GENERAL ARTICLE

Cost-sensitive Learning for Profit-driven Credit Scoring

George Petrides^{a,b}, Darie Moldovan^c, Lize Coenen^a, Tias Guns^a and Wouter Verbeke^{a,d}

^aVrije Universiteit Brussel (VUB), Faculty of Social Sciences and Solvay Business School, Data Analytics Laboratory, Pleinlaan 2, 1050 Brussels, Belgium

^bUniversity of Bergen, Department of Informatics, Thormøhlensgt. 55, 5008, Bergen, Norway

^cUniversitatea Babeş-Bolyai (UBB), Business Information Systems Department, Strada Teodor Mihali 58-60, 400591, Cluj-Napoca, Romania

^dKatholieke Universiteit Leuven, Faculty of Business and Economics, Naamsestraat 69, 3000 Leuven, Belgium

ARTICLE HISTORY

Compiled October 31, 2020

ABSTRACT

Application scorecards allow to assess the creditworthiness of loan applicants and decide on acceptance. The accuracy of scorecards is of crucial importance for minimizing bad debt loss and maximizing returns. In this paper, we extend upon prior benchmarking studies that experimentally compare the performance of classification techniques to discriminate between good and bad applications. We evaluate a range of cost-sensitive learning methods in terms of their ability to boost the profitability of scorecards. These methods allow to take into account the variable misclassification costs that are involved in rejecting good loan applications and accepting bad loan applications.

An approach is proposed to estimate these misclassification costs, and various approaches to handle missing credit bureau scores are evaluated. The results of a case study involving a Romanian nonbanking financial institution (NBFI) indicate that cost-sensitive learning complements the existing state-of-the-art scorecard of the NBFI. The best performing cost-sensitive models are found to increase profitability across the three business channels, with a single-digit improvement for two of the channels and a double-digit increase for the other one. The result is partly explained by the default rate, which is higher for this latter channel and therefore offers greater potential for improving profitability.

KEYWORDS

Credit scoring; Cost-sensitive learning; Application scorecards; Profit-driven analytics

1. Introduction

Credit scoring is concerned with the development of predictive models, i.e., credit scorecards, to support decision making in the retail credit business (J. N. Crook, Edelman, & Thomas, 2007; Lessmann, Baesens, Seow, & Thomas, 2015). Specifically,

CONTACT Wouter Verbeke. Email: wouter.verbeke@vub.be

ORCID ids: G. Petrides: 0000-0002-1125-4343, D. Moldovan: 0000-0001-7268-0453, W. Verbeke: 0000-0002-8438-0535

application credit scoring is concerned with distinguishing good from bad loan applicants (Thomas, Edelman, & Crook, 2002; Verbraken, Bravo, Weber, & Baesens, 2014). Given the importance of accurate credit risk assessments for financial institutions to effectively and efficiently manage risk and as such to optimize profitability, credit scoring has long attracted broad interest both from academia and industry.

The creditworthiness of loan applicants can be assessed by adopting either a judgmental or a statistical approach (Baesens, Roesch, & Scheule, 2016). The statistical or data-driven approach relies on predictive models, i.e., credit scorecards, that are developed by analyzing historical data of past applicants and their repayment behavior. The judgmental approach, on the other hand, relies on the expertise and ability of credit officers to accurately judge loan applications based on their experience and the available information on a new application. The information that is used by credit officers in assessing loan applications is often summarized by the five Cs of credit: capital, character, capacity, collateral and conditions. Whereas the first four Cs relate to the risk involved in a loan, conditions reflect the potential returns or profits for the financial institution. Risk and returns are typically balanced; the higher the risk involved in a loan, the higher the potential returns should be for a financial institution to take on the risk. Here, note that data-driven approaches adopted by financial institutions for assessing creditworthiness, e.g., a logistic regression-based scorecard, currently typically do not account for the profitability of a loan and merely focus on distinguishing good from bad applications.

Cost-sensitive classification methods have been developed to take into account the unequal misclassification costs involved in false positive and false negative errors by classification models, which are inherent to a variety of classification tasks. For instance, in fraud detection, it is much more costly to miss a fraudulent case than to wrongly classify a nonfraudulent case as fraudulent (Baesens, Van Vlasselaer, & Verbeke, 2015); in customer churn prediction, it is less costly to misclassify a nonchurner, involving a cost equal to the cost of the retention effort, than a churner, involving a cost proportional to the customer lifetime value (Verbeke, Dejaeger, Martens, Hur, & Baesens, 2012). Similarly, in credit scoring, incorrectly classified good and bad loan applicants incur different costs (Verbraken et al., 2014). A misclassified good loan applicant will be rejected and hence involves an opportunity cost, i.e., missed revenues, whereas a misclassified bad loan applicant will be accepted, potentially yielding a significant financial loss.

In this paper, we extend upon a comprehensive review and experimental benchmarking study of data-driven approaches for developing credit scorecards by Lessmann et al. (2015). We evaluate a broad selection of state-of-the-art cost-sensitive learning approaches for improving credit scorecard profitability in a case study involving a Romanian nonbanking financial institution (NBFI). An NBFI is usually a company carrying out financial services other than traditional banking, such as loans to low-income clients, leasing or private equity funds. The case presented here involves an NBFI that is oriented towards microfinance for low-income customers. The customers that usually apply for a loan with the NBFI have a low credit score or a credit history that is limited or absent. Such customers are often called *subprime* customers and typically rejected by traditional financial institutions when applying for credit. The loans that are offered are not backed by collateral, are usually for a short term (from a few months up to four years) and involve small amounts (as low as 200 euro). Therefore, credit risk is higher than for a traditional bank loan. As a result, the NBFI typically faces high default rates, which they compensate by charging higher interest rates than for loans extended to prime customers. Finally, note that the involved NBFI is not

subject to the Basel regulation, and the credit risk models that are operated by the NBFIs are not subject to review by a regulator. Hence, the NBFIs can in principle adopt advanced, nonlinear modeling approaches that in a regulatory setting are typically not allowed since the resulting credit risk models are considered black box models that cannot be interpreted and validated.

In our work, we adopt a profit-driven perspective towards developing credit scorecards. Hence, we extend research (Verbeke et al., 2012; Verbraken, Verbeke, & Baesens, 2013) that introduced a profit-driven perspective towards classification model performance evaluation. This perspective was customized to credit scorecard performance evaluation in (Verbraken et al., 2014), where the authors concluded the following: *”Currently, models typically optimize a statistical criterion, such as e.g. maximum likelihood. A focus shifted to profitability may provide further opportunities for improving credit scoring practices”*. In this paper, we effectively shift the focus towards minimizing misclassification costs, i.e., maximizing profitability, by adopting cost-sensitive learning approaches for developing application credit scorecards. We will largely extend a handful of studies in the literature that report on the use of some such approaches, for instance, Bahnsen, Aouada, and Ottersten (2015).

Our research objective is to investigate the use of cost-sensitive learning to improve the profitability of credit scorecards. A range of techniques (both cost-insensitive and cost-sensitive) are applied in conjunction with the state-of-the-art method that is currently employed in practice by the NBFIs and are evaluated with respect to their profitability.

The contributions of this paper are as follows:

- (1) To be able to both assess profitability and apply cost-sensitive learning methods, the costs and benefits associated with each type of prediction, both correct and incorrect and both for the good and bad class, are needed at the individual loan level. In practice, as will be discussed in this paper, these are not always known. Therefore, we introduce a simple yet effective and generic approach for estimating these parameters; this allows us to adopt a profit-driven perspective in learning and evaluating the performance of credit scorecards.
- (2) We provide empirical evidence resulting from a real-world case study, which indicates that cost-sensitive learning consistently yields increased profitability compared to cost-insensitive approaches. It is important to stress that a limited number of studies in the literature adopt a profit-driven perspective to credit scoring.
- (3) In elaborating the case study, we add to the body of knowledge and best practices in credit scoring by comparing various approaches to handle missing values of external credit risk scores, such as FICO or Equifax scores, that are commonly used in the credit application process in the consumer lending industry but that may be missing for a substantial number of loan applicants; additionally, a segmentation approach is evaluated. It is important to note that we also evaluate the added value of these approaches from a profit-driven perspective.
- (4) The dataset that is used in this paper has been made publicly available¹ with the aim of facilitating the reproduction of the presented experimental results and to spur further research on observation-dependent cost-sensitive learning and profit-driven analytics for credit risk modeling.
- (5) Two minor adjustments to the expected maximum profit measure (Verbeke et al., 2012; Verbraken et al., 2014, 2013) are introduced to facilitate an intuitive

¹<https://github.com/vub-dl>

interpretation and comparison across studies, as well as to ease computations; the maximum profit measure is expressed in this study in a relative manner as a percentage of the maximum profit that can be achieved when classification accuracy is 100%; additionally, the stochastic nature of parameters in the profit formula, i.e., the misclassification costs, is accounted for by relying on the representativity of the observation-dependent misclassification costs in the dataset.

The remainder of this paper is structured as follows. In Section 2, we review related work in the field of credit scoring in general and in the field of cost-sensitive learning in particular. Next, in Section 3, we provide details on the adopted methodology. Additionally, the case study is introduced, and details on the dataset are provided. Then, in Section 4, the results of the case study are presented and discussed. We conclude the paper and provide directions for future research in Section 5.

2. Related Work

2.1. Credit Scoring

Credit scoring is concerned with a classification problem, which is a supervised learning task where observations need to be assigned to one of a number of predefined classes based on a number of known characteristics, i.e., that need to be classified (Friedman, Hastie, & Tibshirani, 2009). The predefined classes in credit scoring are typically labelled good and bad, with bad meaning that a customer will default and hence most likely will not repay the loan in full, resulting in bad debt loss. A customer is in default when failing to make the contractually determined payments over a predefined number of days. Typically, default is defined as 90 to 180 days of payment arrears. For past credit applications that were accepted, it can be observed whether the customer defaulted after acceptance. Hence, a structured dataset can be compiled with observations representing past applications, including information on the customer and loan collected at the time of application and related to credit risk, as well as a class variable indicating whether the customer defaulted. Note that default, as indicated by the class variable, is typically observed in application credit scorecard development over the first twelve to eighteen months after acceptance and not over the full lifetime of the loan. This practice allows us to more rapidly include recent observations in the dataset that is used to develop a credit scorecard, with more recent observations typically being more similar to and hence representative of future loan applications.

Typical information collected at the time of application includes sociodemographic information (e.g., age and occupation), financial information (e.g., income and number of outstanding loans), and information on a customer’s credit history (e.g., missed payments and prior default information), which is often acquired from a credit bureau. A number of recent studies have investigated the use of alternative information for credit scoring, which in the big data era is becoming abundantly available. For instance, Óskarsdóttir, Bravo, Sarraute, Vanthienen, and Baesens (2019) demonstrate the added value of social network information, obtained through the analysis of call detail record data, for improving credit scorecard performance. Social network information is shown to be of use when credit history information is missing, which is of crucial importance in a microfinance context to enhance financial inclusion. Oreski, Oreski, and Oreski (2012) explore the opportunities offered by newly accumulated data by proposing a hybrid system involving an artificial neural network and a genetic

algorithm to perform improved variable selection and as such to effectively leverage the abundance of data. Finally, Hurley and Adebayo (2017) review the current credit scoring environment, emphasizing the role of big data in the emancipation of alternative credit scoring methods. They also focus on the legal challenges concerning less transparent methods, suggesting that legislation should adapt to accommodate new opportunities in improving creditworthiness assessment and as such credit risk management by financial institutions.

The development of a credit scorecard hence involves the development of a predictive model to estimate a binary target or dependent variable as a function of a range of predictor or independent variables. For this purpose, a standard methodology can be adopted, such as the cross-industry standard process for data mining (CRISP-DM) (Wirth & Hipp, 2000) or the analytics process model (Verbeke, Baesens, & Bravo, 2017). Note that the primary objective in developing credit scorecards is typically to accurately risk-rank loan applicants, allowing to set a minimum creditworthiness level beyond which applicants are rejected and above which applicants can be accepted. However, the risk score is typically also used to support a risk-sensitive pricing strategy by determining loan conditions in relation to the risk score. For Basel-compliant financial institutions, credit scorecards support capital calculation by allowing to define rating grades and calibrate probability of default estimates as a complement to the scorecard (Bequé, Coussement, Gayler, & Lessmann, 2017; Castermans, Martens, Gestel, Hamers, & Baesens, 2010) and in line with the applicable regulation. More recently, new accounting regulations (i.e., IFRS9 and CECL) were introduced, requiring financial institutions to implement forward-looking estimates with respect to expected losses in their financial statements. Expected losses can also be calculated using the risk estimates produced by credit scorecards.

A broad variety of classification methods have been proposed in the literature (Friedman et al., 2009) and applied for credit scoring. For instance, a recent benchmarking study focusing on ensemble methods for credit scoring (Abellán & Castellano, 2017) shows how the performance of a simple classifier for credit risk assessment improves when used in an ensemble model, with reported results on six relatively small data sets including between 240 and 2435 observations. A comprehensive experimental benchmarking study was recently performed by Lessmann et al. (2015) as an update of, and extension to, an earlier experimental benchmarking study by Baesens et al. (2003). In this new study, the performance of 41 classification methods is compared across eight credit scoring datasets using appropriate statistical testing procedures and adopting various classification performance metrics. As such, they arrived at a ranking of methods based on six different measures. Their conclusions show that the use of ensemble performance measures for model selection may increase accuracy and suggest that the recent evolution in computational power and increase in the sophistication of classification methods can lead to significant improvement in discrimination power in the area of credit scoring. The industry standard approach, logistic regression, is shown to be significantly outperformed by more advanced methods. Finally, the importance of assessing the financial implications of using credit scorecards is discussed, indicating the difficulty of obtaining the required cost and benefit information and supporting efforts towards the profit-driven optimization of credit scorecards and the development and application of profit-driven credit scoring approaches.

Lessmann et al. (2015) highlight a shortage of real financial data that is available for academic research on credit scoring. The authors compared a large number of research papers concerning consumer credit modeling, finding that research on credit scoring suffers not only from the small number of publicly available datasets but also

from the relatively small size of these datasets, with an average number of 6167 observations. This can be explained by financial institutions being reluctant to disclose sensitive information to competitors and to external parties because of privacy concerns. Moreover, publicly available datasets typically do not contain information on the involved costs and benefits, neither averaged across the involved loans nor at the individual loan level. As a consequence, the publicly available datasets cannot be used for a cost-based evaluation of classification models, as would be desired in this context. For these reasons, an extensive benchmarking study is practically infeasible, and the results from a single dataset, which is made publicly available, are provided in this paper.

2.2. Cost-Sensitive Learning

Cost-sensitive learning methods are intended to develop models that focus on minimizing the total cost of errors instead of minimizing the total error or maximizing the accuracy, as typically performed by cost-insensitive learning methods. This approach is useful not only when different types of misclassification errors bear different costs but also when the datasets suffer from *class imbalance*, that is, when the numbers of observations in each class differ considerably. An imbalanced class distribution may cause learners to experience difficulties in developing a predictive model that accurately predicts the minority class. As an alternative solution to cost-sensitive learning, (re-)sampling methods such as oversampling, undersampling or SMOTE are often adopted (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). For an experimental evaluation of the suitability of these approaches for credit scoring, one may refer to (Brown & Mues, 2012; Marqués, García, & Sánchez, 2013).

In a typical credit scoring case, observations of the *positive* class (i.e., bads or defaulters) are much less prevalent than observations of the *negative* class (i.e., goods or nondefaulters). Misclassifying a good loan as a bad one (a *false positive* prediction) results in missed revenues, whereas a *false negative* prediction, i.e., misclassifying a bad loan as a good one, leads to losses due to nonrepayment of the loan and nonrecovery of the outstanding amount at time of default. We denote these by C_{FP} and C_{FN} , respectively.

Unfortunately, the literature on cost-sensitive learning, which offers a plethora of methods, lacks points of reference as to which methods can be regarded as state-of-the-art. This is partly attributed to the absence of publicly available datasets with real-world costs that would allow for a comprehensive comparison.

In a recent study on the available cost-sensitive variants of AdaBoost (Freund & Schapire, 1997), a well-known boosting method, Nikolaou, Edakunni, Kull, Flach, and Brown (2016) concluded that only *Naive CS AdaBoost* (Masnadi-Shirazi & Vasconcelos, 2011; Viola & Jones, 2001; Zadrozny, Langford, & Abe, 2003), *AdaBoost with minimum expected cost criterion* (mecc) (Ting, 2000a, 2000b) and Asymmetric AdaBoost (Viola & Jones, 2001) have solid theoretical foundations and that they perform best when calibrated, using, for example, *Platt Scaling* (Platt, 1999). One non-AdaBoost-based ensemble method is *Costing* (Zadrozny et al., 2003), a variant of Bagging (Breiman, 1996), where data sampling is done by taking costs into account. Other possibilities include specifying costs in terms of data weights and building a *cost-sensitive weighted tree* (Ting, 1998) and using a cost-sensitive threshold instead of the cost-insensitive 0.5 on any probability output, such as the average probability that can be returned by Bagging and Random Forests (Breiman, 2001). Choices for

this threshold are $T_{cs} = \frac{C_{FP}}{C_{FP}+C_{FN}}$ (Elkan, 2001) and that found to yield the least costs on a validation set, a process called *Thresholding* (Sheng & Ling, 2006). Finally, *MetaCost* (Domingos, 1999; Ting, 2000b) involves first relabelling the training data according to how a cost-sensitive model built on them classifies them and then retraining a single, cost-insensitive, decision tree.

Papers considering cost-sensitive models for credit scoring have already appeared in the literature and report a positive effect on profitability; see (Xia, Liu, & Liu, 2017) for an example and an up-to-date review. Their main drawbacks, however, are that comparisons are far from comprehensive with respect to cost-sensitive methods (particularly in view of the absence of what can be considered state-of-the-art), and most relevant to this paper, there is no comparison to state-of-the-art operational credit application evaluation procedures (such as credit scorecards) used in practice by financial institutions.

Whereas a broad variety of cost-sensitive learning approaches have been proposed and can be adopted for profit-driven credit scorecard development, dedicated profit-driven modeling approaches for credit scoring are more recent and much less common (Bahnsen, Aouada, & Ottersten, 2014).

Verbraken et al. (2014) proposed a profit-driven performance measure, which explicitly takes into account expected profits and losses resulting from the credit granting process. The proposed Expected Maximum Profit measure for credit scoring (EMP-CS) finds the optimal trade-off between expected losses in case of default and net revenues of good loans by optimizing the decision threshold or cutoff for a scorecard implementation to maximize the resulting profit. The EMP-CS additionally takes into account variability or uncertainty with respect to the involved parameters.

The EMP-CS measure extended upon previous work (Verbeke et al., 2012; Verbraken et al., 2013), which introduced the Expected Maximum Profit (EMP) evaluation framework with an application to customer churn prediction models. Recently, Stripling (2018) introduced a general EMP learning framework that allows to replace the objective function in learning methods with the application-dependent EMP measure. This framework has thus far been instantiated to two supervised learning methods, i.e., logistic regression (Stripling, vanden Broucke, Antonio, Baesens, & Snoeck, 2018) and decision trees (Höppner, Stripling, Baesens, vanden Broucke, & Verdonck, 2020). These methods are an alternative to the cost-sensitive learning methods that are evaluated in this study. A priority for future research will be to compare the two approaches.

Finally, Maldonado, Pérez, and Bravo (2017) experimented with a cost-based variable selection mechanism to optimize profitability. Their approach was to build mixed-integer linear programming models for constructing classifiers while taking into account the acquisition costs for the variables. In fact, this approach is perpendicular to the profit-driven learning approaches, which account for the costs of the observations (i.e., the rows in a structured dataset) and not the costs of the predictors (i.e., the columns in a structured dataset). Therefore, the two approaches are complementary and ideally should be combined to yield the ultimate profit-maximizing approach. The development and evaluation of such a combined approach is also identified as a topic of prime importance for future research.

Table 1. Overview of the dataset used in our case study.

Dataset	Size	% of Total	Default rate
All data	18918	100%	16.9%
Business Channel 1	14179	75%	15%
Business Channel 2	2831	15%	34%
Business Channel 3	1908	10%	5.6%

3. Methodology

3.1. Case Study

The presented case involves a dataset containing information regarding the accepted loans given to individual customers by a Romanian NBFi during a one-year period from January 2016 to December 2016. The anonymized dataset has been made publicly available². Default is recorded during the twelve months after acceptance. For modeling purposes, the NBFi considers a loan to be in default if the customer is more than 45 consecutive days in payment arrears. This approach was embraced by the NBFi due to the increased risk involved in subprime lending. Note that the Basel regulation for the purpose of capital calculation defines default as 90 days of payment arrears and exceptionally allows a definition of 180 days of payment arrears to be used. The Basel definition is typically also adopted in the industry for developing application credit scorecards and for acceptance policies, although this is strictly speaking not necessary for a nonbanking financial institution to be Basel compliant.

The independent variables in the dataset can be classified into three types: *sociodemographic information*, for example, education, age, and profession; *credit history*, for instance, the credit bureau score; and *economic information* such as loan amount, income, and gearing ratio. The full list of 20 independent variables in the dataset is provided in Table A1 in Appendix A.

Table 1 provides details with respect to the size of the dataset. The full dataset contains 18,918 observations, and the overall default rate is 16.9%. Since the NBFi business is conducted through three main channels, details also are provided per channel. A substantial difference between the default rates across the three channels can be observed in Table 1, ranging from 5.6% for Business Channel 3, over 15% for Channel 1, to 34% for Channel 2. These three channels have been strategically developed to reach different customer segments. The first channel covers the vast majority of clients and is considered the most mature part of the business (> 70% of granted loans). The second channel targets customers carrying higher risk and, accordingly, involving a higher interest rate for the loans. The third channel is dedicated to in-store loans for financing different personal purchases. It has the lowest associated risk among the three channels and is very profitable but accounts for only a small proportion of the company’s business (less than 5%). Since the default rates reported in Table 1 provide a strong indication that the risk profiles of the customers that are reached through the various channels are substantially different, we will evaluate whether segmentation per channel may improve the predictive power and, more important in this case, the profitability of the resulting credit assessments.

²<https://github.com/vub-dl>

3.2. Calculation of the Misclassification Costs

Cost-sensitive learning approaches, as explained in Section 2.2, require the misclassification costs for false positives and false negatives to be specified. Misclassification costs can be specified at the class level by means of a cost matrix (Verbeke et al., 2017), thereby assuming them to be equal for all observations in the same class. Note that cost-insensitive learning approaches implicitly assume the misclassification costs to be equal for all observations in both classes. Obviously, this assumption does not hold in a credit scoring context, since bads predicted to be good are much more costly than goods predicted to be bad. Moreover, since loans involve different amounts and conditions and since defaults involve different losses given default (LGD) and exposures at default (EAD), the assumption of equal misclassification costs at the class level is invalid in credit scoring.

Therefore, in this study, we consider observation-dependent misclassification costs (also called example-, record- or instance-dependent misclassification costs), which are to be specified individually per observation. If necessary, for instance, because a method cannot handle observation-dependent misclassification costs, they can be averaged to obtain class-dependent misclassification costs. For recent observations, both goods and bads, the eventual misclassification costs cannot simply be observed because nondefaulted loans have not yet reached maturity and defaulted loans have not yet reached the end of the workout period. Therefore, in this paper, we propose the following straightforward approach to estimate these costs.

Recall that the positive class represents the bad applicants who defaulted. The misclassification cost for a false positive, i.e., the loss involved in classifying a good applicant as bad, is equal to the profit that is obtained over the lifetime of the loan (Verbraken et al., 2014). When rejecting a good applicant because the scorecard predicts the applicant to be bad, we miss out on future returns. However, we reason that the false positive cost should not be taken equal to the full return that would be obtained when a loan is paid back duly, as implemented, for instance, in the EMP-CS measure (Verbraken et al., 2014), since a good applicant will generate less profit in case of early repayment. Additionally, since goods are defined as non-default during the first twelve months of the loan, a good applicant may nevertheless default later and hence not generate the full return. Therefore, an adjustment both for the possibility of early repayment and for default occurring after the first year is to be made in estimating the misclassification cost for the goods in the dataset.

For this purpose, an average rate of return (ROI) is calculated across all *good* loans in a dataset spanning a five-year time window. The observations in this dataset concern loans that either reached maturity without default, involved early repayment or defaulted. Loans that defaulted during the first twelve months after acceptance comprise observations of the bad class and hence are excluded from the calculation. This dataset as such is a representative sample that allows us to estimate the average ROI of goods, which is to be multiplied by the loan amount of a good loan to obtain the observation-dependent misclassification cost for false positives.

The misclassification cost of a false negative, i.e., the loss involved in classifying a bad applicant as good, is equal to the economic loss resulting from the default as incurred by the financial institution (Verbraken et al., 2014). For recent defaults with incomplete workouts, for which the final loss has not yet been observed, an estimation procedure is required. For this purpose, we calculated the average loss given the default rate, again using a dataset including observations of bad applicants over a five-year time window for which the final loss has been observed. Multiplying the average LGD

value by the involved exposure at default (EAD) provides an estimate of the eventual loss, i.e., the misclassification cost.

In addition to the actual costs for defaulted loans and the actual returns for fully reimbursed loans, on-site discussions with the management of the NBFIs revealed an additional factor known to be important in calculating misclassification costs, i.e., the business channel. Considering the different default rates across the three business channels as reported in Table 1, it is evident that the associated risks are very different. Hence, the calculation of the average ROI, LGD and EAD value has been performed separately for each channel.

By applying the above procedure, we are able to estimate the misclassification costs for all observations in the dataset. Note that a refinement in estimating the eventual loss could be implemented by estimating individual LGD values instead of using the average historical LGD. However, when adopting more advanced methods for estimating LGD, as well as ROI and EAD, we did not obtain sufficiently robust results improving upon the presented procedure, which has the important advantage of being straightforward to implement. A potential topic for future research is to elaborate and evaluate an advanced predictive modeling approach for estimating misclassification cost. However, given the rather poor performance of LGD and EAD models reported in the literature, with industry benchmark R-squared values being between 10% and 30% (Baesens et al., 2016; Bellotti & Crook, 2012), there may be rather limited added value to adopting a more complex approach.

We now show that minimizing the above observation-dependent misclassification costs corresponds to maximizing the profit. Let $\{(X_i, y_i)\}$ be a set of observations with X_i being the independent variables (features) of an observation i , and $y_i \in \{\oplus, \ominus\}$ its known outcome namely positive or negative. Let c_i be the misclassification cost of observation (X_i, y_i) as defined in this section, that is, in case $y_i = \ominus$ it is the false positive misclassification cost, and in case $y_i = \oplus$ it is the corresponding false negative misclassification cost. Let f be a binary classifier such that $f(X_i) \in \{\oplus, \ominus\}$ is the prediction of the classifier for observation i . Minimizing the misclassification costs then corresponds to minimizing $\sum_{(X_i, y_i)} c_i * I[f(X_i) \neq y_i]$ with $I[\cdot]$ the Iverson bracket which represents value '1' in case $f(X_i) \neq y_i$ is true, e.g. a misclassification, and value '0' otherwise.

Let the profit of a negative (non-default) observation $y_i = \ominus$ be p_i , which is the amount of the loan times the average ROI and which is equal to the false positive cost as defined before; hence, $p_i = c_i$. The loss of a negative (default) observation $y_i = \ominus$ is the false negative cost c_i . The profit, Π , is then computed over all observations that are classified as negative and is equal to the profit that results from true negative observations, $p_i * I[f(X_i) = \ominus]$ if $y_i = \ominus$ minus the loss that is due to false negative observations, $c_i * I[f(X_i) = \oplus]$ if $y_i = \ominus$. We can then derive the following, starting

from the formula to calculate the profit:

$$\Pi := \sum_{(X_i, y_i), y_i = \ominus} p_i * I[f(X_i) = \ominus] - \sum_{(X_i, y_i), y_i = \oplus} c_i * I[f(X_i) = \ominus] \quad (1)$$

$$= \sum_{(X_i, y_i), y_i = \ominus} p_i * (1 - I[f(X_i) \neq \ominus]) - \sum_{(X_i, y_i), y_i = \oplus} c_i * I[f(X_i) \neq \oplus] \quad (2)$$

$$= \sum_{(X_i, y_i), y_i = \ominus} p_i - \sum_{(X_i, y_i), y_i = \ominus} p_i * I[f(X_i) \neq \ominus] - \sum_{(X_i, y_i), y_i = \oplus} c_i * I[f(X_i) \neq \oplus] \quad (3)$$

$$= \sum_{(X_i, y_i), y_i = \ominus} c_i - \sum_{(X_i, y_i), y_i = \ominus} c_i * I[f(X_i) \neq y_i] - \sum_{(X_i, y_i), y_i = \oplus} c_i * I[f(X_i) \neq y_i] \quad (4)$$

$$= \sum_{(X_i, y_i), y_i = \ominus} c_i - \sum_{(X_i, y_i)} c_i * I[f(X_i) \neq y_i] \quad (5)$$

Optimizing a classifier f to maximize the profit hence corresponds to minimizing $\sum_{(X_i, y_i)} c_i * I[f(X_i) \neq y_i] - \sum_{(X_i, y_i), y_i = \ominus} c_i$ where the latter is independent of f and can be removed from the minimization; hence, minimizing the misclassification costs $\sum_{(X_i, y_i)} c_i * I[f(X_i) \neq y_i]$ corresponds to maximizing the profit. We can therefore use state-of-the-art cost-sensitive classification methods to obtain profit maximizing classifiers.

3.3. Treating Missing FICO Score Values

Credit bureau scores, such as FICO or Equifax scores, are external credit risk assessments provided by credit bureaus that are often used by retailers as a means to decide on the eligibility of customers to purchase goods on credit, as well as by financial institutions. Credit bureau scores can either be used as the sole criterion to decide on loan application acceptance or alternatively be used as an input variable for the in-house credit scorecard. Credit bureaus gather detailed information on a customer's credit history, which is used to determine a customer's risk profile in terms of a score.

However, customers applying for a loan with a financial institution may not have a credit bureau score, resulting in a missing value for the concerned input variable in the financial institution's in-house credit scorecard. Additionally, for historical observations in the dataset that are used to develop an application scorecard, the external credit bureau score may be missing for a substantial part of the observations, as in the presented case. Note that a missing FICO score could be interpreted as a lack of credit history for those particular customers.

In the experiments, we evaluate three replacement approaches to handle missing values:

- (1) A first approach sets missing FICO scores to zero, expressing as such the idea that a missing FICO score reflects a missing credit history and as such an increased risk profile.
- (2) A second approach replaces the missing FICO scores with a score of 570, representing the equivalent quantile in terms of default rate
- (3) A third approach replaces the original FICO score variable using weights of evidence (WOE) to obtain a more accurate distribution in relation to the target variable. Five discrete groups were identified, resulting in the greatest differences in WOE between adjacent groups, with the aim of maximizing the Information

Value of the predictor. Observations missing the FICO score formed one of these groups.

In addition to replacing the missing values, we added a missing value dummy variable to keep this information in the dataset and available to the modeling technique.

3.4. Experimental Design

The aim of the experimental part of our work was to obtain evidence on (a) which of the three treatments of missing FICO score values (see Section 3.3) is the most appropriate for this case study, (b) whether segmenting the training data by business channel improves predictive power and profitability, and (c) whether cost-sensitive learning offers an advantage when applied in conjunction with the current logistic regression-based approach used by the NBFi, as well as compared to cost-insensitive methods.

To achieve these aims, in the experiments, we applied the cost-sensitive methods mentioned in Section 2.2 and well-known and frequently adopted cost-insensitive methods including XGBoost (Chen & Guestrin, 2016) and those considered by Lessmann et al. (2015), namely, LR, ANN, Random Forests and HCES-Bag. Table 2 provides a summary of these methods. We also considered a second scorecard model developed by the NBFi with the same purpose of being applied in addition to the method currently employed in practice by the NBFi. Moreover, we trained models on data from all business channels together and individually, and all experiments were repeated for each missing FICO score treatment.

We built 50 trees per ensemble, and all results were averaged over three repetitions in each of which the dataset splitting (see below) used different randomness. The number of trees per ensemble could be fine-tuned for optimizing model performance, however, a full-blown hyperparameter optimization is beyond the scope of this paper.

3.4.1. Data Preparation

A stratified 30% of the observations, respecting the original ratios of both the business channels and the classes (see Table 1), was set aside to be used as the test set. The remaining observations were further split, also in a stratified manner, into training and validation sets (80% and 20% respectively, or 56% and 14% of the total number of observations).

3.4.2. Training Costs

For training, we used two false positive and false negative misclassification cost pairs (C_{FP}, C_{FN}) , namely, (exp, exl) and $(\overline{exp}, \overline{exl})$, where (exp) and (exl) are the observation-dependent expected profit and loss variables described in Section 3.2, respectively, and $\overline{exp}$ and $\overline{exl}$ are their respective averages in the training dataset.

3.4.3. Evaluation

For evaluation, we calculated the relative *MaxProfit%*, the percentage of the maximum possible profit (when accepting all nondefaulters and rejecting all defaulters) that a model achieves. The profit was calculated based on the (observation-dependent) exp and exl variables. For reference, we also included some standard cost-insensitive measures, namely, true and false positive rates and AUC. The McNemar test (McNemar,

Table 2. Overview of the methods considered. Cost-sensitive methods are indicated by the column CS. Note that, MetaCost (MC) was applied on methods one to nine.

#	Name	Description	CS	#	Name	Description	CS
1.	AdaMEC	AdaBoost with mecc & Platt scaling	✓	13.	LR-Thres	Logistic Regression with Thresholding	✓
2.	NCSAB	Naive CS AdaBoost & Platt scaling	✓	14.	LR- T_{cs}	Logistic Regression with T_{cs}	✓
3.	AsymAB	Asymmetric AdaBoost & Platt scaling	✓	15.	MLP-Thres	Multi-layer Perceptron with Thresholding	✓
4.	RF-Thres	Random Forests with Thresholding	✓	16.	MLP- T_{cs}	Multi-layer Perceptron with T_{cs}	✓
5.	RF- T_{cs}	Random Forests with T_{cs}	✓	17.	HCESB-Thres	Hill-Climbing Bagged Ensemble Selection with Thresholding	✓
6.	BG-Thres	Bagging with Thresholding	✓	18.	HCESB- T_{cs}	Hill-Climbing Bagged Ensemble Selection with T_{cs}	✓
7.	BG- T_{cs}	Bagging with T_{cs}	✓	19.	XGBRF-Thres	XGBoost RF Classifier with Thresholding	✓
8.	Costing	Costing	✓	20.	XGBRF- T_{cs}	XGBoost RF Classifier with T_{cs}	✓
9.	cswt	CS weighted tree	✓	21.	LR	Logistic Regression	
10.	MC	MetaCost (on methods 1-9)	✓	22.	MLP	Multi-layer Perceptron	
11.	BG	Bagging		23.	HCESB	Hill-Climbing Bagged Ensemble Selection	
12.	AB	AdaBoost Platt scaling		24.	XGBRF	XGBoost RF Classifier	
				25.	RF	Random Forests	

1947) was adopted to compare each proposed model with the base model, thereby evaluating whether differences in performance are significant. The results of the McNemar test for the various methods reported in Table 3 have p-values below the adopted significance level of 0.05, allowing us to reject the null hypothesis that the base model performs equally well.

Note that the MaxProfit% as defined here adjusts the maximum profit measure (Verbeke et al., 2012; Verbraken et al., 2013) by reporting the performance as a percentage of the maximum profit, i.e., in a relative manner, rather than the absolute value of the profit per customer. This facilitates the interpretation and comparison of the results, including across applications and studies. Additionally, whereas in (Verbraken et al., 2014) a theoretical distribution is to be assumed for the class-dependent false negative cost, i.e., the loss rate, in calculating the *expected* maximum profit, here we propose to calculate the MaxProfit% measure by taking into account the observation-dependent misclassification costs as observed in the test set. In other words, the empirical distribution of the misclassification costs is used, which eases computation and interpretation, and if the sample is sufficiently large and representative, it can be expected to account for the stochastic nature of the misclassification costs and, as such, yield a similar result as the EMP-CS measure.

4. Results

In a risky business environment, such as lending to subprime customers, profit optimization represents a constant preoccupation for companies such as the NBFi. Our experiments focused on testing the utility of cost-sensitive learning in improving profitability across the three business channels of the NBFi when applied in complement to the existing approach. In other words, the presented results allow us to assess the

Table 3. The 5 best performing models from Table 2 per business channel used for testing, and the second scorecard (SSC), all when applied in addition to the scorecard (SC) models that the NBFI is currently using (included for reference). Zero, avg and woe denote how the FICO score feature (FicoS) was treated (see Section 3.4). All of these best models used record-dependent training costs. The results are averaged over three runs with different splits of the dataset into training/test sets.

Model	CS	FicoS	Business Channel		MaxProfit%	TPR(%)	FPR(%)	AUC	McNemar χ^2
			Train	Test					
AdaMEC	✓	woe	1	1	71.95	21.57	4.67	0.585	198.6
RF- T_{cs}	✓	avg	All	1	71.84	17.03	3.70	0.566	95.0
AdaMEC	✓	avg	1	1	71.72	21.97	5.07	0.585	224.2
LR-Thres	✓	zero	All	1	71.60	19.13	4.38	0.748	170.5
LR- T_{cs}	✓	zero	All	1	71.59	19.54	4.52	0.748	163.4
SSC		woe	All	1	70.38	28.07	9.10	0.595	17.2
SSC		woe	1	1	70.42	27.47	8.27	0.596	N/A
SC				1	70.16	0	0	-	-
AdaMEC	✓	woe	All	2	52.61	25.37	6.60	0.594	146.7
AdaMEC	✓	avg	All	2	51.97	24.73	6.97	0.589	189.2
AdaMEC	✓	zero	All	2	51.94	26.20	7.13	0.595	130.0
NCSAB	✓	avg	All	2	47.81	6.20	1.27	0.525	4.4
NCSAB	✓	woe	All	2	47.61	5.83	1.3	0.746	88.2
SSC		woe	All	2	35.01	48.10	21.90	0.631	52.4
SSC		woe	2	2	36.99	62.13	31.07	0.655	N/A
SC				2	16.97	0	0	-	-
RF- T_{cs} -MC	✓	avg	All	3	89.07	10.40	0.50	0.549	39.3
BG-Thres-MC	✓	woe	All	3	88.88	20.80	1.83	0.595	43.2
BG- T_{cs} -MC	✓	avg	All	3	88.84	15.67	1.00	0.573	42.0
BG-Thres	✓	avg	3	3	88.82	18.77	1.83	0.585	50.7
XGBRF-Thres	✓	avg	All	3	88.77	14.58	1.05	0.717	20.3
SSC		woe	All	3	85.45	23.97	5.40	0.593	25.3
SSC		woe	3	3	82.83	43.77	10.43	0.667	N/A
SC				3	88.00	0	0	-	-

added value, i.e., the decrease in losses, of a model that is developed on accepted applicants. No reject inference was applied, and as such, the results are to be interpreted as if a second-stage model would be implemented to further filter loan applicants in a second stage, so after a first screening by the current credit scorecard that is operated by the NBFI. In fact, almost all experimental results reported in the literature on credit risk modeling are concerned with the analysis of accepted applicant data. Hence, these results are interpreted as the improvement of a second-stage model.

One may wonder whether developing a credit scoring model on a *full* dataset would yield better performance given that such a dataset would include more observations of the minority class, which may facilitate learning. Although there is a substantial body of literature on reject inference (e.g., J. Crook and Banasik (2004)), which addresses this question, we believe that adopting a profit-driven approach in this respect may be an interesting avenue for future research.

Table 3 shows the test set results averaged over the three runs of the experiments. As seen from the table, the results on Business Channel 2 stand out, showing a dramatic improvement in terms of increased profitability, with an up to 200% relative increase in profitability (i.e., from 16.97% to 52.61%) that is achieved by the best cost-sensitive models over the current scorecard used by the company and an over 42% relative increase over the *second scorecard* they developed (i.e., from 36.99% to 52.61%).

The achieved improvement in performance for Channel 2 is likely due to the high default rate of 34% that is observed for this channel, which offers considerable potential for improving profitability and discrimination power. It is worth mentioning that

Channel 2 is the NBFI’s newest business channel.

The default rates in Channels 1 and 3 are significantly lower, as reported in Table 1, thus providing less opportunity to improve upon the performance of the operational scorecard. The improvements in profitability that are achieved by adopting cost-sensitive models, as reported in Table 3, hence are clearly less prominent than for Channel 2. The limited improvements for Channels 1 and 3 are explained by the business maturity, which is higher for Channels 1 and 3, and the risk profile of the customers in these channels. For customers in Channels 1 and 3, the company has already been optimizing the scoring and acceptance process by leveraging the longer business history, in contrast to the newer Channel 2 with limited historical records available to support the development of the operational scorecard. Nevertheless, for Channels 1 and 3, an improvement in MaxProfit% over the current state-of-the-art scorecard is achieved by the cost-sensitive models. Although modest in a relative sense, in an absolute sense, the increase in profit may be worth the added complexity of the resulting scoring model and process.

The *second scorecard* approach, employing logistic regression, fails to increase the profitability for Channel 1, and for Channel 3, it even yields decreased profitability. However, when evaluating the models in terms of AUC, the second scorecard appears to achieve the best performance. This confirms that optimizing a model to classify as accurately as possible effectively yields a different result than when optimizing to maximize profitability, i.e., to minimize misclassification costs.

This is evident since it makes perfect sense that models trained with different objective functions align with the objectives expressed by these functions and perform better on measures that evaluate how well the model performs in terms of the employed objective. An interesting example is the AUC value for the XGBoost algorithm in Channel 3, which is considerably higher than the other results for this channel; however, the MaxProfit% is slightly less than the top performer’s. Nonetheless, a general appreciation and broad acceptance of the requirement to align the adopted objective function when learning from data with the actual business objective is still missing.

With respect to the different approaches that were adopted for treating the missing values of the FICO score, i.e., the lack of credit history for a subsample of customers, both *weights of evidence* and the *average* proved to be viable approaches, without there being a clear winner. The *zero* approach can be concluded to be less appropriate, both from an expert perspective and from the reported results.

Finally, the results with respect to the adopted segmentation approach do not provide a clear improvement across the segments. For Channel 1, the best model was trained on observations from that channel only, whereas overall for all channels, the majority of the best results (11 out of 15) were obtained by training on the full dataset, including observations from all business channels.

5. Conclusions and Future Research

This paper presents the results of a case study that was elaborated in collaboration with a Romanian NBFI with the aim of investigating the potential for increasing the profitability of a credit scorecard by adopting cost-sensitive learning techniques. To apply cost-sensitive learning, an approach was developed in collaboration with the NBFI to determine the costs that result from both false positive and false negative misclassified cases. In addition, a segmentation approach based on the business channel and three approaches to handle missing credit bureau scores were evaluated. Two

minor adjustments to the EMP-CS measure were proposed to ease computation and interpretation, as well as to facilitate comparison across applications and studies.

The experimental results indicate that cost-sensitive learning is a promising direction for improving a data-driven credit assessment methodology, which results in a better balance between profitability and risk taking. Cost-sensitive learning naturally aligns with the business objective by acknowledging the imbalanced misclassification costs that are involved in the loan approval process. Since the results presented in this paper are obtained on a single dataset, the experiments do not provide a conclusive answer as to which cost-sensitive learning approach overall is the best performing technique; however, given the no-free-lunch theorem, we know that there is no clear, overall winner and that some experimentation will always be required to optimize scorecard performance. Nonetheless, this paper provides a comprehensive overview of prime candidate approaches that are to be considered when developing a profit-driven credit scorecard.

Our aim with this paper was to contribute to a small but growing stream of literature adopting a profit-driven perspective towards the development of predictive models for optimizing business decision-making. In essence, such approaches pertain to the field of prescriptive analytics, which adds to predictive analytics by not merely predicting the future but learning to optimize the outcome of the eventual application that is supported by the predictive model. The presented approach was intended to not only predict which loan applicants are most likely to default but also learn and prescribe which loan applications are to be accepted to maximize profits.

Further research is required to explore alternative approaches to develop prescriptive credit scorecards that maximize profitability. A promising alternative approach is the recently introduced expected maximum profit learning framework for profit-driven model building by Stripling (2018), instantiated to logistic regression (Stripling et al., 2018) and decision trees (Höppner et al., 2020), and previously applied to predicting customer churn. An extensive comparison with the cost-sensitive learning approaches that have been applied in this study, across a broad selection of credit datasets and in terms of a variety of evaluation measures (e.g., (Miguéis, Benoit, & Van den Poel, 2013)), would be a significant next step in adding to the work of Lessmann et al. (2015).

A key issue with respect to developing profit-driven scorecard approaches is the lack of publicly available datasets that include the required information on observation-dependent costs and benefits. Simulated data may offer a solution, since financial institutions are understandably reluctant to allow publication of data. As an exception to the general rule, we are pleased that the NBFİ involved in this case study gave us permission to make the dataset we used publicly available³.

Funding

This work was supported in part by Innoviris, the Brussels Region Research funding agency. It was also partly supported by the Romanian National Authority for Scientific Research and Innovation.

³<https://github.com/vub-dl>

References

- Abellán, J., & Castellano, J. G. (2017). A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Systems with Applications*, 73, 1–10.
- Baesens, B., Roesch, D., & Scheule, H. (2016). *Credit risk analytics: Measurement techniques, applications, and examples in sas*. John Wiley and Sons.
- Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the operational research society*, 54(6), 627–635.
- Baesens, B., Van Vlasselaer, V., & Verbeke, W. (2015). *Fraud analytics using descriptive, predictive, and social network techniques: a guide to data science for fraud detection*. John Wiley and Sons.
- Bahnsen, A. C., Aouada, D., & Ottersten, B. (2014). Example-dependent cost-sensitive logistic regression for credit scoring. In *Machine learning and applications (icmla), 2014 13th international conference on* (pp. 263–269).
- Bahnsen, A. C., Aouada, D., & Ottersten, B. (2015). Example-dependent cost-sensitive decision trees. *Expert Systems with Applications*, 42(19), 6609–6619.
- Bellotti, T., & Crook, J. (2012). Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting*, 28(1), 171–182.
- Bequé, A., Coussement, K., Gayler, R., & Lessmann, S. (2017). Approaches for credit scorecard calibration: An empirical analysis. *Knowledge-Based Systems*, 134, 213–227.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 26(2), 123–140.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39(3), 3446–3453.
- Castermans, G., Martens, D., Gestel, T. V., Hamers, B., & Baesens, B. (2010). An overview and framework for pd backtesting and benchmarking. *Journal of the Operational research society*, 61(3), 359–373.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321–357.
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* (pp. 785–794). (02776)
- Crook, J., & Banasik, J. (2004). Does reject inference really improve the performance of application scoring models? *Journal of Banking & Finance*, 28(4), 857–874.
- Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, 183(3), 1447–1465.
- Domingos, P. (1999). Metacost: A general method for making classifiers cost-sensitive. In *Proc. kddm'99* (p. 155–164).
- Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proc. seventeenth int. conf. on artificial intelligence* (p. 973–978).
- Freund, Y., & Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer System Sciences*, 55(1), 119–139.
- Friedman, J., Hastie, T., & Tibshirani, R. (2009). *The elements of statistical learning*. Springer series in statistics New York, NY, USA:.
- Höppner, S., Stripling, E., Baesens, B., vanden Broucke, S., & Verdonck, T. (2020). Profit driven decision trees for churn prediction. *European Journal of Operational Research*, 284(3), 920–933.
- Hurley, M., & Adebayo, J. (2017). Credit scoring in the era of big data. *Yale Journal of Law and Technology*, 18(1), 5.
- Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124–136.
- Maldonado, S., Pérez, J., & Bravo, C. (2017). Cost-based feature selection for support vector

- machines: An application in credit scoring. *European Journal of Operational Research*, 261(2), 656–665.
- Marqués, A. I., García, V., & Sánchez, J. S. (2013). On the suitability of resampling techniques for the class imbalance problem in credit scoring. *Journal of the Operational Research Society*, 64(7), 1060–1070.
- Masnadi-Shirazi, H., & Vasconcelos, N. (2011). Cost-sensitive boosting. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(2), 294–309.
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157.
- Miguéis, V. L., Benoit, D. F., & Van den Poel, D. (2013). Enhanced decision support in credit scoring using bayesian binary quantile regression. *Journal of the Operational Research Society*, 64(9), 1374–1383.
- Nikolaou, N., Edakunni, N., Kull, M., Flach, P., & Brown, G. (2016). Cost-sensitive boosting algorithms: Do we really need them? *Machine Learning*, 104(2), 359–384.
- Oreski, S., Oreski, D., & Oreski, G. (2012). Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. *Expert systems with applications*, 39(16), 12605–12617.
- Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., & Baesens, B. (2019). The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. *Applied Soft Computing*, 74, 26–39.
- Platt, J. (1999). Probabilistic outputs for support vector machines and comparison to regularised likelihood methods. *Advances in Large Margin Classifiers*, 10(3), 61–74.
- Sheng, V., & Ling, C. (2006). Thresholding for making classifiers cost-sensitive. In *Proc. aaii’06* (Vol. 1, p. 476–481).
- Stripling, E. (2018). Regularized empirical maximization framework for profit-driven model building. *Doctoral Thesis*.
- Stripling, E., vanden Broucke, S., Antonio, K., Baesens, B., & Snoeck, M. (2018). Profit maximizing logistic model for customer churn prediction using genetic algorithms. *Swarm and Evolutionary Computation*, 40, 116–130.
- Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). *Credit scoring and its applications*. SIAM.
- Ting, K. (1998). Inducing cost-sensitive trees via instance weighting. In *Proc. pkdd’98* (Vol. 1510, p. 139–147).
- Ting, K. (2000a). A comparative study of cost-sensitive boosting algorithms. In *Proc. icml’00* (p. 983–990).
- Ting, K. (2000b). An empirical study of metacost using boosting algorithms. In *Proc. ecml’00* (Vol. 1810, p. 413–425).
- Verbeke, W., Baesens, B., & Bravo, C. (2017). *Profit driven business analytics: A practitioner’s guide to transforming big data into added value*. John Wiley and Sons.
- Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, 218(1), 211–229.
- Verbraken, T., Bravo, C., Weber, R., & Baesens, B. (2014). Development and application of consumer credit scoring models using profit-based classification measures. *European Journal of Operational Research*, 238(2), 505–513.
- Verbraken, T., Verbeke, W., & Baesens, B. (2013). A novel profit maximizing metric for measuring classification performance of customer churn prediction models. *IEEE transactions on knowledge and data engineering*, 25(5), 961–973.
- Viola, P., & Jones, M. (2001). Fast and robust classification using asymmetric adaboost and a detector cascade. *Advances in Neural Information Processing Systems*, 14, 1311–1318.
- Wirth, R., & Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (pp. 29–39).
- Xia, Y., Liu, C., & Liu, N. (2017). Cost-sensitive boosted tree for loan evaluation in peer-to-

peer lending. *Electronic Commerce Research and Applications*, 24, 30–49.

Zadrozny, B., Langford, J., & Abe, N. (2003). Cost-sensitive learning by cost-proportionate example weighting. In *Proc. icdm'03* (p. 435-442).

Appendix A. Description of the Dataset

Table A1 lists and describes the features of the dataset.

Table A1. Description of the features of the dataset.

#	Name	Description
1.	Business channel	The company lends money using three distribution channels with significantly different default rates
2.	Loan amount	The amount lent to the customer (standardized value)
3.	Max gearing ratio	The maximum level of debt allowed for the customer (typically from 40% to 60% of the customer's income) (standardized value)
4.	Gearing coefficient	Percentage of the max gearing ratio, considering the loan amount (standardized value)
5.	Age	The age of the customer (standardized value)
6-13.	V1-V8	Several nominal features including information such as education, profession, employment, housing. The actual names and values are coded for privacy reasons
14.	Monthly income	Customer's monthly income (standardized value)
15.	Has FICO	Dummy variable indicating whether the customer has a FICO Score
16.	FICO score	FICO score value (from 300 to 850) (standardized value)
17.	Days late	Number of days the customer is late in repaying the loan
18.	Default 45	The target (binary) variable: 0 if days late is below 45 and 1 otherwise
19.	Expected profit	Expected profit - see Section 3.2 (standardized value)
20.	Expected loss	Expected loss - see Section 3.2 (standardized value)
21-23.	Test set 1..3	Dummy variables indicating the data partitioning used by the authors in their experiments (0 - train, 1 - test)