ExamModuleCode	Question Number	Comments for Students
M45S16	1	Q1. Generally good marks on this question (mean=13). Pleasing to see many students getting part (h) right. However, part (i) proved difficult, as expected, with only a few students getting some marks on this question. In part (b), a broader range of answers were accepted. In part (i), up to 3 marks were allowed for a geometric solution.
M45S16	2	Q2. Marks on this question were weak (mean=9) and around a third with a mark less than 7, suggesting they did not revise this material thoroughly. Most parts were answered except part (h): only a few students got this. Part (h) was designed as a hard question.
M45S16	3	a). More detailed description of statistical aspects and consequence of selection bias sometimes needed B).perhaps a misinterpretation of being asked on occasion c) weights of evidence not especially relevant here (d) some unclear arguments (f) some misunderstanding of odds ratios, and some limited discussion of effect of bias on parameters (g) some misunderstanding of the role of the regulator.
M45S16	4	Generally good attempts. Few consistent errors.

Imperial College
London

M3S16

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)

May-June 2019

This paper is also taken for the relevant examination for the Associateship of the Royal College of Science

Credit Scoring

Date: Friday 10 May 2019
Time: 14.00 - 15.30
Time Allowed: 1 Hour 30 Minutes
This paper has 3 Questions.
Candidates should use ONE main answer book for Questions 1 and 2, and ONE main answer book for Question 3.
Supplementary books may only be used after the relevant main book(s) are full.
All required additional material will be provided.

- DO NOT OPEN THIS PAPER UNTIL THE INVIGILATOR TELLS YOU TO.
- Affix one of the labels provided to each answer book that you use, but DO NOT USE THE LABEL WITH YOUR NAME ON IT.
- Calculators may not be used.

© 2019 Imperial College London

Imperial College
London

M4/5S16

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)

May-June 2019

This paper is also taken for the relevant examination for the Associateship of the Royal College of Science

Credit Scoring

Date: Friday 10 May 2019

Time: 14.00 - 16.00

Time Allowed: 2 Hours

This paper has 4 Questions.

Candidates should use ONE main answer book for Questions 1 and 2, and ONE main answer book for Question 3 and 4.

Supplementary books may only be used after the relevant main book(s) are full.

All required additional material will be provided.

- DO NOT OPEN THIS PAPER UNTIL THE INVIGILATOR TELLS YOU TO.
- Affix one of the labels provided to each answer book that you use, but DO NOT USE THE LABEL WITH YOUR NAME ON IT.
- Calculators may not be used.

© 2019 Imperial College London

1. Use the following definitions in this question:

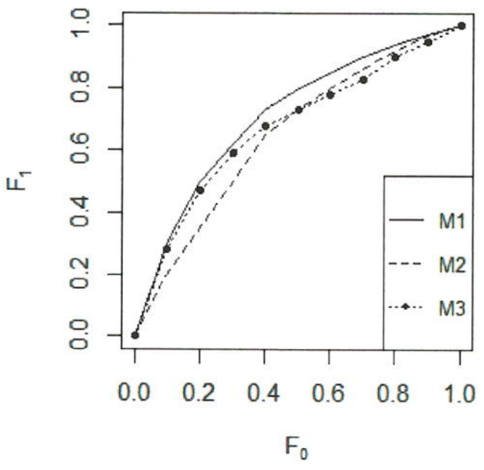
- For a loan scorecard, a loan is predicted as default if and only if the borrower's credit score S is less than a given cut-off score c.
- Let $Y \in \{0, 1\}$ indicate the outcome of a loan: 0=non-default, 1=default.
- Let $F_0(c) = P(S \leq c|Y = 0)$ and $F_1(c) = P(S \leq c|Y = 1)$ for all $c \in \mathbb{R}$.
- Let $f_i(c)$ be the derivative of $F_i(c)$ with respect to c.
- Let $K = \max_c |F_1(c) - F_0(c)|$ be the Kolmogorov-Smirnoff statistic for measuring scorecard performance.

(a) Define the receiver operating characteristic (ROC) curve.

(b) What aspect of model performance does the ROC curve represent?

(c) What position in the ROC curve space represents the performance of a perfect model? Explain why.

(d) Which of the three ROC curves shown below represents the best model, M1, M2 or M3? Explain why.



(e) Why does the diagonal from (0,0) to (1,1) on the ROC graph represent a model with the least predictive power?

QUESTION CONTINUED ON NEXT PAGE.

(f) From the data in the following table, plot the ROC curve and compute the area under the ROC curve (AUC).
Use linear interpolation for values of $F_i(c)$ between the points given in the table.

c	$F_0(c)$	$F_1(c)$
-0.8	0	0
-0.6	0.2	0.4
-0.4	0.6	0.8
0	1	1

(g) Compute K for this data.

For the next two parts, assume $F_0(c) \leq F_1(c)$ for all $c \in \mathbb{R}$.

(h) Show algebraically that, under this assumption, the partial area under the ROC curve given in the range $c \in [a, b]$,

$$\int_a^b F_1(c) f_0(c) \mathrm{d}c \geq \frac{1}{2}\left(F_0(b)^2 - F_0(a)^2\right).$$

(i) Show algebraically that the following relationship holds between AUC A and the Kolmogorov-Smirnoff statistic K, in general:

$$\frac{1}{2}(1 + K^2) \leq A \leq \frac{1}{2}(1 + 2K - K^2).$$

2. Use the following information for this question:-

- Consider historical data of n credit cards with a vector of m application variables $\mathbf{x}_i = (x_{i1}, \cdots, x_{im})$ and response, monetary value of credit card usage in 12 months (standardized), $y_i \in \mathbb{R}$, for credit card i.
- Suppose the data has been standardized so that $\sum_{i=1}^n x_{ij} = 0$, $\sum_{i=1}^n x_{ij}^2 = 1$ for all $j \in \{1, \cdots, m\}$, and $\sum_{i=1}^n y_i = 0$.
- The OLS estimator with LASSO penalty is given by minimizing

$$f(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n} (y_i - \boldsymbol{\beta} \cdot \mathbf{x}_i)^2 + \lambda ||\boldsymbol{\beta}||_1$$

with respect to the vector of coefficients $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_m)$ for some given $\lambda > 0$.

- Define the function

$$S(\hat{\beta}, \lambda) = \begin{cases} \hat{\beta} - \lambda & \text{if } \hat{\beta} > \lambda \\ \hat{\beta} + \lambda & \text{if } \hat{\beta} < -\lambda \\ 0 & \text{otherwise} \end{cases}$$

(a) For the univariate case, when $m = 1$, so $\boldsymbol{\beta} = (\beta_1)$, show that for $\beta_1 \neq 0$,

$$\frac{\mathrm{d}f(\beta_1)}{\mathrm{d}\beta_1} = \beta_1 - \hat{\beta}_{\text{OLS}} + \lambda \,\text{sign}(\beta_1)$$

where $\hat{\beta}_{\text{OLS}}$ is the usual OLS estimator of β_1 (ie when $\lambda = 0$).

(b) Therefore show that in the univariate case, $\beta_1 = S(\hat{\beta}_{\text{OLS}}, \lambda)$ is the minimizer of $f(\boldsymbol{\beta})$.

(c) Briefly explain the pathwise coordinate optimization method and how it can be used to compute the OLS estimator with LASSO penalty for the multivariate case, ie $m \geq 2$.

(d) Using the pathwise coordinate optimization method, derive the iterative update function for each coefficient β_j for the multivariate case, $m \geq 2$.

QUESTION CONTINUED ON NEXT PAGE.

Consider the following output from an OLS regression model with and without the LASSO penalty.

Variable	OLS coefficient estimate	OLS with LASSO coefficient estimate
Income (log)	1.5	1.2
Home owner? (0/1)	0.5	0
Months in current residence (log)	-0.1	0
Outstanding debt on other loans (log)	-3.3	-2.3

(e) Which of the variables have been deselected when using the LASSO penalty?

(f) For those variables selected when using the LASSO penalty, what is the direction of association with the outcome?

(g) Suppose the model is rebuilt on the same training data with three different values of λ and tested on the same test set. Results using the Kolmogorov-Smirnoff (KS) statistic as a performance measure are given in the following table. On the basis of just these results which value of λ gives the best model? Explain why.

λ	KS on training data	KS on test data
0.1	0.55	0.45
0.5	0.52	0.48
1	0.48	0.42

(h) Let λ_1 and λ_2 be two values of λ giving coefficient estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ for β respectively when applying the OLS estimator with LASSO penalty in each case, on exactly the same training data. Show that when $\lambda_1 < \lambda_2$, $||\hat{\beta}_1||_1 \geq ||\hat{\beta}_2||_1$.

3. Use the following definitions in this question:

- For a loan application scorecard, a new applicant is typically rejected if his or her credit score S is less than a cut-off score c.
- Let $Y \in \{0,1\}$ indicate the outcome of a loan: 0=non-default, 1=default.
- Let $\mathbf{X}$ be a random vector of predictor variables for a loan application.
- For historical data, let A denote the event that a loan application was accepted.

(a) Explain the problem of selection bias in the context of models developed for accept/reject decisions on loan applications.

(b) Suppose we assume $P(Y=1|\bar{A},\mathbf{X}=\mathbf{x}) > P(Y=1|A,\mathbf{X}=\mathbf{x})$.

(i) Explain what this means in terms of accepted and rejected loans and why it is a reasonable assumption.

(ii) Show that it implies $P(Y=1|\mathbf{X}=\mathbf{x}) > P(Y=1|A,\mathbf{X}=\mathbf{x})$.

A lender has 10000 loan applications. Its existing scorecard model A rejects 2500 of these applications and accepts the remainder. However, the lender randomly selects 250 of these rejected applications and decides to override model A for just these cases and give them a loan anyway. A year later, the lender builds two new scorecard models based on the outcome from these loans. Model B is built on just those applications accepted by model A, whilst model C is built on *all* loans including the 250 overrides. In model C the 250 overrides are reweighted by 10, whilst the other loans have weight 1.
All models are unsegmented logistic regression models of non-default.

(c) Explain how model C is addressing the selection bias problem.

(d) What is the reason for reweighting the 250 overridden rejects?

(e) Explain how reweighting of the loans is implemented with logistic regression.

QUESTION CONTINUED ON NEXT PAGE.

(f) Models B and C have two predictor variables X_{10}, *annual income (log)*, and X_{11}, *value of outstanding debt (log)*, amongst other variables and they have the following coefficient estimates in the two models:

	Model B			Model C		
Variable	Coefficient estimate β	s.e.	$\exp(\beta)$	Coefficient estimate β	s.e.	$\exp(\beta)$
X_{10}	0.048	0.0042	1.049	0.051	0.0040	1.052
X_{11}	0.18	0.032	1.20	0.69	0.029	1.99

s.e. = *standard error*

(i) What are the odds ratios of non-default for each variable in each model (use 1 unit of change in each variable)?

(ii) From the evidence given by these odds ratios, describe how selection bias has affected model B.

(g) Why might a banking regulator object to the modelling approach taken by the lender?

4. This question is based on the article,

- A Novel Profit Maximizing Metric for Measuring Classification Performance of Customer Churn Prediction Models, by Wouter Verbeke and Bart Baesens *IEEE Transactions on Knowledge and Data Engineering*, January 2012.

(a) What is the motivation of this article?

(b) What is meant by customer churn?

(c) Suppose F_0 and F_1 are normal distributions with different means μ_0 and μ_1 respectively but the same variance σ^2. In this case, show that

$$T = \frac{\sigma^2 \log(\pi_1\theta/\pi_0)}{\mu_0 - \mu_1} + \frac{1}{2}(\mu_0 + \mu_1).$$

(d) Derive Equation (14) from Equation (13).

(e) Why is EMPC expressed as a single integral over just γ in Equation (18), instead of the multiple integral over variables given for EMP in Equation (5)?

(f) Why is the beta distribution used as the parametric form for $h(\gamma)$?

(g) Explain the process of model selection used in Section 6.1 with results given in Table 4.

(h) Explain what is shown in Table 4. How does this demonstrate the value of EMPC?

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/254436157

A Novel Profit Maximizing Metric for Measuring Classification Performance of Customer Churn Prediction Models

Article *in* IEEE Transactions on Knowledge and Data Engineering · January 2012

DOI: 10.1109/TKDE.2012.50

CITATIONS
50

READS
243

3 authors, including:

Wouter Verbeke
Vrije Universiteit Brussel
53 PUBLICATIONS **761** CITATIONS

SEE PROFILE

Bart Baesens
University of Southampton
364 PUBLICATIONS **8,014** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Customer churn prediction research View project

Credit Risk Analytics View project

All content following this page was uploaded by Wouter Verbeke on 23 January 2018.

The user has requested enhancement of the downloaded file.

A Novel Profit Maximizing Metric for Measuring Classification Performance of Customer Churn Prediction Models

Thomas Verbraken, *Student Member, IEEE*, Wouter Verbeke, and Bart Baesens

Abstract—The interest for data mining techniques has increased tremendously during the past decades, and numerous classification techniques have been applied in a wide range of business applications. Hence, the need for adequate performance measures has become more important than ever. In this paper, a cost-benefit analysis framework is formalized in order to define performance measures which are aligned with the main objectives of the end users, i.e., profit maximization. A new performance measure is defined, the expected maximum profit criterion. This general framework is then applied to the customer churn problem with its particular cost-benefit structure. The advantage of this approach is that it assists companies with selecting the classifier which maximizes the profit. Moreover, it aids with the practical implementation in the sense that it provides guidance about the fraction of the customer base to be included in the retention campaign.

Index Terms—Data mining, classification, performance measures

1 Introduction

As a result of the steep growth in computational power, the interest for data mining techniques has increased tremendously the past decades. A myriad of classification techniques has been developed and is being used in a wide range of business applications. As more and more methods are elaborated, the need for adequate performance measures has become more important than ever before. There has been a lot of attention for the receiver operating characteristic (ROC) curve, which is a graphical representation of the classification performance for varying thresholds [1]. However, rather than visually comparing curves, practitioners prefer to capture the performance of a classification method in a single number. A very popular and straightforward concept is the area under the ROC curve (AUC), which is closely related to the Gini coefficient and the Kolmogorov-Smirnov statistic. The problem with these measures is that they implicitly make unrealistic assumptions about misclassification costs. The H measure is a new approach to performance measurement, which overcomes this problem and focuses on misclassification costs [2].

In this paper, not only the misclassification costs, but also the benefits originating from a correct classification are explicitly taken into account. The main rationale is that the most important goal for practitioners is profit maximization. In Section 2, a cost-benefit analysis framework will be worked out, in which two types of performance measures will be defined. The first metric is the maximum profit (MP), whereas the second metric is the expected maximum profit (EMP). The difference between both measures is that MP is a deterministic approach, which assumes that all parameters related to the costs and benefits are accurately known. The EMP measure on the other hand defines a probability distribution for the cost and benefit parameters, and is a probabilistic approach. Analogously to the H measure, also EMP is related to the ROC curve of the classifier, as will be discussed in Section 3.

Due to the multitude of business situations in which classification methods are employed, it is difficult, if not impossible, to define one single profit driven performance measure. Section 4 will focus on the prediction of customer churn, which has become a very important business problem. During the last decade, the number of mobile phone users has increased drastically, with expectations of 5.6 billion mobile phone subscribers in 2011,[1] which is around 80 percent of the world population. Hence, telecommunication markets are getting saturated, particulary in developed countries, and the emphasis is shifting from attraction of new customers to retention of the existing customer base. In this context, customer churn prediction models play a crucial role and they are increasingly being researched (see, e.g., the extensive literature overview given by Verbeke et al. [3]) and new approaches such as the use of social network data are explored (see among others [4]). When investigating and comparing these data mining techniques for customer churn prediction, it is imperative

- T. Verbraken is with the Department of Decision Sciences and Information Management, KU Leuven, Naamsestraat 69, B-3000 Leuven, Belgium. E-mail: thomas.verbraken@kuleuven.be.
- W. Verbeke is with the University of Edinburgh Business School, 29 Buccleuch Place, Edinburgh, EH8 9JS, United Kingdom. E-mail: wouter.verbeke.ac@gmail.com.
- B. Baesens is with the Department of Decision Sciences and Information Management, KU Leuven, Naamsestraat 69, B-3000 Leuven, Belgium, and the School of Management, University of Southampton, Highfield Southampton SO17 1BJ, United Kingdom. E-mail: bart.baesens@kuleuven.be.

Manuscript received 30 June 2011; revised 18 Nov. 2011; accepted 11 Feb. 2012; published online 2 Mar. 2012.
Recommended for acceptance by Z.-H. Zhou.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2011-06-0380.
Digital Object Identifier no. 10.1109/TKDE.2012.50.

[1] www.eito.com.

1041-4347/13/$31.00 © 2013 IEEE Published by the IEEE Computer Society

to have an adequate performance measure. Therefore, with the guidance of the cost-benefit framework, a novel performance measure which consistently incorporates the losses and gains will be developed. The main advantage of this metric is that it not only unambiguously identifies the classifier which maximizes the profit for a given customer retention campaign, but also determines the fraction of the customer base which should be targeted to maximize the profit. This is a crucial help for practitioners, since deviating from the optimal fraction leads to suboptimal profits. Section 5 explores the link with the H measure and shows that the H measure can be used as an approximation for the EMP measure for customer churn. Finally, the developed performance measures will be tested in an extensive case study, of which the findings are reported in Section 6.

2 The Cost-Benefit Analysis Framework

Measuring the performance of classifiers is essential for identifying superior classification techniques [5]. This paper focusses on classification problems with a binary outcome, i.e., it deals with the two-class case, where instances belong either to class 0 or to class 1. Many classifiers produce a continuous score, $s(\mathrm{x})$, which is a function of the attribute vector x and aims to discriminate between two classes. It is assumed that the instances from class 0 have a lower score than the instances from class 1; if this is not true, the score could be multiplied with minus one. To classify an instance, a cutoff t has to be chosen, where all instances with s smaller than t are classified as a zero, whereas instances for which s is larger than t are classified as ones.

Throughout this paper, the following notation will be adopted. The number of instances in the data set is denoted with N, whereas the prior probabilities of instances belonging to class 0 or 1 are denoted with π_0 and π_1, respectively. Two types of instances are considered, those from class 0 and 1, and for both groups of observations, the probability distribution of the classification scores can be defined. The probability density functions are $f_0(s)$ and $f_1(s)$, whereas the cumulative density functions are $F_0(s)$ and $F_1(s)$ for class 0 and 1, respectively. Usually, the outcome of a classification task is used as input to a business process leading to benefits for correct classifications and costs for misclassifications. The cost or benefit related to classifying an instance from class j to class i is denoted with $c(i|j)$. By convention, both costs and benefits will be positive or equal to zero. Hence, in the cost-benefit analysis, a minus sign will be applied for the misclassification costs.

In general, a classification exercise leads to a confusion matrix as shown in Table 1. For example, the upper right cell contains the instances belonging to class 0 which are incorrectly classified into class 1. The number of instances in this category is given by $\pi_0(1-F_0(t))N$, whereas the cost of such a misclassification is $c(1|0)$. For notational convenience, the costs and benefits are put to b_0, c_0, b_1, and c_1, as indicated in Table 1. By multiplying the total number of observations in a certain cell with the classification cost or benefit (between square brackets), and summing this over all cells, one obtains the total classification profit, as will be discussed later.

TABLE 1
Confusion Matrix Containing the Number of Instances Classified in Each Cell

		Classified into	
		Class 0	Class 1
Belongs to	Class 0	$\pi_0 F_0(t)\,N$ $[c(0\|0)=b_0]$	$\pi_0(1-F_0(t))\,N$ $[c(1\|0)=c_0]$
	Class 1	$\pi_1 F_1(t)\,N$ $[c(0\|1)=c_1]$	$\pi_1(1-F_1(t))\,N$ $[c(1\|1)=b_1]$

Between square brackets, the related classification costs and benefits are given.

In the literature, several performance measures have been proposed. An overview of performance measures for classification techniques can be found in [6] and [7]. In the data mining community, the most well-known measures include:

$$\begin{aligned}
&\text{Accuracy} = \pi_0 F_0(t) + \pi_1(1-F_1(t)),\\
&\text{Sensitivity} = F_0(t),\\
&\text{Specificity} = 1 - F_1(t),\\
&\text{AUC} = \int F_0(s) f_1(s)\mathrm{d}s.
\end{aligned}$$

The accuracy of a classifier measures the percentage of correctly classified observations. The sensitivity represents the proportion of zeros which are correctly classified, whereas the specificity measures the proportion of ones correctly predicted. Most of these performance measures do not take into account the misclassification costs, and are only valid when the misclassification costs are equal.

There has been much attention for cost-sensitive learning. For example, Domingos [8] proposed a general method to produce cost-sensitive classifiers, [9] combined ROC curve analysis with cost distribution information, [10] developed an ontology-based approach for cost-sensitive classification, [11] used over- and undersampling and threshold moving (and an ensemble of these methods) for cost-sensitive learning with neural networks, and more recently, [2] introduced the H-measure, which takes into account misclassification costs. However, as pointed out by Elkan [12]:

> "Although most recent research in machine learning has used the terminology of costs, doing accounting in terms of benefits is generally preferable, because avoiding mistakes is easier, since there is a natural baseline from which to measure all benefits, whether positive or negative."

So even though a cost-only approach could be mathematically equivalent, in this study the benefits will be taken explicitly into account since this allows for a more straightforward interpretation by the practitioner.

The central concept in this paper is the cost-benefit framework, i.e., the fact that a different benefit or cost is related to each entry in the confusion matrix. The main purpose is to work out a performance metric which selects the classifier with the highest profit. This metric is used to select the best performing technique, and thus can be used with all classification algorithms. The average classification profit of a classifier is defined as follows:

Definition 1. ***The average classification profit** of a classifier is the profit generated by the employment of this classifier. It is* *the sum of the classification benefits minus the classification costs, divided by the number of instances.*

The average classification profit can be expressed as a function of the classification threshold, t:

$$P(t; b_0, c_0, b_1, c_1) = b_0\pi_0 F_0(t) + b_1\pi_1(1 - F_1(t)) - c_0\pi_0(1 - F_0(t)) - c_1\pi_1 F_1(t), \quad (1)$$

with b_0, c_0, b_1, and c_1 parameters as defined by Table 1. These parameters are positive unless explicitly stated otherwise. Rearranging yields

$$P(t; b_0, c_0, b_1, c_1) = (b_0 + c_0)\pi_0 F_0(t) - (b_1 + c_1)\pi_1 F_1(t) + b_1\pi_1 - c_0\pi_0. \quad (2)$$

The average classification profit depends on the cutoff, t. For a perfect classifier at the optimal threshold, it holds that $F_0 = 1$ and $F_1 = 0$, which results in an average classification profit equal to $b_0\pi_0 + b_1\pi_1$. For a fixed classifier and given costs and benefits, the threshold t is to be optimized, resulting in the maximum profit of a classifier.

Definition 2. *The maximum profit, MP, of a classification technique is the profit resulting from the classification outcome when the optimal cutoff, T, is used.*

The maximum profit is a measure for the effectiveness of the technique, since profit maximization is the ultimate aim. This approach to classification performance measurement has already been explored by Verbeke et al. [13]. MP is analytically expressed as:

$$\mathrm{MP} = \max_{\forall t} P(t; b_0, c_0, b_1, c_1) = P(T; b_0, c_0, b_1, c_1), \quad (3)$$

with T the optimal threshold. The optimal threshold satisfies the first order condition for the maximization of the average profit P:

$$\frac{f_0(T)}{f_1(T)} = \frac{\pi_1(b_1 + c_1)}{\pi_0(b_0 + c_0)} = \frac{\pi_1\theta}{\pi_0}. \quad (4)$$

The parameter $\theta = (b_1 + c_1)/(b_0 + c_0)$, the *cost benefit ratio*, is introduced for notational convenience. It also indicates that the optimal threshold and profit depends on a ratio of costs and benefits, and thus is not dependent on the measurement scale. Since it is assumed that the cost-benefit parameters (b_0, c_0, b_1, c_1) are positive, the value of θ ranges from zero to plus infinity. Equation (4) has an appealing graphical interpretation on the ROC-curve, as will be shown in Section 3.1.

The MP measure is a *deterministic* approach to cost-benefit analysis, since the costs and benefits are supposed to be known. In reality, however, it is not straightforward to estimate precise values for these parameters. Therefore, this paper proposes to adopt a probability distribution for the cost and benefit parameters, leading to a *probabilistic* cost-benefit-based performance measure, the *expected* maximum profit of a classifier.

Definition 3. *The expected maximum profit, EMP, is the expectation of the maximal profit of a classifier with respect to the distribution of classification costs.*

The expected maximum profit measures the effectiveness of a classification technique, taking into account the uncertainty of the classification costs and benefits and is a genuine contribution of this paper.

Let $w(b_0, c_0, b_1, c_1)$ be the joint probability density of the classification costs, then the expected maximal profit is equal to

$$\mathrm{EMP} = \int_{b_0}\int_{c_0}\int_{b_1}\int_{c_1} P(T(\theta); b_0, c_0, b_1, c_1) \cdot w(b_0, c_0, b_1, c_1)\,db_0\,dc_0\,db_1\,dc_1. \quad (5)$$

Equation (5) is the most general expression for the expected maximum profit for a given classifier. Note that for *each* combination of (b_0, c_0, b_1, c_1), the optimal threshold T is determined through (4). The profit is then calculated for the optimal threshold and weighed with the probability density for this particular combination of classification costs. Integrating over all possible benefits and costs leads to the expected maximum classification profit. When the costs are known with certainty, the function $w(b_0, c_0, b_1, c_1)$ is a Dirac impulse at the known values, and EMP simplifies to MP. Although the previous discussion only focussed on the profit related to a classification outcome, there is a link to the ROC curve of the classifier, as will be explored in the next section.

3 The Receiver Operating Characteristic Curve and the H-Measure

The receiver operating characteristic curve is a concept which has been extensively used in the machine learning community [1]. Section 3.1 discusses the ROC curve in the context of cost-benefit analysis, along with a derived performance measure, the Area under the ROC-curve, also known as AUC [14]. However, [2] showed recently that AUC is a flawed measure and proposed an alternative, the H-measure, which will be treated in Section 3.2.

3.1 The ROC-Curve and the Area under the ROC-Curve

A ROC curve is a graphical representation of the classification performance with varying threshold t. It is a plot of the sensitivity versus one minus the specificity, i.e., $F_0(t)$ as a function of $F_1(t)$. The interested reader is referred to [1] or [15] for an extensive discussion on ROC-curves. Because ROC-curves are not convenient to compare classifiers, especially when their ROC-curves intersect, the area under the ROC-curve is often used to quantify the performance. A larger AUC would indicate superior performance. The area under the ROC curve, which is closely related to the Gini coefficient and the Kolmogorov-Smirnov statistic, has the following statistical interpretation: the AUC of a classification method is the probability that a randomly chosen positive instance will be ranked higher than a randomly chosen negative instance [1].

Also in the context of the cost-benefit analysis framework, the ROC-curve has its merit. Since each point on the ROC-curve corresponds to a threshold t, the optimal cutoff T, is located somewhere on the curve.

Theorem 1. *If the ROC-curve is convex*[2] *and continuously differentiable, the optimal cutoff, for given classification costs* c_0 *and* c_1*, and benefits* b_0 *and* b_1*, corresponds to the point on the ROC-curve for which the tangent slope equals* $\pi_1\theta/\pi_0 = [\pi_1(b_1 + c_1)]/[\pi_0(b_0 + c_0)]$.

This theorem is a direct result from the definition of the ROC curve and the first order condition as defined by (4). When the ROC-curve is not convex, and thus the slope is not monotonically decreasing, there may be multiple points on the ROC curve satisfying the first order condition. More specifically, points in the concave region are not optimal for *any* θ, in which case the convex hull needs to be defined.

Theorem 2. *The convex hull of a (nonconvex) ROC-curve defines a set of points where each point corresponds to the optimal cutoff for a certain value of the cost benefit ratio* $\theta \in [0, +\infty)$.

This theorem is based on the second order condition for profit maximization and is discussed in detail by Fawcett [1]. The interesting implication of this fact is that an integration over a range of θ values is equivalent to an integration over the corresponding part of the ROC curve. In fact, every value for $\theta \in [0, +\infty)$, has a corresponding isoperformance line with a slope equal to $\pi_1\theta/\pi_0$. The optimal cutoff for a given θ value is then located where the isoperformance line is tangent to the ROC curve. Thus, by varying θ from zero to infinity, each point on the ROC curve is traversed.

This interpretation of an integration over the ROC curve has another very important consequence, as was revealed by Hand [2]. He showed that the popular AUC is equivalent to an expected maximum profit measure. However, the probability density which is implicitly assumed when calculating the AUC depends on the empirical score distribution of the classifier itself. Therefore, AUC is seriously flawed as a performance measure, and Hand proposed an alternative, the H measure, which will be discussed in the next section.

3.2 The H-Measure

Hand [2] proposed the H measure as a coherent alternative to the area under the ROC curve. This measure resembles an EMP measure, with the difference that the H measure only explicitly states the misclassification costs and not the benefits. The core of the H measure is the average classification *loss*, Q, which is equal to the negative of the average classification profit, P, with $b_0 = b_1 = 0$. Hence, the focus is not on the expected maximum profit, but on the expected minimum loss. Moreover, a variable substitution is carried out:

$$Q(t; c, b) = b \cdot [c\pi_0(1 - F_0(t)) + (1 - c)\pi_1 F_1(t)], \qquad (6)$$

with $c = c_0/(c_0 + c_1)$ and $b = c_0 + c_1$. In this case, the cost benefit ratio on which the optimal threshold T depends, is equal to $\theta = (1 - c)/c$.

To calculate a concrete value for the expected minimum loss, assumptions have to be made regarding the probability density of b and c. A first assumption is the independence of b and c, namely that $w(b, c) = u(c)v(b)$ holds true, where $w(b, c)$ is the joint probability density function of b and c, whereas $u(c)$ and $v(b)$ are the marginal probability density functions of c and b, respectively. The expected minimum loss (L) is then equal to

$$\mathrm{L} = \mathrm{E}[b] \int_0^1 Q(T(c); b, c) \cdot u(c)\mathrm{d}c, \qquad (7)$$

with $\mathrm{E}[b] = 1$ for an appropriate choice for the unit in which b is measured. Furthermore, the probability density function of c is supposed to follow a beta distribution with parameters α and β, which is characterized as follows:

$$u_{\alpha,\beta}(x) = \begin{cases} \dfrac{x^{\alpha-1} \cdot (1 - x)^{\beta-1}}{B(1, \alpha, \beta)} & \text{if } x \in [0, 1], \\ 0 & \text{else,} \end{cases} \qquad (8)$$

with α and β greater than one, and

$$B(x, \alpha, \beta) = \int_0^x t^{\alpha-1} \cdot (1 - t)^{\beta-1}\mathrm{d}t. \qquad (9)$$

Finally, to arrive at the H measure, a normalization is performed to obtain a performance measure bounded by zero and one:

$$H = 1 - \frac{\int_0^1 Q(T(c); b, c) \cdot u(c)\mathrm{d}c}{\pi_0 \int_0^{\pi_1} c \cdot u(c)\mathrm{d}c + \pi_1 \int_{\pi_1}^1 (1 - c) \cdot u(c)\mathrm{d}c}, \qquad (10)$$

with $u(c)$ shorthand notation for $u_{\alpha,\beta}(c)$. The denominator gives the misclassification loss for the worst classifier, i.e., a random classifier. Also, the integration over $c = [0..1]$ corresponds to an integration over $\theta = [0.. + \infty)$, and thus of a ROC curve tangent slope from plus infinity to zero. Thus, the H measure takes into account the entire ROC curve of a classifier. A close inspection of the H measure shows that only a part of the measure changes with varying classification techniques, holding the data set constant:

$$H = 1 + \frac{H_{var} - \pi_0 \int_0^1 c \cdot u(c)\mathrm{d}c}{\pi_0 \int_0^{\pi_1} c \cdot u(c)\mathrm{d}c + \pi_1 \int_{\pi_1}^1 (1 - c) \cdot u(c)\mathrm{d}c}, \qquad (11)$$

with H_{var} defined as:

$$H_{var} = \int_0^1 [\pi_0 c F_0 - \pi_1(1 - c)F_1] \cdot u(c)\mathrm{d}c. \qquad (12)$$

This variable part of H, H_{var}, is an expression which will be useful in Section 5.

4 Cost Benefit-Based Performance Measurement for Customer Churn Prediction

Until now, the performance measures MP and EMP have been discussed in general terms, without any assumptions on the distribution of the cost and benefit parameters other than that their values are assumed to be positive. However, when the cost-benefit framework is to be applied, an

[2] It was already pointed out by Hand [2] that the machine learning community in the context of ROC-curves by convention considers a function $g(x)$ convex if $g(\lambda x + (1 - \lambda)y) \geq \lambda g(x) + (1 - \lambda)g(y)$ for $0 < \lambda < 1$. A twice differentiable function $g(x)$ is convex if $\partial^2 g(x)/\partial x^2 \leq 0$. The mathematical community on the contrary adopts the definition of convexity with the inequality sign reversed.

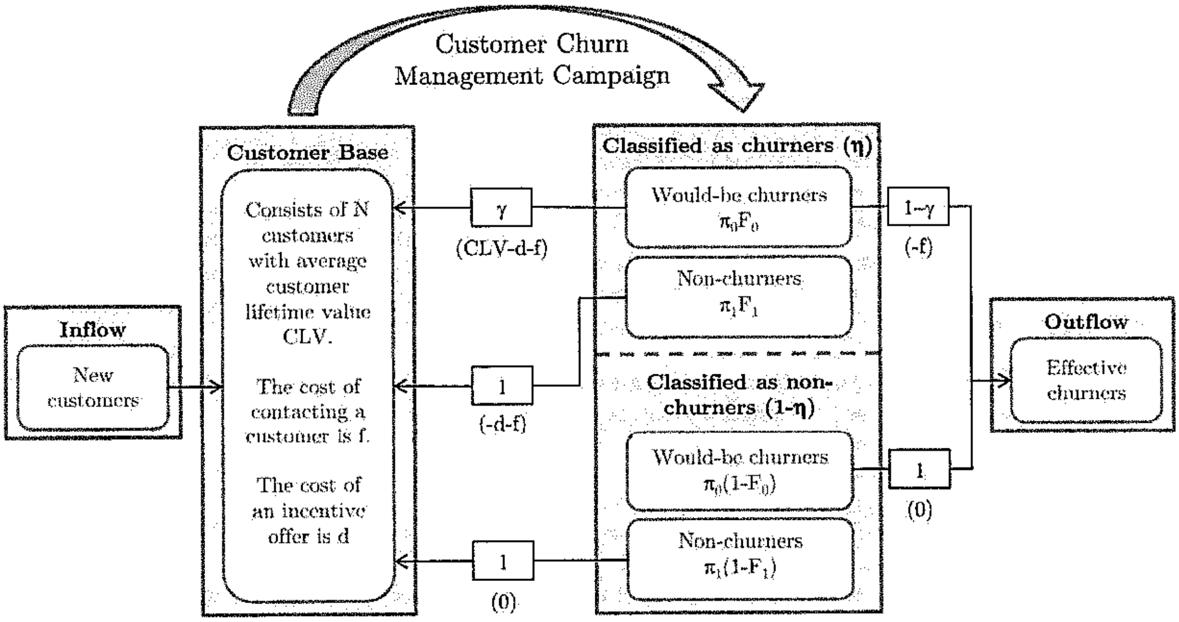

Customer Churn
Management Campaign
Inflow
New customers
Customer Base
Consists of N customers with average customer lifetime value CLV.
The cost of contacting a customer is f.
The cost of an incentive offer is d
γ
(CLV-d-f)
1
(-d-f)
1
(0)
Classified as churners (η)
Would-be churners π0F0
Non-churners π1F1
Classified as non-churners (1-η)
Would-be churners π0(1-F0)
Non-churners π1(1-F1)
1-γ
(-f)
1
(0)
Outflow
Effective churners

Fig. 1. Schematic representation of customer churn and retention dynamics within a customer base.

interpretation has to be given to the parameters, which requires domain specific knowledge. Section 4.1, will discuss the deterministic approach for performance measurement of customer churn prediction. The probabilistic approach will be treated in detail in Section 4.2.

4.1 The Maximum Profit Criterion for Customer Churn

Fig. 1 schematically represents the dynamical process of customer churn and retention within a customer base. New customers flow into the customer base by subscribing to a service of an operator, and existing customers flow out of the customer base by churning. When setting up a churn management campaign, the fraction η of the customer base with the highest propensity to churn is contacted at a cost f per person and is offered an incentive with cost d. In this fraction, there are true would-be churners and false would-be churners. In the latter group everyone accepts the incentive and does not churn, as they never had the intention. From the former group, a fraction γ accepts the offer and thus results in gained customer lifetime value (CLV), whereas the fraction $(1-\gamma)$ effectively churns. In the fraction $(1-\eta)$, which is not targeted, all would-be churners effectively churn, and all nonchurners remain with the company. The benefits per customer related to each flow are shown between brackets in Fig. 1. These are *incremental* benefits, as compared to not undertaking the customer churn management campaign. This process was described by Neslin et al. [16], who established the following expression for the total profit of a retention campaign:

$$\text{Profit} = N\eta[(\gamma CLV + d(1-\gamma))\pi_0\lambda - d - f] - A, \qquad (13)$$

with η the fraction of the customer base that is targeted, CLV the customer lifetime value, d the cost of the incentive, f the cost of contacting the customer, and A the fixed administrative costs. The lift coefficient, λ, is the percentage of churners within the targeted fraction η of customers, divided by the base churn rate, π_0. Lastly, γ is the fraction of the would-be churners accepting the offer, or alternatively it is interpreted as the probability of a targeted churner accepting the offer and thus not churning. It is assumed that CLV, A, f, and d are positive, and that $CLV > d$. Note that η depends on the choice for the threshold t, and thus the company has influence on the size of the targeted fraction.

Equation (13) can be expressed in terms of the score distributions. Moreover, if the average rather than the total profit is considered, and the fixed cost A, irrelevant for classifier selection, is discarded, it is possible to obtain a functional form equivalent to the expression for P in (1). To work out the conversion, note that:

$$\eta(t) = \pi_0 F_0(t) + \pi_1 F_1(t),$$
$$\lambda(t) = \frac{F_0(t)}{\pi_0 F_0(t) + \pi_1 F_1(t)}.$$

Moreover, two dimensionless parameters are introduced, being $\delta = d/CLV$ and $\phi = f/CLV$.

Definition 4. ***The average classification profit of a classifier for customer churn,*** *P_C, is the interpretation of Definition 1 specifically for customer churn:*

$$P_C(t;\gamma,CLV,\delta,\phi) = CLV(\gamma(1-\delta)-\phi)\cdot\pi_0F_0(t) \\ -CLV(\delta+\phi)\cdot\pi_1F_1(t). \tag{14}$$

Comparison with (1) shows that $b_0 = CLV(\gamma(1-\delta)-\phi)$ and $c_1 = CLV(\delta+\phi)$. When all parameter values in (14) are assumed to be known, the deterministic performance measure, the maximum profit defined in Definition 2, can be worked out for a customer churn context.

Definition 5. ***The maximum profit measure for customer churn,*** *MPC, is the interpretation of Definition 2 in a customer churn setting:*

$$\mathrm{MPC} = \max_{\forall t} P_C(t;\gamma,CLV,\delta,\phi). \tag{15}$$

As pointed out by Verbeke et al. [13], the MPC criterion is preferred over the commonly used top-decile lift. Setting the targeted fraction of customers to, e.g., 10 percent is a purely arbitrary choice and most likely leads to suboptimal profits and model selection. Since the ultimate goal of a company setting up a customer retention campaign is to minimize the costs associated with customer churn, it is logical to evaluate and select a customer churn prediction model by using the maximum obtainable profit as the performance measure. Moreover, it has another advantage, which is very appealing to practitioners, in the sense that it is possible to determine the optimal MPC fraction. This quantity represents how many customers should be targeted for profit maximization.

Definition 6. ***The profit maximizing fraction for customer churn,*** *$\bar{\eta}_{\mathrm{mpc}}$, is the fraction of the customer base which should be targeted to maximize the profit generated by the retention campaign:*

$$\bar{\eta}_{\mathrm{mpc}} = \pi_0F_0(T) + \pi_1F_1(T), \tag{16}$$

with:

$$T = \arg\max_{\forall t} P_C(t;\gamma,CLV,\delta,\phi). \tag{17}$$

The combination of the maximum profit MPC and the optimal fraction η_{mpc} provides telecom operators with a rigorous framework for making operational decisions with profit maximization as main goal.

4.2 The Expected Maximum Profit Measure for Customer Churn

In the previous section, it was assumed that accurate estimates for the parameters in the expression for P_C are available, in which case a deterministic performance measure can be employed. Often, however, there is significant uncertainty involved. Specifically for the problem of customer churn, there are four parameters, of which customer lifetime value, the cost of the incentive and the contacting cost can be estimated with sufficient reliability. However, about the probability of a churner accepting the retention offer much less is known. Therefore, this probability, γ, is considered to introduce uncertainty in the estimation of the maximum profit, and thus, a probabilistic performance measure is proposed.

Definition 7. ***The expected maximum profit measure for customer churn, EMPC,*** *is the interpretation of Definition 3 in a customer churn setting:*

$$\mathrm{EMPC} = \int_\gamma P_C(T(\gamma);\gamma,CLV,\delta,\phi)\cdot h(\gamma)\mathrm{d}\gamma, \tag{18}$$

with T, the optimal cutoff for a given γ, and $h(\gamma)$ the probability density function for γ.

Analogously to the H measure, a beta distribution is proposed, because of its flexibility. The parameters are named α' and β', where the prime aims to distinguish these parameters from those of the H measure. A procedure for the estimation of EMPC is given in the mathematical appendix, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TKDE.2012.50. The cost benefit ratio for customer churn, θ_C, is equal to

$$\theta_C = \frac{b_1+c_1}{b_0+c_0} = \frac{\delta+\phi}{\gamma(1-\delta)-\phi}, \tag{19}$$

where γ is the parameter introducing the uncertainty.

At this point, an important remark has to be made regarding the relation between the ROC curve of a classifier and its EMPC. In Section 3.1, it was stated that, theoretically, calculating the EMP considers all points of the convex hull of a ROC curve. With EMPC, the cost and benefit parameters obey certain assumptions, specific for a customer churn context. The value of θ_C now ranges from $-(\delta+\phi)/\phi$ to $-\infty$ for $\gamma\in[0,\phi/(1-\delta))$, and from $+\infty$ to $(\delta+\phi)/(1-\delta-\phi)$ for $\gamma\in(\phi/(1-\delta),1]$. Hence, for a certain interval of γ, θ_C (and also b_0) is negative. This corresponds to the situation when the probability of a churner accepting the retention offer is so low that the expected net gained customer lifetime value, i.e., $\gamma(CLV-d)$, does not offset the cost of contacting the customers, f. In this case, the optimal decision is not to contact any customer, which corresponds to the origin of the ROC curve.

A second observation concerns the fact that θ_C does not extend over the entire positive set of real numbers, $\mathbb{R}^+$, since it only spans $[(\delta+\phi)/(\gamma-\gamma\delta-\phi),+\infty)$. Therefore, points on the convex hull of a ROC curve with a slope smaller than $(\pi_1/\pi_0)\cdot(\delta+\phi)/(1-\delta-\phi)$ are not optimal for *any* value of $\gamma\in[0,1]$. Consequently, when calculating EMPC, not all points of the convex hull are taken into account. Moreover, the more skewed the data set toward nonchurners, the smaller the portion of the convex hull which is taken into account. Intuitively this makes sense since, for very low churn rates, the percentage of churners in a large targeted fraction will quickly become insignificant and hence jeopardize the profitability. The optimal targeted fraction can also be determined when dealing with the expected maximal profit.

Definition 8. ***The expected profit maximizing fraction for customer churn,*** *$\bar{\eta}_{\mathrm{empc}}$, is the fraction of the customer base* *which is targeted when taking into account the uncertainty about* γ*:*

$$\bar{\eta}_{\mathrm{empc}} = \int_{\gamma} [\pi_0 F_0(T(\gamma)) + \pi_1 F_1(T(\gamma))] \cdot h(\gamma)\mathrm{d}\gamma, \tag{20}$$

with $T(\gamma)$ *being the optimal cutoff, as defined in (16).*

Hence, there are two frameworks available to telecom operators. One framework, based on MPC, takes a deterministic approach and considers all parameters in the profit equation to be known. The second framework is based on EMPC and provides practitioners with a probabilistic method to evaluate profits and losses, reflecting the uncertainty about the response rate γ. Due to its specific costs and benefits, performance measurement in a customer churn setting cannot simply be done by applying the H measure. However, it is possible to approximate the EMPC measure to a certain degree, as will be shown in Section 5.

4.3 Sensitivity of the EMPC Measure to Variations in the Fixed Parameters

The EMPC measure incorporates uncertainty about the acceptance rate γ, but for the parameters CLV, δ, and ϕ, fixed values are assumed. In this paragraph, the sensitivity of the EMPC measure to small variations in these fixed parameters will be assessed. Therefore, the first derivative of (18) with respect to CLV is calculated:

$$\frac{\partial \mathrm{EMPC}}{\partial CLV} = \int_{\gamma} \frac{\partial P_C(T(\gamma); \gamma, CLV, \delta, \phi)}{\partial CLV} h(\gamma)\mathrm{d}\gamma.$$

Note that the optimal cutoff, $T(\gamma)$, depends on the value of CLV. In what follows, the partial derivative within the integration is elaborated. Using (14) yields

$$\begin{aligned}\frac{\partial P_C(T)}{\partial CLV} &= (\gamma(1-\delta) - \phi) \cdot \pi_0 F_0(T) - (\delta + \phi) \cdot \pi_1 F_1(T) \\ &\quad + CLV(\gamma(1-\delta) - \phi) \cdot \pi_0 f_0(T) \frac{\partial T}{\partial CLV} \\ &\quad - CLV(\delta + \phi) \cdot \pi_1 f_1(T) \frac{\partial T}{\partial CLV},\end{aligned}$$

which, by using (14) and (19), can be rewritten as follows:

$$\begin{aligned}\frac{\partial P_C(T)}{\partial CLV} = &\frac{P_C(T)}{CLV} + CLV \frac{\partial T}{\partial CLV} (\gamma(1-\delta) - \phi) \\ &\pi_1 f_1(T) \cdot \left(\frac{\pi_0 f_0(T)}{\pi_1 f_1(T)} - \theta_C \right).\end{aligned}$$

Now, (4) can be applied, which leads to the last term between brackets being equal to zero, and thus:

$$\frac{\partial \mathrm{EMPC}}{\partial CLV} = \int_{\gamma} \frac{P_C(T)}{CLV} h(\gamma)\mathrm{d}\gamma = \frac{\mathrm{EMPC}}{CLV}. \tag{21}$$

In other words, this means that when CLV is changed, holding δ and ϕ constant, EMPC changes proportionally. Note that holding δ and ϕ constant in fact means that the cost of the retention offer *as a percentage of CLV*, and the cost of contacting a customer *as a percentage of CLV* remain constant.

Analogously, for variations in ϕ the following equation can be derived:

$$\begin{aligned}\frac{\partial \mathrm{EMPC}}{\partial \phi} &= -CLV \int_{\gamma} [\pi_0 F_0(T) + \pi_1 F_1(T)] h(\gamma)\mathrm{d}\gamma \\ &= -CLV \cdot \bar{\eta}_{\mathrm{empc}},\end{aligned} \tag{22}$$

where $\bar{\eta}_{\mathrm{empc}}$ is the expected profit maximizing fraction. For changes in δ, the following holds true:

$$\begin{aligned}\frac{\partial \mathrm{EMPC}}{\partial \phi} &= -CLV \int_{\gamma} [\gamma \pi_0 F_0(T) + \pi_1 F_1(T)] h(\gamma)\mathrm{d}\gamma \\ &= -CLV \cdot \bar{\rho}_{\mathrm{empc}},\end{aligned} \tag{23}$$

with $\bar{\rho}_{\mathrm{empc}}$, the expected fraction of the customer base which *accepts the retention offer*. In fact, this implies that the larger the optimal targeted fraction (or the fraction accepting the offer), the more sensitive the expected maximum profit for variations in ϕ (or δ). Also note that an increase in CLV, while holding d and f constant instead of δ and ϕ, corresponds to a parallel decrease in δ and ϕ. In Section 6, the sensitivity of the rankings will be analyzed in a real-life case study.

4.4 Parameter Values of the Profit Function and Response Rate Distribution

To employ the MPC and EMPC measures, it is necessary to obtain values for the parameters in the profit function. The values for these parameters, such as CLV, are industry specific and may even vary from company to company. This paper focuses on customer churn prediction in the telecom industry and estimates for the parameters are based on values reported in the literature ([16] and [17]) and information from telecom operators. When a specific company uses MPC or EMPC for selecting the optimal model, they can plug in their own estimates for the parameters. In this study, the values of the parameters CLV, d, and f are taken as €200, €10, and €1, respectively.

The parameter γ, representing the response rate, is much more difficult to estimate. For the MPC criterion, a single point estimate would be required, which corresponds to one degree of freedom. For the EMPC criterion, however, there are two degrees of freedom, i.e., α and β. This enables the practitioner to define an expected value and a standard deviation, where the latter accounts for the uncertainty in the practitioner's estimation. The α and β parameters can be obtained by solving the following system of equations:

$$\begin{cases} \mathrm{E}[\gamma] = \mu = \alpha/(\alpha+\beta) \\ \mathrm{Var}[\gamma] = \sigma^2 = \alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)], \end{cases}$$

which yields

$$\begin{cases} \alpha = \mu((1-\mu)\mu/\sigma^2 - 1) \\ \beta = (1-\mu)((1-\mu)\mu/\sigma^2 - 1). \end{cases}$$

There is only one restriction, namely that α and β need to be strictly greater than one in order to obtain a unimodal beta distribution.

Neslin et al. [16] assumed values ranging from 10 to 50 percent, therefore this paper proposes the expected value and standard deviation of the acceptance rate to be equal to 30 percent and 10 percent respectively. This leads to the parameters α' and β' being 6 and 14, respectively, which corresponds to the probability density function plotted in Fig. 2 (solid line). When MPC is calculated, a point estimate

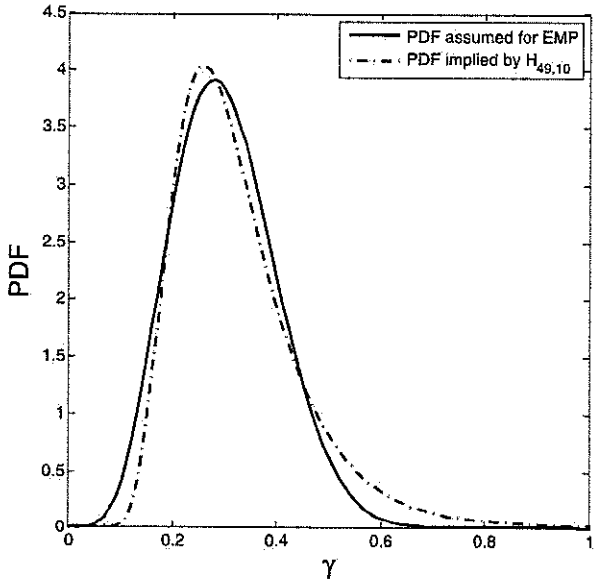


Fig. 2. The graph shows the proposed density function for the response rate in the EMPC measure, and the density function implied by the H-measure with optimized parameters.

for γ is required, in which case the expected value, equal to 30 percent, is taken.

5 Expected Maximum Profit and the H Measure

As mentioned in Section 3.2, the H measure is an expected minimum loss measure, with certain assumptions about the distribution of the cost and benefit parameters. These assumptions imply that the H measure takes into account all points on the convex hull. The EMPC measure, however, only takes into account a part of the convex hull. In this section, it will be investigated whether it is possible to approximate the EMPC measure with the H measure. After all, it would be convenient if the EMPC measure fits in an existing performance measurement framework.

The discussion will focus on the variable part of the H measure, defined in (12). Only this part of the H measure changes with varying classifiers, holding the data set constant, and thus, the ranking of techniques for a data set only depends on H_{var}. At this point, it is assumed that the variable c in H_{var} is a function of γ, δ, and ϕ:

$$c = \frac{\gamma(1-\delta)-\phi}{\gamma(1-\delta)+\delta} = \frac{K\gamma-\phi}{K\gamma+\delta},$$

where $K = 1-\delta$ for notational convenience. In what follows, a variable transformation from c to γ will be carried out in the expression for H_{var}. Expressions for dc and $u_{\alpha,\beta}(c)$ as a function of γ are derived:

$$\mathrm{d}c = \frac{K(\delta+\phi)}{(K\gamma+\delta)^2}\mathrm{d}\gamma, u_{\alpha,\beta}(c(\gamma)) = \frac{1}{B(1,\alpha,\beta)}\left(\frac{K\gamma-\phi}{K\gamma+\delta}\right)^{\alpha-1}\left(\frac{\delta+\phi}{K\gamma+\delta}\right)^{\beta-1}.$$

Hence, the variable substitution in the expression for H_{var} leads to

$$\begin{aligned} H_{var} &= \int_0^1 [\pi_0 c F_0 - \pi_1(1-c)F_1]\cdot u_{\alpha,\beta}(c)\mathrm{d}c \\ &= \int_{\phi/K}^{+\infty} [(K\gamma-\phi)\pi_0 F_0 - (\delta+\phi)\pi_1 F_1] \\ &\quad \cdot \underbrace{\frac{K(\phi+\delta)^\beta}{B(1,\alpha,\beta)}\cdot\frac{(K\gamma-\phi)^{\alpha-1}}{(K\gamma+\delta)^{\alpha+\beta+1}}}_{g(\gamma)}\mathrm{d}\gamma \\ &= \frac{1}{CLV}\int_{\phi/K}^{+\infty} P_C(T(\gamma);CLV,\gamma,\delta,\phi)\cdot g(\gamma)\mathrm{d}\gamma. \end{aligned}$$

For $\gamma < \phi/K$, the benefit associated with a targeted churner is negative. Since the maximal profit, $P(T(\gamma))$, for this range of γ is equal to zero, H_{var} can be expressed as:

$$H_{var} = \frac{1}{CLV}\int_0^{+\infty} P(T(\gamma);CLV,\delta,\phi)\cdot g(\gamma)\mathrm{d}\gamma. \qquad (24)$$

Equation (24) shows that H_{var} essentially is an EMPC measure scaled by CLV, for which the distribution $h(\gamma)$ is replaced with the weight function $g(\gamma)$, and where $\gamma \in \mathbb{R}^+$. It is to be noted that the function $g(\gamma)$ is not a probability density in the strict sense, since generally, the area under g will not be equal to one. However, by simply multiplying H_{var} with a normalization constant, this is resolved. Let the proper density function $\tilde{h}(\gamma)$ be the normalized function $g(\gamma)$, then H_{var} equals

$$\begin{aligned} H_{var} &= \frac{\int_0^{+\infty} g(\gamma)\mathrm{d}\gamma}{CLV} \\ &\quad \cdot \int_0^{+\infty} P(T(\gamma);CLV,\delta,\phi)\cdot\tilde{h}(\gamma)\mathrm{d}\gamma. \end{aligned} \qquad (25)$$

Because the H measure takes into account all points of the convex hull, this corresponds to integrating over $\gamma \in [0, +\infty)$.

As illustrated by (25), the H measure will rank classifiers according to their expected maximum profit, implicitly assuming the distribution $\tilde{h}(\gamma)$, which is specified by the parameters α and β. The previously proposed distribution for γ, $h(\gamma)$, is specified by the parameters α' and β'. It is not possible to analytically find values for α and β so that $\tilde{h}(\gamma)$ equals $h(\gamma)$. Therefore, a distance measure between the two distributions is numerically minimized. Several distance measures have been experimented with, but the *Hellinger distance* leads to the best results. The Hellinger distance between two probability measures, D_h, is defined as the L_2 distance between the square root of their densities [23]. Hence, the optimal parameters for the approximating H measure can be found by solving the following minimization problem:

$$\begin{aligned} &\min_{\forall\alpha,\beta} D_h(h(\gamma),\tilde{h}(\gamma)) \\ &= \min_{\forall\alpha,\beta}\int_0^{+\infty}\left(\sqrt{h(\gamma)}-\sqrt{\tilde{h}(\gamma)}\right)^2\mathrm{d}\gamma. \end{aligned} \qquad (26)$$

Note that $h(\gamma)$ is parameterized by α' and β', which are considered to be known, whereas $\tilde{h}(\gamma)$ is parameterized by α and β, which are to be optimized.

In order to find the optimal parameters for the approximating H measure, the minimization problem in (26) has been numerically solved, yielding $\alpha \approx 49$ and $\beta \approx 10$ as best

TABLE 2
Summary of Data Set Characteristics: ID, Source, Region, Number of Observations, Number of Attributes, Percentage Churners, and References to Previous Studies Using the Data Set

ID	Source	Region	# Obs.	# Att.	%Churn	Reference
O1	Operator	North America	47,761	53	3.69	[18], [13], [19]
O2	Operator	East Asia	11,317	21	1.56	[20], [13]
O3	Operator	East Asia	2,904	15	3.20	[20], [13]
O5	Operator	East Asia	2,180	15	3.21	[20], [13]
O6	Operator	Europe	338,874	727	1.80	[13]
D1	Duke	North America	93,893	197	1.78	[21], [22], [16], [13], [19]
D2	Duke	North America	38,924	77	1.99	[13], [19]
D3	Duke	North America	7,788	19	3.30	[13], [19]
UCI	UCI	-	5,000	23	14.14	[22], [13], [19]
KDD	KDD Cup 2009	Europe	46,933	242	6.98	[13]

values. This density function is plotted as the dashed line in Fig. 2, and has an expected value of 28.9 percent and a standard deviation of 10.4 percent. Although the density theoretically is nonzero for $\gamma > 1$, the accumulated probability for this region is smaller than 10^{-3}. In Section 6, a case study will be carried out and it will be investigated how well the H measure approximates the EMPC measure in terms of ranking several classification techniques.

6 Case Study

In this section, a case study will be presented to illustrate the use of the EMPC and H measure for measuring classification performance in a customer churn prediction context. A selection of 21 techniques has been tested on 10 different customer churn data sets. The data sets have been preprocessed to prepare them for the analysis. Table 2 summarizes the main characteristics of the data sets. The classification techniques used in the benchmarking study are listed in Table 3 and include many commonly applied data mining methods. A more extensive discussion regarding the data sets, the preprocessing and the employed techniques can be found in the original benchmarking study, carried out by Verbeke et al. [13].

Each data set has been split into a training set and a test set, where the former was used to train the classifiers, and the latter to perform an out-of-sample test. Thus, each instance in the test set is assigned a score s, on which classification is based by setting a cutoff value. Through the empirical score distributions, the AUC, H measure, MPC measure, and EMPC measure are then calculated. Per data set, each measure leads to a ranking of the 21 techniques. Since the ranking is the most important output of the classification performance measurement process, the rankings from the different measures will be compared to one another. As the 10 data sets involved in this study are independent, the 10 resulting rankings of techniques are considered independent observations.

6.1 Comparison of EMPC with Other Performance Measures

Thus, each performance measure results in a ranking for every data set. For each single data set, Kendall's τ_b is used as a measure of dependence between the rankings based on different performance measures, yielding 10 values for each correlation. The box plot in Fig. 3 shows the distribution of the values of the correlation between EMPC and the four displayed performance metrics. It is clear that the agreement between AUC and EMPC-based rankings is lowest,

TABLE 3
Overview of the Classification Techniques

Overview of Classification Techniques	
Decision Tree Approaches	**Rule Induction Techniques**
C4.5 Decision Tree (C4.5)	RIPPER (RIPPER)
Classification and Regression Tree (CART)	PART (PART)
Alternating Decision Tree (ADT)	
	Statistical Classifiers
Ensemble Methods	Logistic Regression (Logit)
Random Forests (RF)	Naive Bayes (NB)
Logistic Model Tree (LMT)	Bayesian Networks (BN)
Bagging (Bag)	
Boosting (Boost)	**Nearest Neighbors**
	k-Nearest neighbors $k = 10$ (KNN10)
SVM Based Techniques	k-Nearest neighbors $k = 100$ (KNN100)
SVM with linear kernel (linSVM)	
SVM with radial basis function kernel (rbfSVM)	**Neural Networks**
LSSVM with linear kernel (linLSSVM)	Multilayer Perceptron (NN)
LSSVM with radial basis function (rbfLSSVM)	Radial Basis Function Network (RBFN)
Voted Perceptron (VP)	

More information regarding these data mining techniques can be found in the original benchmarking study [13] or in a standard data mining handbook (e.g., [24], [25]).

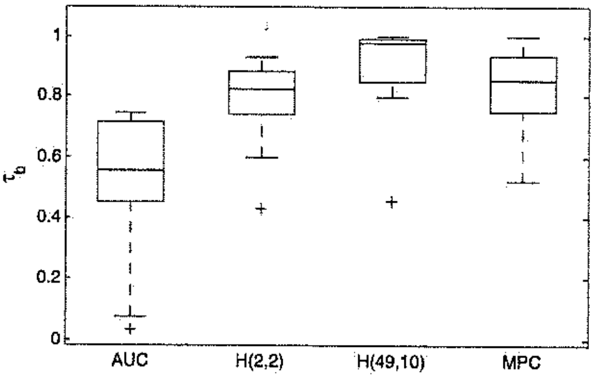


Fig. 3. Box plot indicating the distribution of correlations between EMPC and $H_{2,2}$, $H_{49,10}$, and MPC, respectively.

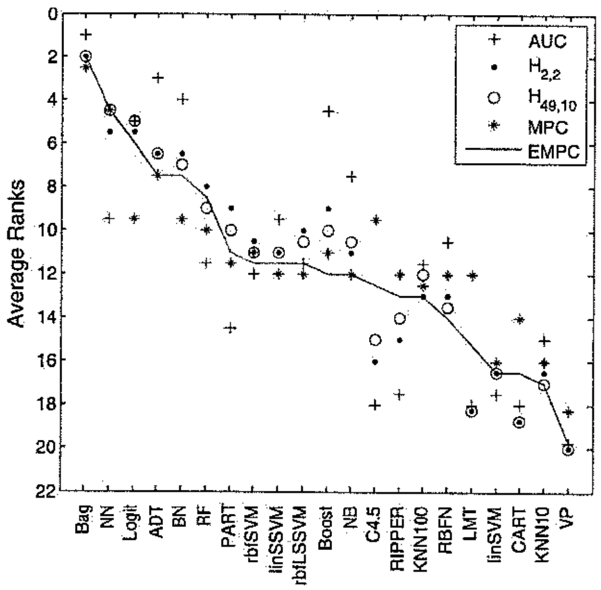


Fig. 4. The average rank over 10 data sets for each technique based on several performance measures (i.e., AUC, $H_{2,2}$, $H_{49,10}$, MPC, EMPC).

and it has the largest variability. The H measure with default parameters ($\alpha = 2$ and $\beta = 2$) shows a higher level of correlation, but still significant variability. The H measure with optimized parameter values, however, shows very high correlation and lowest variability, indicating that both rankings agree to a large extent and that this is a reasonable approximation to the EMPC measure. Finally, the correlation between MPC and EMPC is again lower, which can be attributed to the fact that the underlying cost and benefit assumptions are different, i.e., it is a deterministic versus a probabilistic approach. This also suggests that when the operator has accurate estimates of the response rate γ, it is preferable to use MPC as criterion. When there is more uncertainty involved with the estimation of γ, a probabilistic approach, and thus EMPC, is recommended. Furthermore, the box plot shows an outlier for the correlation between EMPC and AUC, $H_{2,2}$, and $H_{49,10}$. These outliers correspond to the data set D2, where the expected maximum profit is zero for all techniques. As a result, EMPC and MPC consider all techniques to be unprofitable, and they all receive the same rank. AUC, $H_{2,2}$, and $H_{49,10}$ on the other hand will rank techniques differently. Therefore, the correlation is low for this data set.

A further indication of the agreement in ranking is given in Fig. 4, which shows the rank of each technique, averaged over the 10 data sets. The full line represents the ranking according to the EMPC measure, whereas the other points represent other performance metrics. Hence, a point plotted far from the full line shows strong disagreement with the EMPC-based ranking. One can immediately see that the disagreement of AUC with EMPC is larger than any other performance measure. The H measure with optimized parameters follows the EMPC ranking much closer, again indicating that it is a reasonable approximation.

A second point of view is the profitability of a retention campaign. As explained before, EMPC ranks classifiers according to the expected profit based on assumptions about the classification costs and benefits. AUC, as shown by Hand [2], is also a measure of profit, but with invalid assumptions for the distribution of costs and benefits. Therefore, choosing a classifier based on AUC may lead to suboptimal model selection from a profit point of view, which is illustrated in Table 4. For each data set, the optimal technique is selected based on EMPC and on AUC, and the expected profit for that particular choice of classifier is given. As indicated in the last column, selection based on AUC leads in some cases to suboptimal model selection, with losses up to €0.137 per customer, a substantial amount for telecom operators with millions of customers. Moreover, for selection based on EMPC, it is possible to calculate the fraction of the vast customer base which needs to be

TABLE 4
Model Selection Based on EMPC and on AUC, with the Expected Profit for Each Method

	Selection based on EMPC			Selection based on AUC			
Data Set	Technique	EMPC (€)	η_{empc}(%)	Technique	AUC	EMPC (€)	Difference in EMPC (€)
O1	ADT	0.129	2.10	Logit	0.747	0.120	0.009
O2	NN	0.051	0.09	Bag	0.644	0.017	0.034
O3	Boost	0.646	3.77	Bag	0.903	0.521	0.125
O5	Bag	1.652	3.44	Bag	0.964	1.652	0
O6	Bag	0.047	0.32	Bag	0.902	0.047	0
D1	KNN10	0.001	0.03	Boost	0.599	0	0.001
D2	PART	0	0	Bag	0.618	0	0
D3	BN	0.351	2.83	NB	0.809	0.214	0.137
UCI	RF	5.734	13.65	Bag	0.915	5.640	0.093
KDD	ADT	0.379	7.82	Bag	0.713	0.375	0.004

For model selection with EMPC, also the optimal fraction of customers to target is given.

targeted to realize the maximal profit, which is also displayed in Table 4. This is one of the major advantages of the EMPC (and also the MPC) measure, as it gives guidance to practitioners about how many customers to include in a retention campaign. When selecting a model with AUC or the H measure, there is no such guidance, and deviating from the optimal fraction may again lead to suboptimal profits.

6.2 Sensitivity of the EMPC Measure and the Ranking to Variations in the Fixed Parameters

Section 4.3 discusses the impact of variations in CLV, δ, and ϕ on the estimated expected profit and analytically derives first order approximations for this sensitivity. This yields some straightforward rules of thumb for the sensitivity, such as, e.g., the higher the targeted fraction, the more sensitive the profit to changes in ϕ. However, the question arises how the ranking between classification algorithms is affected. Therefore, the results of the case study are used to analyze this impact. First, the techniques have been ranked with the proposed values for CLV, δ, and ϕ. Then, each parameter has been multiplied with a constant (while holding the others equal), and the techniques have been ranked again with this new parameter value. The correlations between the ranking in the base scenario and the ranking in the new scenario have been plotted for varying values of the multiplier, ranging from 1/2 to 2. Fig. 5 shows these results for all three fixed parameters. The median and the first and third quartile (over the 10 data sets) have been plotted. Note that the plot for CLV assumes that CLV is changed while d and f are held constant (not δ and ϕ). It can be seen that variations in both CLV and δ have a similar impact, with median correlations decreasing until 0.8. The impact of ϕ is virtually nonexistent.

Furthermore, when the ranking changes, also the best performing technique (with the highest expected profit) may change. Therefore, Fig. 5 also displays the impact of suboptimal classifier selection due to incorrect parameter values on the profitability. The percentage loss is plotted on the right axis. Again, the impact is most significant for CLV and δ, whereas variations in ϕ do not impact the profitability. But even though there is an impact, it is relatively limited to losses of maximal 20 percent, and this for substantial variations in the parameters (doubling or halving the CLV or δ). These results, both in terms of correlation between rankings and percentage loss due to incorrect classifier selection, indicate that the EMPC measure and corresponding rankings are robust to changes in the fixed parameters. The impact only becomes noticeable for multipliers smaller than 0.75 or larger than 1.5.

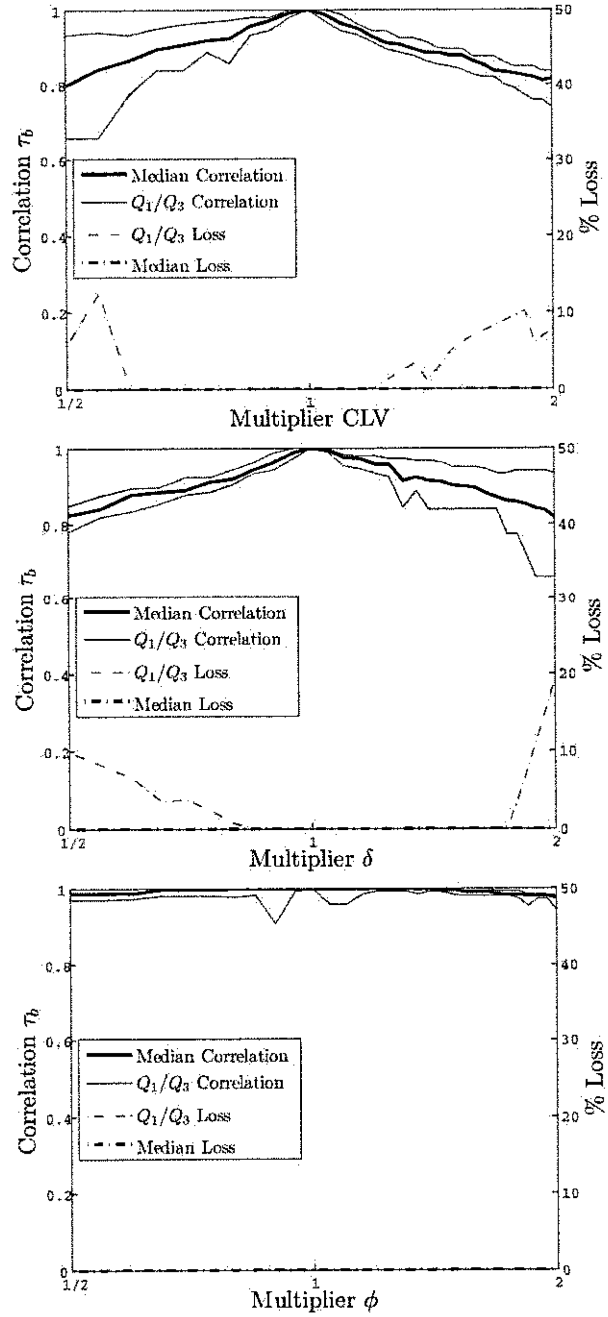


Fig. 5. Sensitivity of the EMPC measure to variations in CLV, δ, and ϕ, respectively. The left axis and full lines indicate the correlation between the ranking of the base scenario and the ranking with the changed parameter. The right axis and dashed lines show the percentage loss due to incorrect classifier selection. For both correlation and percentage loss, the median, first quartile (Q_1) and third quartile (Q_3) have been plotted. The x-axis is plotted on a logarithmic scale with base 2, with the multiplier ranging from 1/2 to 2.

7 Conclusion

The implications of the findings presented in this paper are straightforward but essential. Companies rely more than ever on data mining techniques to support their decision making processes. When evaluating a classification technique which is to be used in a business context, it is imperative to base any evaluation criterion on the goal of the end user. Since companies strive for profit maximization, a performance measure evidently should take this into account. The very commonly used area under the ROC curve does have its merits and an interesting interpretation in the sense that the AUC of a classification method is the probability that a randomly chosen positive instance will be ranked higher than a randomly chosen negative instance. However, as [2] has shown, AUC makes incorrect implicit assumptions about the misclassification costs, and the use of this performance metric in a business environment leads to suboptimal profits. This paper outlines a theoretical framework which incorporates all gains and losses related to the employment of a data mining technique, and defines a probabilistic performance measure, the expected maximum profit.

As each corporate environment has its own specificities, the framework is defined on a general level. To be applicable to a certain business problem, the particularities of its cost and benefit structure need to be incorporated. This process is worked out in detail for the problem of customer churn and an EMP measure for customer churn, EMPC, is derived. Also the link between EMPC and the H measure was investigated and it appears that the latter with appropriately chosen distribution parameters is a good approximation of the former. The performance measure for customer churn is validated in an extensive case study. The results clearly indicate that the use of AUC as a performance metric leads to suboptimal profits. The case study also points to one of the major advantages of the EMPC measure. It does not only select the classifier which maximizes the profit, but it also provides the practitioner with an estimate of the fraction of the customer base which needs to be targeted in the retention campaign. This optimal fraction varies from case to case, and deviating from this fraction again leads to suboptimal profits. Note that the H measure, although it is able to select the most profitable classifier, does not provide guidance on the optimal fraction of the customer base to be included in the retention campaign. Finally, a sensitivity analysis was carried out, to analyze how vulnerable the EMPC measure is to incorrect estimation of the fixed parameters CLV, δ, and ϕ. The outcome shows that the EMPC measure and its resulting ranking is relatively robust with regard to changes in these parameter values.

An obvious though not straightforward direction for further research entails the extension of the framework to multiclass problems. Furthermore, besides the application of the cost-benefit analysis framework to the customer churn problem, there remain many business problems where a profit driven performance measure would add value. Among others, data mining techniques are employed for financial credit scoring [26], direct marketing response models [27], fraud detection [28], and viral marketing in social networks [29]. With the explosive growth of data, and the increasing popularity of social network sites, companies will rely more than ever on data mining techniques to support their decision making process. In such cases, the cost-benefit analysis framework presented in this paper provides directions on how to develop performance measures tailored to specific business problems.

Acknowledgments

The authors extend our gratitude to the Flemish Research Council for financial support (FWO postdoctoral research grant, Odysseus grant B.0915.09), and to the reviewers for their insightful comments.

References

[1] T. Fawcett, "An Introduction to ROC Analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006.

[2] D. Hand, "Measuring Classifier Performance: A Coherent Alternative to the Area under the ROC Curve," *Machine Learning*, vol. 77, no. 1, pp. 103-123, 2009.

[3] W. Verbeke, D. Martens, C. Mues, and B. Baesens, "Building Comprehensible Customer Churn Prediction Models with Advanced Rule Induction Techniques," *Expert Systems with Applications*, vol. 38, pp. 2354-2364, 2011.

[4] A.A. Nanavati, R. Singh, D. Chakraborty, K. Dasgupta, S. Mukherjea, G. Das, S. Gurumurthy, and A. Joshi, "Analyzing the Structure and Evolution of Massive Telecom Graphs," *IEEE Trans. Knowledge and Data Eng.*, vol. 20, no. 5, pp. 703-718, May 2008.

[5] S. Ali and K. Smith, "On Learning Algorithm Selection for Classification," *Applied Soft Computing*, vol. 6, no. 2, pp. 119-138, 2006.

[6] P. Baldi, S. Brunak, Y. Chauvin, C. Andersen, and H. Nielsen, "Assessing the Accuracy of Prediction Algorithms for Classification: An Overview," *Bioinformatics*, vol. 16, no. 5, pp. 412-424, 2000.

[7] N. Chawla, "Data Mining for Imbalanced Datasets: An Overview," *Data Mining and Knowledge Discovery Handbook*, pp. 875-886, Springer, 2010.

[8] P. Domingos, "Metacost: A General Method for Making Classifiers Cost-Sensitive," *Proc. Fifth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, pp. 155-164, 1999.

[9] F. Provost and T. Fawcett, "Robust Classification for Imprecise Environments," *Machine Learning*, vol. 42, no. 3, pp. 203-231, 2001.

[10] A. Bernstein, F. Provost, and S. Hill, "Toward Intelligent Assistance for a Data Mining Process: An Ontology-Based Approach for Cost-Sensitive Classification," *IEEE Trans. Knowledge and Data Eng.*, vol. 17, no. 4, pp. 503-518, Apr. 2005.

[11] Z. Zhou and X. Liu, "Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem," *IEEE Trans. Knowledge and Data Eng.*, vol. 18, no. 1, pp. 63-77, Jan. 2006.

[12] C. Elkan, "The Foundations of Cost-Sensitive Learning," *Proc. Int'l Joint Conf. Artificial Intelligence*, vol. 17, no. 1, pp. 973-978, 2001.

[13] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New Insights into Churn Prediction in the Telecommunication Sector: A Profit Driven Data Mining Approach," *European J. Operational Research*, vol. 218, no. 1, pp. 211-229, 2012.

[14] A. Bradley, "The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145-1159, 1997.

[15] R. Prati, G. Batista, and M. Monard, "A Survey on Graphical Methods for Classification Predictive Performance Evaluation," *IEEE Trans. Knowledge and Data Eng.*, vol. 23, no 11, pp. 1601-1618, Nov. 2011.

[16] S. Neslin, S. Gupta, W. Kamakura, J. Lu, and C. Mason, "Detection Defection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models," *J. Marketing Research*, vol. 43, no. 2, pp. 204-211, 2006.

[17] J. Burez and D. Van den Poel, "CRM at a Pay-TV Company: Using Analytical Models to Reduce Customer Attrition by Targeted Marketing for Subscription Services," *Expert Systems with Applications*, vol. 32, pp. 277-288, 2007.

[18] M. Mozer, R. Wolniewicz, D. Grimes, E. Johnson, and H. Kaushansky, "Predicting Subscriber Dissatisfaction and Improving Retention in the Wireless Telecommunications Industry," *IEEE Trans. Neural Networks*, vol. 11, no. 3, pp. 690-696, May 2000.

[19] T. Verbraken, W. Verbeke, D. Martens, and B. Baesens, "Profit Optimizing Customer Churn Prediction with Bayesian Network Classifiers," accepted for publication in *Intelligent Data Analysis*, 2013.

[20] J. Hur and J. Kim, "A Hybrid Classification Method Using Error Pattern Modeling," *Expert Systems with Applications*, vol. 34, no. 1, pp. 231-241, 2008.

[21] A. Lemmens and C. Croux, "Bagging and Boosting Classification Trees to Predict Churn," *J. Marketing Research*, vol. 43, no. 2, pp. 276-286, 2006.

[22] E. Lima, C. Mues, and B. Baesens, "Domain Knowledge Integration in Data Mining Using Decision Tables: Case Studies in Churn Prediction," *J. Operational Research Soc.*, vol. 60, no. 8, pp. 1096-1106, 2009.

[23] A. Van der Vaart, *Asymptotic Statistics*. Cambridge Univ Press, 2000.

[24] P. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Pearson Education, 2006.

[25] I.H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers, Inc., 2000.

[26] B. Baesens, R. Setiono, C. Mues, and J. Vanthienen, "Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation," *Management Science*, vol. 49, no. 3, pp. 312-329, 2003.

[27] G. Cui, M. Wong, and H. Lui, "Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming," *Management Science*, vol. 52, no. 4, pp. 597-612, 2006.

[28] S. Viaene, R.A. Derrig, B. Baesens, and G. Dedene, "A Comparison of State-of-the-Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection," *J. Risk and Insurance*, vol. 69, no. 3, pp. pp. 373-421, 2002.

[29] P. Domingos, "Mining Social Networks for Viral Marketing," *IEEE Intelligent Systems*, vol. 20, no. 1, pp. 80-82, 2005.

Thomas Verbraken received the MSc degree in civil engineering from the K.U. Leuven, Belgium, in 2007, and in 2011, he completed all three levels of the CFA program. Since 2010, he has been working toward the PhD degree at the Faculty of Business and Economics, Department of Decision Sciences and Information Management at the K.U. Leuven, Belgium. Being a member of the Research Center for Management Informatics (LIRIS), his main research focuses on data mining for business applications. More specifically, he has been working on customer churn prediction, Bayesian network classifiers, profit-based classification performance measurement, and classification in network environments. He is a student member of the IEEE.

Wouter Verbeke received the PhD degree in applied economics at the University of Leuven, Belgium. He is a lecturer in management science at the University of Edinburgh Business School, and his research focuses on the development and application of data mining and network modeling techniques in business settings, such as customer churn prediction models in the telco sector, and credit rating migration models in the financial sector.

Bart Baesens is an associate professor at K.U. Leuven, Belgium, and a lecturer at the University of Southampton, United Kingdom. He has done extensive research on predictive analytics, data mining, customer relationship management, web analytics, fraud detection, and credit risk management. His findings have been published in well-known international journals (e.g., *Machine Learning, Management Science, IEEE Transactions on Neural Networks, IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Evolutionary Computation, Journal of Machine Learning Research,*) and presented at international top conferences. He is also a coauthor of the book *Credit Risk Management: Basic Concepts*, published in 2008. He regularly tutors, advises, and provides consulting support to international firms with respect to their data mining, predictive analytics, and credit risk management policy.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.

M345S16 SOLUTIONS

seen ⇓

1. (a) Plot $F_1(c)$ against $F_0(c)$ for all c. $\boxed{1}$

   (b) Classification (or discrimination). $\boxed{1}$

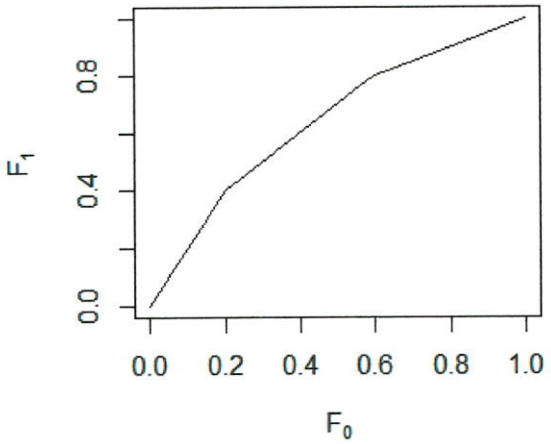
   (c) Point $(0, 1)$ since that is when $F_0(c) = 0$ and $F_1(c) = 1$; ie no Type 1 or Type 2 errors. $\boxed{1}$

   (d) Model M1 is best because it dominates the others across the whole range of the ROC curve. $\boxed{2}$

   (e) The diagonal expresses $F_1(c) = F_0(c)$ for all c, which is the case when $S \leq c$ is independent of Y; ie the score is unrelated to outcome. $\boxed{2}$

part seen ⇓

   (f) ROC curve is:



   Sum of three trapezoids: $0.4/2 \times 0.2 + (0.8 + 0.4)/2 \times 0.4 + (1 + 0.8)/2 \times 0.4 = 0.04 + 0.24 + 0.36 = 0.64$. $\boxed{3+2}$

   (g) Maximum difference is $K = 0.2$. $\boxed{1}$

unseen ⇓

(h)

$$\begin{aligned}\int_a^b F_1(c)f_0(c)\mathrm{d}c &= \int_a^b (F_1(c)-F_0(c))f_0(c)\mathrm{d}c + \int_a^b F_0(c)f_0(c)\mathrm{d}c \\ &\geq \tfrac{1}{2}[F_0(c)^2]_a^b \qquad \text{since} F_1(c) \geq F_0(c) \\ &= F_0(b)^2 - F_0(a)^2\end{aligned}$$

2

(i) Firstly the lower bound:-
There exists a c^* such that $K = F_1(c^*) - F_0(c^*))$.
Now consider segments of the ROC curve as follows, where $\lambda = F_0^{-1}(K + F_0(c^*))$ using result from part (i) and knowing that F_1 is an increasing function:-

$$\begin{aligned}\int_{-\infty}^{c^*} F_1(c)f_0(c)\mathrm{d}c &\geq \tfrac{1}{2}F_0(c^*)^2, \\ \int_{c^*}^{\lambda} F_1(c)f_0(c)\mathrm{d}c &\geq \int_{c^*}^{\lambda} F_1(c^*)f_0(c)\mathrm{d}c \\ &= F_1(c^*)[F_0(\lambda) - F_0(c^*)] \\ &= F_1(c^*)[K + F_0(c^*) - F_0(c^*)] \\ &= KF_1(c^*), \\ \int_{\lambda}^{\infty} F_1(c)f_0(c)\mathrm{d}c &\geq \tfrac{1}{2}[1 - F_0(\lambda)^2] \\ &= \tfrac{1}{2}[1 - (K + F_0(c^*))^2] \\ &= \tfrac{1}{2}[1 - K^2 - 2KF_0(c^*) - F_0(c^*)^2].\end{aligned}$$

Summing the terms to get AUC,

$$\begin{aligned}A &\geq \tfrac{1}{2}\left[F_0(c^*)^2 + 2KF_1(c^*) + 1 - K^2 - 2KF_0(c^*) - F_0(c^*)^2\right] \\ &= \tfrac{1}{2}\left[1 + K^2\right].\end{aligned}$$

Secondly the upper bound:-
Since K is the maximum, $F_1(c) - F_0(c) \leq K$ for all $c \in \mathbb{R}$
$\Rightarrow F_1(c) \leq K + F_0(c)$ and $F_1(c) < 1$ for $c = \gamma = F_0^{-1}(1-K)$
Now consider segments of the ROC curve as follows:-

$$\begin{aligned}\int_{-\infty}^{\gamma} F_1(c)f_0(c)\mathrm{d}c &\leq \int_{-\infty}^{\gamma} Kf_0(c) + F_0(c)f_0(c)\mathrm{d}c \\ &= K[F_0(c)]_{-\infty}^{\gamma} + \tfrac{1}{2}[F_0(c)^2]_{-\infty}^{\gamma} \\ &= K(1-K) + \tfrac{1}{2}(1-K)^2 \\ &= \tfrac{1}{2}[2K - 2K^2 + 1 - 2K + K^2] \\ &= \tfrac{1}{2}[1 - K^2], \\ \int_{\gamma}^{\infty} F_1(c)f_0(c)\mathrm{d}c &\leq \int_{\gamma}^{\infty} f_0(c)\mathrm{d}c \\ &= 1 - F_0(\gamma) = K\end{aligned}$$

Summing the terms to get AUC, $A \leq \frac{1}{2}[1 + 2K - K^2]$.

2+3

Note: for parts (i) and (j), marks are for algebraic solutions. A geometric solution is quicker (and will motivate the algebraic solution). However if only a geometric solution is given then this will yield a maximum of 2 marks.

2. (a) Objective function $\boxed{\text{seen} \Downarrow}$

$$\begin{aligned} f(\beta_1) &= \tfrac{1}{2}\textstyle\sum_{i=1}^{n}(y_i - \beta_1 x_{i1})^2 + \lambda||\beta_1||_1 \\ &= \tfrac{1}{2}\textstyle\sum_{i=1}^{n} y_i^2 - \beta_1 \sum_{i=1}^{n} x_{i1}y_i + \tfrac{1}{2}\beta_1^2 \sum_{i=1}^{n} x_{i1}^2 + \lambda||\beta_1||_1 \\ &= \tfrac{1}{2}\textstyle\sum_{i=1}^{n} y_i^2 - \beta_1\widehat{\beta}_{\mathrm{OLS}} + \tfrac{1}{2}\beta_1^2 + \lambda||\beta_1||_1 \end{aligned}$$

since x_{ij} are standardized. Therefore when $\beta_1 \neq 0$,

$$\frac{\mathrm{d}f(\beta_1)}{\mathrm{d}\beta_1} = -\widehat{\beta}_{\mathrm{OLS}} + \beta_1 + \lambda\,\mathrm{sign}(\beta_1).$$

$\boxed{3}$

(b) For $\beta_1 \neq 0$, setting the derivative to zero to find the stationary point, $\beta_1 = \widehat{\beta}_{\mathrm{OLS}} - \lambda\,\mathrm{sign}(\beta_1)$.

* Consider $\widehat{\beta}_{\mathrm{OLS}} > \lambda$, then $\beta_1 > 0$, hence $\beta_1 = \widehat{\beta}_{\mathrm{OLS}} - \lambda$.
* Consider $\widehat{\beta}_{\mathrm{OLS}} < -\lambda$, then $\beta_1 < 0$, hence $\beta_1 = \widehat{\beta}_{\mathrm{OLS}} + \lambda$.
* Consider $-\lambda \leq \widehat{\beta}_{\mathrm{OLS}} \leq \lambda$ and $\beta_1 < 0$, then $\beta_1 = \widehat{\beta}_{\mathrm{OLS}} + \lambda \geq 0$ which is a contradiction.
* Consider $-\lambda \leq \widehat{\beta}_{\mathrm{OLS}} \leq \lambda$ and $\beta_1 > 0$, then $\beta_1 = \widehat{\beta}_{\mathrm{OLS}} - \lambda \leq 0$ which is also a contradiction.

Since $-\lambda \leq \widehat{\beta}_{\mathrm{OLS}} \leq \lambda$ leads to no solution, it must be that $\beta_1 = 0$. This is demonstrated by observing that in this case, $\frac{\mathrm{d}f(\beta_1)}{\mathrm{d}\beta} < 0$ for $\beta_1 < 0$ and $\frac{\mathrm{d}f(\beta_1)}{\mathrm{d}\beta_1} > 0$ for $\beta_1 > 0$ (ie f is decreasing when β_1 increasing to 0, then increasing again after 0).

$\boxed{4}$

(c)
* For multivariate optimization problems, say $\min f(\beta)$, solve by iteratively moving in the direction of the optima one dimension/coordinate at a time.
* Start with an initial proposal estimate $\beta_1^{[0]}$ and set $t = 1$.
* For $j = 1$ to m,
$$\beta_j^{[t]} = \arg\min_{\beta_j} f(\beta_j; \beta_1^{[t-1]}).$$
* Repeat for $t \leftarrow t + 1$ until convergence.

$\boxed{3}$

(d) For multivariate OLS regression, write

$$f(\beta_1) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{k\neq j} x_{ik}\beta_k - x_{ij}\beta_j\right)^2 + \lambda\sum_{k\neq j}|\beta_k| + \lambda|\beta_j|.$$

Then for one iteration of the pathwise coordinate descent for j, when $\beta_j \neq 0$,

$$\frac{\mathrm{d}f(\beta_1)}{\mathrm{d}\beta_j} = -\sum_{i=1}^{n} x_{ij}\left(y_i - \sum_{k\neq j} x_{ik}\beta_k - x_{ij}\beta_j\right) + \lambda\mathrm{sign}(\beta_j).$$

Set this to zero to find stationary point and update to β_j estimation ($\beta_j^{[t]}$) by rearranging terms:

$$\beta_j^{[t]} = \sum_{i=1}^{n} x_{ij}\left(y_i - \sum_{k\neq j} x_{ik}\beta_k\right) - \lambda\mathrm{sign}(\beta_j)$$

since x_{ij} is standardized. This has the same form as part (b) and so the same argument can be applied to get the update function

$$\beta_j^{[t]} = S\left(\sum_{i=1}^{n} x_{ij}\left(y_i - \sum_{k\neq j} x_{ik}\beta_k^{[t-1]}\right), \lambda\right)$$

or, equivalently it can be written as

$$\beta_j^{[t]} = S\left(\beta_j^{[t]} + \sum_{i=1}^{n} x_{ij}\left(y_i - \beta_1^{[t-1]} \cdot x_{i1}\right), \lambda\right)$$

$\boxed{3}$

$\boxed{\text{unseen} \Downarrow}$

(e) Variables Home Owner and Months in current residence have been deselected. $\boxed{1}$

(f) Increase in income is associated with more credit card usage. Increase in outstanding debt is associated with less credit card usage. $\boxed{1}$

(g) Choose $\lambda = 0.5$ since this maximizes KS on the independent test data set.

$\boxed{1}$

(h) Write $g(\beta_1)$ for $\frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_1 \cdot x_{i1})^2$.
Since $\widehat{\beta}_{1i}$ minimizes the objective function for each λ_i,

$$g(\widehat{\beta}_{11}) + \lambda_1||\widehat{\beta}_{11}||_1 \leq g(\widehat{\beta}_{12}) + \lambda_1||\widehat{\beta}_{12}||_1$$

and

$$g(\widehat{\beta}_{12}) + \lambda_2||\widehat{\beta}_{12}||_1 \leq g(\widehat{\beta}_{11}) + \lambda_2||\widehat{\beta}_{11}||_1.$$

Substituting $g(\widehat{\beta}_{12})$ from the second equation into the first,

$$g(\widehat{\beta}_{11}) + \lambda_1||\widehat{\beta}_{11}||_1 \leq g(\widehat{\beta}_{11}) + \lambda_2||\widehat{\beta}_{11}||_1 - \lambda_2||\widehat{\beta}_{12}||_1 + \lambda_1||\widehat{\beta}_{12}||_1.$$

Then $g(\widehat{\beta}_{11})$ cancels out and rearranging terms,

$$(\lambda_2 - \lambda_1)(||\widehat{\beta}_{11}||_1 - ||\widehat{\beta}_{12}||_1) \geq 0.$$

Since $\lambda_2 > \lambda_1$, it follows that $||\widehat{\beta}_{11}||_1 \geq ||\widehat{\beta}_{12}||_1$ as required.

$\boxed{4}$

3. (a)

- ∗ Population that model is planned to be applied to is all applicants. $\boxed{\text{seen} \Downarrow}$
- ∗ However, only historic applicants who were previously accepted will have an outcome.
- ∗ Therefore the model can only be built on historically *accepted* loans.
- ∗ This can be thought of as a non-ignorable missing data problem on the response variable. $\boxed{4}$

(b;i)

- ∗ We expect the probability of default to be greater for rejects than accepts, controlling for the known predictor variables.
- ∗ This is reasonable since we would expect if an observation was rejected there must be an additional reason to suppose the individual is high risk, compared to if it is accepted, that is not known.

$\boxed{2}$

(b;ii)

$$\begin{aligned}
P(Y=1|\mathbf{X}=\mathbf{x}) &= P(Y=1,A|\mathbf{X}=\mathbf{x})+P(Y=1,\bar{A}|\mathbf{X}=\mathbf{x})\\
&= P(Y=1|A,\mathbf{X}=\mathbf{x})P(A|\mathbf{X}=\mathbf{x})+P(Y=1|\bar{A},\mathbf{X}=\mathbf{x})P(\bar{A}|\mathbf{X}=\mathbf{x})\\
&> P(Y=1|A,\mathbf{X}=\mathbf{x})P(A|\mathbf{X}=\mathbf{x})+P(Y=1|A,\mathbf{X}=\mathbf{x})P(\bar{A}|\mathbf{X}=\mathbf{x})\\
&> P(Y=1|A,\mathbf{X}=\mathbf{x})\left[P(A|\mathbf{X}=\mathbf{x})+P(\bar{A}|\mathbf{X}=\mathbf{x})\right]\\
&> P(Y=1|A,\mathbf{X}=\mathbf{x})
\end{aligned}$$

$\boxed{2}$

(c)

- ∗ Model C is using experimentation on loan applicants to gather data. $\boxed{\text{part seen} \Downarrow}$
- ∗ Some "rejects" are allowed through to enable estimation of $P(Y=1|\bar{A},\mathbf{X}=\mathbf{x})$.

$\boxed{2}$

(d) The reweighting ensures that the few "reject" overrides that are accepted are weighted to represent the total number of rejects (ie 250 × 10 = 2500). This is using an Augmentation approach to reweight to approximate the population distribution (ie all rejects and accepts). $\boxed{\text{unseen} \Downarrow}$

$\boxed{2}$

(e) Reweighting can be implemented in the context of any maximum likelihood estimator. If the standard log-likelihood is given for n observations in the form $\boxed{\text{part seen} \Downarrow}$

$$\sum_{i=1}^{n} L(\theta;\mathbf{x}_i)$$

then the log-likelihood adjusted by weights w_i on each observation is given by

$$\sum_{i=1}^{n} w_i L(\theta;\mathbf{x}_i).$$

$\boxed{2}$

unseen ⇓

(f;i) Odds ratios (OR) given by $\exp(k\beta)$ for unit change k, hence for $k = 1$, OR for X_{10} is 1.049 and 1.052 for models B and C respectively, and OR for X_{11} is 1.20 and 1.99 for models B and C respectively.

2

(f;ii)

- $*$ There is very little change in OR for annual income between the two models so little evidence of a selection bias effect on that variable.
- $*$ The change in OR is substantial for value of outstanding debt and the estimates suggest that model B is under-estimating the size of association of that variable due to selection bis.

Note that the standard errors are important since they demonstrate that the coefficient estimates are efficient, and hence the point estimates shown are accurate, but no marks for this observation.

2

seen ⇓

(g) The regulator would be concerned that as part of the experiment some high risk applicants will be given loans which could constitute irresponsible lending.

2

4. *Note that students have been given explicit instructions that there will be no questions about the H-measure.*

unseen ⇓

(a) ∗ Existing metrics such as AUC and KS are limited because they make unrealistic assumptions about misclassification costs.

∗ This article is motivated to improve on these existing metrics by developing a novel profit metric for use with credit risk models.

∗ Also, the intention is to demonstrate the new method in application specifically to the customer churn problem.

For the second point, the student does not have to specifically identify the new metrics as profit based.

$\boxed{3}$

(b) Customer churn is the phenomena of existing customers leaving the company; eg someone transferring a mortgage to another lender.

$\boxed{1}$

(c) Equation (4) gives the condition for maximizing MP with respect to t. Substituting the normal PDFs for f_0 and f_1 then gives

$$\frac{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(T-\mu_0)^2}{2\sigma^2}\right)}{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(T-\mu_1)^2}{2\sigma^2}\right)} = \frac{\pi_1\theta}{\pi_0}$$

$$\Rightarrow \exp\left(\frac{-(T-\mu_0)^2+(T-\mu_1)^2}{2\sigma^2}\right) = \frac{\pi_1\theta}{\pi_0}$$

$$\Rightarrow -T^2+2T\mu_0-\mu_0^2+T^2-2T\mu_1+\mu_1^2 = 2\sigma^2\log\left(\frac{\pi_1\theta}{\pi_0}\right)$$

$$\Rightarrow 2T(\mu_0-\mu_1)-\mu_0^2+\mu_1^2 = 2\sigma^2\log\left(\frac{\pi_1\theta}{\pi_0}\right)$$

$$\Rightarrow T = \frac{\sigma^2\log\left(\frac{\pi_1\theta}{\pi_0}\right)}{\mu_0-\mu_1}+\frac{1}{2}(\mu_0+\mu_1)$$

$\boxed{4}$

(d) Substitute 0 for A, $\eta(t)$ for η, $\lambda(t)$ for λ, δCLV for d and ϕCLV for f in Equation (13) following the remarks in the article:

$$\text{Profit} = N(\pi_0F_0(t)+\pi_1F_1(t))\left[(\gamma CLV+\delta CLV(1-\gamma))\frac{\pi_0F_0(t)}{\pi_0F_0(t)+\pi_1F_1(t)}-\delta CLV-\phi CLV\right].$$

Divide through by N to get average profit and re-arrange terms:

$$\begin{aligned} P_C &= CLV\left[(\gamma+\delta(1-\gamma))\pi_0F_0(t)-(\delta+\phi)(\pi_0F_0(t)+\pi_1F_1(t))\right] \\ &= CLV\left[\gamma(1-\delta)\pi_0F_0(t)-\phi(\pi_0F_0(t)+\pi_1F_1(t))-\delta\pi_1F_1(t)\right] \\ &= CLV\left[\gamma(1-\delta)-\phi\right]\pi_0F_0(t)-CLV\left[\phi+\delta\right]\pi_1F_1(t) \end{aligned}$$

$\boxed{3}$

(e) ∗ Equation (5) is an integral over b_0, c_0, b_1 and c_1 based on Equation (1). Equation (14) expresses Equation (1) with a change of variables to CLV, γ, δ and ϕ.

∗ In Section 4.2, the authors argue that four parameters (including T) can be estimated with reliability and only γ remains uncertain. Hence when the expected profit is given in Equation (18), this only needs to be integrated over the uncertain variable γ. $\boxed{2}$

(f) ∗ Firstly, because $\gamma \in [0, 1]$

∗ Secondly, the beta distribution has flexibility in expressing densities in that range. $\boxed{2}$

(g) For several data sets, models are selected on the basis of maximal EMPC or AUC. $\boxed{2}$

(h) ∗ For all data sets (except one), the selection method based on EMPC provides a performance gain over the selection method based on AUC (the exception has the same performance).

∗ This shows that if average profit is the performance we are interested in, then the EMPC gives a performance gain and using AUC, the industry standard, under-performs.

∗ This is only a test against AUC, so is limited. It says more about the deficiencies of AUC than the benefit of EMPC. $\boxed{3}$