Correa Bahnsen *et al. Decision Analytics* (2015) 2:5
DOI 10.1186/s40165-015-0014-6

Decision Analytics
a SpringerOpen Journal

RESEARCH Open Access

A novel cost-sensitive framework for customer churn predictive modeling

Alejandro Correa Bahnsen*, Djamila Aouada and Björn Ottersten

*Correspondence:
alejandro.correa@uni.lu
Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, Luxembourg City, Luxembourg

Abstract

Customer churn predictive modeling deals with predicting the probability of a customer defecting using historical, behavioral and socio-economical information. This tool is of great benefit to subscription based companies allowing them to maximize the results of retention campaigns. The problem of churn predictive modeling has been widely studied by the data mining and machine learning communities. It is usually tackled by using classification algorithms in order to learn the different patterns of both the churners and non-churners. Nevertheless, current state-of-the-art classification algorithms are not well aligned with commercial goals, in the sense that, the models miss to include the real financial costs and benefits during the training and evaluation phases. In the case of churn, evaluating a model based on a traditional measure such as accuracy or predictive power, does not yield to the best results when measured by the actual financial cost, ie. investment per subscriber on a loyalty campaign and the financial impact of failing to detect a real churner versus wrongly predicting a non-churner as a churner.

In this paper, we present a new cost-sensitive framework for customer churn predictive modeling. First we propose a new financial based measure for evaluating the effectiveness of a churn campaign taking into account the available portfolio of offers, their individual financial cost and probability of offer acceptance depending on the customer profile. Then, using a real-world churn dataset we compare different cost-insensitive and cost-sensitive classification algorithms and measure their effectiveness based on their predictive power and also the cost optimization. The results show that using a cost-sensitive approach yields to an increase in cost savings of up to 26.4 %.

Keywords: Predictive modeling; Classification; Cost-sensitive; Churn; Customer lifetime value

Background

The two main objectives of subscription-based companies are to acquire new subscribers and retain those they already have, mainly because profits are directly linked with the number of subscribers. In order to maximize the profit, companies must increase the customer base by incrementing sales while decreasing the number of churners. Furthermore, it is common knowledge that retaining a customer is about five times less expensive than acquiring a new one (Farris et al. 2010), this creates pressure to have better and more effective churn campaigns.

A typical churn campaign consists in identifying from the current customer base which ones are more likely to leave the company, and make an offer in order to avoid that behavior.

With this in mind the companies use intelligence to create and improve retention and collection strategies. In the first case, this usually implies an offer that can be either a discount or a free upgrade during certain span of time. In both cases the company has to assume a cost for that offer, therefore, accurate prediction of the churners becomes important. The logic of this flow is shown in Fig. 1.

The typical churn campaign process starts with the sales that every month increase the customer base, however, monthly there is a group of customers that decide to leave the company for many reasons. Then the objective of a churn model is to identify those customers before they take the decision of defecting.

Using a churn model, those customers more likely to leave are predicted as churners and an offer is made in order to retain them. However, it is known that not all customers will accept the offer, in the case when a customer is planning to defect, it is possible that the offer is not good enough to retain him or that the reason for defecting can not be influenced by an offer. Using historical information, it is estimated that a customer will accept the offer with probability γ. On the other hand, there is the case in which the churn model misclassified a non-churner as churner, also known as false positives, in that case the customer will always accept the offer that means and additional cost to the company since those misclassified customers do not have the intentions of leaving.

In the case were the churn model predicts customers as non-churners, there is also the possibility of a misclassification, in this case an actual churner is predicted as non-churner, since these customers do not receive an offer and they will leave the company, these cases are known as false negatives. Lastly, there is the case were the customers are actually non-churners, then there is no need to make a retention offer to these customers since they will continue to be part of the customer base.

It can be seen that a churn campaign (or churn model) have three main points. First, avoid false positives since there is a financial cost of making an offer were it is not needed. Second, to the true positives, give the right offer that maximize γ while maximizing the profit of the company. And lastly, to decrease the number of false negatives.

From a machine learning perspective, a churn model is a classification algorithm. In the sense that using historical information, a prediction of which current customers are more likely to defect, is made. This model is normally created using one of a number of well establish algorithms (Logistic regression, neural networks, random forests, among others) (KhakAbi et al. 2010; Ngai et al. 2009). Then, the model is evaluated using measures such

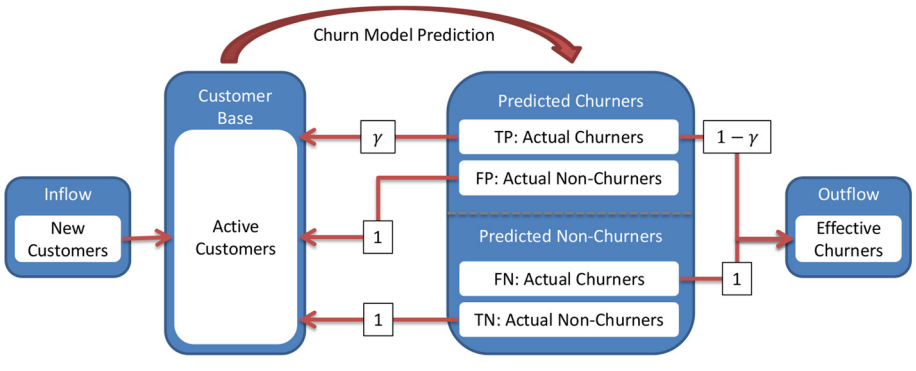

Fig. 1 Flow analysis of a churn campaign (Verbraken 2012)

as misclassification error, receiver operating characteristic (*ROC*), Kolmogorov–Smirnov (*KS*) or F_1Score statistics (Verbeke et al. 2012). However these measures assume that misclassification errors carry the same cost, which is not the case in churn modeling, since failing to identify a profitable or unprofitable churner have significant different financial costs (Glady et al. 2009).

In this paper we propose a new financial based measure for evaluating the effectiveness of a voluntary churn campaign taking into account the available portfolio of offers, their individual financial cost and probability of acceptance depending on the customer profile. Moreover, we compare state-of-the-art classification algorithms, against recently proposed cost-sensitive algorithms such as Bayes minimum risk (Correa Bahnsen et al. 2014b), cost-sensitive logistic regression (Correa Bahnsen et al. 2014a), and cost-sensitive decision trees (Correa Bahnsen et al. 2015). Then using a real-world churn dataset we compare different cost-insensitive and cost-sensitive predictive analytics models, using the traditional and proposed statistics. The results will show that using a cost-sensitive approach results in an increase in profitability of up to 26.4 %. Furthermore, the source code used for the experiments is publicly available as part of the *CostSensitiveClassification* (Correa Bahnsen 2015) library.

The remainder of the paper is organized as follows: The first section, we propose a new financial based measure for evaluating the effectiveness of a churn campaign. Then, we describe the different cost-insensitive and cost-sensitive predictive analytics models. Afterwards, the experimental setup is given. Here the dataset, and its partitioning are presented. Finally the results and the conclusions of the paper are presented in the last two sections.

Evaluation of a churn campaign

Traditionally, a churn model is evaluated as a standard binary classification model, using measures such as misclassification error, receiver operating characteristic (*ROC*), Kolmogorov–Smirnov (*KS*) or F_1Score statistics (Verbeke et al. 2012). Most of these measures are extracted by using a confusion matrix as shown in Table 1.

From this table several statistics are extracted. In particular:

- Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$
- Recall = $\frac{TP}{TP+FN}$
- Precision = $\frac{TP}{TP+FP}$
- $F_1Score = 2\frac{Precision \cdot Recall}{Precision+Recall}$

However, these measures may not be the most appropriate evaluation criteria when evaluating a churn model, because they tacitly assume that misclassification errors carry the same cost, similarly with the correct classified examples. This assumption does not

Table 1 Classification confusion matrix

	Actual positive $y = 1$	Actual negative $y = 0$
Predicted Positive $c = 1$	True Positive (*TP*)	False Positive (*FP*)
Predicted Negative $c = 0$	False Negative (*FN*)	True Positive (*TN*)

hold in many real-world applications such as churn modeling, since when misidentifying a churner the financial losses are quite different than when misclassifying a non-churner as churner (Glady et al. 2009). Furthermore, the accuracy measure also assumes that the class distribution among examples is constant and balanced (Provost et al. 1998), and typically the distributions of a churn data set are skewed (Verbeke et al. 2012).

Different studies have proposed measures to deal with these cost-sensitivity related to evaluating a churn model. In (Neslin et al. 2006), a profit-based measure was proposed by starting with the confusion matrix and multiplying it with the expected profit of each case.

$$Profit_1 = (TP + FP)\left[(\gamma CLV + C_o(1-\gamma)(-C_a))\,\pi_1\gamma - C_o - C_a\right] - A, \tag{1}$$

with A being the fixed administrative cost of running the campaign, C_o the average cost of the retention offer, C_a the cost of contacting the customer, π_1 the prior churn rate and CLV the average customer lifetime value. Moreover, as discussed in (Verbraken et al. 2013), if the average instead of the total profit is considered and the fixed cost A is discarded since is irrelevant for classifier selection, the profit can be expressed as:

$$Profit_2 = TP\,(\gamma(CLV - C_o - C_a) + (1-\gamma)C_a) + FP(-C_o - C_a). \tag{2}$$

Nevertheless, equations (1) and (2), assume that every customer has the same CLV and C_o, whereas this is not true in practice. In fact, different customers have a very different CLV, and not all offers can be made to every customer, neither do they have the same impact across customers. In order to obtain a more business oriented measure, we first analyze the financial impact of the different decisions, ie. false positives, false negatives, true positives and true negatives, for each customer. In Fig. 2, the financial impact of a churn model is shown. Note than we take into account the costs and not the profit in each case.

When a customer is predicted to be a churner, an offer is made with the objective of avoiding the customer defecting. However, if a customer is actually a churner, he may or not accept the offer with a probability γ_i. If the customer accepts the offer, the financial impact is equal to the cost of the offer (C_{o_i}) plus the administrative cost of contacting the customer (C_a). On the other hand, if the customer declines the offer, the cost is the expected income that the clients would otherwise generate, also called customer lifetime

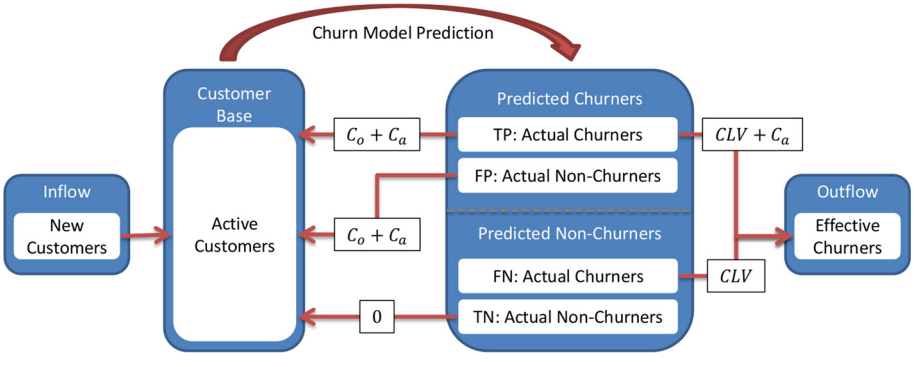

Fig. 2 Financial impact of the different decisions ie. False positives, false negatives, true positives and true negatives

value (CLV_i), plus C_a. Lastly, if the customer is not actually a churner, he will be happy to accept the offer and the cost will be C_{o_i} plus C_a.

In the case that the customer is predicted as non-churner, there are two possible outcomes. Either the customer is not a churner, then the cost is zero, or the customer is a churner and the cost is CLV_i. In Table 2, we summarize the different costs in a cost matrix (Elkan 2001).

Using the cost matrix, and following the example-dependent cost-sensitive framework defined in (Correa Bahnsen et al. 2014a), an example-dependent cost statistic is defined as:

$$\begin{aligned} Cost_i &= y_i(c_i C_{TP_i} + (1-c_i)C_{FN_i}) + (1-y_i)(c_i C_{FP_i} + (1-c_i)C_{TN_i}) \\ &= y_i(c_i(\gamma_i C_{o_i} + (1-\gamma_i)(CLV_i + C_a)) + (1-c_i)CLV_i) \\ &\quad + (1-y_i)(c_i(C_{o_i} + C_a) + (1-c_i)(0)) \\ &= y_i(c_i\left(\gamma_i(C_{o_i} - CLV_i - C_a) - C_{o_i}\right) + CLV_i) + c_i(C_{o_i} + C_a), \end{aligned} \tag{3}$$

leading to a total cost of:

$$Cost = \sum_{i=1}^{N} Cost_i. \tag{4}$$

Furthermore, with the objective of having a measure that is comparable between databases, the savings are defined as:

$$Savings = \frac{Cost_l - Cost}{Cost_l}, \tag{5}$$

where $Cost_l = \min\{Cost(f_0), Cost(f_1)\}$, or minimum between the cost of classifying all the examples as negatives f_0, or the cost of classifying all the examples as positives f_1. In almost cases the costless class will be the negative class, as typically the distribution of a churn dataset is skewed towards the non-churners (Verbeke et al. 2012). Given that $Cost_l$ can be expressed as $Cost(f_0)$, or simply $Cost$ with $c_i = 0\ \forall i$:

$$Cost_l = \sum_{i=0}^{N} y_i CLV_i. \tag{6}$$

This is consistent with the notion that if no model is used, the total cost would be the sum of the customer lifetime values of the actual churners, which gives the insight that the *Savings* measure is comparing the financial impact of the campaign of using a classification model against no using a model at all.

Customer lifetime value

Lastly, one of the key values to calculate the *Savings*, as described in (5), is the customer lifetime value. Within marketing there exists a common misconception between

Table 2 Proposed churn modeling example-dependent cost matrix

	Actual positive $y_i = 1$	Actual negative $y_i = 0$
Predicted Positive $c_i = 1$	$C_{TP_i} = \gamma_i C_{o_i} + (1-\gamma_i)(CLV_i + C_a)$	$C_{FP_i} = C_{o_i} + C_a$
Predicted Negative $c_i = 0$	$C_{FN_i} = CLV_i$	$C_{TN_i} = 0$

customer profitability and customer lifetime value. The two terms are usually used in an interchangeable way, creating confusion of what the actual objective of a churn modeling campaign should be. Several studies have proposed models providing a unique definition of both terms (Milne and Boza 1999; Neslin et al. 2006; Pfeifer et al. 2004; van Raaij et al. 2003). Customer profitability indicates the difference between the income and the cost generated by a customer i during a financial period t. It is defined as:

$$CP_{i,t} = \mu \cdot s_{i,t}, \tag{7}$$

where $s_{i,t}$ refers to the consumption of customer i during time period t, and μ refers to the average marginal profit by unit product usage.

Moreover, we are interested to see what is the expected income that a particular customer will generate in the future, in other words, calculating the expected sum of discount future earnings (Neslin et al. 2006). Therefore, the CLV_i is defined as:

$$CLV_i = \sum_{t=1}^{T} \frac{\mu \cdot s_{i,t}}{(1+r)^t}, \tag{8}$$

where r is the discount rate, and T the number of time period. Typically T should be considered large enough since without prior knowledge a customer is expected to keep being a customer for the foreseeable future. In practice T is set up to be ∞ (Glady et al. 2009). Also, for simplicity it can be assumed that $s_{i,t+1} = s_{i,t} \cdot (1+g)\ \forall i,t$, which means that there is a constant growth g in the customer consumption. Given that, the customer lifetime value can be re-written as

$$CLV_i = \sum_{t=1}^{\infty} \frac{(1+g)^t}{(1+r)^t} \cdot \mu \cdot s_{i,1}, \tag{9}$$

which in the case of $g < r$, this is a geometric series meaning that it can be expressed as

$$CLV_i = \frac{\mu \cdot s_{i,1}}{(r-g)}. \tag{10}$$

Cost-sensitive classification

Classification in the context of machine learning, deals with the problem of predicting the class y_i of a set of examples S, given their k variables, i.e. $X_i = \left[x_i^1, x_i^2, \ldots, x_i^k\right]$. The objective is to construct a function $f(S)$ that makes a prediction c_i of the class of each example using its variables X_i. Traditionally, predictive modeling methods are designed to minimize some sort of misclassification measure such as the $F_1 Score$ (Hastie et al. 2009). However, this means assuming that the different misclassification errors carry the same cost, and as discussed before this is not the case in many real-world applications specifically in churn modeling.

Methods that use different misclassification costs are known as cost-sensitive classifiers. In particular we are interested in methods that are example-dependent cost-sensitive, in the sense that the costs vary among examples and not only among classes. Example-dependent cost-sensitive classification methods can be grouped according to the step where the costs are introduced into the system. Either the costs are introduced prior the training of the algorithm, after the training or during training (Wang 2013). In Fig. 3, the different algorithms are grouped according to the stage in a classification system where they are used.

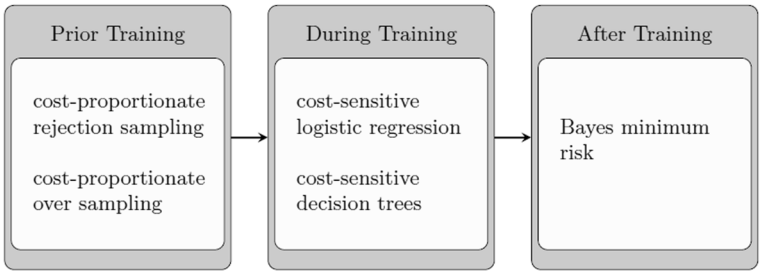

Fig. 3 Different example-dependent cost-sensitive algorithms grouped according to the stage in a classification system where they are used

The first set of methods that were proposed to deal with the cost-sensitivity, consist in re-weighting the training examples based on their costs, either by cost-proportionate rejection-sampling (Zadrozny et al. 2003), or over-sampling (Elkan (Elkan 2001)). The rejection-sampling approach consists in selecting a random subset S_r by randomly selecting examples from S, and accepting each example i with probability $w_i / \max_{1,\dots,N}\{w_i\}$, where w_i is defined as the expected misclassification error of example i

$$w_i = y_i \cdot C_{FN_i} + (1 - y_i) \cdot C_{FP_i}. \tag{11}$$

On the other hand, the over-sampling method consists in creating a new set S_o, by making w_i copies of each example i. However, cost-proportionate over-sampling increases the training since $|S_o| >> |S|$, and it also may result in over-fitting (Drummond and Holte 2003). Furthermore, none of these methods uses the the full cost matrix but only the misclassification costs, which as described in the previous section, is not the case in churn modeling.

However, the aforementioned methods, only introduce the cost by modifying the training set. In (Correa Bahnsen et al. 2013, 2014b), a cost-sensitive model called Bayes minimum risk classifier (*BMR*) was proposed.

The BMR classifier is a decision model based on quantifying tradeoffs between various decisions using probabilities and the costs that accompany such decisions. This is done in a way that for each example the expected losses are minimized. In what follows, we consider the probability estimates p_i as known, regardless of the algorithm used to calculate them. The risk that accompanies each decision is calculated. In the specific framework of binary classification, the risk of predicting the example i as negative is

$$R(c_i = 0|X_i) = C_{TN_i}(1 - \hat{p}_i) + C_{FN_i} \cdot \hat{p}_i, \tag{12}$$

and

$$R(c_i = 1|X_i) = C_{TP_i} \cdot \hat{p}_i + C_{FP_i}(1 - \hat{p}_i), \tag{13}$$

is the risk when predicting the example as positive, where $\hat{p}_i$ is the estimated positive probability for example i. Subsequently, if

$$R(c_i = 0|X_i) \leq R(c_i = 1|X_i), \tag{14}$$

then the example i is classified as negative. This means that the risk associated with the decision c_i is lower than the risk associated with classifying it as positive. However, when using the output of a binary classifier as a basis for decision making, there is a need for a

probability that not only separates well between positive and negative examples, but that also assesses the real probability of the event (Cohen and Goldszmidt 2004), given that, the estimated probabilities are usually calibrated either by an isotonic regression, Platt regression, or the ROC convex hull methodologies (Hernandez-Orallo et al. 2012).

In a recent paper a cost-sensitive logistic regression algorithm (Correa Bahnsen et al. 2014a), was proposed. This method not only uses the costs before or after the training phase, it also introduces the example-dependent costs into a logistic regression, by changing the objective function of the model to one that is cost-sensitive. The the new cost function is defined as:

$$J^c(\theta) = \frac{1}{N}\sum_{i=1}^{N}\Big(y_i(h_\theta(X_i)C_{TP_i} + (1-h_\theta(X_i))C_{FN_i}) + (1-y_i)(h_\theta(X_i)C_{FP_i} + (1-h_\theta(X_i))C_{TN_i})\Big), \tag{15}$$

where $h_\theta(X_i) = g\left(\sum_{j=1}^{k}\theta^j x_i^j\right)$ refers to the hypothesis of i given the parameters θ, and $g(\cdot)$ is the logistic sigmoid function, defined as $g(z) = 1/(1+e^{-z})$. To find the coefficients of the regression θ, the cost function is minimized by using binary genetic algorithms (Haupt and Haupt 2004).

Following the same objective of modifying an existing algorithm by introducing the different cost into its calculation, a cost-sensitive decision tree algorithm (Correa Bahnsen et al. 2015) was recently proposed. In this method a new splitting criteria is used during the tree construction. In particular instead of using a traditional splitting criteria such as Gini, entropy or misclassification, the *Cost* as defined in (4), of each tree node is calculated, and the gain of using each split evaluated as the decrease in total *Savings* of the algorithm.

Experimental setup

In this section we describe the dataset used to evaluate the different cost-insensitive and cost-sensitive classification algorithms. Afterwards, we show the procedure used to estimate the probability of acceptance (γ_i) of each customer. Lastly, the partitioning of the dataset is shown.

Database

For this paper we used a dataset provided by a TV cable provider. The dataset consists of active customers during the first semester of 2014. The total dataset contains 9,410 individual registries, each one with 45 attributes, including a churn label indicating whenever a customer is a churner. This label was created internally in the company, and can be regarded as highly accurate. In the dataset only 455 customers are churners, leading to a churn ratio of 4.83 %.

Offer acceptance calculation

In practice companies have a set of offers to make to a customer as a part of the retention campaign, they vary from discounts, to upgrades among others. In the particular case of a TV cable provided, the offers include adding a new set of channels, changing the TV receiver to one with new technology (ie. high definition, video recording, 4K), or to offer a

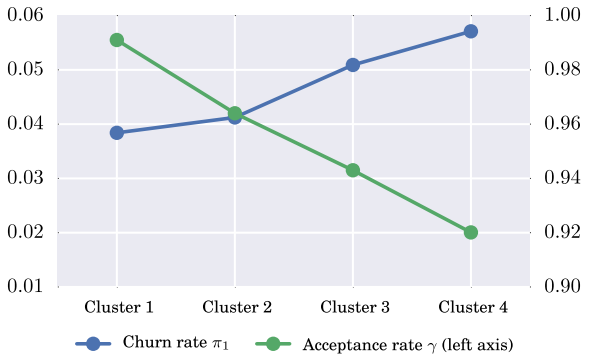

Fig. 4 Acceptance rate (γ) of the best offer for each customer profile. As expected, the higher the churn rate the lower the acceptance rate, as it is more difficult to make a good offer to a customer which is more likely to defect

discount on the monthly bill. Unsurprisingly, not all offers apply to all clients. For instance a customer that already has all the channels can not be offered a new set of channels. Moreover, an offer usually means an additional cost to the company and not all offers have not the same cost or the same impact in reducing churn.

Taking into account the cost and the implication of the offers, the problem can be resumed in making each customer the offer that will maximize the acceptance rate and more important reducing the overall cost.

In order to calculate the acceptance probability γ_i a champion-challenger process was made. First, the customers were grouped into clusters according to their behavioral and socio-economical characteristics. In particular the K-means algorithm was used (Marslan 2009). Then for a period of two months, randomly selected offers were made to the customers and their response was evaluated. Unfortunately, for confidentiality reasons we can not describe the different clusters, neither the actual offer made to each customer. Nevertheless, in Fig. 4, the average churn rate and acceptance rate γ_i per cluster is shown. As expected, the higher the churn rate the lower the acceptance rate, as it is more difficult to make a good offer to a customer which is more likely to defect.

Database partitioning

From the initial dataset, three different datasets are extracted: training, validation and testing. Each one containing 50 %, 25 % and 25 % of the examples, respectively. Afterwards, with the objective of having a more balanced dataset, ie. same distribution of churners and not churners, an under-sampling of the churners is made. Lastly, we also

Table 3 Description of datasets

Set	N	π_1	C_0
Total	9,410	.0483	580,884
Training (t)	3,758	.0505	244,542
Validation	2,824	.0477	174,171
Testing	2,825	.0442	162,171
Under-sampling (u)	374	.5080	244,542
CS Rejection-sampling (r)	428	.4135	431,428
CS Over-sampling (o)	5,767	.03124	2,350,285

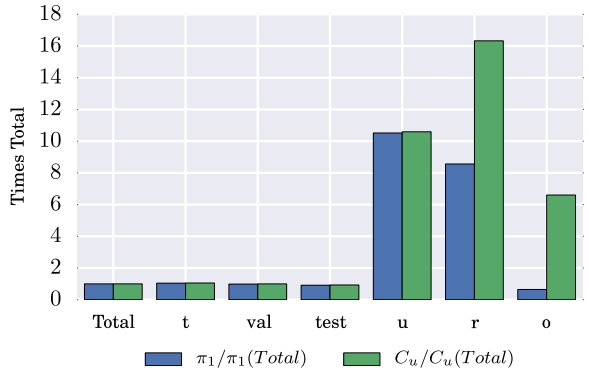

Fig. 5 Comparison of the distribution of the churn rate and the average unitary cost ($C_u = C_0/N$) of each dataset with respect to the Total set. It is observed that the cost-sensitive sampling procedures have an average unitary cost much higher than the other sets, also showing why a simple under-sampling procedure of the dataset does not take into account the costs

applied the cost-sensitive re-balancing techniques cost-proportionate rejection-sampling (Zadrozny et al. 2003) and cost-proportionate over-sampling (Elkan 2001), described in cost-sensitive classification Section. Table 3, summarizes the different datasets, where N, π_1 and C_0 represents the number of customers, the percentage of churners and the total losses if no model is used, respectively. Moreover, in Fig. 5, a comparison of the churn rate and the average unitary cost ($C_u = C_0/N$) of each dataset with respect of the Total set is shown. It is observed that the cost-sensitive sampling procedures have an average unitary cost much higher than the other sets, also showing why a simple under-sampling procedure of the dataset does not take into account the costs.

Results

For the experiments we first used three classification algorithms, decision tree (*DT*), logistic regression (*LR*) and a random forest (*RF*). Each algorithm is trained using the different training sets: training (*t*), under-sampling (*u*), cost-proportionate rejection-sampling (*r*)

Table 4 Results of the decision tree (*DT*), logistic regression (*LR*) and random forest (*RF*) algorithms, estimated using the different training sets: training (*t*), under-sampling (*u*), cost-proportionate rejection-sampling (*r*) and cost proportionate over-sampling (*o*)

Algorithm	Set	*Savings*	F_1 *Score*
DT	t	-0.0001 ± 0.0193	0.0750 ± 0.0199
	u	-0.0370 ± 0.0603	0.1177 ± 0.0108
	r	0.0018 ± 0.0549	0.1200 ± 0.0129
	o	0.0249 ± 0.0203	0.1019 ± 0.0189
LR	t	-0.0001 ± 0.0002	0.0000 ± 0.0000
	u	0.0062 ± 0.0487	0.1227 ± 0.0097
	r	0.0500 ± 0.0372	0.1260 ± 0.0112
	o	0.0320 ± 0.0225	0.1088 ± 0.0199
RF	t	-0.0026 ± 0.0081	0.0245 ± 0.0148
	u	0.0424 ± 0.0547	0.1342 ± 0.0113
	r	0.1033 ± 0.0402	0.1443 ± 0.0127
	o	0.0205 ± 0.0161	0.0845 ± 0.0204

Table 5 Results of the decision tree (*DT*), logistic regression (*LR*) and random forest (*RF*) algorithms, estimated using the different training sets

Algorithm	Set	*Savings*	F_1 *Score*
DT - BMR	t	0.0303 ± 0.0148	0.0946 ± 0.0158
	u	0.0574 ± 0.0387	0.1095 ± 0.0203
	r	0.0652 ± 0.0365	0.1151 ± 0.0169
	o	0.0306 ± 0.0149	0.0924 ± 0.0184
LR - BMR	t	0.1058 ± 0.0319	0.1361 ± 0.0154
	u	0.0963 ± 0.0388	0.1319 ± 0.0166
	r	0.0823 ± 0.0364	0.1240 ± 0.0153
	o	0.0986 ± 0.0287	0.1333 ± 0.0149
RF - BMR	t	0.0835 ± 0.0349	0.1252 ± 0.0151
	u	0.1300 ± 0.0368	0.1429 ± 0.0127
	r	0.1336 ± 0.0348	0.1429 ± 0.0132
	o	0.0907 ± 0.0359	0.1275 ± 0.0136

and cost-proportionate over-sampling (*o*). Unless otherwise stated, the random selection of the training set was repeated 50 times, and in each time the models were trained and results collected, this allows us to measure the stability of the results. On Table 4 the results are shown. First, when observing the results on the *t* and *u* sets, the *RF* algorithm produces the best result by savings. Nevertheless, it is observed that the model with the highest savings is not the same as the one with the highest F_1*Score*, corroborating the conclusions from (Correa Bahnsen et al. 2013), as selecting a method by a traditional statistic does not gives the same result as selecting it using a business oriented measure such as financial savings. Lastly, when observing the results from the algorithms estimated using the cost-proportionate *r* and *o* sets, in all cases the results measured by savings notably increase. The *RF* estimated with the *r* set, arises to savings of 10.33 %, more than twice the best model without using the cost-proportionate sampling sets.

Furthermore, using each algorithm's estimated probabilities we evaluate the result of the Bayes minimum risk (BMR) model. On Table 5, the results are given. The best model measured by savings is the *RF* trained with the *r* set. As is shown in Fig. 6, for all algorithms there is an increase in savings when using the *BMR*. The *BMR* algorithm generates

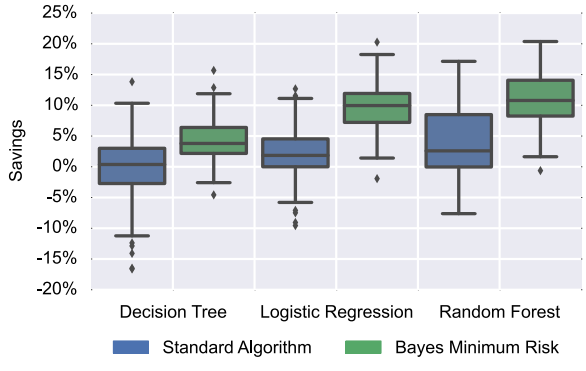

Fig. 6 Savings of the algorithms with and without using the *BMR* algorithm, estimated using the five training sets. The *BMR* algorithm generates an increase in savings regardless of the algorithm used

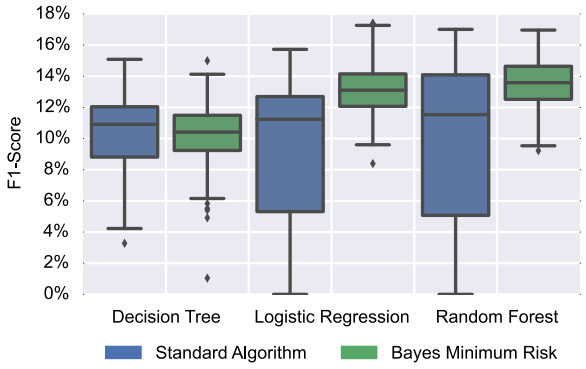

Fig. 7 F_1 *Score* of the algorithms with and without using the *BMR* algorithm, estimated using the five training sets. There is no significant impact of using the *BMR* algorithm

an increase in savings regardless of the algorithm used. However, as can be observed in Fig. 7, there is no significant impact of using the *BMR* algorithm.

Subsequently, we evaluate the cost-sensitive logistic regression (*CSLR*), estimated using the default parameters as suggested in (Correa Bahnsen et al. 2014a). The results are shown on Table 6. The *CSLR* method produces the significantly better results measured by savings than the previously analyzed models. This method arise to an increase in savings of 10.82 % compared with the $RF - BMR$ model. Similarly, with the $F_1 Score$ statistic, the best results are found when using the training set. However, this model is consistently more unstable, since the standard deviation of the savings is more than twice than the one of the previous models. This may be because, this model is estimated using genetic algorithms, which is a random based algorithm.

Finally, the cost-sensitive decision trees (*CSDT*) is evaluated. The results are shown on Table 7. Overall, the *CSDT* method produces highest results measured by savings. As shown in Fig. 8, the *CSDT* algorithm has not only the highest savings but also is much more stable than the *CSLR*. Moreover, it is interesting that both cost-sensitive methods that include the costs during the training phase have the best results when trained using the full training set, and that is because the algorithms need to learn the actual population and costs distributions. Lastly, in Fig. 9, a comparison of the $F1 - Score$ of the different algorithms is made. It is observed that the best model selected by savings is not the same as the one selected by the $F1 - Score$ confirming the intuition that a model should be both trained and evaluated taking into account the actual financial costs of the application, in this case of the churn campaign process.

Table 6 Results of the cost-sensitive logistic regression (*CSLR*) algorithm, estimated using the different training sets

Algorithm	Set	*Savings*	F_1 *Score*
CSLR	t	0.2418 ± 0.0859	0.1079 ± 0.0318
	u	0.1933 ± 0.0879	0.0908 ± 0.0055
	r	0.1971 ± 0.0897	0.0911 ± 0.0057
	o	0.2042 ± 0.0914	0.0917 ± 0.0060

Table 7 Results of the cost-sensitive decision tree (*CSDT*) algorithm, estimated using the different training sets

Algorithm	Set	*Savings*	F_1 *Score*
CSDT	t	0.3062 ± 0.0338	0.1254 ± 0.0210
	u	0.1674 ± 0.0942	0.0922 ± 0.0063
	r	0.1931 ± 0.1002	0.0935 ± 0.0071
	o	0.2716 ± 0.1157	0.1002 ± 0.0102

Conclusions

In this paper a new framework for a cost-sensitive churn predictive modeling was presented. First we show the importance of using the actual financial costs of the churn modeling process, since there are significant differences in the results when evaluating a churn campaign using a traditional such as the $F1Score$, than when using a measure that incorporates the actual financial costs such as the savings. Moreover, we also show the importance of having a measure that differentiates the costs within customers, since different customers have quite different financial impact as measured by their lifetime value. Also, this framework can be expanded by using an additional classifier to predict the offer response probability by customer.

Furthermore, our evaluations confirmed that including the costs of each example and using an example-dependent cost-sensitive methods leads to better results in the sense of higher savings. In particular, by using the cost-sensitive decision tree algorithm, the financial savings are increased by 153,237 Euros, as compared to the savings of the cost-insensitive random forest algorithm which amount to just 24,629 Euros.

Additionally, by testing the different example-dependent cost-sensitive classification methods, we observed that when the costs are included during the pre-processing stage, by using the cost-proportionate sampling methods, the savings are 60,005 Euros. On the other hand, when the costs are included after the training with the Bayes minimum risk algorithm, the savings are 77,606 Euros. Finally, by using the cost-sensitive decision tree algorithm, which include the costs during the training phase, the savings increase quite

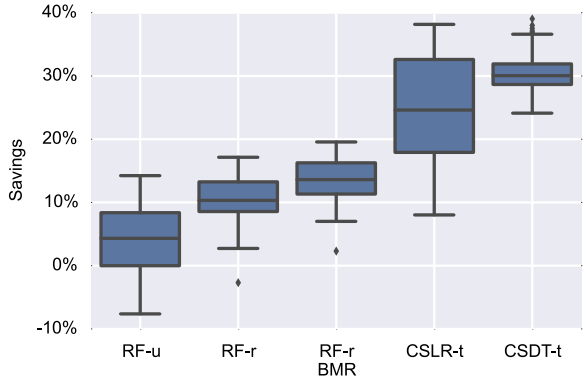

Fig. 8 Comparison of the random forest, logistic regression, Bayes minimum risk and cost-sensitive logistic regression algorithms. Overall, the *CSDT* is the method that produces significant higher savings. Moreover, it is observed that the cost-sensitive methods *CSLR* and *CSDT* gave the best results when estimated using the full training dataset, thats because these models does not appear to be impacted by the class imbalance, and themselves deal with the cost-sensitivity of each customer

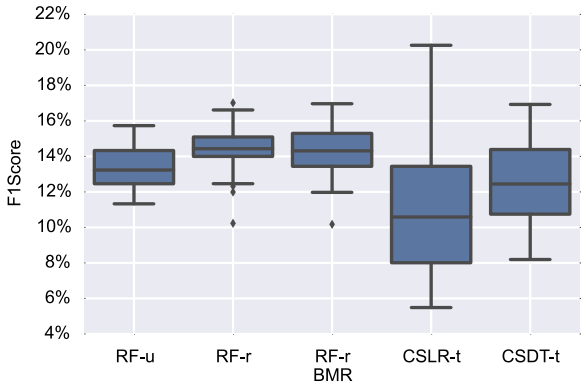

Fig. 9 Comparison of the random forest, logistic regression, Bayes minimum risk and cost-sensitive logistic regression algorithms. The best model selected by savings is not the same as the one selected by the $F1 - Score$

significantly to 177,867 Euros, hence, confirming the importance of using an algorithm that take into account the different example-dependent costs during the training phase.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
ACB carried out the research and wrote the manuscript. DA supervised the research and reviewed the manuscript. BO co-supervised the research and gave recommendations for improvements. All authors read and approved the final manuscript.

Authors' information
Alejandro Correa Bahnsen receive his MSc from Universidad de los Andes, Bogota, Colombia. He is currently working towards a Ph.D in Machine Learning at Luxembourg University. His research area relates to cost-sensitive classification and its application in a variety of real-world problems such as fraud detection, credit risk, direct marketing and churn modeling. Moreover, he has several years of experience in analytics, applying data mining models in a variety of areas from advertisement to credit risk.
Djamila Aouada was born in Blida, Algeria, on November 10, 1982. She received the State Engineering degree (Ingeniorat d' Etat) in electronics in June 2005, from the Ecole Nationale Polytechnique (ENP), Algiers, Algeria, and the Ph.D. degree in electrical engineering in May 2009, from North Carolina State University (NCSU), Raleigh, NC. From June to August 2007, she participated in the data sciences summer school at Los Alamos National Laboratory (LANL), Los Alamos, NM, as part of the Geometric Measure Theory group (GMT). From July to September 2008, Dr. Aouada worked as a consultant for Alcatel-Lucent Bell Laboratories, Murray Hill, NJ. Since November 2009, Dr. Aouada has been performing and supervising research as a Research Associate at the Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg . Her research interests span the areas of signal and image processing, computer vision, pattern recognition and data modeling. Dr. Aouada is member of the IEEE Signal Processing Society and the Eta Kappa Nu honor society (HKN). She is a co-author of a paper awarded Best Student Paper Award at IEEE 7th International Symposium on Image and Signal Processing and Analysis (ISPA'11).
Björn Ottersten was born in Stockholm, Sweden, 1961. He received the M.S. degree in electrical engineering and applied physics from Linkoping University, Linkoping, Sweden, in 1986. In 1989 he received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA. Dr. Ottersten has held research positions at the Department of Electrical Engineering, Linkoping University, the Information Systems Laboratory, Stanford University, the Katholieke Universiteit Leuven, Leuven, and the University of Luxembourg. During 96/97 Dr. Ottersten was Director of Research at ArrayComm Inc, a start-up in San Jose, California based on Ottersten's patented technology. He has coauthored journal papers that received the IEEE Signal Processing Society Best Paper Award in 1993, 2001, and 2006 and 3 IEEE conference papers receiving Best Paper Awards. In 1991 he was appointed Professor of Signal Processing at the Royal Institute of Technology (KTH), Stockholm. From 1992 to 2004 he was head of the department for Signals, Sensors, and Systems at KTH and from 2004 to 2008 he was dean of the School of Electrical Engineering at KTH. Currently, Dr. Ottersten is Director for the Interdisciplinary Centre for Security, Reliability and Trust at the University of Luxembourg. Dr. Ottersten has served as Associate Editor for the IEEE Transactions on Signal Processing and on the editorial board of IEEE Signal Processing Magazine. He is currently editor in chief of EURASIP Signal Processing Journal and a member of the editorial board of EURASIP Journal of Applied Signal Processing. Dr. Ottersten is a Fellow of the IEEE and EURASIP. In 2011 he received the IEEE Signal Processing Society Technical Achievement Award. He is a first recipient of the European Research Council advanced research grant. His research interests include security and trust, reliable wireless communications, and statistical signal processing.

Acknowledgments
Funding for this research was provided by the Fonds National de la Recherche, Luxembourg.

Received: 26 February 2015 Accepted: 13 May 2015
Published online: 12 June 2015

References

Correa Bahnsen, A, Aouada, D, Ottersten, B (2015). Example-Dependent Cost-Sensitive Decision Trees, Expert Systems with Application, *in press*. http://doi.org/10.1016/j.eswa.2015.04.042.

Cohen, I, & Goldszmidt, M (2004). Properties and Benefits of Calibrated Classifiers, In *Knowledge Discovery in Databases: PKDD 2004* (pp. 125–136). Pisa, Italy.

Correa Bahnsen, A, Stojanovic, A, Aouada, D, Ottersten, B (2013). Cost Sensitive Credit Card Fraud Detection Using Bayes Minimum Risk, In *2013 12th International Conference on Machine Learning and Applications* (pp. 333–338). Miami, USA: IEEE.

Correa Bahnsen, A, Aouada, D, Ottersten, B (2014a). Example-Dependent Cost-Sensitive Logistic Regression for Credit Scoring, In *2014 13th International Conference on Machine Learning and Applications* (pp. 263–269). Detroit, USA: IEEE.

Correa Bahnsen, A, Stojanovic, A, Aouada, D, Ottersten, B (2014b). Improving Credit Card Fraud Detection with Calibrated Probabilities, In *Proceedings of the Fourteenth SIAM International Conference on Data Mining* (pp. 677–685). Philadelphia, PA.

Drummond, C, & Holte, R (2003). C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling, In *Workshop on Learning from Imbalanced Datasets II, ICML*. Washington, DC, USA.

Elkan, C (2001). The Foundations of Cost-Sensitive Learning, In *Seventeenth International Joint Conference on Artificial Intelligence* (pp. 973–978).

Farris, PW, Bendle, NT, Pfeifer, PE, Reibstein, DJ (2010). *Marketing Metrics: The Definitive Guide to Measuring Marketing Performance*, 2nd, (p. 432). New Jersey, USA: Pearson FT Press.

Glady, N, Baesens, B, Croux, C (2009). Modeling churn using customer lifetime value. *European Journal of Operational Research*, *197*(1), 402–411.

Hastie, T, Tibshirani, R, Friedman, J (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction.

Haupt, RL, & Haupt, SE (2004). *Practical Genetic Algorithms*, Second edition, (p. 261). New Jersey: John Wiley & Sons, Inc.

Hernandez-Orallo, J, Flach, P, Ferri, C (2012). A Unified View of Performance Metrics : Translating Threshold Choice into Expected Classification Loss. *Journal of Machine Learning Research*, *13*(July), 2813–2869.

Correa Bahnsen, A (2015). CostSensitiveClassification Library in Python. http://dx.doi.org/10.5281/zenodo.17789.

KhakAbi, S, Gholamian, MR, Namvar, M (2010). Data Mining Applications in Customer Churn Management, In *2010 International Conference on Intelligent Systems, Modelling and Simulation* (pp. 220–225). Liverpool, UK.

Marslan, S (2009). *Machine Learning: An Algorithmic Perspective*. New Jersey, USA: CRC Press.

Milne, GR, & Boza, ME (1999). Trust and Concern in Consumers' Perception of Marketing Information Management Practices. *Journal of Interactive Marketing*, *13*(1), 5–24.

Neslin, SA, Gupta, S, Kamakura, W, Lu, J, Mason, CH (2006). Defection Detection : Measuring and Understanding the Predictive Accuracy of Customer Churn Models. *Journal of Marketing Research*, *43*(2), 204–211.

Ngai, EWT, Xiu, L, Chau, DCK (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, *36*(2), 2592–2602.

Pfeifer, PE, Haskins, ME, Conroy, RM (2004). Customer lifetime value, customer profitability, and the treatment of acquisition spending. *Journal of Managerial Issues*, *17*(1), 11–25.

Provost, F, Fawcett, T, Kohavi, R (1998). The case against accuracy estimation for comparing induction algorithms, In *Proceedings of the Fifteenth International Conference on Machine Learning* (pp. 445–453).

van Raaij, EM, Vernooij, MJA, van Triest, S (2003). The implementation of customer profitability analysis: A case study. *Industrial Marketing Management*, *32*, 573–583.

Verbeke, W, Dejaeger, K, Martens, D, Hur, J, Baesens, B (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, *218*(1), 211–229.

Verbraken, T (2012). Toward profit-driven churn modeling with predictive marketing analytics, In *Cloud Computing and Analytics: Innovations in E-business Services. Workshop on E-Business (WEB2012)*. Orlando, US.

Verbraken, T, Verbeke, W, Baesens, B (2013). A novel profit maximizing metric for measuring classification performance of customer churn prediction models. *IEEE Transactions on Knowledge and Data Engineering*, *25*(5), 961–973.

Wang, T (2013). *Efficient Techniques for Cost-Sensitive Learning with Multiple Cost Considerations*. Sydney: PhD thesis, University of Technology.

Zadrozny, B, Langford, J, Abe, N (2003). Cost-sensitive learning by cost-proportionate example weighting, In *Third IEEE International Conference on Data Mining* (pp. 435–442).