Information Sciences xxx (xxxx) xxx

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Why you should stop predicting customer churn and start using uplift models

Floris Devriendt*, Jeroen Berrevoets, Wouter Verbeke

Data Analytics Laboratory, Faculty of Economic and Social Sciences and Solvay Business School, Vrije Universiteit Brussel, Pleinlaan 2, Brussels B-1050, Belgium

ARTICLE INFO

Article history:
Received 22 March 2018
Revised 8 November 2019
Accepted 27 December 2019
Available online xxx

Keywords:
Prescriptive analytics
Uplift modeling
Customer churn prediction
Customer retention
Maximum profit

ABSTRACT

Uplift modeling has received increasing interest in both the business analytics research community and the industry as an improved paradigm for predictive analytics for data-driven operational decision-making. The literature, however, does not provide conclusive empirical evidence that uplift modeling outperforms predictive modeling. Case studies that directly compare both approaches are lacking, and the performance of predictive models and uplift models as reported in various experimental studies cannot be compared indirectly since different evaluation measures are used to assess their performance.

Therefore, in this paper, we introduce a novel evaluation metric called the maximum profit uplift (MPU) measure that allows assessing the performance in terms of the maximum potential profit that can be achieved by adopting an uplift model. This measure, developed for evaluating customer churn uplift models, extends the maximum profit measure for evaluating customer churn prediction models. While introducing the MPU measure, we describe the generally applicable liftup curve and liftup measure for evaluating uplift models as counterparts of the lift curve and lift measure that are broadly used to evaluate predictive models. These measures are subsequently applied to assess and compare the performance of customer churn prediction and uplift models in a case study that applies uplift modeling to customer retention in the financial industry. We observe that uplift models outperform predictive models and lead to improved profitability of retention campaigns.

1. Introduction

1.1. Business analytics and predictive modeling

Business analytics is a catch-all term covering a broad variety of what essentially are data-processing techniques [1]. Business analytics are applied to an increasingly diverse range of well-specified tasks across a broad range of industries. Popular examples include applications in credit risk management [2], fraud detection [3], and customer relationship management, e.g., customer churn prediction [4], the latter being the application of interest in this article. In its broadest sense,

* Corresponding author.
E-mail addresses: floris.devriendt@vub.be (F. Devriendt), jeroen.berrevoets@vub.be (J. Berrevoets), wouter.verbeke@vub.be (W. Verbeke).

https://doi.org/10.1016/j.ins.2019.12.075
0020-0255/

Please cite this article as: F. Devriendt, J. Berrevoets and W. Verbeke, Why you should stop predicting customer churn and start using uplift models, Information Sciences, https://doi.org/10.1016/j.ins.2019.12.075

the field of business analytics overlaps significantly with data science, statistics, and related fields such as *artificial intelligence* (AI) and machine learning [5]. Analytics are a toolbox containing a variety of instruments and methodologies allowing one to analyze data to support evidence-based decision-making for the purpose of enhancing efficiency, efficacy, and thus, ultimately, profitability. Various types of analytical tools, in the increasing order of complexity, are descriptive, diagnostic, predictive, and prescriptive analytics. While descriptive and diagnostic analytics offer insight into past and current situations, predictive analytics allow one to detect complex patterns and relations between variables and to predict future trends.

Customer churn prediction models, for instance, are designed to predict which customers are about to churn and to facilitate an accurate segmentation of the customer base to allow organizations to target the customers who are most likely to churn with a retention campaign. Doing so permits an efficient use of limited marketing budgets to reduce churn, i.e., to increase the *return on marketing investment* (ROMI) [6]. Generally, customer retention has been shown to be highly profitable to companies because (1) attracting new clients costs five to six times more than retaining existing customers [7], (2) long-term customers are more profitable, tend to be less sensitive to competitive marketing activities, tend to be less costly to serve, and may generate new referrals through positive word-of-mouth, whereas dissatisfied customers might spread negative word-of-mouth [8], and (3) losing customers leads to opportunity costs due to a reduction in sales [9]. Therefore, even a small improvement in customer retention may yield significant returns [10].

1.2. Prescriptive analytics and uplift modeling

The first challenge to traditional customer churn prediction models is that they do not fully align with their business objective, as they only predict the gross outcome, i.e., whether a customer will churn. Models estimating the net effect, however, focus on whether a customer is intent on churning AND will be retained when targeted with the campaign. The true business objective is to reduce customer churn. Customers who are about to churn but cannot be retained should be excluded from the campaign, as targeting them will be a waste of scarce resources. Moreover, it has been reported that retention efforts may also provoke customers to churn [11]. For instance, a retention offer may remind a customer about the imminent expiration of a contractual agreement and cause churn as a result. As noted in Radcliffe and Simpson [12], churn risk is highly correlated with customer dissatisfaction, and the goal in turn becomes that of preventing a dissatisfied customer from actually churning. Any attempt made to contact and retain a dissatisfied customer may actually provoke the customer to churn and thus contribute to a negative net effect of the campaign. Such customers are definitely to be excluded from a retention campaign and are to be distinguished those who will not churn regardless of whether they are targeted. Note that classic methods only differentiate customers who are about to churn from non-churners, while uplift modeling differentiates customers whose targeting will benefit the company from other customers. Targeting customers prescribed by an uplift model will not only reduce churn but do so with a lower resource expenditure, effectively resolving this first issue associated with traditional techniques.

The second issue is that traditional customer churn prediction models are subject to feedback loops [13]. When an organization operates a customer churn prediction model to select customers for retention campaigns, it factually alters customer behavior. The data collected *during operation* of a customer churn prediction model is therefore *biased*. Hence, if such data is used in developing a new or updated customer churn prediction model, a biased predictive model will be learned, as illustrated in Fig. 1. Note that the observed customer behavior Y_2, i.e., churn or no-churn during period 2, as captured by the target variable in period 1 during which churn prediction model 1 is used, is potentially influenced by the retention campaign. We can further analyze the customers that have been observed to churn and that have been observed not to churn, depending on whether they were targeted, as follows:

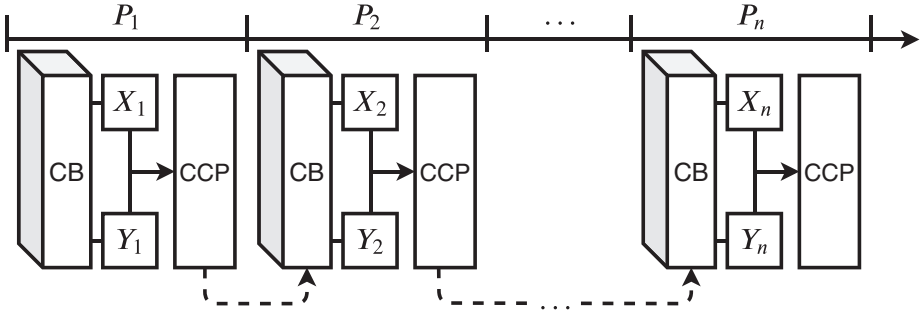


Fig. 1. Illustration of feedback loops causing future customer churn prediction models to be biased due to retention campaigns altering customer behavior. Data for the initial customer churn prediction model P_1 are collected during period 1, with the predictors gathered during the beginning of period 1 to predict churn at the end of period 1. Data for a novel customer churn prediction model P_2 are collected during period 2 while the first model involving retention campaigns altering customer behavior is used. In the figure, CB and CCP denote customer base and customer churn prediction, respectively.

Please cite this article as: F. Devriendt, J. Berrevoets and W. Verbeke, Why you should stop predicting customer churn and start using uplift models, Information Sciences, https://doi.org/10.1016/j.ins.2019.12.075

1. Customers observed to churn in period 1 may be either (1) customers who have been predicted to churn and were targeted with a campaign but churned nonetheless and were hence a lost cause or (2) customers who were not targeted in the retention campaign since they were not predicted to churn by the customer churn prediction model in use at the time and hence are very difficult to model correctly. Thus, the novel customer churn prediction model learned from data of period 1 and prior periods will learn to predict lost causes or churners whose decisions are *very difficult to predict*.
2. Customers observed not to churn in period 1 are either (1) those who were not about to churn and were not targeted or (2) those who were about to churn but were retained by the campaign. The second group here are customers who were correctly predicted to churn by the operational predictive model. However, these customers are labeled as no-churners ($Y_2 = nochurn$), and hence a novel churn prediction model learned from data collected during period 2 will learn to predict such customers to be non-churners in the future even though in fact they represent the very type of customers who should be targeted in the future.

Uplift modeling aims at establishing the net difference in customer behavior that results from applying a specific *treatment* to customers, e.g., a reduction in the likelihood of churn when customers are targeted with the retention campaign. Generally, uplift models allow estimation of an outcome in function of a specific input variable configuration, including decision variables. These variables represent decisions that are under control of the organization and hence can be optimized. Uplift models thus allow prediction of an outcome under various decision-making scenarios and can therefore be used to prescribe optimal decisions. Decision variables may capture any decision, action, setting, or setup that has a potential impact on organizational performance and requires decision-making. Uplift modeling is a type of prescriptive analytics that essentially allows learning simulation models from historical data [14]. Such modeling requires the availability of data that captures the behavior of the modeled system under varying conditions and decision-making scenarios. Such data may need to be actively collected by setting up experiments but often results from data analytics applications being used to support decision-making. Hence, we claim that uplift modeling offers a way to address the second issue listed above, i.e., feedback loops. Recording and including information on all potentially customer behavior-altering actions undertaken towards customers based on analytical models, and setting up control groups and collecting unbiased data allow feedback loops to be internalized when uplift models are developed.

1.3. Contributions

Despite a steadily growing body of literature on uplift modeling and theoretical evidence of the appropriateness of this approach, little conclusive empirical evidence of improved performance of uplift modeling over predictive modeling has been presented to date. Only [15] highlights the ineffectiveness of using standard churn prediction in two fields studies where uplift models are found to be more effective at retaining customers. Therefore, in this paper we further contrast *customer churn prediction* (CCP) and *customer churn uplift* (CCU) modeling for customer retention by comparing these approaches' performance when they are applied to an experimental case study in the financial industry. CCP and CCU models, however, are evaluated using different evaluation approaches since they produce different kinds of output. For evaluating CCP models, the receiver operating characteristic (ROC) curve and the lift curve are often used. Performance is typically reported in terms of the area under the ROC curve (AUC) and top-decile lift. In evaluating uplift models, Qini curves and uplift-per-decile plots are typically used [16], and performance is commonly reported in terms of the Qini coefficient or top-decile uplift.

In this article, we develop a common evaluation procedure by extending the *maximum profit* (MP) measure to facilitate evaluation of the performance of CCU models. As the goal of customer churn prediction is to maximize ROMI, Verbeke et al. [17] introduced the MP measure for evaluating CCP models. This measure calculates the profit generated by a retention campaign when targeting the optimal proportion of top-ranked customers determined by a CCP model, given the cost and benefit parameters characterizing the retention campaign [18]. This measure allows selecting the optimal model and proportion of customers to be included in the campaign, potentially yielding a significant increase in profitability relative to the profit earned if statistical measures are used [19]. After the introduction of the extension of the MP, called the *maximum profit uplift* (MPU), we describe the generally applicable liftup curve and liftup measure that characterizes the performance of an uplift model. These serve as counterparts of the lift curve and lift measure that are broadly used for evaluating predictive models. Both the MP measure and the MPU measure introduced in this paper will be used to compare the performance of CCP and CCU models developed using logistic regression and random forest-based methods in an experimental case study.

In summary, our main contributions are threefold:

1. We extend the maximum profit measure for evaluating uplift models by introducing the maximum profit uplift measure.
2. We introduce the generally applicable liftup curve and liftup measure for evaluating uplift models that correspond to the lift curve and lift measure used for evaluating predictive models.
3. We provide empirical evidence of the merits of uplift modeling versus predictive modeling by performing an experimental case study.

This paper is structured as follows. In Section 2, we first introduce customer churn prediction modeling before discussing uplift modeling as an alternative approach to predictive modeling. Next, in Section 3 the MP measure for CCP models is

Please cite this article as: F. Devriendt, J. Berrevoets and W. Verbeke, Why you should stop predicting customer churn and start using uplift models, Information Sciences, https://doi.org/10.1016/j.ins.2019.12.075

discussed and extended to facilitate its application to evaluation of customer churn's uplift models. In Section 4, we describe the experimental design of the case study and subsequently discuss the results of our experiments. Finally, in Section 5, conclusions are presented.

2. Literature

In Section 2.1, customer churn prediction is introduced along with the current standard approaches described in the literature and adopted in the industry. Afterwards, in Section 2.2 we describe uplift modeling and discuss prominent uplift modeling techniques and performance measures for evaluating uplift models.

2.1. Customer churn prediction

Customer churn, also referred to as customer attrition or customer defection, is defined as the *loss* or outflow of customers from the customer base [1]. In saturated markets, limited opportunities exist or significant investments are required to attract new customers. Hence, retaining existing customers is considered essential to maintaining profitability. Established customers have been shown to be more profitable due to the lower cost of serving them, and the sense of brand loyalty they have developed over time renders them on average less likely to churn. Loyal customers tend to be satisfied customers who also serve as word-of-mouth advertisers, referring new customers to a given company. Within retention, two types of defections are considered: (1) total defection, the customer completely interrupts the relationship with the company [20], and (2) partial defection, the customer changes behaviour over time and interacts less and less with the company (e.g. purchase some goods from other companies). Some type of relationship with the customer still exists [20]. In the context of a financial institution, as described in the case study provided in Section 4, a definition of churn, i.e., active contract termination upon customer request, is naturally present in the data.

Customer retention efforts are typically supported by a customer churn prediction model, which is a classification model such as a logistic regression or a decision tree model [17]. Such a model estimates for each customer the probability of that customer churning during a subsequent period of time. The customers with the highest likelihood of churning can then be offered an incentive, e.g., a discount or another promotional offer, to encourage them to extend their subscription or to keep their account active. In summary, customers who are susceptible to churn are typically identified by a customer churn prediction model and by extension targeted with a retention campaign. Accurate predictions are perhaps the most apparent goal of developing a customer churn prediction model, but acquiring insights into what motivates customers to churn is valuable to an organization and is often the second objective of developing such a model. Comprehensible models can offer novel insights into correlations between customer characteristics and behavior, and the propensity to churn [6]. Such insights allow addressing factors leading to customers' defections and taking preventive measures to avoid rather than to cure churn.

Numerous classification techniques, including traditional statistical methods such as logistic regression [18], nonparametric statistical models such as k-nearest neighbor models [21], decision trees [22], ensemble methods [4], support vector machines [23] and neural networks [24], have been used for churn prediction. Moreover, social network analysis has been successfully used to predict customer churn [25] and for survival analysis that can be used to analyze and estimate the time until a customer's defection. Analyses of the latter type allow focusing on the profitability of a customer over the entire customer lifetime [26]. An extensive literature review of customer churn prediction modeling is provided by Verbeke et al. [6]. In Verbeke et al. [17], results of an extensive benchmarking study are reported that confirm the no-free-lunch theorem's applicability to customer churn prediction, with no modeling technique consistently achieving the best performance across various datasets. A recent study of customer churn prediction is covered in [27].

2.2. Uplift modeling

In Section 2.2.1, a brief introduction to uplift modeling is provided. In Section 2.2.2, an overview of the most prominent uplift modeling techniques is presented. Finally, in Section 2.2.3, evaluation measures for assessing the performance of uplift models are discussed.

2.2.1. Definition

Uplift models aim to estimate the net effect of applying *a treatment* to *an outcome*. Hence, uplift modeling is about estimating the individual treatment effect and therefore is equivalent to individual causal modeling under a strong ignorability assumption [28]. In marketing, a treatment may concern any action taken towards a customer, such as a discount offered to retain a customer.

Conceptually, a customer base can be divided into four categories along two dimensions, as shown in Table 1 [5,29]:

1. *Sure Things*. Customers who would never churn. Targeting *sure things* does not generate additional returns yet in fact leads to additional costs, i.e., the fixed costs of contacting a customer and possibly a cost related to a financial incentive offered to targeted customers.

Please cite this article as: F. Devriendt, J. Berrevoets and W. Verbeke, Why you should stop predicting customer churn and start using uplift models, Information Sciences, https://doi.org/10.1016/j.ins.2019.12.075

Table 1
Four theoretical classes.

Churn when targeted		Churn when not targeted: No	Churn when not targeted: Yes
Yes		Do-Not-Disturbs	Lost Causes
No		Sure Things	Persuadables

2. *Lost Causes*. Customers to churn regardless of campaign used. *Lost causes* will not generate additional revenues yet in fact generate additional costs, although the latter are lower than the costs of *sure things*, as *lost causes* in contrast to *sure things* do not take advantage of financial incentives being offered.
3. *Persuadables*. Customers who do not churn *only because* they have been exposed to a retention campaign. They do not churn only if contacted and cause a campaign to generate additional revenues and, as such, a net profit after the subtraction of costs stemming from the inclusion of other types of customers.
4. *Do-Not-Disturbs*. Customers who would churn *only because* they were exposed to a retention campaign. They will not churn if not targeted but will churn if they are. Populations targeted for retention efforts can have an adverse reaction, e.g., discontinuing the purchase of the delivered product or service. Including *do-not-disturbs* in a campaign thus generates no additional revenues but leads to considerable additional costs. This category is sometimes referred to as *sleeping dogs* since, as long as these customers are undisturbed, they will continue to provide benefits to the company.

The aim of uplift modeling is to allow targeting of *persuadables* only. Note that this classification of customers is dependent on the campaign. It is possible for a customer to be a *lost cause* if a campaign offers a 5% discount on the next purchase while being a *persuadable* if offered a 20% discount. If the data analytics perspective is adopted, uplift modeling involves the determination of the optimal settings for *decision variables*, such as a dummy treatment variable reflecting whether a customer is targeted with a retention campaign, that optimize a certain outcome, e.g., customer retention. Although in almost all studies of uplift modeling for marketing applications the decision variables are dummy variables that indicate whether a customer is targeted or not, such variables may also be continuous or multivalued categorical variables, e.g., representing the amount of discount or the channel used to contact a customer. Uplift modeling may be applied in various settings and for a broad variety of purposes beyond retention and marketing (e.g., in personalized medicine [30] and persuasion modeling in political campaigns [31]).

Few cases of uplift modeling for customer retention have been documented. Radcliffe and Simpson [12] applied uplift modeling to data from two retention campaigns in telecommunications. One campaign was highly effective and profitable, whereas the other was observed to be counterproductive and yielding a net loss. Both campaigns' outcomes in terms of reducing churn improved as a result of uplift modeling. Guelman et al. [32] applied uplift modeling in insurance. Although on average the treatment had an almost neutral impact on retention for the entire sample, the authors showed that a positive impact of treatment could be attained if specific subgroups of the customer base were selected using uplift modeling.

We assume that a sample of customers is divided into two groups defined as the treatment and control groups. A customer is either part of the treatment group, i.e., is exposed to the campaign, or part of the control group, i.e., is not exposed to the campaign. The sample should be stratified, i.e., it should have the same distribution of churners and non-churners, and the same distribution of customers in the treatment and control groups as the full set of samples. As a formal definition, let X be a vector of inputs or predictor variables, $X = \{x_1, \ldots, x_n\}$, and let Y be the binary outcome variable, $Y \in \{0, 1\}$, which indicates whether a customer churned. Let the treatment variable T denote whether a customer belongs to the treatment group ($T = 1$) or to the control group ($T = 0$). P denotes the probability of churn. Uplift is then defined for customer i with characteristics x_i as the probability of churn (i.e., $y_i = 1$) if the customer is not treated (i.e., $t_i = 0$) minus the probability of churn if the customer is treated (i.e., $t_i = 1$):

$$U(x_i) := P(y_i = 1|x_i; t_i = 0) - P(y_i = 1|x_i; t_i = 1) \tag{1}$$

In essence, uplift is the difference in outcome, e.g., customer behavior, resulting from a treatment. Uplift modeling aims at estimating uplift as a function of treatment and customer characteristics. Individuals associated with high uplift are then targeted by the campaign; i.e., they are classified as persuadable for treatment $t_i = 1$.

2.2.2. Techniques

Uplift modeling techniques can be classified into data preprocessing and data processing approaches. Methods of the first class apply traditional predictive analytics within an adapted setup for learning an uplift model, whereas those of the second class apply adapted predictive analytical methods to estimate uplift.

Data preprocessing approaches include transformation approaches [30,33] that redefine a target variable, and approaches that allow one to estimate uplift by defining additional predictor variables that are incorporated within a standard predictive model [11].

Approaches of the first group define a transformed target variable that is estimated. Historical data containing information on observed customer behavior never allows classifying customers in the four groups shown in Table 1, as the net

Please cite this article as: F. Devriendt, J. Berrevoets and W. Verbeke, Why you should stop predicting customer churn and start using uplift models, Information Sciences, https://doi.org/10.1016/j.ins.2019.12.075

effect of treatment cannot be observed for an individual customer due to the fundamental problem of causal inference [34]. We can only know for sure whether a customer belonged to the treatment or control group and whether the customer churned. Hence, customers can be classified in the following four groups: treatment responders, treatment non-responders, control responders and control non-responders. Techniques such as Lai's approach [33,35] and *pessimistic uplift modeling* [36] use this classification to define a transformed target variable, transforming the uplift modeling problem into a binary classification problem. A standard classification technique can subsequently be applied to learn an uplift model.

Data preprocessing approaches of the second group extend the set of predictor variables to allow estimation of uplift. In Lo [11], Kane et al. [35], an uplift modeling approach that groups the treatment and control groups into a single sample for developing a predictive model is proposed. A dummy variable is introduced to denote the group of origin for each customer. Afterwards, a model is developed from (1) the original predictor variables, (2) the added dummy variable, and (3) additional interaction variables between the predictor and dummy variables. Subsequently, a standard classification method can be used, yielding an uplift model that allows predicting the probability of churn in treatment and no-treatment scenarios, the difference being the uplift.

Data preprocessing approaches. A further distinction among data processing approaches can be made between indirect and direct estimation approaches.

Indirect estimation approaches include the two-model or naive approach, which is a straightforward approach to uplift modeling. Two separate predictive models can be developed: one for the treatment group, M_T, and one for the control group, M_C. Both models estimate the probability of churn. The aggregated uplift model M_U subtracts the probabilities resulting from both models to determine the uplift:

$$M_U = M_T - M_C. \tag{2}$$

From a practitioner's perspective, this approach has the advantage of being easy to implement since it allows reuse of the existing customer churn prediction modeling procedure. It is to some extent similar to the second group of data preprocessing approaches. The main disadvantage of the two-model approach is that the two models are built independently of each other; as such, they are not necessarily consistent in terms of included predictor variables, and the errors of independent estimates may reinforce each other, leading to significant resulting errors in uplift estimates [37]. Therefore, this approach only appears to apply to the simplest of cases [38], and the reported performance is often observed to be weak [14].

Alternatively, uplift can be modeled directly. Given the group-based nature of the uplift modeling problem, the most frequently adopted direct estimation approaches are tree-based methods that subsequently split the population into smaller segments. Uplift tree approaches adapt well-known algorithms such as *classification and regression trees* (CART) [39] or *chi-squared automatic interaction detection* (CHAID) methods [40], applying modified splitting criteria and pruning approaches. Examples of tree-based uplift modeling approaches include significance-based uplift trees proposed in Radcliffe and Surry [37], decision trees using information theory-inspired splitting criteria presented in Rzepakowski and Jaroszewicz [41], and uplift random forest and causal conditional trees introduced in Guelman et al. [42].

2.2.3. Evaluation

In predictive modeling, evaluation metrics typically assess the error of pointwise estimates obtained by a model on each observation in a holdout test set. The observed and predicted outcomes can be compared, and the errors can subsequently be summarized or aggregated to obtain the overall performance measure. In uplift modeling, however, the actual outcome that is estimated is unobserved, and therefore the error made by the model cannot be measured at the individual customer level. Uplift or, in other words, whether a customer is persuadable, a lost cause, do-not-disturb or a sure thing, cannot be observed due to the fundamental problem of causal inference [34]. A customer cannot be treated and not-treated simultaneously. Therefore, evaluation measures adopted in predictive modeling cannot be used to appropriately evaluate uplift models. One approach to evaluating uplift models is to observe uplift for equivalent segments of the treatment and control groups [37], which allows assessing the use and power of the model to identify groups of customers for whom treatment results in a substantial net effect.

Performance metrics in predictive modeling. Examples of popular classification measures are the area under the receiver operating characteristic curve (AUC), the Gini coefficient, Kolmogorov-Smirnov distance and the top decile lift. The AUC assesses the behavior of a classifier disregarding class distribution, classification cutoff and misclassification costs [17]. The top-decile lift only considers the 10% of customers with the highest predicted probabilities of churn, which to a certain extent may be more consistent with the actual use of a classifier such as a customer churn prediction model.

Performance metrics in uplift modeling In the literature on uplift modeling, an adapted version of the Gini coefficient, i.e., the Qini coefficient [16,35] is often used to assess performance. Additionally, charts are often used for visual evaluation [11,43]. The performance of an uplift model can be visualized by plotting the cumulative difference in response rates between treatment and control groups as a function of the selected proportion x of customers ranked by the uplift model from high to low values of estimated uplift. This curve is referred to as the cumulative uplift, as cumulative incremental gains, or as the Qini curve [16]. The cumulative difference in the response rate is measured as the absolute or relative number of additional favorable responders, i.e., expressed as the additional number of favorable responders or as a proportion of the total population, respectively. An example is shown in Fig. 2. Note that performance is evaluated by comparing groups of observations rather than assessing the accuracy of predictions for individual customers. As Fig. 2 shows, the Qini curve does not always have to be increasing. The curve will not increase if it fails to capture persuadables and will even decrease if it

Please cite this article as: F. Devriendt, J. Berrevoets and W. Verbeke, Why you should stop predicting customer churn and start using uplift models, Information Sciences, https://doi.org/10.1016/j.ins.2019.12.075

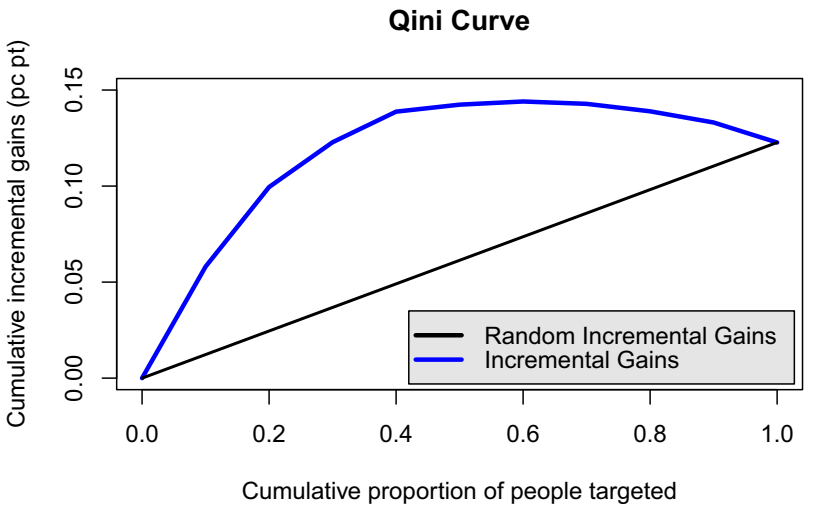


Fig. 2. Incremental gains, also referred to as the Qini curve.

captures do-not-disturbs. This further indicates that targeting the entire population is not an optimal strategy, as unsuitable customers would be included.

The Qini metric is a measure related to the Qini curve. It measures the area between the Qini curve of the uplift model and the Qini curve of the baseline random model (see Fig. 2). This measure is an adapted version of the Gini metric, which in turn is related to the Gini or cumulative gains curve [16].

Although uplift models are developed and used to enhance the efficiency and returns of retention campaigns, few research papers assess the costs and benefits of adopting such a model. Hansotia and Rukstales [44] compute the incremental return on investment at the gross margin level. The respective gross profits are then regarded as a contribution to overhead and to net profits [44]. In Radcliffe [16], the incremental profit is calculated by multiplying the incremental response rate by the total profit. Similarly, in Rzepakowski and Jaroszewicz [41] the gain in profit is calculated by subtracting from the profit of a certain percentage p of the highest-scoring individuals in the treatment group, the profit associated with the same percentage p of the highest-scoring individuals in the control group. In the next section, we analyze the involved costs and benefits in detail and develop a profit-driven approach to evaluating customer churn uplift models.

3. Maximum profit measure

The first part of this section discusses the *maximum profit* (MP) measure for customer churn prediction, introduced in Verbeke et al. [17]. In the second part, we extend this measure to evaluating customer churn uplift models, which will subsequently allow us to meaningfully compare customer churn prediction and uplift models in Section 4.

3.1. Customer churn prediction models

To maximize the efficiency and returns of a retention campaign, a limited proportion of the customer base is typically targeted and given an incentive to remain loyal. Therefore, customer churn prediction models are often evaluated using, e.g., the top-decile lift measure that only accounts for the performance of the model for 10% of customers with the highest predicted probabilities of churn. Recently, Verbeke et al. [17] demonstrated that, from a profit-centric point of view, using the top-decile lift can be expected to result in suboptimal model selection. The maximum profit measure that calculates the profit resulting from targeting with a retention campaign the optimal proportion of top-ranked customers determined by the CCP model has been proposed as a business-oriented alternative. In essence, this measure evaluates a customer churn prediction model at the cutoff leading to the maximum profit rather than at an arbitrary cutoff such as 10%. Performance is expressed as the profit in monetary units that can be earned by using the model to select customers to be targeted in a retention campaign. This approach, as shown by the authors, may lead to different model selection and yield a significant increase in profitability over that of using statistical measures and that of choosing the proportion of customers to be targeted arbitrarily or based on expert opinions [17].

Fig. 3 visualizes the dynamic customer churn and retention process and allows deriving the retention campaign profit formula introduced by Neslin et al. [18]:

$$\Pi = N\alpha[\beta\gamma(b - c_{contact} - c_{incentive}) + \beta(1-\gamma)(-c_{contact}) + (1-\beta)(-c_{contact} - c_{incentive})] - A \tag{3}$$

where Π denotes the profit generated by the campaign, N is the number of customers in the customer base, α is the proportion of the customer base targeted in the retention campaign, β is the proportion of true would-be churners among

Please cite this article as: F. Devriendt, J. Berrevoets and W. Verbeke, Why you should stop predicting customer churn and start using uplift models, Information Sciences, https://doi.org/10.1016/j.ins.2019.12.075

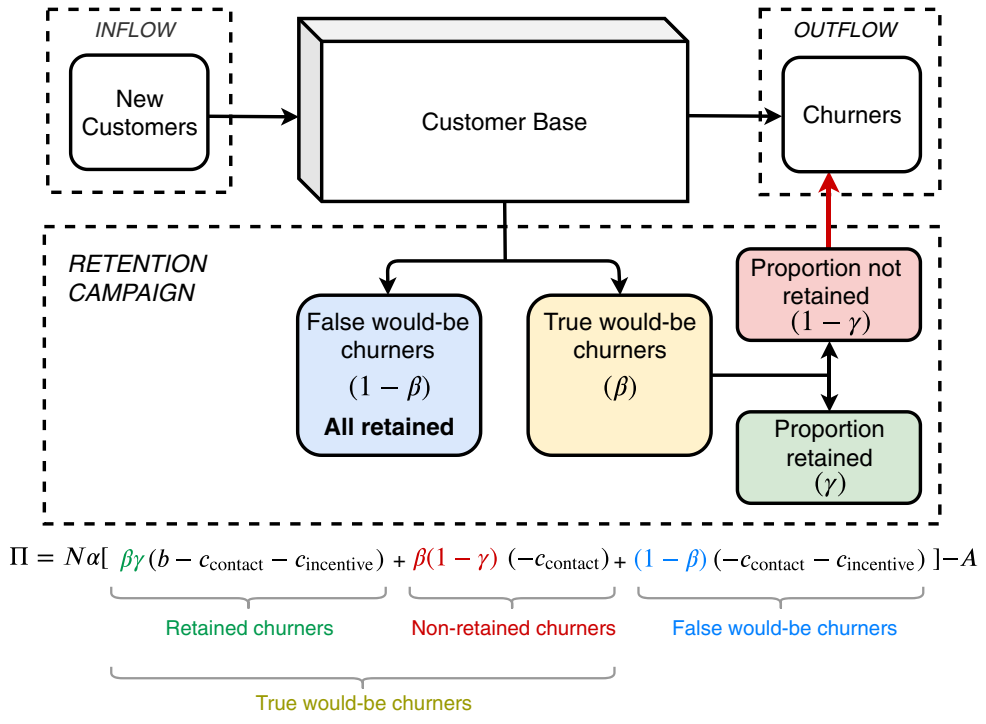


Fig. 3. Visual representation of the formula of Neslin et al. [18]. Colors indicate matching parts of the formula and schematics.

customers targeted by the retention campaign, γ is the retention rate, i.e., the proportion of targeted would-be churners that are retained (or, in other words, the success rate of the retention campaign in persuading churners to stay), b is the benefit of retaining a customer (e.g., the average customer lifetime value), $c_{contact}$ is the cost of contacting a customer, $c_{incentive}$ is the cost of the incentive if a customer accepts the offer, and A is the fixed administrative cost of running the churn management program.

An intuitive interpretation of the profit formula is obtained by separating Π into five parts:

(a) $N\alpha$ represents the number of customers targeted in the campaign; except for the fixed administrative costs A, only the targeted customers induce costs and benefits related to the campaign.
(b) $\beta\gamma(b - c_{contact} - c_{incentive})$ represents the net profit generated by the campaign that equals the reduction in lost revenues due to churn less the costs of the campaign, $b - c_{contact} - c_{incentive}$, multiplied by the proportion γ of would-be churners among the proportion of correctly identified would-be churners β targeted by the campaign.
(c) $\beta(1-\gamma)(-c_{contact})$ represents the cost related to including correctly identified would-be churners who were not retained.
(d) $(1-\beta)(-c_{contact} - c_{incentive})$ reflects the cost resulting from targeting non-churners with the campaign; these customers accept and take advantage of the incentive offered to them.
(e) A reflects the fixed administrative cost that reduces the overall profitability of a retention campaign.

As noted in Neslin et al. [18], β reflects the capacity of the predictive model to identify would-be churners and can be expressed as

$$\beta = \lambda\beta_0 \tag{4}$$

where β_0 denotes the overall churn rate, i.e., the proportion of all customers who will churn, and λ denotes the lift (i.e., the lift at the top-α percentile; cfr. top-decile lift). Rearranging the terms in Eq. (3) yields

$$\Pi = N\alpha\{[\gamma b + c_{incentive}(1-\gamma)]\beta_0\lambda - c_{incentive} - c_{contact}\} - A \tag{5}$$

Neslin et al. [18] use the direct link between lift and profitability as a means to motivate the use of lift as a performance measure for evaluating customer churn prediction models. Verbeke et al. [17], however, show that using the lift at an arbitrary cutoff as a performance measure may lead to suboptimal model selection and, from a business perspective, a significant reduction of profitability. Therefore, the authors propose a profit-centric performance measure called the maximum profit measure, defined as [17]

$$MP = \max_{\alpha}(\Pi) \tag{6}$$

Please cite this article as: F. Devriendt, J. Berrevoets and W. Verbeke, Why you should stop predicting customer churn and start using uplift models, Information Sciences, https://doi.org/10.1016/j.ins.2019.12.075

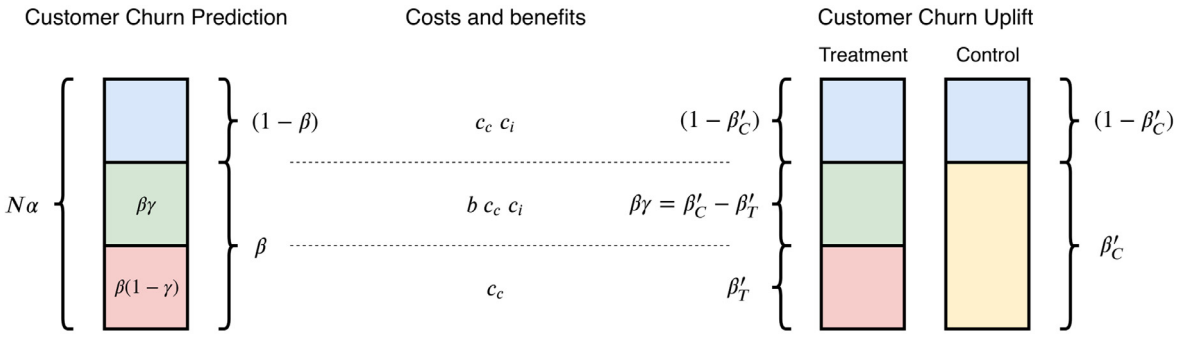


Fig. 4. The left panel visualizes Neslin's formula, focusing on the campaign-targeted population; the right panel provides the equivalent visualization for uplift modeling.

To calculate the maximum profit measure, a pragmatic approach is typically adopted [45], and two assumptions are made: (1) the retention rate γ is constant and independent of the number of customers targeted by the retention campaign, and (2) the benefit b is constant and independent of the number of customers targeted by the retention campaign. Given the lift curve of the classification model that represents the relation between the lift and α, the maximum of Eq. (5) considered as a function of α can then be calculated straightforwardly [17] and interpreted as a measure of performance of a CCP model.

3.2. Customer churn uplift models

To evaluate customer churn uplift models as well as to compare CCP and CCU models, we can extend the profit formula in Eq. (3) to the uplift modeling case.

To this end, the left panel of Fig. 4 visualizes the set of $N\alpha$ top-ranked customers in the CCP modeling case, which is regarded in the CCP profit formula of Eq. (3) as consisting of three subgroups: (1) the proportion $1-\beta$ (represented by the blue part) of customers who will not churn and are erroneously included in the campaign, (2) the proportion $\beta\gamma$ (represented by the green part) of true would-be churners who accept the offer, and (3) the proportion $\beta(1-\gamma)$ (represented by the red part) of true would-be churners who do not accept the offer. Recall that γ is the campaign retention rate that is assumed to be constant and is to be estimated.

The right panel of Fig. 4 visualizes the set of $N\alpha$ top-ranked customers in the CCU case. When an uplift model is developed, this set of customers will contain those from both the control group and the treatment group. For the control group, this set consists of two subgroups: (1) a proportion β'_C of true would-be churners, i.e., the churn rate in the control group (none of the respective customers have been retained since the control group was not targeted), and (2) a proportion $1-\beta'_C$ of false would-be churners. Note that the proportion β of true would-be churners in the CCP profit formula equals the observed proportion β'_C of churners in the control group, i.e., the churn rate in the control group is the proportion of true would-be churners. For the treatment group, again three subgroups can be identified: (1) a proportion of false would-be churners, equal to the proportion of false would-be churners in the control group, $1-\beta'_C$ (represented by the blue part), (2) a proportion β'_T of true would-be churners that are not retained, i.e., the churn rate in the treatment group (represented by the red part), and (3) a proportion $\beta'_C - \beta'_T$ of true would-be churners that are retained, i.e., the difference in churn rates between the control and treatment groups (represented by the green part).

Comparing the right and left panels in Fig. 4, we observe that the retained proportion of true would-be churners in the CCP profit formula equals the difference in churn rates between the control and treatment groups in the CCU case:

$$\beta = \beta'_C \tag{7}$$

$$\beta(1-\gamma) = \beta'_T \tag{8}$$

from which follows:

$$\begin{aligned}\beta\gamma &= \beta - \beta'_T \\ &= \beta'_C - \beta'_T\end{aligned} \tag{9}$$

The difference between the churn rates in the treatment and control groups by definition is equal to the uplift achieved at cutoff α:

$$\beta\gamma = \beta'_C - \beta'_T = \upsilon \tag{10}$$

Please cite this article as: F. Devriendt, J. Berrevoets and W. Verbeke, Why you should stop predicting customer churn and start using uplift models, Information Sciences, https://doi.org/10.1016/j.ins.2019.12.075

Hence, in CCU modeling, the retention rate can be ascertained by comparing the churn rates in the control and treatment groups, as

$$\gamma = \frac{\upsilon}{\beta} = \frac{\upsilon}{\beta'_C} \tag{11}$$

In CCP modeling, γ was estimated and assumed to be constant. This assumption clearly does not hold in uplift modeling, where the objective is to achieve increased uplift for top-ranked customers. Uplift υ is explicitly indicated to be a function of cutoff α, as uplift is typically visualized using decile graphs or Qini curves. In the remainder of this paper, the variable retention rate is represented as υ/β'_C, i.e., the uplift at cutoff α divided by the base churn rate, which will be used in the newly introduced MPU measure for evaluating CCU models. The retention rate used in the MP measure for evaluating CCP models and assumed to be constant will always be referred to as γ.

Hence, Eq. (3) can be rewritten to evaluate the profitability of a retention rate under an uplift model for selecting a proportion α of customers as follows:

$$\Pi_u = N\alpha\Big[(\beta'_C - \beta'_T)(b - c_{contact} - c_{incentive}) + \beta'_T(-c_{contact}) + (1-\beta'_C)(-c_{contact} - c_{incentive})\Big] - A \tag{12}$$

Note that β'_C, β'_T and, as a result, υ, are a function of α. If α is equal to one, then β'_C and β'_T are equal to churn rates observed in the full control and treatment groups, respectively. As in Eq. (5), we may define a measure similar to lift that characterizes the performance of an uplift model by comparing the performance of the model in terms of the achieved uplift at proportion α to the overall uplift achieved when targeting the entire population.

We introduce and formally define liftup λ_u as follows:

$$\beta'_C - \beta'_T = \upsilon = \upsilon_0 * \lambda_u \tag{13}$$

where υ_0 is the baseline uplift achieved when targeting the full customer base. Note that similarly to a lift curve, a liftup curve can be plotted to provide a visual evaluation of an uplift model. Additionally, the top-decile liftup may be reported, as well as the liftup at any other percentile cutoff, consistently with the practice of, e.g., reporting the top-decile lift. In the experimental section below, liftup curves are provided for illustration. Rewriting Eq. (12) yields

$$\Pi_u = N\alpha\Big[\upsilon_0 * \lambda_u * b - c_{contact} - (1-\beta'_T) * c_{incentive}\Big] - A \tag{14}$$

As to CCP modeling, the final objective of an uplift model is to maximize the profit earned as a result of a retention campaign, as expressed by Eq. (14), yielding the maximum profit uplift (MPU) measure:

$$MPU = \max_{\alpha}(\Pi_u) \tag{15}$$

The MPU measure expresses the performance of a CCU model in terms of profit earned per customer in the customer base when targeting the optimal proportion of customers ranked according to the estimated uplift determined by the CCU model. While MPU gives us the maximum achievable profit, this is easily extended to the argmax of *alpha* to give us the most optimal proportion of the population to target. This leads to new managerial implications as to the setup of a marketing campaign.

4. Experiments

The experiments reported in this section allow comparing and contrasting customer churn prediction and customer churn uplift modeling. In the first part of this section, Section 4.1, information about the experimental setup, i.e., the dataset and experimental methodology, is provided. The results of experiments are presented in Section 4.2, and are discussed and analyzed in detail in Section 4.3.

4.1. Experimental design

4.1.1. Dataset

The dataset used in the experiments was obtained from a financial institution. It consists of records containing customer information, including a churn indicator and a variable determining whether a customer was targeted with a retention campaign. In this context, a customer who churns corresponds to an active total defection due to contract termination. Table 2 provides detailed information about the dataset. The retention campaign was targeted at a treatment group for which, in the subsequent three-month period, a churn rate of 13.25 % was observed. For the control group, which was not targeted by the retention campaign, a significantly higher churn rate of 25.52 % was observed. The baseline uplift υ_0, introduced in the previous section, that was achieved by the retention campaign thus equals 12.27 %. The dataset includes 162 variables, including sociodemographic data as well as usage and activity indicators.

4.1.2. Methodology

Random stratified sampling was applied to the treatment and control groups to obtain training and test sets including 2/3 and 1/3 of records, respectively. To establish the CCP model, only the training set that was sampled from the control group is used; in contrast, the training set that was sampled from both the treatment and control group is used in learning

Please cite this article as: F. Devriendt, J. Berrevoets and W. Verbeke, Why you should stop predicting customer churn and start using uplift models, Information Sciences, https://doi.org/10.1016/j.ins.2019.12.075

Table 2
Information on the dataset obtained from a European financial institution.

Data	
Type of organization	Financial institution
Total number of observations	200,903
Total number of variables	162
Number of control group observations	118,809
Control group churn rate	25.52%
Number of treatment group observations	82,094
Treatment group churn rate	13.25%
Overall Uplift	12.27%

the CCU model. To compare the performance of the resulting CCP and CCU models, two scenarios are considered. In the first, the *classic* MP profit measure is used (cfr. Eq. (3)) for evaluating both CCP and CCU models, which requires using the test set of the control group. In the second scenario, the novel MPU measure is adopted (cfr. Eq. (12)) for testing the CCP and CCU models, which requires both test sets that were sampled from the treatment and control groups.

Two modeling techniques—logistic regression and random forests—are used to develop CCP and CCU models. Both techniques can be used straightforwardly to develop predictive models and have been adapted to developing uplift models [11,32]. The decision to apply these two techniques in the experiments is motivated as follows. Logistic regression is the standard predictive modeling approach used in industrial settings across various applications and is a typical benchmark approach used in experimental studies and scientific research. Additionally, logistic regression facilitates the interpretation of the resulting model and typically performs well [14,46]. Random forests, on the other hand, represent the state-of-the-art in the field of business analytics, are widely applied in the industry as well as in scientific research, and typically achieve strong performance [14,46]. Note that a full-scale benchmarking analysis including a broad range of predictive and uplift modeling techniques for various datasets is beyond the scope of this study.

To perform the experiments, open source R software is used [47]. CCP modeling relies on implementations in R package *Caret* [48]. For CCU modeling, adapted implementations were used to take into account and contrast customer behavior of the treatment and control groups. Nonetheless, the underlying learning approach of the CCU methods is similar to that adopted by their predictive modeling counterparts. For logistic regression, Lo's approach was used [11] in our experiments to facilitate comparison with the standard logistic regression, whereas the uplift random forests method proposed in Guelman et al. [32] was used via R package *uplift* [49] and compared with the standard random forests method introduced by Breiman [50].

4.2. Results

4.2.1. Scenario 1 - Evaluation with maximum profit

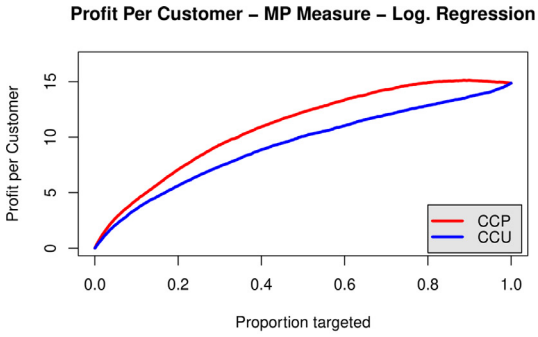
This section reports results of experiments for the first scenario in which the MP measure is used to evaluate the performance of logistic regression and random forests CCP and CCU models, as detailed above. Fig. 5 shows the profit curves for the experiments for scenario 1. As no information was provided by the financial institution regarding the actual values of the cost and benefit parameters of the MP measure, three different sets of parameters are used to calculate MP that are based on values reported in the literature [17,18] and represent cases of low, medium and high returns resulting from retaining a customer. A full sensitivity analysis of the impact of the adopted cost and benefit parameters is beyond the scope of this study and is recognized as a topic for further research. However, the results of experiments performed using the three sets of parameters are fully consistent, and thus conclusions drawn from the experiments appear to hold irrespective of the assumed parameter values.

The profit curves presented in Fig. 5 show the profits earned per customer of the customer base for a proportion α (shown on the x-axis) of customers targeted by the retention campaign. These values may be ranked based on the estimated probability of churn according to CCP models (red profit curves) or based on the estimated uplift score of CCU models (blue profit curves). Note that the profit earned per customer of the customer base, rather than the total profit, is plotted because the former is independent of the size of the customer base and proportional to the total profit. Therefore, the profit curves illustrate the optimal proportion of customers to be targeted by the retention campaign, giving rise to the maximum profit.

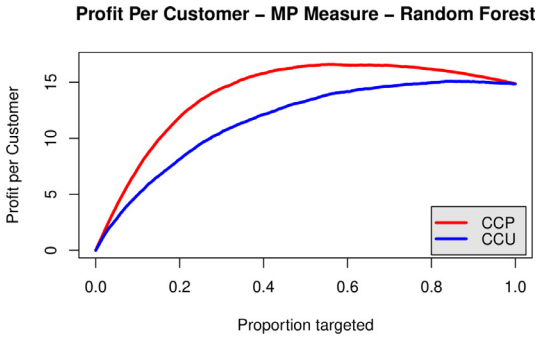
4.2.2. Scenario 2 - Evaluation with maximum profit uplift

Fig. 6 shows the profit curves for the experiments for the second scenario detailed in the previous section, with the results of the CCP and CCU models evaluated using the novel *maximum profit uplift* performance measure. The MPU measure includes both treatment and control group observations in the test set of the evaluation.

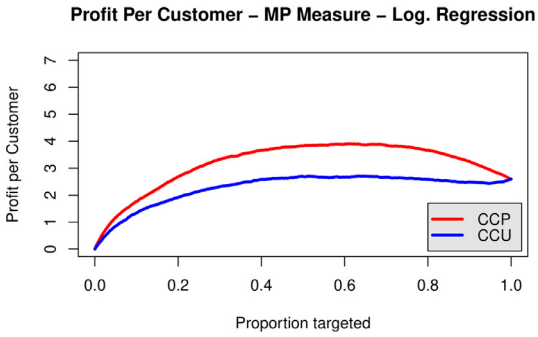
Please cite this article as: F. Devriendt, J. Berrevoets and W. Verbeke, Why you should stop predicting customer churn and start using uplift models, Information Sciences, https://doi.org/10.1016/j.ins.2019.12.075



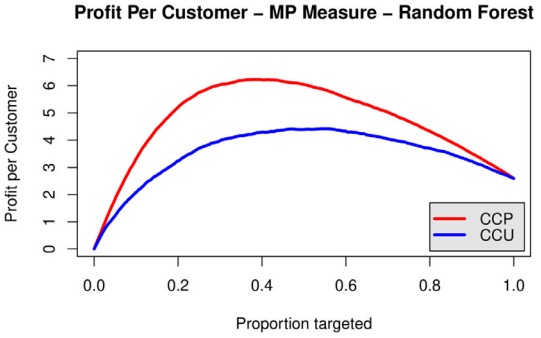
(a) Logistic Regression
$b = 200$, $c_{incentive} = 10$, $c_{contact} = 1$



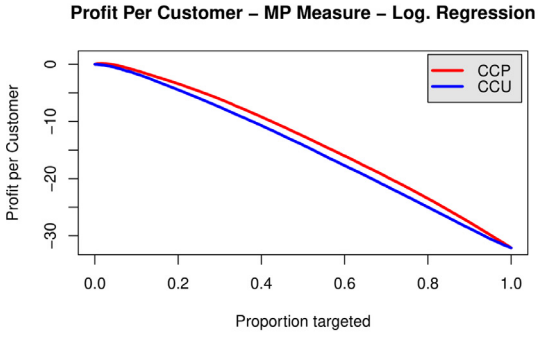
(b) Random Forest
$b = 200$, $c_{incentive} = 10$, $c_{contact} = 1$



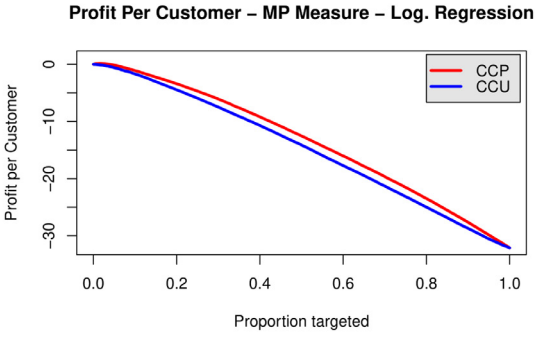
(c) Logistic Regression
$b = 100$, $c_{incentive} = 10$, $c_{contact} = 1$



(d) Random Forest
$b = 100$, $c_{incentive} = 10$, $c_{contact} = 1$



(e) Logistic Regression
$b = 100$, $c_{incentive} = 50$, $c_{contact} = 1$



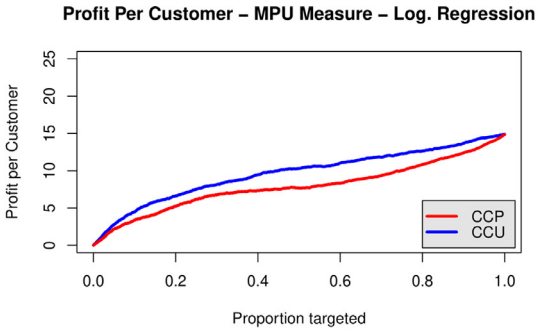
(f) Random Forest
$b = 100$, $c_{incentive} = 50$, $c_{contact} = 1$

Fig. 5. Profit curves for logistic regression (left) and random forests (right) CCP (red curves) and CCU (blue curves) models for the first scenario using the MP measure for three sets of cost and benefit parameters. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

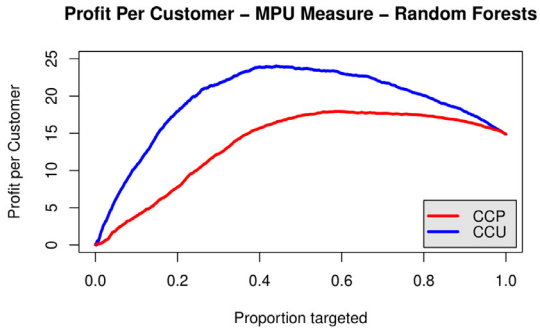
4.3. Discussion

We consider our results from three different perspectives. First, in Section 4.3.1 we examine the profit curves according to MP and MPU formulas. Second, in Section 4.3.2 we assess the churn rate and the corresponding lift and liftup curves. Finally, in Section 4.3.3 we focus on the rank correlation between different setups.

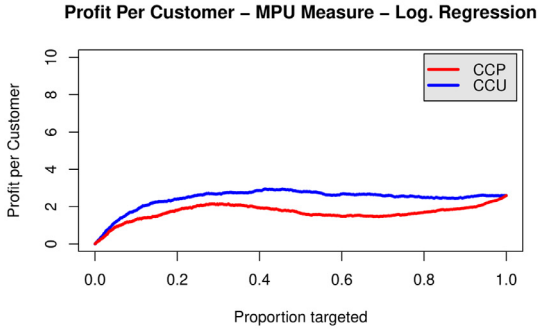
Please cite this article as: F. Devriendt, J. Berrevoets and W. Verbeke, Why you should stop predicting customer churn and start using uplift models, Information Sciences, https://doi.org/10.1016/j.ins.2019.12.075



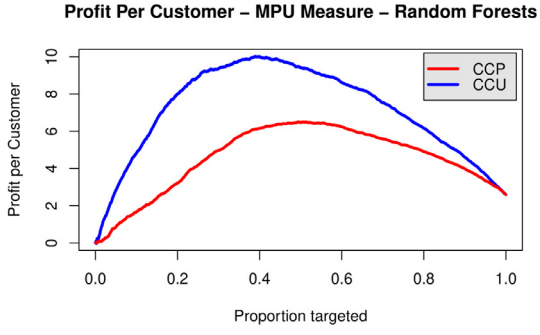
(a) Logistic Regression
$b = 200$, $c_{incentive} = 10$, $c_{contact} = 1$



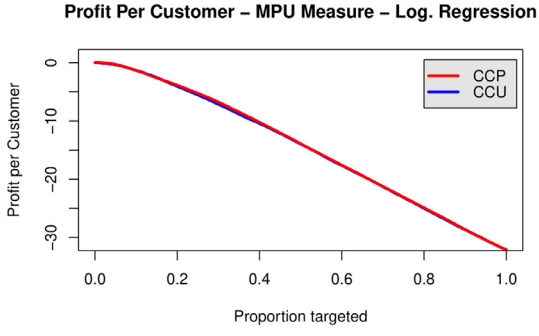
(b) Random Forest
$b = 200$, $c_{incentive} = 10$, $c_{contact} = 1$



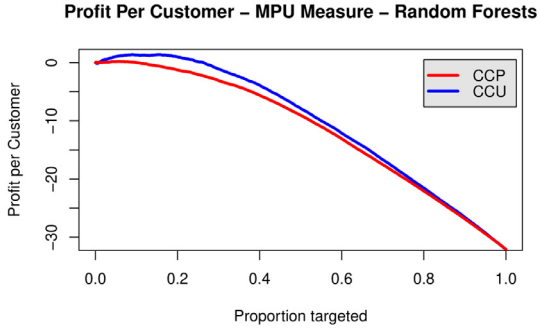
(c) Logistic Regression
$b = 100$, $c_{incentive} = 10$, $c_{contact} = 1$



(d) Random Forest
$b = 100$, $c_{incentive} = 10$, $c_{contact} = 1$



(e) Logistic Regression
$b = 100$, $c_{incentive} = 50$, $c_{contact} = 1$



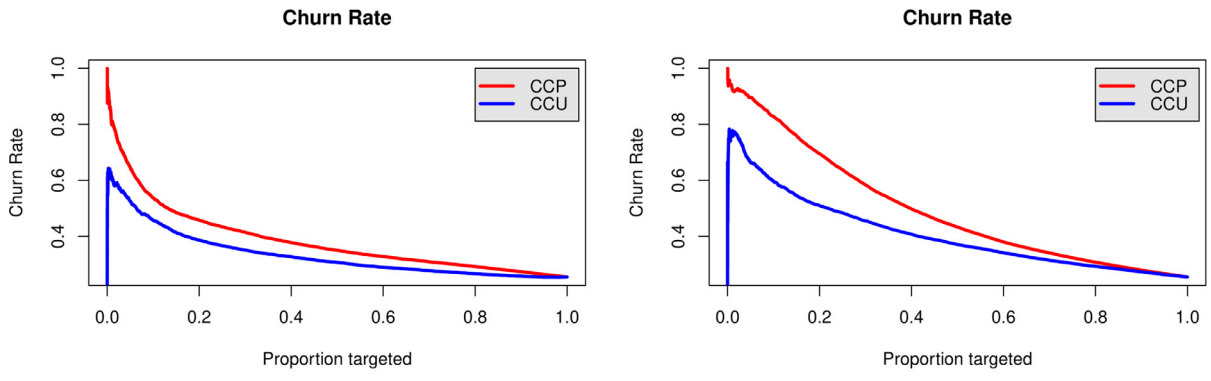
(f) Random Forest
$b = 100$, $c_{incentive} = 50$, $c_{contact} = 1$

Fig. 6. Profit curves for logistic regression (left) and random forests (right) CCP (red curves) and CCU (blue curves) models for the second scenario using the MPU measure for three sets of cost and benefit parameters. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.3.1. Profit curves

The evaluation based on the MP measure for scenario 1 clearly shows that CCP modeling yields higher profits than does CCU modeling for logistic regression and random forests. The profit curves shown in Fig. 5 for CCP models dominate the profit curves of CCU models. We conclude that CCP models are superior in predicting which customers will churn. This is reasonable, as CCP models are trained with the objective of predicting churn, whereas uplift models are developed to predict uplift. Many of the churners identified by the CCP model can be hypothesized to be lost causes. Many churners can

Please cite this article as: F. Devriendt, J. Berrevoets and W. Verbeke, Why you should stop predicting customer churn and start using uplift models, Information Sciences, https://doi.org/10.1016/j.ins.2019.12.075



(a) Logistic Regression (b) Random Forest

Fig. 7. Churn rate as a function of the selected proportion of customers for CCP and CCU models using (a) logistic regression and (b) random forest.

also be expected to have made up their minds and to not be persuaded by a retention campaign. A successful uplift model should rank these customers *at the bottom of the ranking*; i.e., it should estimate their uplift as being close to zero since the retention campaign does not yield any effect. Then, if MP is used as the evaluation measure, it is straightforward to observe that CCP performs better than CCU since CCP models better predict churn, and MP assumes a constant retention rate for targeted churners. In particular, MP does not acknowledge that the retention rate will be as low, if lost causes are targeted, as the true retention rate can be observed if both control and treatment groups are available. As the MP measure is heavily influenced by the predictive power of a model, MP will indicate that the CCP model is superior to CCU models.

This is confirmed by the analysis of results of experiments for scenario 2. Using the MPU measure to evaluate the performance of CCP and CCU models, we observe that CCU models outperform CCP models. This can be attributed to the fact that uplift models effectively succeed in predicting uplift, which is accounted for in the MPU measure, as discussed in Section 3.2, by having the variable retention rate be a function of the targeted proportion of customers α. When ranking both the treatment and control groups in the test set according to the predicted probabilities of churn determined by CCP models and according to the predicted uplift determined by CCU models, we can compare the reduction in churn rate as a function of the selected proportion α plotted along the x-axis for both models. As CCP models assign high rank to customers who are likely to churn but may not necessarily be retained (which is exactly what the CCU model aims to do), CCP models are observed to be less effective in reducing churn and hence to be less profitable than CCU models. The objective of CCU models is to ascribe high scores to customers who are likely to both churn and be retained, and as such, they achieve higher degrees of uplift and profitability.

Note that it is only possible to calculate the MPU measure if both control and treatment groups are present, which is not the case in traditional customer churn prediction setups. The MP measure is still useful in such settings, but the results of experiments indicate that uplift modeling is a superior paradigm with respect to developing a data-driven customer retention program.

In addition, although it is of less importance here, the obtained profit curves clearly show that random forests outperform logistic regression in this case. Random forest models can generate higher profits per customer and higher profits from a smaller proportion of customers targeted by a retention campaign. This result is unsurprising and is fully consistent with the results of benchmarking experiments performed across various business domains and reported in the literature [2,17].

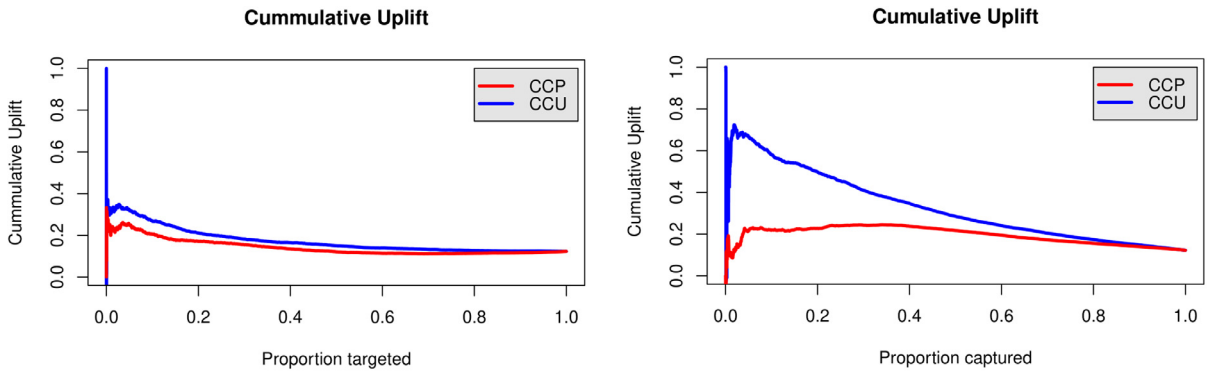
4.3.2. Churn rates

To further analyze and gain insight into the results of experiments, we plot the churn rate as a function of the proportion of targeted customers α for CCP and CCU logistic regression and random forest models in the left and right panels of Fig. 7, respectively. These figures show that the cumulative churn rate for CCP models for all cutoffs exceeds the churn rate for the CCU model. This indicates that the CCP model captures more churners than does the CCU model for the same proportion of selected customers. Additionally, the uplift as a function of the proportion of targeted customers α for CCP and CCU logistic regression and random forest models is plotted in the left and right panels of Fig. 8, respectively. Here, the CCU model detects a causal effect and exploits it to achieve a larger reduction in the churn rate for the treatment group that for the control group.

Figs. 7 and 8 confirm the above analysis and support the conclusion that CCP models tend to detect numerous *lost causes*, whereas CCU models aim and succeed at avoiding targeting *lost causes* and instead allow selecting and targeting *persuadables*, and as a result lead to a larger decrease in the churn rate and yield increased returns. This conclusion holds for both logistic regression and random forest models.

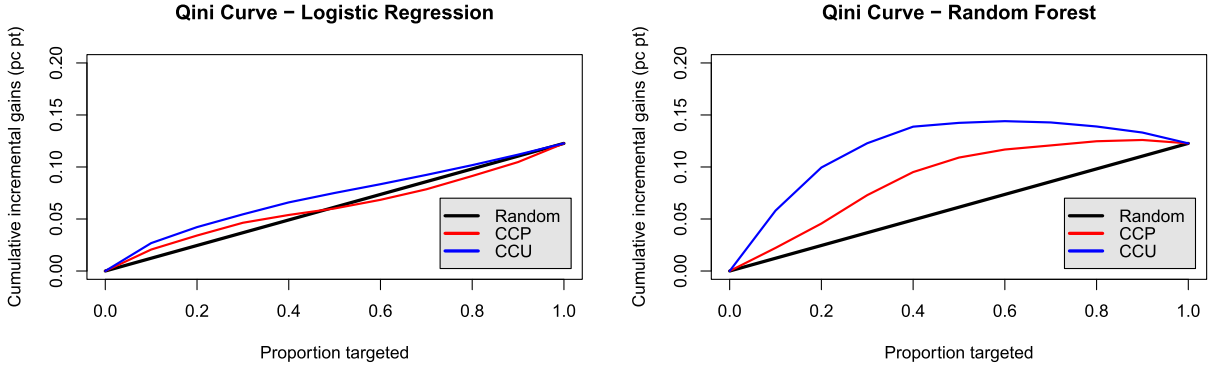
The Qini curves in Fig. 9 show the incremental gains of both the CCP and CCU models. Recall that a Qini curve plots the cumulative difference in response rates between treatment and control groups as a function of the selected proportion

Please cite this article as: F. Devriendt, J. Berrevoets and W. Verbeke, Why you should stop predicting customer churn and start using uplift models, Information Sciences, https://doi.org/10.1016/j.ins.2019.12.075



(a) Logistic Regression (b) Random Forest

Fig. 8. Cumulative uplift as a function of the proportion of customers selected for CCP and CCU models using (a) logistic regression and (b) random forests.



(a) Logistic Regression (b) Random Forest

Fig. 9. Qini curve showing incremental gains as a function of the selected proportion of customers for CCP and CCU models using (a) logistic regression and (b) random forests. The black line corresponds to random targeting.

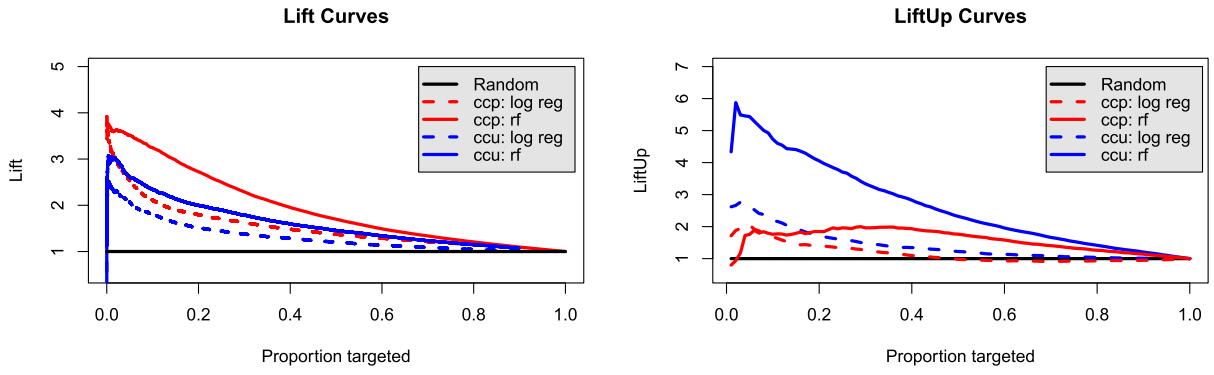
α of customers ranked by an uplift model or a predictive model from high to low values of estimated uplift or probability of churn. The black line in Fig. 9 is the baseline performance achieved when randomly targeting customers. In Fig. 9a, both CCP and CCU logistic regression models barely outperform the baseline. The CCP model performs even worse than the baseline from $\alpha = 50\%$ onwards. A comparison of this to random forest-based models, shown in Fig. 9b, indicates a significant difference in terms of incremental gains. The CCP model performs well; however, most incremental gains are captured at $\alpha = 40\%$ and onwards, whereas the CCU model clearly captures more incremental gains for smaller selected proportions of customers.

In Eqs. (5) and (14), lift and liftup, respectively, are introduced to compare the models' performance at a certain cutoff relative to the overall baseline. Fig. 10 shows the lift curves corresponding to Scenario 1 and the liftup curves corresponding to Scenario 2. Both plots lead to the conclusion similar to that for the profit and Qini curves, where uplift modeling is shown to be very valuable in this churn prediction case study. The CCP random forest model in Fig. 10b shows a non-optimal targeting of the population as the liftup curve rises around 20% to 40% instead of exhibiting a steady decline when an optimal targeting has been determined. This further demonstrates that CCP models tend to target so-called *sure things* over *persuadables*.

4.3.3. Similarity in ranking

The next step in the analysis of the experimental results involves an assessment of similarities in the rankings of customers determined using various models. To this end, Spearman's rank order correlation and Kendall's tau are calculated for the first and second scenarios and are reported in Tables 3 and 4. We observe that, overall, the rankings resulting from various models differ substantially. For the first scenario, the strongest similarity is observed between the logistic regression-based CCP and CCU models and between the random forest-based CCP and CCU models, both having the maximum observed value of the Spearman's rank order correlation of 0.52. The weakest similarities are observed between the CCP logistic regression model and the CCU random forest model, with the Spearman's rank order correlation value of 0.31 in the first

Please cite this article as: F. Devriendt, J. Berrevoets and W. Verbeke, Why you should stop predicting customer churn and start using uplift models, Information Sciences, https://doi.org/10.1016/j.ins.2019.12.075



(a) Lift Curves (b) LiftUp Curves

Fig. 10. (a) Lift and (b) liftup compare the performance of the model in terms of (a) lift and (b) uplift achieved over random targeting.

Table 3
Spearman's rank order correlation and Kendall's tau, scenario 1.

	Scenario 1 - Spearman's rank order correlation				Scenario 1 - Kendall's tau			
	CCP Log Reg	CCP RF	CCU Log Reg	CCU RF	CCP Log Reg	CCP RF	CCU Log Reg	CCU RF
CCP Log Reg	1	0.46	0.52	0.31	1	0.32	0.39	0.21
CCP RF	0.46	1	0.23	0.52	0.32	1	0.15	0.37
CCU Log Reg	0.52	0.23	1	0.39	0.39	0.15	1	0.27
CCU RF	0.31	0.52	0.39	1	0.21	0.37	0.27	1

Table 4
Spearman's rank order correlation and Kendall's tau, scenario 2.

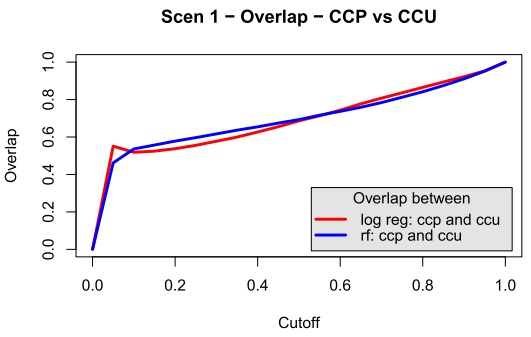
	Scenario 2 - Spearman's rank order correlation				Scenario 2 - Kendall's tau			
	CCP Log Reg	CCP RF	CCU Log Reg	CCU RF	CCP Log Reg	CCP RF	CCU Log Reg	CCU RF
CCP Log Reg	1	0.47	0.59	0.17	1	0.32	0.44	0.12
CCP RF	0.47	1	0.24	0.35	0.32	1	0.16	0.25
CCU Log Reg	0.59	0.24	1	0.35	0.44	0.16	1	0.24
CCU RF	0.17	0.35	0.35	1	0.12	0.25	0.24	1

scenario and a value of only 0.17 in the second scenario, when assessing both the treatment and control groups' test sets. For the CCP random forest and CCU logistic regression models, the Spearman's rank order correlation is observed to be 0.23 in the first scenario and 0.24 in the second scenario. These models are most dissimilar since both are different in terms of predictive versus uplift paradigm and the use of logistic regression versus the random forest modeling method. In the second scenario that considers both the control and treatment groups' test sets, we observe that the rankings of CCP and CCU logistic regression models are more similar, whereas the Spearman's rank order correlations for the rankings of CCP and CCU random forest models decreases to 0.35, which is equal to the correlation between the CCU logistic regression and CCU random forest models. Overall, these results confirm that CCU and CCP models will lead to significantly different selections of customers to be targeted in retention campaigns.

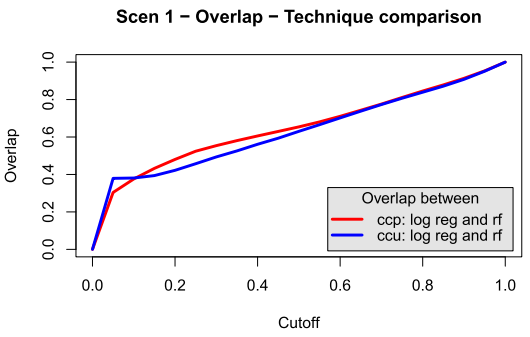
The results of the previous analysis regarding the similarities in the rankings of customers are confirmed by considering the proportions, referred to as overlap, of the same customers selected by various models at a certain cutoff α. Fig. 11 shows the overlap between various techniques and methodologies. In the first scenario, both the logistic regression and random forest models exhibit an overlap of 55% and 46% at $\alpha = 5\%$ for the CCP and CCU settings, respectively (Fig. 11a). When comparing logistic regression and random forest models for each setting, we observe a lower overlap of 30% and 38% at $\alpha = 5\%$, respectively (Fig. 11b), confirming the significantly different selections of targeted customers resulting from different models. In the second scenario, logistic regression models exhibit a 52% overlap between the CCP and CCU setups at the cutoff of $\alpha = 5\%$. The largest difference is observed between the random forest models in the CCP and CCU setups with an overlap of 21% at $\alpha = 5\%$ (Fig. 11c). Finally, Fig. 11d shows overlaps of 31% and 29% at $\alpha = 5\%$ for both CCP models and for both CCU models, respectively. This indicates a significant difference between the rankings in a comparison of logistic regression and random forest models for either CCP or CCU.

In previous studies of uplift modeling, performance of uplift models has been reported to be unstable, i.e., to vary significantly across test folds in an n-fold cross validation setup [14]. Therefore, the experiments discussed above were repeated

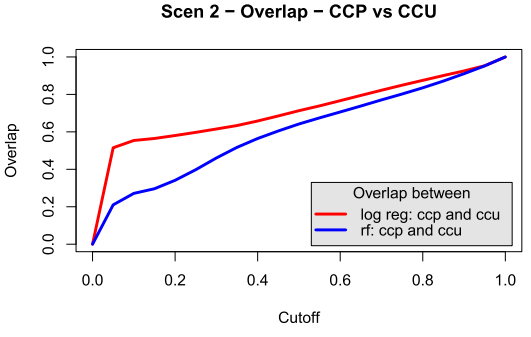
Please cite this article as: F. Devriendt, J. Berrevoets and W. Verbeke, Why you should stop predicting customer churn and start using uplift models, Information Sciences, https://doi.org/10.1016/j.ins.2019.12.075



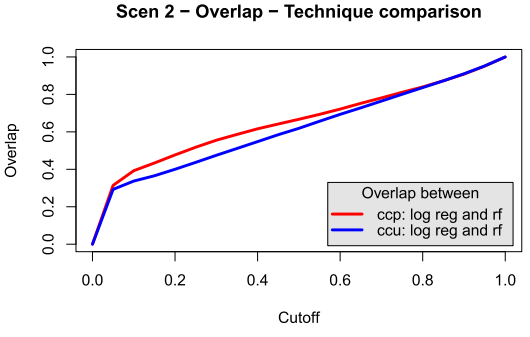
(a) Overlap comparison across methodologies for scenario 1.



(b) Overlap comparison across techniques for scenario 1.



(c) Overlap comparison across methodologies for scenario 2.



(d) Overlap comparison across techniques for scenario 2.

Fig. 11. Overlap in customers observed when comparing different cutoffs of the ranking of setups (11a and 11c) and techniques (11b and 11d) for scenarios 1 and 2.

five times to assess the impact of randomly splitting the dataset into training and test sets. The results obtained in the five repetitions and reported above were observed to be very stable, which supports the validity of the presented findings.

5. Conclusions and directions for future research

Prescriptive analytics, and more specifically uplift modeling, have been proposed as a paradigm that improves predictive analytics. However, no empirical evidence of uplift models outperforming predictive models has been presented in the literature. No comparative studies have been performed, and an indirect comparison based on the performance of predictive models reported in empirical studies of predictive modeling and the performance of uplift models reported in the literature is prevented by the use of different evaluation measures used to evaluate both types of models. In uplift modeling, the predicted outcome is not observed, hence the need for alternative evaluation measures that can be used to evaluate uplift models compared to predictive models. Of course, predictive models can be evaluated in terms of the uplift they achieve, whereas uplift models can be evaluated in terms of their ability to predict. However, since the above would entail evaluating a model in terms of an evaluation measure that is inconsistent with the objective adopted in learning the model, such comparisons seem unfair and inappropriate. Therefore, in this article we adopt a profit-driven evaluation approach that aligns the evaluation of analytical models with their business objective. Specifically, we focus on the use of analytics to support customer churn retention efforts, which is a standard marketing analytics application in the industry.

We introduce a novel, profit-driven evaluation measure called the maximum profit uplift measure for assessing the performance of customer churn uplift models. The proposed MPU measure extends the maximum profit measure for customer churn prediction models and allows assessing the performance of a customer churn uplift model in terms of profit per customer in the customer base that is earned when targeting the optimal proportion of customers with the highest uplift scores by a retention campaign. The optimal proportion of customers to be targeted is determined by maximizing the profit generated by the retention campaign, which is shown in this article to be directly related to the ability of the uplift model to identify the so-called persuadables, i.e., customers who are about to churn who will be retained if targeted by the campaign.

Please cite this article as: F. Devriendt, J. Berrevoets and W. Verbeke, Why you should stop predicting customer churn and start using uplift models, Information Sciences, https://doi.org/10.1016/j.ins.2019.12.075

The MP measure for customer churn prediction is directly related to the ability of a predictive model to predict customer churn, as represented by the lift that is a function of the proportion of targeted customers. To define the MPU measure, we introduce the equivalent liftup measure that is a function of the proportion of targeted customers and expresses the uplift achieved at a certain cutoff relative to the overall baseline uplift achieved when targeting all customers. The introduction of the liftup measure is a genuine contribution of this article, and presents a generally applicable evaluation approach to assessing the use of an uplift model. The liftup is shown to be directly related to profit, and therefore, as argued by Neslin et al. [18], is an appropriate approach to evaluation. If liftup is used in the profit formula for retention campaigns, it becomes clear that the assumption of a constant retention rate in the profit formula adopted by the MP measure is inappropriate for evaluating either a CCU model or a CCP model, as shown in the empirical part of the study.

This article presents the results of a real-life case study in the financial industry. An experimental study was developed and performed to assess the added value of prescriptive (i.e., uplift modeling) over predictive analytics. The results indicate that customer churn uplift models outperform customer churn prediction models. Uplift models effectively appear to be able to identify the so-called persuadables and therefore yield higher returns than do customer churn prediction models. CCP models seem to assign the highest rank and thus lead to the selection of lost causes, i.e., customers who are about to churn but will not be retained if targeted by the retention campaign. These results strongly imply that uplift modeling is a superior paradigm to predictive modeling in the context of supporting customer retention efforts and likely beyond.

Future studies will focus on generalizing the newly introduced MPU measure to facilitate its adoption in different fields, as there is a need for powerful, application-oriented evaluation measures for assessing the performance of uplift models. This study also opens doors to the development of profit-driven uplift modeling approaches that aim at maximizing profitability. Finally, further empirical evidence—obtained by examining a broader variety of application types—of the added value of uplift versus predictive modeling is necessary to validate the findings of this paper.

Declarations of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] B. Baesens, Analytics in a Big Data World: The Essential Guide to Data Science and its Applications, John Wiley & Sons, 2014.
[2] S. Lessmann, B. Baesens, H.-V. Seow, L.C. Thomas, Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research, Eur. J. Oper. Res. 247 (1) (2015) 124–136.
[3] B. Baesens, V. Van Vlasselaer, W. Verbeke, Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection, John Wiley & Sons, 2015.
[4] K. Coussement, K.W. De Bock, Customer churn prediction in the online gambling industry: the beneficial effect of ensemble learning, J. Bus. Res. 66 (9) (2013) 1629–1636.
[5] W. Verbeke, B. Baesens, C. Bravo, Profit Driven Business Analytics: A Practitioner's Guide to Transforming Big Data into Added Value, John Wiley & Sons, 2017.
[6] W. Verbeke, D. Martens, C. Mues, B. Baesens, Building comprehensible customer churn prediction models with advanced rule induction techniques, Expert Syst. Appl. 38 (3) (2011) 2354–2364.
[7] M.R. Colgate, P.J. Danaher, Implementing a customer relationship strategy: the asymmetric impact of poor versus excellent execution, J. Acad. Mark. Sci. 28 (3) (2000) 375–387, doi:10.1177/0092070300283006.
[8] J. Ganesh, M.J. Arnold, K.E. Reynolds, Understanding the customer base of service providers: an examination of the differences between switchers and stayers, J. Mark. 64 (3) (2000) 65–87, doi:10.1509/jmkg.64.3.65.18028.
[9] R.T. Rust, A.J. Zahorik, Customer satisfaction, customer retention, and market share, J. Retail. 69 (2) (1993) 193–215, doi:10.1016/0022-4359(93)90003-2.
[10] D.V. den Poel, B. Larivière, Customer attrition analysis for financial services using proportional hazard models, Eur. J. Oper. Res. 157 (1) (2004) 196–217, doi:10.1016/S0377-2217(03)00069-9. Smooth and Nonsmooth Optimization.
[11] V.S.Y. Lo, The true lift model: a novel data mining approach to response modeling in database marketing, SIGKDD Explor. Newsl. 4 (2) (2002) 78–86, doi:10.1145/772862.772872.
[12] N.J. Radcliffe, R. Simpson, Identifying who can be saved and who will be driven away by retention activity., J. Telecommun. Manag. 1 (2) (2008).
[13] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, D. Dennison, Hidden technical debt in machine llearning systems, in: C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems 28, Curran Associates, Inc., 2015, pp. 2503–2511. http://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf.
[14] F. Devriendt, D. Moldovan, W. Verbeke, A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics, Big Data 6 (1) (2018) 13–41, doi:10.1089/big.2017.0104. PMID: 29570415.
[15] E. Ascarza, Retention futility: targeting high-risk customers might be ineffective, J. Mark. Res. 55 (1) (2018) 80–98, doi:10.1509/jmr.16.0163.
[16] N.J. Radcliffe, Using control groups to target on predicted lift: building and assessing uplift models, Direct Mark. J. Direct. Mark. Assoc. Anal. Council. 1 (2007) 14–21.
[17] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, B. Baesens, New insights into churn prediction in the telecommunication sector: a profit driven data mining approach, Eur. J. Oper. Res. 218 (1) (2012) 211–229, doi:10.1016/j.ejor.2011.09.031.
[18] S.A. Neslin, S. Gupta, W. Kamakura, J. Lu, C.H. Mason, Defection detection: measuring and understanding the predictive accuracy of customer churn models, J. Mark. Res. 43 (2) (2006) 204–211, doi:10.1509/jmkr.43.2.204.
[19] T. Verbraken, C. Bravo, R. Weber, B. Baesens, Development and application of consumer credit scoring models using profit-based classification measures, Eur. J. Oper. Res. 238 (2) (2014) 505–513, doi:10.1016/j.ejor.2014.04.001.
[20] W. Buckinx, D.V. den Poel, Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual fmcg retail setting, Eur. J. Oper. Res. 164 (1) (2005) 252–268, doi:10.1016/j.ejor.2003.12.010.
[21] P. Datta, B. Masand, D.R. Mani, B. Li, Automated cellular modeling and prediction on a large scale, Artif. Intell. Rev. 14 (6) (2000) 485–502, doi:10.1023/A:1006643109702.
[22] E. Lima, C. Mues, B. Baesens, Domain knowledge integration in data mining using decision tables: case studies in churn prediction, J. Oper. Res. Soc. 60 (8) (2009) 1096–1106, doi:10.1057/jors.2008.161.

Please cite this article as: F. Devriendt, J. Berrevoets and W. Verbeke, Why you should stop predicting customer churn and start using uplift models, Information Sciences, https://doi.org/10.1016/j.ins.2019.12.075

[23] Z.-Y. Chen, Z.-P. Fan, M. Sun, A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data, Eur. J. Oper. Res. 223 (2) (2012) 461–472.
[24] S.-Y. Hung, D.C. Yen, H.-Y. Wang, Applying data mining to telecom churn management, Expert Syst. Appl. 31 (3) (2006) 515–524, doi:10.1016/j.eswa.2005.09.080.
[25] W. Verbeke, D. Martens, B. Baesens, Social network analysis for customer churn prediction, Appl. Soft Comput. 14 (2014) 431–446.
[26] A. Backiel, B. Baesens, G. Claeskens, Predicting time-to-churn of prepaid mobile telephone customers using social network analysis, J. Oper. Res. Soc. 67 (9) (2016) 0, doi:10.1057/jors.2016.8.
[27] K. Coussement, S. Lessmann, G. Verstraeten, A comparative analysis of data preparation algorithms for customer churn prediction: a case study in the telecommunication industry, Decis. Support Syst. 95 (2017) 27–36, doi:10.1016/j.dss.2016.11.007.
[28] D.B. Rubin, Causal inference using potential outcomes: design, modeling, decisions, J. Am. Stat. Assoc. 100 (469) (2005) 322–331.
[29] N. Radcliffe, Generating incremental sales: maximizing the incremental impact of cross-selling, up-selling and deep-selling through uplift modelling, Stochast. Solut. Limit. (2007).
[30] M. Jaskowski, S. Jaroszewicz, Uplift modeling for clinical trial data, ICML Workshop on Clinical Data Analysis, 2012.
[31] D. Porter, Pinpointing the persuadables: Convincing the right voters to support barack obama, 2012, (http://www.predictiveanalyticsworld.com/patimes/video-dan-porter-clip/). Accessed: 2016-04-12.
[32] L. Guelman, M. Guillen, A.M. Perez-Marin, Random forests for uplift modeling: an insurance customer retention case, in: K.J. Engemann, A.M. Gil-Lafuente, J. Merigo (Eds.), Modeling and Simulation in Engineering, Economics and Management, Lecture Notes in Business Information Processing, 115, Springer Berlin Heidelberg, 2012, pp. 123–133, doi:10.1007/978-3-642-30433-0_13.
[33] L. Lai, S.F.U. (Canada)., Influential Marketing: A New Direct Marketing Strategy Addressing the Existence of Voluntary Buyers, Canadian theses on microfiche, Simon Fraser University (Canada), 2006. https://books.google.be/books?id=5EvSuAAACAAJ.
[34] P.W. Holland, Statistics and causal inference, J. Am. Stat. Assoc. 81 (396) (1986) 945–960.
[35] K. Kane, V.S.Y. Lo, J. Zheng, True-lift modeling: comparison of methods, J. Mark. Anal. 2 (4) (2014) 218–238.
[36] A. Shaar, T. Abdessalem, O. Segard, Pessimistic uplift modeling, ACM SIGKDD (2016). 10.475/123_4.
[37] N.J. Radcliffe, P.D. Surry, Real-world uplift modelling with significance-based uplift trees, White Paper TR-2011-1, Stochastic Solutions (2011).
[38] D.M. Chickering, D. Heckerman, A decision theoretic approach to targeted advertising, in: Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence, in: UAI'00, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000, pp. 82–88. http://dl.acm.org/citation.cfm?id=2073946.2073957.
[39] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, Classification and regression trees, CRC press, 1984.
[40] G.V. Kass, An exploratory technique for investigating large quantities of categorical data, Appl. Stat. (1980) 119–127.
[41] P. Rzepakowski, S. Jaroszewicz, Decision trees for uplift modeling with single and multiple treatments, Knowl. Inf. Syst. 32 (2) (2012) 303–327, doi:10.1007/s10115-011-0434-0.
[42] L. Guelman, M. Guillen, A.M. Pérez-Marín, Optimal personalized treatment rules for marketing interventions: a review of methods, a new proposal, and an insurance case study, Working Papers, Universitat de Barcena, UB Riskcenter, 2014. http://ideas.repec.org/p/bak/wpaper/201406.html.
[43] B. Hansotia, B. Rukstales, Incremental value modeling, J. Interact. Mark. 16 (3) (2001) 35–46, doi:10.1002/dir.10035.
[44] B. Hansotia, B. Rukstales, Direct marketing for multichannel retailers: issues, challenges and solutions, J. Database Mark. 9 (3) (2002) 259–266.
[45] T. Verbraken, W. Verbeke, B. Baesens, A novel profit maximizing metric for measuring classification performance of customer churn prediction models, IEEE Trans. Knowl. Data Eng. 25 (5) (2013) 961–973.
[46] K. Dejaeger, W. Verbeke, D. Martens, B. Baesens, Data mining techniques for software effort estimation: a comparative study, IEEE Trans. Softw. Eng. 38 (2) (2012) 375–397, doi:10.1109/TSE.2011.55.
[47] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2013. http://www.R-project.org/.
[48] M.K.C. from Jed Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, the R Core Team, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca, Y. Tang, C. Candan, T. Hunt., caret: Classification and Regression Training, 2018. https://CRAN.R-project.org/package=caret r package version 6.0–80.
[49] L. Guelman, uplift: Uplift Modeling, 2014. R package version 0.3.5.
[50] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

Please cite this article as: F. Devriendt, J. Berrevoets and W. Verbeke, Why you should stop predicting customer churn and start using uplift models, Information Sciences, https://doi.org/10.1016/j.ins.2019.12.075