The adjusted Qini coefficient $\hat{q}_{\text{adj}}$ is given for each model.

<table>
<thead>
<tr>
<th>Method</th>
<th>$\hat{q}_{\text{adj}}$</th>
<th>Top 20%</th>
<th>Bottom 20%</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td>0.85</td>
<td>4.98%</td>
<td>-5.87%</td>
</tr>
<tr>
<td>RF</td>
<td>1.07</td>
<td>6.41%</td>
<td>-5.36%</td>
</tr>
<tr>
<td>Q+LHS</td>
<td>1.03</td>
<td>5.94%</td>
<td>-6.02%</td>
</tr>
<tr>
<td>Q+lasso</td>
<td>1.02</td>
<td>5.76%</td>
<td>-5.93%</td>
</tr>
<tr>
<td>MLE+lasso</td>
<td>0.39</td>
<td>1.41%</td>
<td>-4.38%</td>
</tr>
</tbody>
</table>

fixing the other covariates at fixed values, such as their mean, which is what we have done here. Since the company is not interested in all the variables included in the model, we will analyze a subset with relevant interpretation for the business. In addition, for confidentiality reasons, we do not show the analysis of variables related to the insurance premium. The following variables are chosen by our model: client’s credit score, age, gender and marital status (single or not); client’s products: whether it is a single line (home) or a multi-line (automobile and home) account; client’s number of automobile policies, mortgages and residences; and whether the client’s has extra options (additional endorsements) in his/her account. For a model with $p$ variables, the odds ratio $\text{OR}_{X_j}(t)$ associated with one-unit change for a specific variable $X_j$ is given by

$$\frac{\text{Pr}(Y = 1 \mid X_j = x_j + 1, T = t) / \text{Pr}(Y = 0 \mid X_j = x_j + 1, T = t)}{\text{Pr}(Y = 1 \mid X_j = x_j, T = t) / \text{Pr}(Y = 0 \mid X_j = x_j, T = t)} = \exp(\hat{\beta}_j(x_j + 1) + \hat{\delta}_jt(x_j + 1)) / \exp(\hat{\beta}_jx_j + \hat{\delta}_jt x_j) = \exp(\hat{\beta}_j) \exp(\hat{\delta}_jt),$$

where $T$ is the treatment indicator, and where for a binary variable, such as extra options, $x_j$ is set to 0 in the above expression. When the company does not call a client ($T = 0$), the odds ratio is $\text{OR}_{X_j}(0) = \exp(\hat{\beta}_j)$ and when the company calls a client ($T = 1$), the odds ratio is $\text{OR}_{X_j}(1) = \exp(\hat{\beta}_j) \exp(\hat{\delta}_jt) = \text{OR}_{X_j}(0) \exp(\hat{\delta}_jt)$. Table 6 gives the estimated odds ratios $\text{OR}_{X_j}(0)$ and $\text{OR}_{X_j}(1)$ with 95% confidence intervals. We can see, for example, that when the company does not call a client which has extra options in his/her policy, his/her odds ratio of renewing the policy is 0.35 while when the company calls that same client, the odds ratio becomes 1.28. It is worth noting that the interaction terms between the treatment indicator and the variables Single, Residences Count and Extra Options are significant.

Next, we use the $Q+LHS$ model predictions to describe in more detail the two extreme groups found by the model (top 20% and bottom 20%
Table 6
Odds ratios and 95% confidence intervals estimated by the Qini-based uplift regression model (Q+LHS) for some of the selected variables. The symbol * indicates significant coefficients; the symbol † indicates significant interaction terms.

<table>
<thead>
<tr>
<th>Variable</th>
<th>( \exp(\beta_j) )</th>
<th>CI (95%)</th>
<th>( \exp(\beta_j + \delta_j) )</th>
<th>CI (95%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Credit Score</td>
<td>0.998</td>
<td>(0.994; 1.003)</td>
<td>1.001</td>
<td>(0.999; 1.001)</td>
</tr>
<tr>
<td>Age (Years)</td>
<td>0.995</td>
<td>(0.969; 1.023)</td>
<td>0.998</td>
<td>(0.989; 1.006)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1.297</td>
<td>(0.778; 2.163)</td>
<td>0.962</td>
<td>(0.814; 1.135)</td>
</tr>
<tr>
<td>Marital Status</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single†</td>
<td>2.759</td>
<td>(0.956; 7.963)</td>
<td>0.697</td>
<td>(0.452; 1.077)</td>
</tr>
<tr>
<td>Products</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Auto and Home</td>
<td>1.619</td>
<td>(0.586; 4.472)</td>
<td>*1.418</td>
<td>(1.017; 1.977)</td>
</tr>
<tr>
<td>Auto Policies Count</td>
<td>2.106</td>
<td>(0.653; 6.790)</td>
<td>*1.996</td>
<td>(1.368; 2.918)</td>
</tr>
<tr>
<td>Mortgage Count</td>
<td>1.381</td>
<td>(0.789; 2.418)</td>
<td>*1.366</td>
<td>(1.137; 1.642)</td>
</tr>
<tr>
<td>Residences Count†</td>
<td>0.505</td>
<td>(0.172; 1.489)</td>
<td>*1.788</td>
<td>(1.122; 2.848)</td>
</tr>
<tr>
<td>Extra Options†</td>
<td>*0.350</td>
<td>(0.141; 0.874)</td>
<td>1.276</td>
<td>(0.949; 1.715)</td>
</tr>
</tbody>
</table>

predicted uplifts). This furnishes the insurance company with typical profiles of clients that are persuadables (top 20%), and clients that should not be targeted (bottom 20%). Table 7 shows descriptive statistics of some selected predictors for both groups. A MANOVA comprising only these two groups for the selected variables, followed by ANOVA tables involving individual selected variables separately, show that all mean differences were statistically significant (p-value < 0.0005). For example, there are 30% more single clients in the do-not-disturb group than in the persuadable group. Also, 87% of persuadable clients have additional coverage in their policy while less than half of the do-not-disturb clients opt for extra coverage options.

Looking at the average profiles of persuadable and do not disturb clients, we can say that a persuadable client has a higher credit score and is slightly older than a client that should not be targeted. A persuadable client is less likely to be single and more likely to hold both company insurance products (i.e., home and auto policies). Also, this type of client holds more auto policies in his/her account, more mortgages, more residences in his/her name and is more likely to have extra coverage options. Thus, it seems that a persuadable client is a customer with many products to insure.

The correlation matrices associated with these two groups are displayed in image format in Figure 9. There are some obvious patterns that distinguish the two groups. For example, credit score is slightly correlated with client age for persuadable clients, but not for do-not-disturb clients. Client age is negatively correlated with marital status for do-not-disturb clients, but only slightly correlated for persuadables. Indeed, there are several differences in
Table 7
Profiles of the persuadables and do not disturb groups predicted by the Qini-based uplift regression model (Q+LHS) for some of the selected variables. Note that all mean differences are statistically significant (p-value < 0.0005).

<table>
<thead>
<tr>
<th></th>
<th>Persuadables</th>
<th>Do Not Disturb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of observations</td>
<td>4199</td>
<td>4200</td>
</tr>
<tr>
<td>Observed Uplift</td>
<td>5.94%</td>
<td>−6.02%</td>
</tr>
<tr>
<td>Predicted Uplift (± S.E.)</td>
<td>4.94% (±0.10%)</td>
<td>−5.45% (±0.07%)</td>
</tr>
<tr>
<td>Credit Score</td>
<td>771 (±60)</td>
<td>736 (±77)</td>
</tr>
<tr>
<td>Age (Years)</td>
<td>46.1 (±11.6)</td>
<td>41.4 (±11.8)</td>
</tr>
<tr>
<td>Gender</td>
<td>Male</td>
<td></td>
</tr>
<tr>
<td></td>
<td>50% (±50%)</td>
<td>61% (±49%)</td>
</tr>
<tr>
<td>Marital Status</td>
<td>Single</td>
<td></td>
</tr>
<tr>
<td></td>
<td>13% (±33%)</td>
<td>43% (±49%)</td>
</tr>
<tr>
<td>Products</td>
<td>Auto and Home</td>
<td></td>
</tr>
<tr>
<td></td>
<td>83% (±37%)</td>
<td>66% (±47%)</td>
</tr>
<tr>
<td>Auto Policies Count</td>
<td>1.04 (±0.64)</td>
<td>0.76 (±0.63)</td>
</tr>
<tr>
<td>Mortgage Count</td>
<td>0.63 (±0.63)</td>
<td>0.51 (±0.55)</td>
</tr>
<tr>
<td>Residences Count</td>
<td>1.15 (±0.41)</td>
<td>1.02 (±0.19)</td>
</tr>
<tr>
<td>Extra Options</td>
<td>87% (±33%)</td>
<td>41% (±49%)</td>
</tr>
</tbody>
</table>

The differences between these two groups can also be observed through the odds ratios. For any specific variable $X_j$ which takes average values $x_j^{(p)}$ (for persuadable clients), and $x_j^{(d)}$ (for do-not-disturb clients), consider the odds ratio $OR_{X_j}^{(group)}(t)$ between persuadable and do-not-disturb clients

$$
OR_{X_j}^{(group)}(t) = \frac{\Pr(Y = 1 \mid X_j = x_j^{(p)}, T = t) / \Pr(Y = 0 \mid X_j = x_j^{(p)}, T = t)}{\Pr(Y = 1 \mid X_j = x_j^{(d)}, T = t) / \Pr(Y = 0 \mid X_j = x_j^{(d)}, T = t)} = \left( \exp(\hat{\beta}_j) \exp(\hat{\delta}_j) \right)^{x_j^{(p)} - x_j^{(d)}} = \left( OR_{X_j}(t) \right)^{x_j^{(p)} - x_j^{(d)}},
$$

where $T$ is the treatment indicator. Table 8 shows these odds ratios for the two values of $T \in \{0, 1\}$. For example, if we only consider extra options, when the insurance company calls a client (i.e., $T = 1$), the odds ratio between a persuadable client (Extra Options= 87%) and a do-not-disturb client (Extra Options= 41%) is about 1.12 with a 95% confidence interval of [0.98; 1.28]. On the other hand, when the company does not call a customer
Fig 9. Correlations of selected variables of interest for Persuadable (left panel) and Do-not-disturb (right panel) clients. Positive correlations are displayed in black and negative correlations in grey color. Color intensity and the size of the circle are proportional to the correlation coefficients.

(i.e., $T = 0$), the odds ratio becomes 0.62 with a 95% confidence interval of [0.41; 0.94]. These results are quite logical in the sense that the odds of renewing the insurance policy are higher for the persuadable clients if the company calls, while the same odds are higher for the do-not-disturb clients if the company does not call.

Overall, we observe that when calling a client, the odds of renewing the insurance policy of persuadable clients are almost twice (1.81) the odds of do-not-disturb clients with a 95% confidence interval of [1.43; 2.29]. Conversely, when the company does not call a client, the odds of renewing the insurance policy of persuadable clients are half (0.52) the odds of do-not-disturb clients with a 95% confidence interval of [0.26; 1.03]. Hence, based on our model, by calling identified persuadable clients and not calling identified do-not-disturb clients in future marketing campaigns should result in increased retention rates for the company.

**Uplift prediction.** The main objective in analyzing the insurance data is to estimate the parameters of the parametric uplift model which maximizes the Qini. Based on these estimates, we were able to provide useful insights to the company. In order to prevent overfitting, we made use of 5-fold cross validation in the fitting process. However, since uplift models can also be used for predicting future clients behaviour, it is important to evaluate out-of-sample performance. In Table 5, we showed the in-sample performance.
Table 8

Odds ratios $\text{OR}^{(\text{group})}_{j}(t)$ of the persuadable compared to the do not disturb clients (Eq. 13) and 95% confidence intervals estimated by the Qini-based uplift regression model ($Q+LHS$) for some of the selected variables. The $\Delta = x^{(p)}_j - x^{(d)}_j$ column represents the difference of group means from Table 7.

<table>
<thead>
<tr>
<th></th>
<th>$\Delta$</th>
<th>Control CI (95%)</th>
<th>Called CI (95%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td></td>
<td>0.52 (0.26; 1.03)</td>
<td>1.81 (1.43; 2.29)</td>
</tr>
<tr>
<td>Credit Score</td>
<td>35</td>
<td>0.95 (0.82; 1.09)</td>
<td>1.02 (0.98; 1.07)</td>
</tr>
<tr>
<td>Age (Years)</td>
<td>4.7</td>
<td>0.98 (0.86; 1.11)</td>
<td>0.99 (0.95; 1.03)</td>
</tr>
<tr>
<td>Gender</td>
<td>Male</td>
<td>-11% 0.97 (0.92; 1.03)</td>
<td>1.00 (0.99; 1.02)</td>
</tr>
<tr>
<td>Marital Status</td>
<td>Single</td>
<td>-30% 0.73 (0.54; 1.01)</td>
<td>1.11 (0.98; 1.27)</td>
</tr>
<tr>
<td>Products</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Auto and Home</td>
<td>17%</td>
<td>1.09 (0.91; 1.29)</td>
<td>1.06 (1.00; 1.12)</td>
</tr>
<tr>
<td>Auto Policies Count</td>
<td>0.28</td>
<td>1.23 (0.89; 1.71)</td>
<td>1.21 (1.09; 1.35)</td>
</tr>
<tr>
<td>Mortgage Count</td>
<td>0.12</td>
<td>1.04 (0.97; 1.11)</td>
<td>1.04 (1.02; 1.06)</td>
</tr>
<tr>
<td>Residences Count</td>
<td>0.13</td>
<td>0.92 (0.80; 1.05)</td>
<td>1.08 (1.02; 1.15)</td>
</tr>
<tr>
<td>Extra Options</td>
<td>46%</td>
<td>0.62 (0.41; 0.94)</td>
<td>1.12 (0.98; 1.28)</td>
</tr>
</tbody>
</table>

Since we do not have a test sample, in order to evaluate the out-of-sample performance, we proceeded in the following way. We ran 30 experiments. For each experiment, we reserved 25% randomly drawn observations for out-of-sample performance (test-set). We used the remaining observations to fit the models. These observations were further randomly divided into training-set, which comprised $\frac{2}{3}$ of the remaining observations, and validation-set. We use the training data to fit the models and compute the test-set performance through the adjusted Qini coefficient to measure performance from a predictive point of view. In order to mitigate the overfitting, the model parameters and/or coefficients are chosen so as to find the best fit for the validation-set (cross-validation). For each experiment, the training-set size was 10,394 observations, the validation-set size was 5,354 observations, and the test-set size was 5,249 observations. To fit the RFs models, we searched for the hyper-parameters that maximize the adjusted Qini coefficient following the same procedure that was applied in the first part of the insurance data analysis. The average results are displayed in Table 9.
Based on these experiments, we see that the RF model shows the highest performance in the training-set. However, there are strong signs of overfitting. The Q+LHS model clearly outperform the other methods and gives the best results in terms of prediction. We are not surprised by the performance of the RF model because we had experimented with these RF models in the past, and we have not able to get better predictive performance in other marketing campaign initiatives.

Next, we explain the performance of each predictive model in terms of clients retained thanks to the call. We hope this gives the reader a better idea about the test-set \( \hat{q}_{\text{adj}} \) values presented in Table 9. To do so, a comparison of targeting strategies based on the models is presented in Table 10. We normalize the results to 10,000 potential customers.

As shown in Table 1, overall, the marketing campaign had a negative effect, that is, an uplift of \(-0.41\%\). This means that the proportion of renewals in the treatment group is lower than the proportion of renewals in the control group. This implies that based on 10,000 clients, we would expect a difference of 41 renewals between the following strategies: contacting all clients to convince them to stay, or not contacting any client at all. Now, using the models’ results on the test-set, we can compute this difference as a function of the number of clients contacted. It is common to present these numbers by quantiles of predicted uplifts (according to each model, much like uplift barplots, or Qini curves). We present the results in Table 10. We refer to this difference as the expected retained customers.

Let us look at the result associated with the Q+LHS model when 2,000 clients are contacted. By this, we mean the results of the marketing campaign had the company contacted the 2,000 clients with the highest predicted uplifts (according to the Q+LHS model). By contacting these clients we would expect a gain of 55 more clients renewing their policy with respect to the number of clients renewing their policy had these clients not been contacted at all. Similarly, if the 4,000 customers with the highest predicted uplifts were contacted, the company would retain 57 more customers. This means that we expect that the call does not have such a large impact for
Table 10
Expected retained customers as a function of the number of targeted customers following the predictive models from Table 9. The numbers are averaged and rounded to the nearest unit.

<table>
<thead>
<tr>
<th>Method</th>
<th>Clients contacted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2,000</td>
</tr>
<tr>
<td>RF</td>
<td>10</td>
</tr>
<tr>
<td>Q+LHS</td>
<td>55</td>
</tr>
<tr>
<td>Q+lasso</td>
<td>8</td>
</tr>
<tr>
<td>MLE+lasso</td>
<td>6</td>
</tr>
<tr>
<td>Random</td>
<td>−8</td>
</tr>
</tbody>
</table>

customer 2,001 to customer 4,000 when compared with only calling the first 2,000 clients. In terms of financial impact, depending on the cost of the call and the strategy adopted by a business, it is possible to prefer to contact only the first 2,000 customers of the ordered list. Also, not contacting customers 2,001 to 4,000 saves the costs associated with each extra call.

Note that if the RF model is used instead of the Q+LHS model, the company would lose many clients. Indeed, if we call the 2,000 customers with the highest predicted uplifts according to the RF model, we would expect a gain of only 10 more customers renewing their policy with respect to the number of clients renewing their policy had these clients not been contacted at all. If we increase the number of contacted clients to the 4,000 clients with highest uplifts, the expected number of retained customers would decrease to 2. Furthermore, for the remaining clients in the list, the effect of the call would become negative (with respect to the RF model), and therefore, there would be more chance of retaining customers if they were not called at all.

Finally, if the company decides to follow the Q+lasso model’s predictions, 19 clients would be retained by contacting the top 4,000 of the list (in comparison to not contacting them). If we compare this strategy to the strategy of calling 4,000 customers at random, that is, without relying on any uplift model, we would expect to drive away 16 customers because of the call. So, it is better to follow the Q+LHS model strategy than targeting clients at random.

6. Conclusion. Our goal was to analyze the data of a marketing campaign conducted by an insurance company to retain customers at the end of their contract. A random group of policyholders received an outbound courtesy call made by one of the company’s licensed insurance advisors, with the objective to reinforce the customers confidence in the company, to review their coverage and address any questions they might have about their renewal. In the database at our disposal, an independent group of clients
was observed and serves as control. In order to evaluate the causal effect of
the courtesy call on the renewal or cancellation of the insurance policy of
its clients, an uplift model needed to be applied.

We have developed a methodology for estimating parameters of a logistic
regression in the context of uplift models. This is based on a new statis-
tic specially conceived to evaluate uplift models. The statistic, the adjusted
Qini, is based on the Qini coefficient. It takes into account the correlation
between the observed uplift and the predicted uplift by a model. Maximizing
the adjusted Qini to choose an adequate model for uplift acts as a regular-
izing factor to select parsimonious models, much as lasso does for regression
models.

Since the Qini is a difficult statistic to compute, maximizing the adjusted
Qini directly is not an easy task. Instead, we proposed to use lasso-type
likelihood penalization to search the space of appropriate uplift models, so
as to only consider relevant variables for uplift. Since the usual lasso is
not designed for uplift models, we adapted it, by selecting the value \( \lambda \) of
the lasso penalization constant that maximizes the adjusted Qini. At first,
this ensures that the selected variables (i.e., those associated with non-zero
regression coefficients) are important variables for estimating uplift. Then, in
a second step, we estimate the parameters that maximize the adjusted Qini
by searching a Latin hypercube sampling (LHS) surface around the lasso
estimates. A variant of this procedure consists of estimating the parameters
as those that maximize the likelihood associated with the model selected by
\( \hat{\lambda} \), using only the selected variables.

Experimental evaluation showed that for the first stage of the Qini opti-
mized uplift regression, choosing the penalization constant from the logistic-
lasso sequence by maximizing the adjusted Qini dramatically improves the
performance of uplift models. This is the \( Q+\text{lasso} \) model. In addition, using
a LHS search on the second stage leads to a direct maximization of the
adjusted Qini coefficient, and to a further boost in the performance of the
model. The resulting model is the \( Q+\text{LHS} \) model. In addition, our empirical
studies clearly show that the performance of a Qini-based regression model
is much better than the performance of the usual lasso penalized logistic
regression model.

Concerning the particular marketing data available to us from the in-
surance company, we selected two final models and compared them to the
usual lasso regression approach as well as the uplift random forest. The re-
sults show that our method clearly surpasses the usual approach in terms of
performance. We argue that this is due to the Qini-based methods perform-
ing variable selection explicitly build for optimizing uplift. Although, even
if overall, the marketing campaign of the insurance company did not appear to be successful, the uplift models with the selection of the right variables identify a group of customers for which the campaign worked very well. Indeed, the results show that a persuadable client is a customer with many products to insure. Also, notice there is a subgroup of clients for whom the call had a negative impact. This could be explained by the fact that some customers are already dissatisfied with their insurance policies and have already decided to change them before receiving the call. This call could also trigger a behavior that encourages customers to look for better rates. For future campaigns, the company can target only those customers for whom the courtesy call will be useful and remove and investigate more the clients for whom the marketing campaign had a negative effect.

APPENDIX A: THE OBSERVED UPLFIT CAN BE SEEN AS THE SLOPE OF THE QINI CURVE

This can be shown as follows. The slope is defined as

$$\nabla g_k = \frac{g(\phi_{k+1}) - g(\phi_k)}{\phi_{k+1} - \phi_k}$$

By construction, the denominator is simplified to $\phi_{k+1} - \phi_k = 1/J$. Now, because of randomization, we have the following approximations,

$$\sum_{i \in N_{\phi_k}} t_i / \sum_{i \in N_{\phi_k}} (1 - t_i) \approx \sum_{i=1}^{n} t_i / \sum_{i=1}^{n} (1 - t_i),$$

and,

$$\sum_{i \in B_k} t_i / \sum_{i=1}^{n} t_i \approx \sum_{i \in B_k} (1 - t_i) / \sum_{i=1}^{n} (1 - t_i) \approx 1/J.$$

Also, we have

$$\sum_{i \in N_{\phi_{k+1}}} y_i t_i - \sum_{i \in N_{\phi_k}} y_i t_i = \sum_{i \in B_k} y_i t_i,$$
so the numerator simplifies as follows

\[ g(\phi_{k+1}) - g(\phi_k) \approx \left( \sum_{i \in B_k} y_i t_i - \left\{ \sum_{i=1}^n \frac{t_i}{i=1} \sum_{i=1}^n (1 - t_i) \right\} \sum_{i \in B_k} y_i (1 - t_i) \right) / \sum_{i=1}^n t_i \]

\[ = \sum_{i \in B_k} y_i t_i / \sum_{i=1}^n t_i - \sum_{i \in B_k} y_i (1 - t_i) / \sum_{i=1}^n (1 - t_i) \]

\[ = \left\{ \sum_{i \in B_k} t_i / \sum_{i=1}^n t_i \right\} \left\{ \sum_{i \in B_k} y_i t_i / \sum_{i \in B_k} t_i \right\} - \left\{ \sum_{i \in B_k} (1 - t_i) / \sum_{i=1}^n (1 - t_i) \right\} \left\{ \sum_{i \in B_k} y_i (1 - t_i) / \sum_{i \in B_k} (1 - t_i) \right\} \]

\[ \approx (1/J) \left\{ \sum_{i \in B_k} y_i t_i / \sum_{i \in B_k} t_i - \sum_{i \in B_k} y_i (1 - t_i) / \sum_{i \in B_k} (1 - t_i) \right\} \]

\[ = \bar{u}_k / J \]

Therefore,

\[ \nabla g_k \approx \bar{u}_k. \]

Maximizing the adjusted Qini coefficient maximizes the Qini coefficient and simultaneously promotes grouping the individuals in decreasing uplift bins, which in turn result in concave Qini curves. Suppose that for a given model, \( \rho = 1 \). In this case, we have

\[ \rho = 1 \iff \bar{u}_1 \geq \bar{u}_2 \geq \ldots \geq \bar{u}_J \]

\[ \iff \bar{u}_k - \bar{u}_{k+1} \geq 0 \quad \text{for all } k \in 1, \ldots, J - 1 \]

\[ \iff \nabla g_k - \nabla g_{k+1} \geq 0 \quad \text{for all } k \in 1, \ldots, J - 1 \]

which implies concavity of the model’s Qini curve.

APPENDIX B: K-FOLD CROSS VALIDATION

In the following, we give details about our implementation of the Q+LHS K-fold cross-validation used in the parameter estimation part of Section 5. This implementation is available in the R package tools4uplift [Belbahri et al., 2020].

- Using all data, obtain the lasso sequence \( \{\lambda_j\}_{j=1}^J \), for example by a 3-fold cross-validation with the \texttt{glmnet.cv} function available in R through the package \texttt{glmnet}.
• Divide the data into $K$ disjoint equal size parts $D_k$, $k = 1, \ldots, K$.

• For each $k \in \{1, \ldots, K\}$ do:
  
  – Set the $k$th fold $D_k$ as the validation set, and the rest of the observations $D_{-k}$ as the training set.

  – For each $j \in \{1, \ldots, J\}$ do:
    
    * Fit the penalized model with $\lambda = \lambda_j$ on the training set $D_{-k}$ to get $\hat{\theta}_{-k}(\lambda_j)$.

    * Generate $L$ LHS points in the neighbourhood of $\hat{\theta}_{-k}(\lambda_j)$, $\{\hat{\theta}_{-k}^{(l)}(\lambda_j)\}_{l=1}^L$.

    * For each $l \in \{1, \ldots, L\}$ do:
      
      • Using the training set $D_{-k}$, compute $\hat{q}_{adj,-k}^{(l)}(\lambda_j)$ associated with $\hat{\theta}_{-k}^{(l)}(\lambda_j)$.

      * Keep the coefficients associated with the maximum in the set $\{\hat{q}_{adj,-k}^{(l)}(\lambda_j)\}_{l=1}^L$, which we denote by $\tilde{\theta}_{-k}(\lambda_j)$.

      * Using the $k$th fold $D_k$ as the validation set, compute $\hat{q}_{adj,k}(\lambda_j)$ associated with $\tilde{\theta}_{-k}(\lambda_j)$.

  – Return the lasso-LHS sequence $\{\tilde{\theta}_{-k}(\lambda_j)\}_{j=1}^J$ and associated $\{\hat{q}_{adj,k}(\lambda_j)\}_{j=1}^J$.

• For $j = 1, \ldots, J$, compute the $K$-fold cross-validated average adjusted Qini coefficients, that is

\[
\tilde{q}_{adj}(\lambda_j) = \frac{1}{K} \sum_{k=1}^K \hat{q}_{adj,k}(\lambda_j)
\]

• Return $\hat{\lambda} = \arg \max_{\lambda_j} \tilde{q}_{adj}(\lambda_j)$, the value of $\lambda_j$ that maximizes the $K$-fold cross-validated average adjusted Qini, and the $K$ associated coefficients $\left(\tilde{\theta}_{-1}(\hat{\lambda}), \ldots, \tilde{\theta}_{-K}(\hat{\lambda})\right)$.

Once the procedure is complete, in a second stage, we fit a model using all the data by either

(i) calling again the LHS search starting with the coefficient vector $\hat{\theta}(\hat{\lambda})$, that is, the parameters estimated by the penalized likelihood when $\lambda = \hat{\lambda}$; or

(ii) performing $K$ different LHS searches starting from each of the $K$ associated coefficients $\left(\tilde{\theta}_{-1}(\hat{\lambda}), \ldots, \tilde{\theta}_{-K}(\hat{\lambda})\right)$. 
QINI-BASED UPLIFT REGRESSION

ACKNOWLEDGEMENTS

Mouloud Belbahri and Alejandro Murua were partially funded by The Natural Sciences and Engineering Research Council of Canada grant 2019-05444. Vahid Partovi Nia was supported by the Natural Sciences and Engineering Research Council of Canada grant 418034-2012. We would like to thank the editor, associate editor, and the reviewers for their valuable comments and suggestions that helped us to improve this manuscript.

REFERENCES


Pierre Gutierrez and Jean-Yves Gérardy. Causal inference and uplift modelling: A review


