PROFIT MAXIMIZING LOGISTIC REGRESSION MODELING FOR CREDIT SCORING

Arnout Devos[1,2], Jakob Dhondt[3], Eugen Stripling[2], Bart Baesens[2,4], Seppe vanden Broucke[2], Gaurav Sukhatme[1]

[1]Department of Computer Science, University of Southern California, Los Angeles, USA
[2]Department of Decision Sciences and Information Management, KU Leuven, Leuven, Belgium
[3]Switch, Zurich, Switzerland
[4]School of Management, University of Southampton, Southampton, UK

ABSTRACT

Multiple classification techniques have been employed for different business applications. In the particular case of credit scoring, a classifier which maximizes the total profit is preferable. The recently proposed expected maximum profit (EMP) measure for credit scoring allows to select the most profitable classifier. Taking the idea of the EMP one step further, it is desirable to integrate the measure into model construction, and thus obtain a profit maximizing model. Therefore, in this work we propose a method based on the ProfLogit classifier, which optimizes the coefficients of a logistic regression model using a genetic algorithm. The proposed implemented technique shows a significant improvement compared to regular maximum likelihood based logistic regression models on real-life data sets in terms of total profit, which is the ultimate goal for most businesses.

Index Terms— Credit, Profit, Logistic, EMP, Genetic

1. INTRODUCTION

Credit scoring is an application in statistical modeling that is concerned with classifying applicants for credit into *good* and *bad* (default) risk classes [1]. The main goal of such models is to estimate the probability of default, i.e. when a customer does not pay back a loan in a given period. Each customer gets assigned a score which, depending on the decided cutoff value, will result in either a loan being granted or rejected.

Predictive classification techniques for credit scoring are increasingly researched [2]. However, most models developed do not directly focus on the most important business requirement: *profit maximization*. Often the best credit scoring model is selected based on *accuracy* related performance measures, which do not strive for profit maximization directly. The expected maximum profit measure (EMP) for credit scoring does exactly this, and allows to select the most profitable model [3]. Although the EMP for credit scoring allows for a profit-based model evaluation, the profit-related insights are not directly integrated into model construction.

Therefore, based on the ProfLogit classifier of [4], we propose a profit maximizing classifier, which optimizes the EMP for credit scoring when the model is constructed, rather than only evaluating classifier performance after construction. Conforming with Basel II/III regulations regarding credit scoring and transparency, the ProfLogit classifier is a good choice since it employs a logistic regression model structure to compute credit scores, which are required for the profit measure, but the regression coefficients are optimized with regard to the EMP employing a genetic algorithm (GA), contrary to regular maximum likelihood models.

The remainder of this paper is structured as follows: Section II discusses in-depth how credit scoring can be viewed as a classification problem and how the classification performance of a model can be evaluated. Section III explains how we have implemented the maximum profit classifier. In Section IV an empirical study is done with multiple data sets, verifying the classifier's performance. The research findings and possible research outlooks are summarized in Section V.

2. CREDIT SCORING CLASSIFICATION AND EVALUATION

2.1. Credit Scoring as a Binary Classification Problem

Credit scoring problems are usually defined as binary classification problems where the goal is to assign instances, i.e. loan applicants, to one of two classes $Y = \{0, 1\} = \{\text{default, no default}\}$. These assignments are done by the model based on p descriptive features $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ associated with each instance i, for example loan amount or credit history. In binary classification, instances are mapped to a score s that is often transformed to fit within the interval $[0, 1]$, enabling them to be interpreted as probabilities to belong to either class. By comparing the score s of an instance with a classification threshold t the predicted class can be found. All instances with $s < t$ are classified as *defaulters* (class 0), whereas instances for which $s \geq t$ are classified as

Correspondence address: devos@usc.edu. A. Devos was supported by a Belgian American Educational Foundation Graduate Fellowship.

Table 1. Confusion matrix with associated costs and benefits. Class 0 represents the defaulters (bad loans), whereas Class 1 represents the non-defaulters (good loans).

True label	Predicted	
	Class 0	Class 1
Class 0	$\pi_0 F_0(t)$	$\pi_0(1 - F_0(t))$
	$[c(0 \mid 0) = b_0]$	$[c(1 \mid 0) = 0]$
Class 1	$\pi_1 F_1(t)$	$\pi_1(1 - F_1(t))$
	$[c(0 \mid 1) = c_1]$	$[c(1 \mid 1) = 0]$

non-defaulters (class 1). Note that multiple conventions have been used, such as to assign class 1 for defaulters (contrary to this paper which assigns class 0 to defaulters). We opted for this convention, since it offers the advantage that it simplifies notation and is also used by [5], among others.

Based on the credit scores produced by a classifier, and given a threshold t, the confusion matrix can be constructed (Table 1). In Table 1, π_k denotes the prior class probability of class k with $k \in \{0, 1\}$, $f_k(s)$ and $F_k(s)$ represent the probability density function and the cumulative probability density function respectively of class k calculated based on the default scores s.

In order to assign an instance to either one of the two classes, all instances with $s < t$ are classified as defaulters (class 0), whereas instances with $s \geq t$ are classified as non-defaulters (class 1). Each element of the confusion matrix has a cost or benefit $c(i \mid j)$ associated with classifying an instance of class j into class i, with $i, j \in \{0, 1\}$. It is advised to measure these costs and benefits against a base scenario, the situation where no classification is done at all (for credit scoring: granting all loans) [6]. Comparing to a base scenario ensures consistency when evaluating different credit scoring models. Starting from this base scenario, where everyone is predicted to be in class 1, also has an important implication on the costs and benefits: only costs and benefits corresponding to predicted Class 0 (default) are relevant. Therefore, $c(1 \mid 0) = c(1 \mid 1) = 0$, and we define $c(0 \mid 0) = b_0 \geq 0$ and $c(0 \mid 1) = c_1 \geq 0$ to be the benefit and cost of correctly or wrongly classifying an instance to class 0 respectively. The total profit of the model is calculated by subtracting all the costs from all the benefits.

2.2. Binary Classification Performance Evaluation

A classification performance measure is necessary to determine the quality of the credit scoring model for its purpose. In this subsection popular performance measures for binary classification problems are discussed.

In order to compare classifiers, several performance evaluation methods exist. The receiver operating characteristic (ROC) is a popular graphical performance evaluation method which often serves as a basis for other performance metrics [7]. However, in order to make a hard decision whether one classifier is better than another, classification performance metrics that can be represented by a single number are preferred. Examples of such metrics are *accuracy* ($\pi_0 F_0(s) + \pi_1(1 - F_1(s))$) and *error rate* ($\pi_0(1 - F_0(s)) + \pi_1 F_1(s)$). A popular metric closely related to the ROC is *area under the ROC curve* (AUC), which takes the entire range of possible cutoff values into account [7]. It is important to note that these metrics do not take into account any specific misclassification costs, and therefore can only be used when those costs are equal.

2.3. Profit-based Classification Performance Evaluation

In contrast to most classification problems, credit scoring problems are usually highly imbalanced, since often there are a lot less defaulters than non-defaulters [2]. This renders the assumption of equal misclassification costs invalid. A big drawback of the AUC, despite being so popular, is that it implicitly treats the relative severities of misclassification differently between different classifiers [5]. Therefore [5] proposes the H measure that fixes the distribution of relative severities and explicitly accounts for misclassification costs.

However, it is recommended to incorporate both costs and benefits into a performance measure [8]. Correspondingly, a cost benefit analysis framework for credit scoring, which incorporates the costs associated with classifying good loan applicants as defaulters and the benefit of early enough detection of true defaulters, was proposed in [3]. More specifically, the benefit of correctly identifying a defaulter is equal to the fraction of the loan amount which would be lost after default:

$$b_0 = \frac{LGD.EAD}{A} = \lambda, \tag{1}$$

with $\lambda \in [0, 1]$, A the amount still owned on a loan, LGD the Loss Given Default, and EAD the Exposure At Default.

The cost c_1, associated with wrongly classifying a good loan applicant as a defaulter, is equal to the return on investment (ROI) of the loan. The ROI in credit scoring applications can be assumed to be constant [3]. It is also assumed that there is no cost for the action of rejecting a customer. Since most credit scoring models are applied to massive data sets and the costs associated with building the model itself are not related to particular individuals (i.e. variable costs), the cost of building the model can be assumed to be marginal - and thus can be omitted. The parameter λ, which represents the recovery rate, can vary between 0% and 100% of the total loan amount, and several distributions can arise [9]. Most commonly, its cumulative distribution $H(\lambda)$ has three sections: the biggest part of the probability mass is situated around $\lambda = 0$ and a smaller one around $\lambda = 1$, whereas the

rest is spread out almost evenly between zero and one. Therefore, the following assumptions can be made regarding λ: (1) $\lambda = 0$ (customers pay back everything) has a probability of p_0, (2) $\lambda = 1$ (customers pay nothing back) has a probability of p_1, and (3) λ follows a uniform distribution between zero and one, i.e. $h(\lambda) = 1 - p_0 - p_1$ for $\lambda \in]0, 1[$ [3].

Empirical ROC curves are mostly stepwise constant. The EMP measure for credit scoring can be calculated based on the convex hull of the ROC curve that is built up out of m segments with end points (r_{1i}, r_{0i}) with $i = 1, \ldots, m$ and $(r_{10}, r_{00}) := (0, 0)$[3]:

$$\begin{aligned} EMP =& (1 - p_0 - p_1) \sum_{i=0}^{k} \Big[\frac{\pi_0 r_{0i}}{2} (\lambda_{i+1}^2 - \lambda_i^2) - \\ & ROI \pi_1 r_{1i} (\lambda_{i+1} - \lambda_i) \Big] \\ & + \left[\pi_0 r_{0(k+1)} p_1 - ROI \pi_1 r_{1(k+1)} p_1 \right] \end{aligned} \tag{2}$$

where $\lambda_0 = 0$ and

$$\lambda_{i+1} = \frac{\pi_1 (r_{1(i+1)} - r_{1i})}{\pi_0 (r_{0(i+1)} - r_{0i})} ROI \quad (i = 0, \ldots, m-1). \tag{3}$$

3. PROFIT-MAXIMIZING LOGISTIC REGRESSION

In this section, we propose a technique for credit scoring applications, based on the ProfLogit classifier of [4], which employs a standard logistic regression model structure but its parameters are optimized to maximize the EMP for credit scoring. Logistic regression is one of the most commonly used methods for credit scoring [10]. Besides that, it also conforms Basel II/III regulations which require transparency in the loan granting process, unlike more black-box nature models [11]. In the next subsections we explain the implementation of our classifier. The classifier is implemented using the R programming language, exploiting the features available in the `EMP` package [12] and the `GA` package [13].

3.1. Logistic Regression Classification

The logistic regression model is an attractive choice for binary classification problems, and credit scoring specifically, because it offers high interpretability, is easy to use, and performs well in many settings [2]. The logistic regression model calculates, for each instance i, a probability estimate $s_i \in [0, 1]$ based on the vector of descriptive features $\mathbf{x}_i$. With regard to credit scoring, the probability of an instance i, i.e. loan applicant, with features $\mathbf{x}_i$ being a defaulter (class 0) is determined by the conditional probability:

$$P(Y = 0 \mid \mathbf{x}_i) = \frac{e^{\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i}}{1 + e^{\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i}} \tag{4}$$

where $\beta_0 \in \mathbb{R}$ is the intercept, and $\boldsymbol{\beta} \in \mathbb{R}^p$ is the p-dimensional vector of regression coefficients. Note that, following our convention, a lower score indicates a higher likelihood of defaulting. The regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ and intercept β_0 optimize an objective function, which is achieved through numerical optimization. In the regular logistic regression model, this is achieved by using a *maximum likelihood estimation* algorithm in which the objective function that needs to maximized is the likelihood function. However, as we favor *total profit maximization* over likelihood, our classifier optimizes for the EMP measure for credit scoring using a genetic algorithm (GA).

3.2. Fitness function

In the GA, the parameter vector $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}) \in \mathbb{R}^{p+1}$ represents a chromosome in which the regression coefficients and intercept are the genes, and equation (2) acts as the objective function to be maximized. $\boldsymbol{\theta}$ completely defines a logistic regression classification model that can be evaluated to extract credit scores, which can directly serve as an input for the EMP measure to compute the classification profit. The EMP measure for credit scoring requires a constant ROI parameter and the parameters p_0 and p_1 which are part of a bimodal LGD function with point masses p_0 and p_1 representing no loss and total loss , respectively. These parameters vary depending on the used data set.

3.3. Related work

A performance measure based on the same cost benefit framework exists for customer churn models (EMPC), where classification costs and benefits are based on the cost of offer and the customer lifetime value of retained customers [6]. Similar to this work, the EMPC has also been used in the model creation step of the ProfLogit classifier for customer churn which shows to outperform regular logistic regression in terms of total profit [4]. However, contrary to the transparency requirement that logistic regression matches for credit scoring, customer churn models may incorporate more advanced machine learning techniques such as neural networks and support vector machines. Those models can easily surpass the profit performance of the logistic based model [4].

4. EMPIRICAL EVALUATION

4.1. Home equity loan and credit data sets

In our experiments, we use two real-life data sets (see Table 2) that are often used in credit scoring applications [14].

The target is a binary variable `GOOD` that indicates whether an applicant has defaulted (`GOOD` $= 0$) or not (`GOOD` $= 1$).

The first one is the HMEQ data set which consists of 5,960 home equity loans. After removing instances with missing values, 4,408 loans are creditworthy, while 1,128 of them are

Table 2. Real-life Credit Scoring data sets [14]

ID	Vars	# Observations		Default rate [%]	
		Train	Test	Train	Test
HMEQ	12	4,428	1,108	20.37	20.40
GERM	11	800	200	30.00	30.00

not. For each loan there are 12 input variables (10 continuous and 2 nominal) which all are statistically significant.

The second data set GERM consists of 1,000 German loan applications where 700 of them are creditworthy. For each loan application there are 20 input variables (7 numerical, 13 categorical) of which 11 are statistically significant.

4.2. Empirical Setup

The model proposed in Section III is compared to a typical maximum likelihood optimization model regarding its performance in terms of total profit with real data sets. Each data set is randomly divided into an 80% training and 20% test set (see Table 2). The division between training and test set is stratified with respect to the target variable, i.e. the default variable, such that the default distributions are similar.

In terms of preprocessing the data, continuous variables are standardized as follows: each variable is subtracted by its mean and divided by its standard deviation and instances with missing values are omitted from the data set.

Regarding the parameters of the genetic algorithm (GA), several choices need to be made. The population size is set equal to ten times the parameter vector length. Making the population size linearly dependent on the input dimension can be explained by the fact that the size of the search space grows exponentially with respect to the input dimension [4]. The search boundaries for the coefficients are set to -6 and 6 in the GA, since standardization has taken place on the input variables already. The GA is terminated when either: (1) the number of generations has reached 1000, or (2) the best current fitness value has seen no change for the last 100 generations [4].

4.3. Results of the Experiment

In terms of total profit, the proposed technique is overall the most profitable credit scoring model (see Tables 3 and 4). The experimental results for the regular logistic regression model (GLM) and the proposed genetic algorithm EMP technique (GA EMP) are compared based on the AUC, total profit, and extra profit on top of the baseline model where all loans are granted (see Tables 3 and 4). For both data sets, the advantage of the GA EMP technique is clearly visible. On the HMEQ data set, 35.1% of extra profit is achieved compared to 31.2% of extra profit for the GLM, both compared to the base case of granting all loans. Due to its high default rate, the GERM data set has a negative total profit (loss) when all loans are granted, and here the GA EMP technique doubles the positive profit compared to the GLM. Although the GLM model provides a higher AUC in both cases, the GA EMP model is always more profitable, and should therefore be preferred in business situations where profit maximization is an important objective.

Table 3. Profit and ROC results on the HMEQ data set

Model	AUC	Total Profit ($)	Extra Profit ($)
no model	0.5	1,851,022	0
GLM	**0.8077**	2,427,683	576,661
GA EMP	0.8068	**2,501,595**	**650,573**

Table 4. Profit and ROC results on the GERM data set

Model	AUC	Total Profit ($)	Extra Profit ($)
no model	0.5	-53,015	0
GLM	**0.7819**	6,883	59,898
GA EMP	0.7705	**13,890**	**66,905**

5. CONCLUSIONS AND FUTURE WORK

In this work we proposed a profit-maximizing logistic regression modelling technique for credit scoring applications, based on the ProfLogit classifier. Contrary to regular logistic regression, which is designed to maximize likelihood, the proposed technique maximizes the total profit of a credit scoring problem by optimizing for the Expected Maximum Profit (EMP) measure. Additionally, the proposed classifier explicitly takes costs and benefits into account, unlike pure cost-sensitive learners. Evaluation on two credit risk data sets has shown a significant profit improvement of the proposed algorithm over regular logistic regression. In conclusion, the proposed classification algorithm aligns best with the most important business requirement in a credit scoring setting: profit maximization. Concerning future research, it would be interesting to compare the proposed technique against other models (e.g. tree based classification) in terms of profit and common performance measures such as AUC, H measure and others.

6. REFERENCES

[1] David J Hand and William E Henley, "Statistical classification methods in consumer credit scoring: a review,"

Journal of the Royal Statistical Society: Series A (Statistics in Society), vol. 160, no. 3, pp. 523–541, 1997.

[2] Bart Baesens, Tony Van Gestel, Stijn Viaene, Maria Stepanova, Johan Suykens, and Jan Vanthienen, "Benchmarking state-of-the-art classification algorithms for credit scoring," *Journal of the operational research society*, vol. 54, no. 6, pp. 627–635, 2003.

[3] Thomas Verbraken, Cristián Bravo, Richard Weber, and Bart Baesens, "Development and application of consumer credit scoring models using profit-based classification measures," *European Journal of Operational Research*, vol. 238, no. 2, pp. 505–513, 2014.

[4] Eugen Stripling, Seppe vanden Broucke, Katrien Antonio, Bart Baesens, and Monique Snoeck, "Profit maximizing logistic model for customer churn prediction using genetic algorithms," *Swarm and Evolutionary Computation*, 2017.

[5] David J Hand, "Measuring classifier performance: a coherent alternative to the area under the roc curve," *Machine learning*, vol. 77, no. 1, pp. 103–123, 2009.

[6] T. Verbraken, W. Verbeke, and B. Baesens, "A novel profit maximizing metric for measuring classification performance of customer churn prediction models," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 5, pp. 961–973, May 2013.

[7] Tom Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[8] Charles Elkan, "The foundations of cost-sensitive learning," in *International joint conference on artificial intelligence*. Lawrence Erlbaum Associates Ltd, 2001, vol. 17, pp. 973–978.

[9] Mark Somers and Joe Whittaker, "Quantile regression for modelling distributions of profit and loss," *European Journal of Operational Research*, vol. 183, no. 3, pp. 1477–1487, 2007.

[10] Lyn C Thomas, David B Edelman, and Jonathan N Crook, *Credit scoring and its applications*, SIAM, 2002.

[11] "Basel iii: Finalising post-crisis reforms," `https://www.bis.org/bcbs/publ/d424.htm`, Accessed: 2018-01-15.

[12] Cristian Bravo, Seppe vanden Broucke, and Thomas Verbraken, *EMP: Expected Maximum Profit Classification Performance Measure*, 2017, R package version 2.0.2.

[13] Luca Scrucca et al., "Ga: a package for genetic algorithms in r," *Journal of Statistical Software*, vol. 53, no. 4, pp. 1–37, 2013.

[14] Bart Baesens, Daniel Roesch, and Harald Scheule, *Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS*, John Wiley & Sons, 2016.