Technological and Economic Development of Economy
ISSN: 2029-4913 / eISSN: 2029-4921
2022 Volume 28 Issue 6: 1954–1990
https://doi.org/10.3846/tede.2022.17045

FROM CREDIT SCORING TO REGULATORY SCORING: COMPARING CREDIT SCORING MODELS FROM A REGULATORY PERSPECTIVE

Yufei XIA, Zijun LIAO, Jun XU, Yinguo LI*

Business School, Jiangsu Normal University, Xuzhou, PR China

Received 13 September 2021; accepted 08 April 2022

Abstract. Conventional credit scoring models evaluated by predictive accuracy or profitability typically serve the financial institutions and can hardly reflect their contribution on financial stability. To remedy this, we develop a novel regulatory scoring framework to quantify and compare the corresponding regulatory capital charge errors of credit scoring models. As an application of RegTech, the proposed framework considers the characteristic of example-dependence and cost-sensitivity in credit scoring, which is expected to enhance the ability of risk absorption of financial institutions and thus benefit the regulators. Validated on two real-world credit datasets, empirical results reveal that credit scoring models with good predictive accuracy or profitability do not necessarily provide low capital charge requirement error, which further highlights the importance of regulatory scoring framework. The family of gradient boosting decision tree (GBDT) provides significantly better average performance than industry benchmarks and deep multilayer perceptron network, especially when financial stability is the primary focus. To further examine the robustness of the proposed regulatory scoring, sampling techniques, cut-off value modification, and probability calibration are employed within the framework and the main conclusions hold in most cases. Furthermore, the analysis on the interpretability via TreeSHAP algorithm alleviates the concerns on transparency of GBDT-based models, and confirms the important roles of loan characteristics, borrowers' solvency and creditworthiness as powerful predictors in credit scoring. Finally, the managerial implications for both financial institutions and regulators are discussed.

Keywords: credit scoring, RegTech, regulatory scoring, probability of default, financial regulation, gradient boosting decision tree.

JEL Classification: G32, C53, C61.

Introduction

Credit scoring, defined as utilizing quantitative models to guide decision-making in retail credit products, is one of the most successful applications of operational research (Crook et al., 2007). Recent decades witness a rapid development of retail credit business: the con-

*Corresponding author. E-mail: *yinguo.li@jsnu.edu.cn*

sumer credit outstanding in U.S. reached 4,009.7 billion dollars in Q4 2018[1]. In China, consumer credit outstanding exceeded 37.8 billion RMB by the end of 2018 (People's Bank of China [PBC], 2019). These figures imply that effective tools are urgently required to make proper loan decision-makings. To discriminate between risky and non-risky applications, financial institutions have built a variety of internal credit scoring models that transform the problem into a binary classification task and predict the probability of default (PD) of every loan application. The regulation authority in each country also instructs financial institutions to develop credit risk models for measuring risk of their loan portfolio (Florez-Lopez & Ramon-Jeronimo, 2015).

In a thorough review of Lessmann et al. (2015), recent developments in credit scoring comprise three dimensions, namely (1) novel classification algorithms, (2) novel performance measures, and (3) statistical tests to compare performance on different algorithms. Among these research directions, designing novel classification algorithms becomes a dominating one partially due to the boom of computational power and emerging machine learning algorithms. On the one hand, data source of credit scoring has also been enlarged. Mobile phone data (L. Ma et al., 2018), social network (Óskarsdóttir et al., 2019), macroeconomic variables (Xia et al., 2021b), and narrative data (Xia et al., 2020a) provide supplementary source for modelling. On the other hand, advanced machine learning algorithm has been employed to develop scorecards (e.g., deep neural network and gradient boosting decision tree (GBDT) approaches). Furthermore, hybrid model which integrates different types of algorithms is found to provide promising results in several academic papers (Ala'raj & Abbod, 2016b; Pławiak et al., 2019; Xia et al., 2018a, 2020b).

Despite the important role of performance measure, it receives comparatively limited attention in existing literature. Performance measure evaluates the effectiveness of credit scoring models and supports the adoption of new models. Conventional performance measures (e.g., accuracy, type I and II error rates, area under the receiver operating characteristic curve (AUC), H measure, brier score) are, however, originated from statistics and are possibly inadequate to make proper evaluations. The potential reasons are as follows:

(1) Credit scoring is cost-sensitive (Bahnsen et al., 2014). Cost-sensitivity herein implies that the different costs of misclassifying risky and non-risky applications. Specifically, if credit scoring models make mistakes on non-risky borrowers and therefore reject their loan applications, financial institutions only bear minor opportunity costs. On the opposite, financial institutions are likely to suffer from considerable loss if lending to risky borrowers. Consequently, credit scoring models with similar accuracy may incur different performance in real-world application. To remedy this, cost-sensitive learning (CSL) has been recently employed in credit scoring domain (Shen et al., 2019; Xia et al., 2017a; Xiao et al., 2020);

(2) Credit scoring is example-dependent. Example-dependence means that the loss of extended credit, which affected by funded amount and interest rate, may vary among different applications. For example, ceteris paribus, for a loan application with specific PD, the higher exposure at default (EAD) is, the larger expected loss incurs. However, the conventional performance measures seldom consider example-dependence.

[1] https://www.federalreserve.gov/releases/g19/current/

(3) Conflicts of interest between financial institutions and regulatory authorities. Credit scoring are inherently designed for financial institutions, which mainly aims to enhance the profitability of loan portfolio (Finlay, 2010). As a result, profit scoring that evaluates credit scoring model from a profitable perspective is proposed (Crook et al., 2007). One major difference between profit scoring and conventional credit scoring lies on that profitability-related measures such as misclassification cost and expected return are employed, whereas these profitable measures can hardly meet the requirements of regulatory authorities since the regulators target at maintaining financial stability by implementing microprudential and macroprudential policies (Hanson et al., 2011). Moreover, strong evidences have shown that information asymmetry and conflicts exist between financial institutions and regulatory authorities (Duarte et al., 2008; Kadan et al., 2009). Thus, a considerable credit scoring model for financial institutions may be unsuitable for regulatory usage.

Since the 2008 Global Financial Crisis, the regulators have been long criticized as being over-optimistic on internal credit risk management tools and outsourcing most parts of financial regulation on the largest market participators (Moosa, 2010). The recent development of RegTech, a contraction of terms "regulatory" and "technology", spurs the trends of "know your data" by which more effective risk assessment tools are available for market participators (Baxter, 2016). RegTech, initially proposed by Financial Conduct Authority in 2016, refers to the use of information technology in the context of regulatory process such as compliance, reporting and monitoring (Arner et al., 2016). In RegTech era, regulators have the opportunity to build and evaluate credit scoring models from a regulatory perspective: automatic reporting systems decrease the cost of data collection. Furthermore, the applications of big data, cloud computing and distributed learning lower the computational cost.

This paper extends research filed of RegTech by developing a novel regulatory scoring framework. Concretely, the proposed framework is designed for financial institutions and regulators by quantifying and evaluating model performance by capital charge error, rather than the simple predictive accuracy nor profitability of portfolio. To calculate the error on capital requirement, asymptotic single risk factor (ASRF) model is employed. Industry benchmarks and advanced models are compared over a variety of performance measures using two real-world credit datasets. The comparison results demonstrate that model performance may vary over different performance measures. In other words, best credit scoring model in terms of predictive accuracy or profitability may lead to poor estimation on capital charge error. This conclusion remains robust under a variety of sensitivity analysis. The issue of interpretability is also analyzed via an explainable AI algorithm.

The main contribution of this paper is summarized as follows: first, a novel regulatory scoring framework is proposed. In this framework, the mean absolute error (MAE) and mean squared error (MSE) of capital requirement are calculated and compared. Moreover, asymmetric cost of regulatory charge error is further considered. By these means, the corresponding performance on financial stability of each credit scoring model can be evaluated from a regulatory perspective. To the best of our knowledge, no prior studies have built a nexus between credit scoring models and regulatory capital requirement. Second, a variety of performance measures is considered in comparison phase. Popular predictive accuracy measures, profitability-related measures, as well as the proposed capital requirement measures

are employed to evaluate classifiers. Few studies have made a comprehensive comparison on all the aforementioned measures. Finally, the proposed regulatory scoring is inherently an empirical model for RegTech. Although RegTech has attracted researchers' attention, current studies are mostly carried out from theoretical perspective, including discussing the concertation, potential applications, and effects on regulators and financial institutions (Anagnostopoulos, 2018; Baxter, 2016; Kavassalis et al., 2018). As far as we are concerned, no prior studies have built empirical RegTech models regarding credit risk assessment.

The remaining of this paper is organized as follows: in Section 1, we provide a literature review on modelling approaches and performance measures in credit scoring. Section 2 explains the conventional comparison methods and regulatory scoring framework. In Section 3, we introduce the setup of comparison and the results are analyzed in Section 4. Section 5 discusses managerial implications and finally we conclude the main results and summarize potential future research in the last section.

1. Literature review

1.1. Modelling approaches of credit scoring

Exploring new modelling approaches of credit scoring has recently become a research hotspot. The aim of this research stream is to apply various empirical models to predict PD as accurately as possible. The empirical models can be roughly divided into two subsets, namely statistical and machine learning methods. The former mainly consists of linear discriminant analysis (Altman, 1968), logistic regression (LR) (Wiginton, 1980), and survival analysis (Bellotti & Crook, 2009). Although statistical models usually have a strong assumption on data distribution, which may hinder their application in practice, they still become the mainstream of modelling approaches. Machine learning approaches of credit scoring mainly include decision tree (DT) (Bensic et al., 2005), support vector machine (SVM) (Yu et al., 2010), artificial neural network (ANN) (Yu et al., 2008), fuzzy set (Maldonado et al., 2020). Partially due to the ability to recognize non-linear dependence, machine learning methods typically provide more accurate predictions than statistical ones as suggested by comprehensive compressions in Lessmann et al. (2015) and Chen et al. (2016).

To further enhance model performances, hybrid approaches have recently received much attention (Yu et al., 2015). Common hybrid approaches integrate feature selection and classification (Ala'raj & Abbod, 2016b; Chi & Hsu, 2012), or combine multiple classifiers together (i.e., ensemble models). The famous no free lunch theorem (Wolpert & Macready, 1997) spurs the development of ensemble learning. Ensemble credit scoring models are mainly established by bagging, boosting or stacking algorithms (He et al., 2018; Xia et al., 2018a). Bagging artificial neural network (Tsai & Wu, 2008), random forests (RF) (Tang et al., 2019), GBDT (Xia et al., 2021a; Xia et al., 2017b), and heterogeneous ensemble models (Schotten & Morais, 2019) are typical applications of ensemble credit scoring models. The empirical results also demonstrate the advantages of ensemble models (Lessmann et al., 2015). The goal of profit-seeking motivates the research on profit scoring (Thomas, 2000), which aims to maximize the profit of financial institutions using scorecards. Since conventional credit scoring is puzzled by cost-sensitivity, data imbalance and example-dependence when reaching profit-seeking goal, profit scoring attempts to handle part of all these issues.

To deal with cost-sensitivity, CSL is mainly used in existing studies. CSL assigns different costs to different types of error and aims to train a classifier that minimizes total misclassification cost. Specifically, the misclassification cost of a default application is much larger than that of non-risky application for credit scoring. Ling and Sheng (2011) categorized CSL as direct and indirect ones. Regarding direct CSL, the loss function of classifier is modified to make cost-sensitive predictions. Prior studies have modified LR (Shen et al., 2019) and GBDT (Xia et al., 2017a) as solutions to direct CSL. In terms of indirect CSL, cut-off strategy is one of the most popular techniques applied in credit scoring. Researchers believe that selecting an optimal cut-off point (or acceptance threshold) can lead to minimization of misclassification costs or maximization of expected profit (Herasymovych et al., 2019; Papouskova & Hajek, 2019; Verbraken et al., 2014).

Data imbalance is quite severe in real-world credit datasets: it is common to find that default applications account for lower than 10% of samples in datasets (Marqués et al., 2013). As a result, conventional credit scoring models, which aim to minimize misclassification rate, tend to predict most applications as non-risky ones and thus, hinders the discriminative ability of models (Brown & Mues, 2012). To handle data imbalance, sampling techniques and reject inference methods have been commonly adopted. The rationale behind sampling techniques is to adjust the ratio of risky and non-risky applications to a balanced level. Concretely, over-sampling, under-sampling and Synthetic Minority Oversampling Technique (SMOTE) algorithms are commonly-used sampling methods (Marqués et al., 2013; Sun et al., 2018). When further exploring the reasons that data imbalance occur, one will find that some high-risky applications have been rejected before extending credit. The rejected loans have no outcomes (i.e., no labels) and therefore been discarded in most credit scoring studies. However, these unlabeled samples may contain valuable information for classification. Crook et al. (2007) even pointed out that the missing of rejected samples resulted in biased parameter estimation and ambiguous distribution of credit datasets. As a remedy, the potential status of rejected loans (i.e., reject inference) must be inferred and considered in modelling. Consequently, semi-supervised learning, an emerging machine learning field, provides insight to solve these problems. semi-supervised learning can enhance model performance using both labeled and unlabeled samples. In credit scoring domain, scholars have applied semi-supervised SVM (Li et al., 2017), semi-supervised GBDT (Xia, 2019) and semi-supervised heterogeneous ensemble model (Xiao et al., 2020).

Since loan applications usually vary in amount and interest rate, example-dependence is also an important property of credit scoring. Bahnsen and his colleagues are pioneers of building example-dependent credit scoring models (Bahnsen et al., 2014, 2015). The modified example-dependent LR and decision tree models assume a varying misclassification cost and achieve the best performances over benchmarks for all datasets.

1.2. Performance measures of credit scoring

Once credit scoring models are developed, the sequential issue lies on the evaluation on the models. Conventional credit scoring models are usually evaluated by their predictive accuracy. In general, predictive accuracy measures are split into three subtypes: discriminative ability measures (e.g., AUC and GINI coefficient), probability measures (brier score) and

label prediction measures (misclassification rate and accuracy). H measure (Hand, 2009), an advanced measure that fixes the inherent misclassification costs in AUC, is adopted by several recent research (Feng et al., 2019; Xia et al., 2018b; Xu et al., 2019). The aforementioned measures evaluate credit scoring models from a statistical perspective and focus primarily on the minimization of default rates, which only accomplish one of the goals of credit scoring (Eisenbeis, 1977).

The cost-sensitive and example-dependent characteristics of credit scoring trigger novel performance measures from profitability perspective. Misclassification cost (Lohmann & Ohliger, 2019) and expected return (Serrano-Cinca & Gutiérrez-Nieto, 2016) are two representative profitability measures. Relative to statistical performance measures, these profit-based measures enhance economic interpretability and cater to profit-seeking goal. Misclassification cost assigns varying costs for different types of error. Specifically, misclassifying risky borrowers is assigned to higher weights than non-risky borrowers. The expected misclassification cost of each borrower is therefore summed to acquire the total misclassification costs. The critical drawback of misclassification cost, however, lies on the fact that it is highly dependent on number of samples. Consequently, the misclassification costs of different datasets are incomparable. On the contrary, expected return is usually evaluated by a certain return rate, such as internal rate of return (IRR) (Serrano-Cinca & Gutiérrez-Nieto, 2016) or annualized rate of return (ARR) (Xia et al., 2017a). These return rates make comparison feasible among different datasets, but still reflect only the interest of financial institutions and consider little on the regulatory requirement.

2. Regulatory scoring: comparing credit scoring models from a regulatory perspective

In this section, we briefly introduce the conventional comparison methods of credit scoring models applied in academic research and financial institutions. Subsequently, we present the regulatory scoring comparison framework.

2.1. Conventional comparison methods of credit scoring

Given a credit dataset, credit scoring aims to train empirical models using the training set, which is composed of n_t samples, to make predications on the PDs of n_e samples in test set. In the first step, for a set of m credit scoring models indexed by $i = 1, 2, ..., m$, model i predicts on the j-th samples in test set, where $j = 1, 2, ..., n_e$. The predicted PD is denoted as $\widehat{PD}_{i,j}$ and the true value is represented as PD_j.

In the second step, the estimated PDs are transformed into label prediction in order to function as decision-making of financial institutions. Formally, for $\widehat{PD}_{i,j}$, it is assigned to different classes of loan decision, namely

$$\hat{c}_{i,j} = \begin{cases} 1, \text{ if } \widehat{PD}_{i,j} \geq \pi \\ 0, \text{ otherwise} \end{cases}, \tag{1}$$

where $\hat{c}_{i,j} = 1$ denotes rejecting the loan and $\hat{c}_{i,j} = 0$ implies accepting the application. π herein is the cut-off value (or threshold). In some prior studies (Ala'raj & Abbod, 2016a,

2016b), the threshold is manually determined as 0.5 since it is the default setting of classifiers when making label prediction. Some other researchers advocated an optimized cut-off value. Concretely, Bequé and Lessmann (2017) set the classification threshold as the fraction of non-default and default applications in training set. Herasymovych et al. (2019) transformed the optimization of cut-off value into a reinforcement learning issue.

Furthermore, it is worth mentioning that two additional practices, namely sampling and probability calibration, are probably considered in credit scoring modelling. As illustrated previously, sampling techniques are typically used to tackle data imbalance issues. Due to the comparatively high ratio of non-default samples, sampling techniques can adjust the ratio between risky and non-risky applications to a balanced level. As a result, sampling is often performed before training the models. On the opposite, probability calibration is performed after the predictions are made. Probability calibration aims to build a well-calibrated credit scoring model in which it provides PD forecasts that in line with empirical probabilities. For example, if a 10% empirical default rate is observed, a well-calibrated model is likely to predicted 10% of all loan applications as default.

In third step, the evaluation of credit scoring models is inherently performed by a certain loss function defined as

$$\mathcal{L}_i^p = \mathcal{L}\left(PD_j, \widehat{PD}_{i,j}\right) = \mathbb{E}\left(L\left(PD_j, \widehat{PD}_{i,j}\right)\right) \text{ or} \tag{2}$$

$$\mathcal{L}_i^c = \mathcal{L}\left(c_j, \hat{c}_{i,j}\right) = \mathbb{E}\left(L\left(c_j, \hat{c}_{i,j}\right)\right), \tag{3}$$

where $\mathcal{L}$ implies the expected loss, and $L(\cdot,\cdot)$ is an integrable loss function. The performance measures that generally considered in prior studies are accuracy, type I error rate, type II error rate, brier score, AUC, and H measure. Accuracy measures the ratio of correctly classified samples in the whole samples. For the i-th model, the accuracy is defined as

$$\text{Accuracy} = \frac{\sum_{j=1}^{n_e}\left[c_j = 1 \,|\, \hat{c}_{i,j} = 1\right] + \sum_{j=1}^{n_e}\left[c_j = 0 \,|\, \hat{c}_{i,j} = 0\right]}{n_e}, \tag{4}$$

where $[\cdot]$ is the Iverson bracket. Type I error and type II error rates evaluate the capability of models on predicting risky and non-risky loans, respectively. The two types of error rates are calculated as

$$\text{Type I error rate} = \frac{\sum_{j=1}^{n_e}\left[c_j = 1 \,|\, \hat{c}_{i,j} = 0\right]}{\sum_{j=1}^{n_e}\left[c_j = 1\right]}; \tag{5}$$

$$\text{Type II error rate} = \frac{\sum_{j=1}^{n_e}\left[c_j = 0 \,|\, \hat{c}_{i,j} = 1\right]}{\sum_{j=1}^{n_e}\left[c_j = 0\right]}. \tag{6}$$

The aforementioned measures are dependent on label prediction whereas brier score and AUC are computed based on probability prediction. Concretely, brier score evaluates the correctness of probability prediction and is defined as

$$\text{Brier score} = \frac{1}{n_e}\sum_{j=1}^{n_e}\left(\widehat{PD}_{i,j} - c_j\right)^2. \tag{7}$$

AUC measures the entire two-dimensional area under the receiver operating characteristic curve. Following Huang and Ling (2005) is calculated as follows for a binary classification:

$$AUC = \frac{S_{e0} - n_{e0}\left(n_{e0}+1\right)/2}{n_{e0}n_{e1}}, \tag{8}$$

where n_{e0} and n_{e1} denote the number of non-risky and risky loans in test set, respectively. $S_{e0} = \sum rank_j$ is the rank of probability predications of j-th default loans. Moreover, we have considered an extra evaluation measure, namely H measure (Hand, 2009), that overcomes the inherent drawback of inconsistent misclassification cost in AUC to assess the performance of models on discriminative capability. Since performance measures may reflect different aspects of predictive capability and vary among different datasets, multiple performance measures and datasets are typically used in existing literature. However, performance measures may overlap among different models or datasets. Significance tests (including parametric and non-parametric ones) are, therefore, performed to examine whether a certain model provide significantly better results than the others.

The conventional comparison method of credit scoring suffers from two major shortcomings. First, it lacks economic interpretability. For example, a model with higher accuracy does not necessarily brings more profit since the economic benefits are highly dependent on the predictive capability of risky applications and the selection of cut-off value. To enhance economic interpretability, profit-based metrics, such as misclassification cost and expected return, are further applied as performance measures. Misclassification cost assigns different costs to two types of error. Concretely, the loss function of misclassification cost is defined as

$$\text{Misclassification cost} = C \cdot \sum_{j=1}^{n_e}\left[c_j = 1 \,|\, \hat{c}_{i,j} = 0\right] + \sum_{j=1}^{n_e}\left[c_j = 0 \,|\, \hat{c}_{i,j} = 1\right], \tag{9}$$

where C is the cost parameter. Due to the relatively high cost of type I error, C is assume to be larger than 1 but it is a tough work to derive an accurate estimate on it. Thus, we follow the settings of German dataset in University of California Irvine (UCI) machine learning repository to determine $C = 5$. Regarding expected return, the corresponding return of credit scoring models are compared and ranked. Specifically, the loss function of expected return is calculated as

$$\text{Expected return} = \sum_{j=1}^{n_e} r_j \cdot \left[\hat{c}_{i,j} = 0\right], \tag{10}$$

where r_j is the ARR or IRR of the j-th loan. In this paper, r_j is measured as the ARR of the j-th loan.

The profit-based performance measures further lead to the second pitfall: they seldom consider the capital charge from the regulatory perspective. Such a drawback has also been debated in Hurlin et al. (2018), who focused on the comparison of LGD models. As shown in panel A of Figure 1, current credit scoring models are compared independently of the corresponding capital charge. Concretely, the best model in terms of predictive accuracy or profit is held to predict PD of every sample and therefore the capital charge can be computed using ASRF model given other risk parameters (e.g., EAD, LGD, maturity, etc.). The aban-

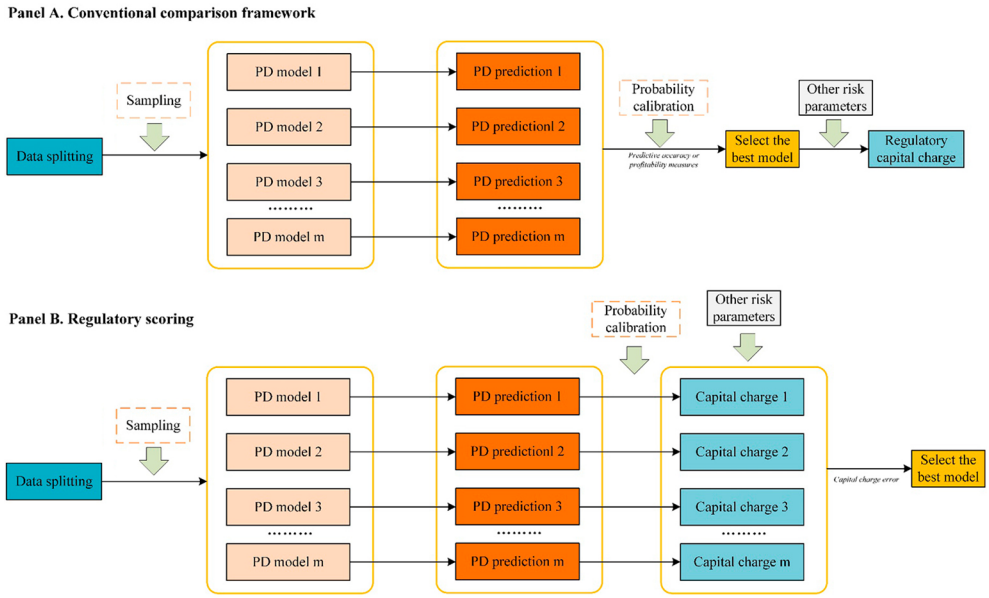

Figure 1. A comparison of convention comparison framework and the proposed regulatory scoring (the dashed boxes indicate the optional process)

doned models, however, have no opportunities to examine their corresponding performance on capital charge. The dashed edges of sampling and probability calibration imply that they are optional steps. Such a conventional comparison method may hinder the exploration of best models in terms of regulatory capital charge.

2.2. Regulatory scoring

In Basel III framework, PD, LGD, EAD, and maturity are four key risk parameters of ASRF model in calculating the capital charge for credit risk. Given that credit scoring models predict PDs of loan applications, the economic loss from the regulatory perspective is seldom considered when evaluating credit scoring models, whereas there is a clear causal effect of credit scoring models on financial stability. The global financial crisis has clearly showed that the ascent of credit can hinder asset quality and lead to financial fragility, which finally results in systemic risk when exogeneous shocks emerge (Gorton & Ordonez, 2014). Efficient credit scoring models, however, contribute to the management of credit risk, enhance financial institution's screening ability, and eventually improve asset quality (Demma, 2017). Consequently, we propose a new comparison framework (i.e., regulatory scoring) to compare model performance in predicting PDs to bridge the gap between financial stability and credit scoring. The panel B of Figure 1 illustrates the proposed regulatory scoring: the predictions of PD for each credit scoring model, along with other risk parameters (i.e., EAD, LGD, and maturity), are used to compute the capital charge. Then, the models are compared in terms of predictability on capital charge, rather than the predictive accuracy or profit. By these means, regulatory scoring takes cost-sensitivity and example-dependence into consideration and offers direct economic interpretability.

Analogy to loss functions described in Eqs (2) and (3), the capital charge loss function is defined as the difference between real capital charge and predicted capital charge, which is computed as

$$\mathcal{L}_i^{cc} = \mathcal{L}\left(cc_j, \widehat{cc}_{i,j}\right) = \mathbb{E}\left(L\left(cc_j, \widehat{cc}_{i,j}\right)\right), \tag{11}$$

where $L(\cdot,\cdot)$ is an integrable capital charge loss function. cc_j is the real capital charge for j-th observation in test set, and $\widehat{cc}_{i,j}$ denote the and predicted capital charge for j-th observation in test set provided by the i-th model. Before explaining the technical details on calculating capital charge using ASRF model (Gordy, 2003), we must introduce some basic concepts, namely expected loss (EL), unexpected loss (UL), and Value at Risk (VaR). Though it is hard to forecast the exact amount of potential loss that a financial institution may encounter in the following year, financial institutions can forecast the average level of credit loss, which is represented as EL. Financial institutions are urged to cover the EL by provisions. Following the specification of Basel Committee on Banking Supervision (BCBS, 2005), given a portfolio containing n credit exposures indexed by $j = 1, 2, ..., n$, the EL is calculated as follows

$$EL = \sum_{j=1}^{n} EAD_j \times LGD_j \times PD_j, \tag{12}$$

where EL herein represents expected loss in currency amounts. The values of LGD_j and PD_j typically range between 0 and 1. It is noteworthy that the PDs are estimated under normal economic scenario, whereas "economic-downturn" LGDs, which reflect the comparatively high values when encountering recession business cycle, are used to calculate EL, based on the specifications in BCBS (2005). Meanwhile, financial institutions may incur UL, which means that losses above expected level whereas the timing and severity cannot be predicted. Provisions are unable to fully absorb the UL. As a result, the UL should be covered by regulatory capital. However, UL cannot be calculated directly. Thus, we introduce VaR, which measures the risk of potential loss for the portfolio at a confidence level of α. By these means, ASRF model defines regulatory capital requirement as the VaR in excess of EL, namely

$$\text{Capital} = VaR(\alpha) - EL. \tag{13}$$

To compute the VaR of credit portfolio, the ASRF model assumes the value of portfolio is modeled with a single common factor (Z), which represents the systemic credit risk in the market.

$$R_j = \sqrt{\rho_j} Z + \sqrt{1-\rho_j}\,\epsilon, \tag{14}$$

where R_j is the value of the j-th credit in the portfolio, Z is the single systematic risk factor, ρ is the correlation coefficient. ϵ herein denotes the idiosyncratic or specific risk factor. According to BCBS (2005), ASRF model assumes Z and ϵ follow mutually independent standard normal distribution. R_j is therefore a standard normal variable. Although Gaussian distribution of risk factor maybe doubted as a strong assumption and some scholars strive to a generalized non-Gaussian ASRF model (Hoese & Huschens, 2013), it is not surprising to see that Gaussian ASRF model is still adopted by Basel Committee on Banking Supervision, partially due to the close-form solution and easy-to-implementation. Thus, we follow BCBS

(2005) to use a conventional Gaussian ASRF model in this paper. Under this model, default losses L can be calculated as follows based on the Merton's model (Merton, 1974):

$$L = EAD \times I \times LGD, \tag{15}$$

where I is the default indicator. Specifically, $I = 0$ if $R_j < \Phi_R^{-1}\left(PD_j\right)$, which means the value of credit portfolio has fallen below a certain threshold for default, and a value of 0 otherwise. $\Phi(\cdot)$ herein represents the probability density function of the standard normal distribution. The expected value of the default indicator I conditional on the common systemic risk factor Z is calculated as

$$E(I_j \mid Z) = \Phi_{\in}\left(\frac{\Phi_R^{-1}\left(PD_j\right) - \sqrt{\rho_j}Z}{\sqrt{1-\rho_j}}\right). \tag{16}$$

ASRF model further assumes that the portfolio is well diversified and has perfect granularity. The expected loss conditional on a value of the common factor is

$$E(L \mid Z) = \sum_{j=1}^{n} EAD_j \times LGD_j \times \Phi_{\in}\left(\frac{\Phi_R^{-1}\left(PD_j\right) - \sqrt{\rho_j}Z}{\sqrt{1-\rho_j}}\right). \tag{17}$$

The VaR of credit portfolio can therefore be directly computed as particular percentiles of the distribution of losses using the previous function:

$$VaR(\alpha) = \sum_{j=1}^{n} EAD_j \times LGD_j \times \Phi_{\in}\left(\frac{\Phi_R^{-1}\left(PD_j\right) - \sqrt{\rho_j}\Phi_R^{-1}(1-\alpha)}{\sqrt{1-\rho_j}}\right). \tag{18}$$

Recall that α herein indicates the significance level. Since regulatory capital requirement is defined as the VaR in excess of EL, we can derive the regulatory capital requirement for a given confidence level of α as follows

$$\text{capital}(\alpha) = \sum_{j=1}^{n} EAD_j \times LGD_j \times \delta\left(PD_j\right), \tag{19}$$

where $\delta\left(PD_j\right) = \left[\Phi_{\in}\left(\frac{\Phi_R^{-1}\left(PD_j\right) - \sqrt{\rho_j}\Phi_R^{-1}(1-\alpha)}{\sqrt{1-\rho_j}}\right) - PD_j\right]$. By determining a confidence level α = 99.9% and we can get the internal rating-based formula without maturity adjustment. The Basel Accord II suggested maturity adjustment depending on the type of exposure. Regarding the corporate, sovereign, and large financial institutions exposure, maturity adjustment $\gamma(M)$ is defined as

$$\gamma(M) = \frac{1 + (M - 2.5) \times b(PD)}{1 - 1.5 \times b(PD)}, \tag{20}$$

where the smoothed maturity adjustment $b(PD) = \left(0.11852 - 0.05478 log(PD)\right)^2$ and M indicates the maturity duration in years. Concerning retailing exposure, no maturity adjustment is performed, that is, $\gamma(M) = 1$.

Subsequently, we introduce the correlation function, which shows the dependence of the value of portfolio and the economy. The correlation function offers a simple method to capture an element of default correlation. The correlation differs for different types of credit exposure. For corporate and sovereign exposure, the correlation function ρ(*PD*) is defined as

$$\rho\left(PD\right)=0.12\times\frac{1-e^{-50PD}}{1-e^{-50}}+0.24\left(1-\frac{1-e^{-50PD}}{1-e^{-50}}\right), \tag{21}$$

where *e* herein is the natural logarithm. Regarding the correlation function of large financial institutions, a multiplier of 1.25 is applied to the corporate and sovereign correlation function. For small and medium enterprises, the correlation is adjusted as follows

$$\rho\left(PD\right)=0.12\times\frac{1-e^{-50PD}}{1-e^{-50}}+0.24\left(1-\frac{1-e^{-50PD}}{1-e^{-50}}\right)-0.04\left(1-\frac{\max\left(S-5,0\right)}{45}\right), \tag{22}$$

where *S* is the enterprise's annual sales turnover in millions and max (·,·)is the max function that returns the largest value in the two arguments. Regarding residential mortgage and revolving retail exposure, the correlation function is relatively simple, being a constant of 0.15 and 0.04, respectively. Since the credit datasets used in this paper belong to the retailing exposure, we therefore determine the value of maturity adjustment as 1 and the value of correlation function as 0.04 in this paper.

To summarize, the real capital charge cc_j and predicted capital charge $\widehat{cc}_{i,j}$ are defined as follows, respectively:

$$cc_j = EAD_j \times LGD_j \times \delta\left(PD_j\right)\times\gamma\left(M_j\right); \tag{23}$$

$$\widehat{cc}_{i,j} = EAD_j \times \widehat{LGD}_j \times \delta\left(\widehat{PD}_{i,j}\right)\times\gamma\left(M_j\right). \tag{24}$$

The two equations, namely Eqs (23) and (24), shows that the actual and predicted capital charge depends on EAD, LGD, PD, and maturity. Although the EADs and maturities are example-dependent, they remain constant given the *j*-th loan. The key factor affecting capital charge then centers on the LGD and PD, which must be determined to compare the credit scoring models from a regulatory perspective. Though abundant research has explored the determinants and the forecasting models of LGD, it is not the main focus of this paper. To make the performances of different credit scoring models comparable, we determine LGD_j as the real LGD of the *j*-th loan and $\widehat{LGD}_j$ as a constant (i.e., 0.3) in this paper. $\widehat{PD}_{i,j}$ is provided by the credit scoring model *i*, and PD_j denotes the real PD of the *j*-th loan, where PD_j = 0 implies non-default and PD_j = 1 indicates default.

The differences between cc_j and $\widehat{cc}_{i,j}$ can be defined by mean absolute error (MAE) or mean squared error (MSE). For example,

$$\text{Capital charge } MAE\ \mathcal{L}_i^{cc-MAE}=\frac{1}{n_e}\sum_{j=1}^{n_e}\left|cc_j-\widehat{cc}_{i,j}\right|; \tag{25}$$

$$\text{Capital charge } MSE\ \mathcal{L}_i^{cc-MSE}=\frac{1}{n_e}\sum_{j=1}^{n_e}\left(cc_j-\widehat{cc}_{i,j}\right)^2. \tag{26}$$

Moreover, the loss functions described in Eqs (25) and (26) do not consider the cost-sensitive characteristic. Since capital charge is employed to absorb UL, an under-estimation of it may incur instability of financial institutions whereas an over-estimation can only lead to a profit decline. The costs of under- and over-estimation are obviously different and we therefore develop an asymmetric cost (AC) of regulatory charge as follows:

$$\mathcal{L}_i^{cc-AC} = \frac{1}{n_e}\left[\theta \sum \left| cc_j - \widehat{cc}_{i,j} \right| \times I_{\left(cc_j \geq \widehat{cc}_{i,j}\right)} + \sum \left| cc_j - \widehat{cc}_{i,j} \right| \times I_{\left(cc_j < \widehat{cc}_{i,j}\right)}\right], \tag{27}$$

where $I(\cdot)$ is the indicator function. θ is the asymmetric cost parameter, which is calculated as the ratio of the under-estimation cost to the over-estimation cost. In this paper, we determine $\theta = 5$. The loss functions based on capital charge provide direct economic interpretability for regulators. Specifically, the MAE and MSE of capital charge measure the gap between corresponding capital charges of predicted PDs and real PDs. Models with low errors are preferred from a regulatory perspective. The AC of capital charge even considers the imbalanced costs of capital charge prediction. The regulators can adjust the asymmetric cost parameter to derive an overall cost of credit scoring model. As a result, the proposed loss functions are directly related to the concerns of regulators and have unit of currency. Thus, the proposed loss functions sharpen the decision-making tools of credit risk management.

3. Experimental setup

We aim to introduce the experimental setup, including the description of datasets, competing models, and comparison details in this section.

3.1. Dataset description

Data is obviously a barrier for research on credit scoring. The existing literature mainly used public credit datasets from UCI machine learning repository or Kaggle community. However, most of these datasets provide several features such as loan characteristics, borrowers' creditworthiness and historical record on payment, along with a binary status variable that represents whether the loan is default or not. Though public credit datasets make models proposed in different research comparable, they still suffer from two inherent pitfalls: first, the lack of information on EAD, interest rate and LGD hinders the calculation of expected return, which partially explains the relatively scarce research on profit scoring. The second drawback is related to the number of samples. Regarding the popular credit datasets, Japanese, German, and Australian datasets of UCI machine learning repository, the number of samples for all the datasets is no larger than 1000. The limited examples are far from the reality and thus may affect the effectiveness of credit scoring models.

The emerging peer-to-peer (P2P) lending provide us a supplementary data source for retailing loans. As a typical component of FinTech, P2P lending can match the demand and supply of money via online platforms. Due to the business is operated online, P2P lending usually incurs lower transaction cost relative to traditional financial institutions. Inspired by the spirit of the Internet, P2P lending is characterized by data transparency. Several mainstream P2P lending platforms even disclosed the real transaction record excluding privacy

online. Though P2P lending platforms are often regarded as information intermediaries, they may also affect the prudential regulation since security business model of P2P lending is adopted in U.S. In such a business model, the loans are initially issued by the commercial banks. The banks subsequently sell the loans to individual investors. As a result, the issuing banks are subject to regulations concerning capital charge.

We use two real-world datasets in this paper. The first one is derived from the transaction records of Lending club that issued between January 2009 and December 2012. After excluding the samples containing missing values, the whole dataset includes 91,335 samples (77,313 non-default and 14,022 default). The second dataset is acquired from loan transactions of Prosper ranging from January 2009 and December 2013. The Prosper dataset contains 11,230 loans, with 8,501 non-default contracts and 2,729 default ones. For each contract in both datasets, we can observe the loan status (default or non-default), loan characteristics (e.g., EAD, maturity, interest rate, and issuing date), borrowers' creditworthiness (FICO score, internal rating grade) and solvency (annual income, debt-to-income (DTI) ratio, revolving utilization rate and etc.). The summary statistics of the two datasets are shown in Tables 1 and 2, respectively. All the variables are used for modeling since this will provide an opportunity of a comprehensive analysis on SHapley Additive exPlanations (SHAP) value (Lundberg & Lee, 2017), a type of feature importance scores on prediction, for the whole feature set. We will elaborate this in Subsection 4.4.

Table 1. Summary statistics of Lending Club dataset

Feature	Type	All samples (N = 37968)				Non-default samples (N = 32590)				Default samples (N = 5378)			
		Min	Max	Mean	S.D.	Min	Max	Mean	S.D.	Min	Max	Mean	S.D.
EAD	Numerical	1000	35000	11462.98	7829.53	1000	35000	10938.10	7129.71	1000	35000	11880.10	7831.61
Term	Numerical	36	60	41.41	10.79	36	60	42.07	10.43	36	60	46.90	11.95
Interest rate (%)	Numerical	5.42	24.59	12.99	4.21	5.42	24.59	11.77	3.71	5.42	24.40	13.95	3.69
LC grade	Numerical	1	35	10.95	6.85	1	35	10.40	6.63	1	35	14.25	7.20
Employment length	Numerical	0	10	4.88	3.60	0	10	4.87	3.59	0	10	4.91	3.66
Annual income	Numerical	4000	600000	69115.58	63687.96	4000	600000	70241.31	65912.00	4080	1250000	62293.81	47494.68
Verification status	Categorial			0.86				0.85				0.92	
Delinquency	Numerical	0	30	13.39	6.67	0	30	13.28	6.68	0	30	14.07	6.58
DTI (%)	Numerical	0	1100	14.44	48.83	0	1100	14.03	48.02	0	800	16.92	53.41
FICO	Numerical	660	825	715.22	35.77	660	825	717.10	36.03	660	820	703.84	31.90
Inquire	Numerical	0	8	0.86	1.06	0	8	0.83	1.04	0	8	1.04	1.14

End of Table 1

Feature	Type	All samples (N = 37968)				Non-default samples (N = 32590)				Default samples (N = 5378)			
		Min	Max	Mean	S.D.	Min	Max	Mean	S.D.	Min	Max	Mean	S.D.
Revolving utilization (%)	Numerical	0	100	49	28	0	100	28.24	0.26	0	100	56.01	27.84
Total ACC	Numerical	3	90	22.18	11.40	3	90	22.30	11.40	3	74	21.45	11.41

Table 2. Summary statistics of Prosper dataset

	Type	All samples (N = 11230)				Non-default samples (N = 8501)				Default samples (N = 2729)			
		Min	Max	Mean	S.D.	Min	Max	Mean	S.D.	Min	Max	Mean	S.D.
Loan amount	Numerical	1000	35000	6373.13	4813.07	1000	35000	6400.25	4926.21	1000	25000	6288.66	4441.98
Term	Numerical	12	60	37.99	10.64	12	60	37.13	10.65	12	60	40.66	10.14
Interest rate	Numerical	0.04	0.36	0.22	0.08	0.04	0.35	0.21	0.09	0.06	0.36	0.26	0.07
Prosper Score	Numerical	1	11	6.27	2.30	1	11	6.53	2.30	1	10	5.48	2.12
Homeownership	Categorial			0.51				0.52				0.47	
Lending group	Categorial			0.05				0.06				0.04	
Listing category	Numerical	0	20	3.75	4.09	0	20	3.70	4.03	1	20	3.89	4.24
Employment length	Numerical	0	755	91.58	89.26	0	732	91.78	87.89	0	755	90.94	93.42
Credit score	Numerical	600	880	701.37	52.85	600	880	705.58	53.96	600	860	688.27	46.92
Current credit lines	Numerical	0	48	9.55	5.28	0	48	9.73	5.24	0	40	8.96	5.35
Total credit lines	Numerical	2	120	26.53	13.94	2	118	26.97	13.93	2	120	25.14	13.87
Revolving accounts	Numerical	0	35	6.40	4.23	0	33	6.52	4.18	0	35	6.01	4.35
Revolving payment	Numerical	0	5720	347.72	390.08	0	5720	354.22	393.52	0	3234	327.49	378.51
Inquiries-6m	Numerical	0	27	1.17	1.67	0	27	1.10	1.58	0	22	1.40	1.89
Total inquiries	Numerical	0	74	4.57	4.26	0	74	4.52	4.20	0	42	4.73	4.42
Current delinquency	Numerical	0	24	0.38	1.21	0	22	0.33	1.12	0	24	0.53	1.45
Bank card utilization	Numerical	0	1.26	0.52	0.33	0	1.26	0.51	0.33	0	1.23	0.52	0.34
Total trades	Numerical	1	118	22.76	12.20	1	102	23.26	12.21	1	118	21.19	12.03
Good trade rate	Numerical	0.16	1	0.90	0.13	0.16	1	0.90	0.13	0.22	1	0.88	0.14
Stated income	Numerical	0	483333.33	5536.98	7209.77	0	483333.33	5848.95	8064.11	0	34813.25	4565.17	3176.99

3.2. Competing models

We employ seven mainstream models in our regulatory scoring comparison, namely (1) LR, (2) RF, (3) GBDT, (4) extreme gradient boosting (XGBoost), (5) LightGBM, (6) CatBoost, and (7) deep multi-layer perceptron network (DMLP).

LR is a classical linear model for classification problems, which describes the PD of a single sample using a logistic function. LR has been considered as industry benchmark of credit scoring and widely used in academic literature (Bahnsen et al., 2014; Bensic et al., 2005; Shen et al., 2019). RF is a tree-based ensemble model combining bagging and random subspace algorithms. RF trains multiple decision trees using only parts of features and training samples, which enhances the bias slightly whereas decreasing variance meanwhile and thus, lead to an overall better model. RF is even advocated by Lessmann et al. (2015) as a new industry benchmark model. RF has been used in credit scoring domain by Malekipirbazari and Aksakalli (2015) and Ala'raj and Abbod (2016a), among others. GBDT is a member of boosting algorithms, which trains models in a sequential manner. Specifically, a variety of weak models (i.e., decision trees) are added to derive a strong model. As a typical ensemble model, GBDT is regarded as a promising algorithm with good generalization capability. This model has been adopted in He et al. (2018) for PD prediction. XGBoost, LightGBM and CatBoost are advanced versions of GBDT. In these models, engineering optimizations have been performed to enhance the efficiency and scalability of the prototype GBDT model. Xia, Liu, Li, and Liu (2017b), X. Ma et al. (2018), and Xia et al. (2020a) have illustrated the sound performance of advanced GBDT-based models in PD prediction. Moreover, the training speed is also accelerated by multi-thread processing and parallel learning. Recent development of deep learning inspires us to consider it as benchmark model in this paper. Following Gunnarsson et al. (2021), we employ two representative models of deep learning, namely a DMLP with three hidden layers (DMLP3) and a DMLP with five hidden layers (DMLP5).

3.3. Comparison details

An out-of-sample validation is applied to perform a careful comparison of the competing models. Following the work of Lessmann et al. (2015) and Shen et al. (2019), we adopt k-fold cross-validation since it ensures the stability of experimental results and makes full use of dataset. For k-fold cross-validation, the dataset is divided into k folds, namely, equal-sized groups of samples. The prediction is made using k−1 fold, and the remaining one is used as a test set. Specifically, this paper uses a 5-fold cross-validation and the experiment loops 50 times. The performance measures, including the conventional measures and the proposed regulatory charge ones, during 50 times experiments are averaged and reported. Since the regulatory capital charge measures are relatively new evaluation metrics, we further examine their robustness by exploring the effects of sampling techniques, cut-off value modification, and probability calibration.

GBDT-based techniques and DMLPs have several hyper-parameters that control the complexity of the model. Following Xia, Liu, Li, and Liu (2017b), we employ a Bayesian hyper-parameter optimization method which achieves a balance between efficiency and com-

putational cost. A 5-fold cross validation AUC is used as the fitness function of the optimization method. The Bayesian method is applied to determine the selection of several common hyper-parameters for GBDT, XGBoost, LightGBM, and CatBoost, namely maximum tree depth (D), number of iterations (N), subsampling rate (sub) and feature subsampling rate ($fsub$). The search spaces are set as $D \in \{2, 3, 4, 5, 6\}$, $N \in N+ [50, 300]$, $sub \in [0.6, 1]$, and $fsub \in [0.6, 1]$. The learning rate of the four types of models are preset as 0.05. For DMLPs, the learning rate is determined as 0.001, and the possible numbers of hidden neurons in each layer are determined as {5,10,15,20}, and the alternative drop rates are {0, 0.25,0.5}. The settings of other hyper-parameters not mentioned remain the default of the package.

Since the performance of competing models may overlap among different measures, one cannot easily determine a specific winner. For example, models that provide good accuracy may perform poorly in the capital charge errors. Therefore, a significance test is useful to determine whether a certain model significantly outperforms the others. In this paper, we employ a non-parametric significance test to compare model performance over different datasets and measures. Specifically, a rank-based Friedman's test is first utilized to examine if a significant difference exists among model performances. Once the null hypothesis of Friedman's test is rejected, a post-hoc test adjusted by Finner procedure is employed to make pair-wise comparison. The adjusted p-value of post-hoc test finally determines whether a significant difference exists.

4. Experimental results

4.1. Results of out-of-sample validation

Tables 3 and 4 reveal the results of out-of-sample validation for Lending club and Prosper datasets over eleven performance measures, consisting predictive accuracy, profitability, and capital charge errors. It is noteworthy that in original out-of-sample validation, the cut-off value remains the default setting (i.e., 0.5) in this subsection. We also examine another possible cut-off value in Subsection 4.3.2. From the two tables, we can derive the following conclusions:

Table 3. Results of out-of-sample validation for Lending club dataset

Model	Accuracy	Brier score	AUC	Type I error rate	Type II error rate	H measure	Misclassification cost	Expected return	Capital charge MAE	Capital charge MSE	AC of capital charge ($\alpha = 5$)
LR	**0.84654**	0.12327	0.67351	**0.00013**	0.99887	0.09678	70040.72	0.91107	0.015325	37.25534	2854.08398
RF	0.84559	0.12477	0.66584	0.00305	0.98898	0.08669	**69573.34**	0.90967	0.015581	39.30519	2855.06471
GBDT	0.84629	0.12240	0.68473	0.00148	0.99306	0.10829	69737.30	**0.91135**	0.015109	36.81905	**2818.46470**
XGBoost	0.84643	0.12231	0.68498	0.00037	0.99830	0.10875	70019.50	0.91086	0.015145	**36.72264**	2825.24873
LightGBM	0.84638	**0.12230**	**0.68542**	0.00071	0.99676	**0.10907**	69937.54	0.91089	**0.015105**	36.77669	2818.98086
CatBoost	0.84644	0.12232	0.68501	0.00042	0.99790	0.10872	69995.36	0.91109	0.015174	36.73894	2830.87713
DMLP3	0.83815	0.12423	0.66260	0.01592	0.95290	0.08374	69810.62	–0.25678	0.015267	36.99830	2882.68171
DMLP5	0.84022	0.12439	0.66064	0.02173	**0.93440**	0.08553	69941.50	–0.30412	0.015370	37.22686	2881.64460

Note: the best-performing model for each performance measure is highlighted in bold.

Table 4. Results of out-of-sample validation for Prosper dataset

Model	Accuracy	Brier score	AUC	Type I error rate	Type II error rate	H measure	Misclassification cost	Expected return	Capital charge MAE	Capital charge MSE	AC of capital charge ($\alpha = 5$)
LR	0.76153	0.16275	0.72630	0.03383	0.87593	0.16926	12239.68	–0.01687	0.07018	99.80793	1613.75796
RF	0.76910	0.16183	0.72789	0.03665	0.83599	0.17216	11718.66	0.21692	0.06961	97.91500	1617.59843
GBDT	0.77100	0.15846	0.74538	0.05135	**0.78237**	0.19684	**11112.04**	0.45110	0.06607	92.94996	1535.60947
XGBoost	**0.77200**	**0.15788**	**0.74700**	0.04290	0.80461	**0.19907**	11343.58	0.37221	0.06671	93.62930	1545.15975
LightGBM	0.77152	0.15813	0.74560	0.04555	0.79833	0.19739	11280.44	**0.38999**	**0.06600**	93.64532	**1535.28124**
CatBoost	0.76871	0.15890	0.74397	0.02649	0.86926	0.19421	12086.18	0.08692	0.06827	**92.66337**	1587.16814
DMLP3	0.75142	0.17787	0.63721	0.02694	0.93900	0.06742	13041.72	–0.00301	0.07733	105.64085	1841.24817
DMLP5	0.75096	0.18011	0.61024	**0.01878**	0.96633	0.04681	13345.22	–0.00357	0.07845	108.69811	1857.18981

Note: the best-performing model for each performance measure is highlighted in bold.

First, model's performances vary over different evaluation measures. Models with good predictive accuracy or profitability do not guarantee a sound performance on capital charge requirement error. Such a finding sheds light on the proposed regulatory scoring framework: the conventional comparison method and performance measures can only select the best PD model in terms of statistical accuracy or profit, rather than determine the optimal model from the perspective of regulation. As a result, the proposed regulatory scoring framework should be further considered when the stability of financial institutions becomes the main issue.

Second, although LR and RF has been regarded as industry benchmarks, their performances beat other models only in very limited metrics. Moreover, LR shows an imbalanced label prediction since it assigns most of test samples into non-default ones, which leads to a comparatively high accuracy and low type I error rate. The imbalanced prediction, however, is achieved at the cost of predictability of default loans. As a result, LR performs poorly in misclassification cost measure.

Third, GBDT-based models provide an overall better performance than industry benchmark. Concretely, they perform the best in most evaluation measures. Among GBDT-based models, XGBoost and LightGBM are two promising alternatives, outperforming the remaining models over several performance measures especially in terms of discriminative ability and capital charge errors. Such a result is in accordance with Xia, Liu, Li, and Liu (2017b) and X. Ma et al. (2018) and encourages further adoption of GBDT-based models in PD modelling.

Fourth, DMLPs provide poor performance on the two datasets and do not show competitiveness in the comparison. Concretely, they are inferior to other models in terms of accuracy, AUC, and H measure and only outperform LR on very limited evaluation metrics. This finding is consistent with those revealed in Gunnarsson et al. (2021). A possible explanation is that deep learning has shown great power on discovering intricate structures when given a large dataset, whereas the credit datasets used in this paper are typically far from a big dataset.

Finally, data imbalance diminishes the effectiveness of some performance measures. For highly imbalanced Lending club dataset, most models tend to predict all the samples as the

majority class (i.e., non-default). Consequently, the predictive accuracy measures of models are quite similar and close to the ratio of majority class samples to all samples. Under this circumstance, the PD models are invalid since they can rarely recognize default samples. Cut-off value modification and sampling techniques are, therefore, become alternative solutions to handle data imbalance issue and we will further report the effects of them in Subsection 4.3.

4.2. Results of significance test

In this subsection, we conduct a significance test to examine whether a specific competing model shown in Tables 3 and 4 significantly outperforms the other ones among the datasets and performance measures. We first rank the competing models according to their performance on each dataset and performance measure. The best-performing model is ranked as 1 and the worst model is sorted as 8. We therefore derive a sum of 22 ranking series (2 datasets × 11 evaluation measures). A Friedman's test is performed on the ranking series and the statistics of Friedman's test is 70.333, rejecting the null hypothesis at 99% confidence level. We therefore perform a post-hoc test and the critical difference is shown in Figure 2. The line segment in Figure 2 indicates the average rank of competing models among all the performance measures. The models within the bold lines exhibit no significant difference in average rank on the evaluation measures. From the figure we can draw that GBDT-based methods provide better overall performance, sweeping the top four places in terms of average rank. Concretely, LightGBM becomes the best-performing model in terms of average rank, significantly outperforming other non-GBDT-based methods. This is in line with the findings of X. Ma et al. (2018) and Xia, Yang, and Zhang (2018b). Among the GBDT-based models, no significant difference is found. However, even inferior GBDT-based method, namely CatBoost, still significantly outperforms current industry benchmark LR and RF, as well as DMLP variants. Thus, our finding may challenge the adoption of LR, RF, or DMLP as modeling approaches when taking all the evaluation measures into consideration.

One may be curious on whether there is statistically significant difference in the financial stability of the credit scoring models. To answer this, we conduct an extra non-parametric

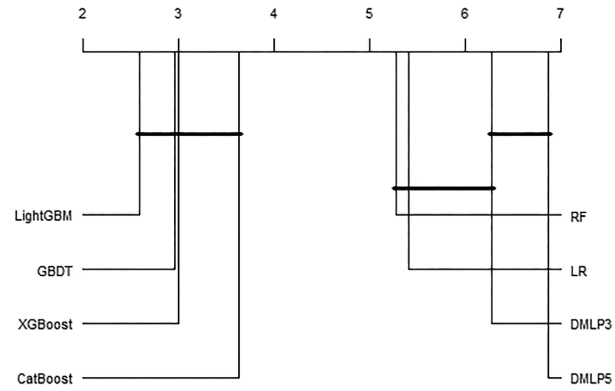

Figure 2. Critical difference plot of competing models over the average rank of all evaluation measures (significance level = 0.05). The models within the bold lines exhibit no significant difference in average rank

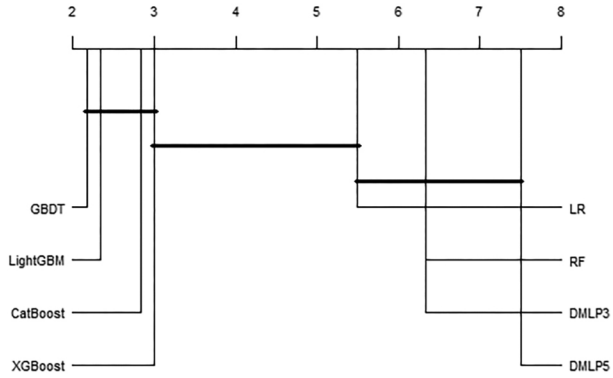

Figure 3. Critical difference plot of competing models over the average rank of capital charge error measures (significance level = 0.05). The models within the bold lines exhibit no significant difference in average rank

significance test that consider the performance of competing models on the three capital requirement error measures over the two datasets. The statistics of Friedman's test is 31.889, which rejects the null hypothesis of Friedman's test at 99% significance level and implies a significant difference among competing models on the capital charge errors. The result of post-hoc test is shown in Figure 3 and it is clearly illustrated that GBDT-based methods account for the top four places regarding the average rank of the three capital charge error measures. GBDT-based models except XGBoost significantly outperform the non-GBDT-based models as shown in Figure 3. This finding supports the adoption of GBDT-based models for modeling in credit scoring when financial stability is the main concern.

4.3. Sensitivity analysis

To examine whether the proposed comparison framework remains robust to varying settings, we conduct a series of sensitivity analyses. These sensitivity analyses concentrate on the effects of sampling techniques, cut-off value modification, and probability calibration.

4.3.1. The effects of sampling techniques

Due to the severe imbalance of the two datasets, we initially carry out a certain type of sampling technique before training the models to deal with imbalanced class issue. After considering the popularity of the aforementioned techniques used in academic studies such as Crone and Finlay (2012), Moscato et al. (2021) and Xiao et al. (2021), we select random under-sampling, random over-sampling, SMOTE, one-side selection (OSS), and adaptive synthetic (ADASYN) for sampling. The class distribution in the training set after processing is determined as 1:1 for positive/negative observations. The results of out-of-sample validation for the two datasets are displayed in Tables 5 and 6, respectively. In each panel of the two tables, different sampling techniques are employed. The following findings are presented from the two tables.

Table 5. Results of out-of-sample validation for Lending club dataset (with sampling techniques)

Model	Accuracy	Brier score	AUC	Type I error rate	Type II error rate	H measure	Misclassification cost	Expected return (%)	Capital charge MAE	Capital charge MSE	AC of capital charge ($\alpha = 5$)
Panel (a) Random over-sampling											
LR	0.76153	0.16275	0.72630	0.03383	0.87593	0.16926	12239.68	–0.01687	0.07018	99.80793	1613.75796
RF	0.76910	0.16183	0.72789	0.03665	0.83599	0.17216	11718.66	0.21692	0.06961	97.91500	1617.59843
GBDT	0.77100	0.15846	0.74538	0.05135	**0.78237**	0.19684	**11112.04**	**0.45110**	0.06607	92.94996	1535.60947
XGBoost	**0.77200**	**0.15788**	**0.74700**	0.04290	0.80461	**0.19907**	11343.58	0.37221	0.06671	93.62930	1545.15975
LightGBM	0.77152	0.15813	0.74560	0.04555	0.79833	0.19739	11280.44	0.38999	**0.06600**	93.64532	**1535.28124**
CatBoost	0.76871	0.15890	0.74397	0.02649	0.86926	0.19421	12086.18	0.08692	0.06827	**92.66337**	1587.16814
DMLP3	0.75142	0.17787	0.63721	0.02694	0.93900	0.06742	13041.72	0.03010	0.07733	105.64085	1841.24817
DMLP5	0.75096	0.18011	0.61024	**0.01878**	0.96633	0.04681	13345.22	0.03569	0.07845	108.69811	1857.18981
Panel (b) Random under-sampling											
LR	**0.63367**	0.22632	0.67589	**0.36314**	0.38389	0.09930	54989.78	0.79335	0.02516	93.20077	2847.55575
RF	0.62477	0.22642	0.67387	0.37547	0.37394	0.09746	55245.08	0.80837	0.02502	92.37506	2839.09342
GBDT	0.63234	**0.22054**	0.68374	0.36723	0.37000	0.10723	54332.50	**0.85183**	**0.02462**	**90.37718**	2807.35625
XGBoost	0.62692	0.22283	0.68440	0.37532	0.36071	0.10806	54306.80	0.84655	0.02478	91.32978	2815.01263
LightGBM	0.63123	0.22104	0.68442	0.36918	0.36652	0.10797	54239.02	0.85013	0.02463	90.53828	**2806.77313**
CatBoost	0.62622	0.22300	**0.68531**	0.37677	**0.35724**	**0.10891**	**54175.82**	0.83585	0.02480	91.44898	2815.83580
DMLP3	0.59425	0.23337	0.65712	0.41225	0.36993	0.08148	57807.40	0.53855	0.02543	94.79807	2882.61070
DMLP5	0.47043	0.24712	0.53693	0.55167	0.40776	0.01583	71239.00	0.14174	0.02664	97.03853	3076.19762
Panel (c) SMOTE											
LR	**0.63205**	0.22725	0.67242	**0.36436**	0.38779	0.09561	55357.56	0.78347	0.02523	93.71715	2856.62309
RF	0.61292	0.22867	0.66936	0.39066	0.36733	0.08944	55956.72	**0.84812**	0.02491	92.38322	2856.62309
GBDT	0.62114	0.22555	0.68265	0.38317	0.35507	0.10541	54518.06	0.84781	**0.02481**	92.04726	**2810.52555**
XGBoost	0.62225	**0.22518**	**0.68387**	0.38186	0.35510	**0.10718**	54418.80	0.84436	0.02486	92.15260	2815.75094
LightGBM	0.62175	0.22551	0.68313	0.38248	0.35493	0.10607	54454.76	0.84582	0.02482	**92.02716**	2811.68738
CatBoost	0.62133	0.22519	0.68372	0.38339	**0.35265**	0.10705	**54365.28**	0.82183	0.02496	92.23157	2827.00988
DMLP3	0.59393	0.23408	0.65680	0.41307	0.36753	0.08067	57702.90	0.15051	0.02561	95.84126	2897.33140
DMLP5	0.52465	0.24855	0.53135	0.47229	0.49224	0.01282	71025.04	0.06062	0.02677	97.60124	3091.58927
Panel (d) OSS											
LR	0.61151	0.22996	**0.67316**	0.39411	**0.35750**	**0.09605**	**55534.20**	0.72948	0.02546	93.55985	2883.16941
RF	0.83506	**0.13434**	0.65116	0.02687	0.92622	0.07218	67015.06	0.73416	**0.01791**	**49.69298**	2883.16941
GBDT	0.83346	0.14365	0.66088	0.03267	0.90466	0.08305	65951.00	0.71901	0.02019	58.18096	**2856.76680**
XGBoost	0.81675	0.15622	0.65416	0.06377	0.84203	0.07596	63964.94	**0.77885**	0.02138	64.58850	2877.91574
LightGBM	0.83715	0.14184	0.65913	**0.02410**	0.92783	0.08123	66913.58	0.74383	0.01997	57.04296	2861.96590
CatBoost	**0.83801**	0.14018	0.67003	0.02486	0.91814	0.09295	66292.22	0.73333	0.01997	56.37138	2913.32194
DMLP3	0.60110	0.23407	0.65039	0.39985	0.39365	0.07649	58512.54	0.58834	0.02554	93.84898	3049.73177
DMLP5	0.66768	0.24332	0.56297	0.27454	0.65088	0.02736	66858.80	0.34473	0.02640	95.86473	2857.47943
Panel (e) ADASYN											
LR	0.84602	0.12371	0.67197	0.00146	0.99493	0.09459	69867.18	0.81399	0.01576	39.21210	2837.09672
RF	0.84428	0.12584	0.66482	0.00775	0.97157	0.08628	68715.98	0.81380	0.01590	41.28582	2837.09672
GBDT	0.84578	0.12282	0.68329	0.00414	0.98173	0.10677	69149.44	0.84459	**0.01552**	38.75366	**2796.51585**
XGBoost	0.84610	**0.12265**	**0.68396**	0.00219	0.99038	**0.10762**	69604.70	0.86976	0.01555	38.65436	2800.79953
LightGBM	0.84592	0.12273	0.68367	0.00299	0.98713	0.10709	69439.02	0.86004	**0.01552**	38.75915	2796.92506
CatBoost	**0.84639**	0.12269	0.68342	**0.00091**	0.99555	0.10665	69868.40	**0.89028**	0.01557	**38.53463**	2809.83677
DMLP3	0.82265	0.12448	0.66262	0.05545	0.84947	0.08544	63843.70	0.84391	0.01569	38.74524	2860.20224
DMLP5	0.82216	0.12462	0.66159	0.05629	**0.84805**	0.08472	**63808.52**	0.85036	0.01574	38.92068	2863.37821

Note: the best-performing model for each performance measure is highlighted in bold.

Table 6. Results of out-of-sample validation for Prosper dataset (with sampling techniques)

Model	Accuracy	Brier score	AUC	Type I error rate	Type II error rate	H measure	Misclassification cost	Expected return (%)	Capital charge MAE	Capital charge MSE	AC of capital charge ($\alpha = 5$)
Panel (a) Random over-sampling											
LR	0.65087	0.20498	0.74408	0.37448	**0.27014**	0.19425	**6869.56**	1.87014	0.08655	157.33992	1385.78252
RF	**0.68214**	**0.19471**	0.74289	**0.30959**	0.34364	0.19279	7320.74	1.74793	**0.08189**	**146.12897**	1367.28315
GBDT	0.67465	0.19725	0.74356	0.32549	0.32491	0.19415	7200.36	1.79474	0.08221	149.03876	**1360.97987**
XGBoost	0.64684	0.21283	0.72648	0.37250	0.29293	0.16897	7163.62	1.63263	0.09042	174.37186	1425.70144
LightGBM	0.64939	0.21231	0.72511	0.36803	0.29632	0.16699	7171.94	1.58781	0.08923	170.73510	1417.92425
CatBoost	0.66580	0.20126	**0.74567**	0.34576	0.29820	**0.19692**	7008.22	**1.88873**	0.08433	154.58300	1364.56726
DMLP3	0.60635	0.22648	0.68339	0.42046	0.31014	0.11559	7806.10	1.34290	0.09421	183.80401	1502.54206
DMLP5	0.48384	0.24878	0.51385	0.53864	0.44612	0.00767	10666.28	-0.27689	0.10479	219.00539	1657.07035
Panel (b) Random under-sampling											
LR	0.64349	0.20842	0.74282	0.38743	**0.26020**	0.19207	**6843.96**	1.86934	0.08791	160.91132	1393.55164
RF	0.65451	0.21033	0.73843	**0.36354**	0.28925	0.18554	7037.26	1.84855	0.08519	**159.58311**	1355.60395
GBDT	0.65196	0.21006	0.73915	0.36919	0.28218	0.18646	6988.78	1.88907	**0.08508**	160.82452	**1354.24616**
XGBoost	0.64271	0.21402	0.72454	0.37862	0.29088	0.16610	7187.62	1.57662	0.09094	175.35277	1432.39294
LightGBM	0.64806	0.21136	0.73044	0.36954	0.29711	0.17266	7195.56	1.72700	0.08907	164.33337	1432.39294
CatBoost	**0.65486**	**0.20808**	**0.74331**	0.36667	0.27809	**0.19285**	6911.52	**1.90506**	0.08621	160.59106	1364.90836
DMLP3	0.57772	0.24410	0.63604	0.44128	0.36309	0.07530	8705.74	0.80913	0.10385	213.81809	1644.46533
DMLP5	0.50114	0.25002	0.50031	0.49661	0.50589	0.00166	11124.56	-0.42177	0.10519	220.60582	1661.66353
Panel (c) SMOTE											
LR	0.75440	0.16601	0.73937	0.11135	0.66382	0.18529	10004.38	0.75397	0.07613	114.24698	1509.58183
RF	0.76026	0.16222	0.73899	0.09293	0.69705	0.18610	10301.20	0.67879	0.07140	**105.22026**	1506.16669
GBDT	**0.76146**	**0.16209**	**0.73949**	**0.08846**	0.70602	**0.18676**	10385.72	0.66148	**0.07114**	105.73671	1501.46889
XGBoost	0.75834	0.16593	0.72919	0.09038	0.71292	0.17155	10496.12	0.59924	0.07399	110.54255	1425.70144
LightGBM	0.74932	0.16957	0.72489	0.11044	0.68753	0.16381	10320.24	0.60137	0.07668	117.30740	**1417.92425**
CatBoost	0.75513	0.16487	0.73884	0.11106	0.66169	0.18493	9972.80	**0.78846**	0.07410	111.93437	1493.99925
DMLP3	0.59364	0.22508	0.68349	0.44579	**0.28353**	0.11601	**7658.42**	0.01417	0.09327	180.82363	1496.54258
DMLP5	0.47465	0.24866	0.51973	0.55933	0.41952	0.01047	10479.16	-0.24222	0.10468	218.62487	1654.30793
Panel (d) OSS											
LR	0.62514	0.21963	0.71078	0.40475	**0.28176**	0.14711	**7285.34**	**1.54428**	0.09338	186.48648	**1462.60116**
RF	0.75023	0.17197	0.71280	0.09528	0.73102	0.14891	10784.78	0.32551	0.07906	123.06909	1471.58775
GBDT	0.75890	0.16549	0.72203	0.07949	0.74451	0.16360	10834.66	0.41141	0.07429	110.99084	1558.73547
XGBoost	0.75467	0.16813	0.72058	0.09029	0.72830	0.16088	10705.18	0.42249	0.07725	118.26805	1556.91286
LightGBM	**0.76037**	**0.16531**	0.72193	**0.07203**	0.76173	0.16331	11006.12	0.35702	**0.07418**	**110.98777**	1562.09058
CatBoost	0.75102	0.17141	**0.72392**	0.10690	0.69156	**0.16498**	10345.06	0.51102	0.08050	125.88974	1553.70333
DMLP3	0.59687	0.22736	0.67816	0.43589	0.30110	0.11077	7813.98	1.31284	0.09555	187.48108	1519.42719
DMLP5	0.50956	0.24570	0.55545	0.51269	0.42113	0.02800	10104.72	0.11844	0.10373	214.66067	1644.45547
Panel (e) ADASYN											
LR	0.75720	0.16629	0.71187	0.05105	0.84011	0.14891	11897.26	0.03598	0.07254	108.24459	1605.89209
RF	0.76634	0.16423	0.72014	0.04948	0.80738	0.16288	11437.36	0.22611	0.07204	104.59246	1783.76985
GBDT	0.76498	0.16280	0.72905	0.07144	**0.74460**	0.17453	**10767.30**	**0.48385**	**0.06926**	101.09694	**1544.48865**
XGBoost	0.76685	**0.16178**	**0.73118**	0.06202	0.76620	**0.17731**	10982.06	0.41054	0.06993	101.32340	1553.35530
LightGBM	0.76669	0.16227	0.72964	0.06522	0.75693	0.17549	10882.76	0.44226	0.06941	101.30620	1548.25881
CatBoost	**0.76789**	0.16230	0.72781	**0.04201**	0.82427	0.17238	11604.34	0.19519	0.07121	**100.59622**	1587.96309
DMLP3	0.74119	0.17767	0.63798	0.06477	0.86325	0.06822	12329.60	-0.18006	0.07900	109.82531	1820.84096
DMLP5	0.74129	0.18050	0.60753	0.04711	0.91787	0.04581	12924.86	-0.29399	0.08019	113.55700	1839.38305

Note: the best-performing model for each performance measure is highlighted in bold.

First, regarding the performance of PD models, LR and RF provide sound performance on label predictions in some cases. On the contrary, GBDT-based models tend to provide better probability predictions due to their comparatively good results on AUC and H measure. Importantly, GBDT-based models also outperform industry benchmarks in terms of capital charge errors in most cases, which is in line with the conclusions drawn in Subsection 4.1 and shows the robustness of the comparison results. DMLPs perform poorly on almost every evaluation measure, which is also consistent with the results shown in Subsection 4.1 and again shows the unsuitability of DMLPs as modeling approach for credit scoring.

Second, when making a vertical comparison, sampling techniques lead to a balanced distribution of error rates, which can be explained by the relatively balanced class distribution after processing. In other words, the predictability of potential default transactions is enhanced at the cost of accuracy. The balanced distribution of error rates further causes a lower misclassification cost than models without sampling techniques. Furthermore, credit scoring models with sampling technique perform poorly in terms of probability estimation and discriminative ability since they provide worse results than the original models on Brier score, AUC, and H measure. Moreover, sampling techniques contribute little on capital charge error measures. Such a phenomenon can be explained by the inherent drawbacks of sampling techniques. Concretely, to achieve the goal of balanced class in a dataset, over-sampling techniques, such as random over-sampling, SMOTE, and ADASYN, iteratively select samples from the minority class and is prone to noisy data. Under-sampling techniques, such as random under-sampling and OSS, drop the samples of majority class and may abandon valuable information during the process. Due to the negative effects of noisy data and scarce training samples, the models provide inferior predictions on PDs and thus achieve unsatisfactory performance in probability estimation and discriminative ability, as well as capital charge errors.

Finally, when making a horizontal comparison, a clear trade-off can be observed among different kinds of performance measures and datasets. Although it is hard to confirm a winning-solution for a certain performance measure since the results may vary in the two datasets, we can still make some simple recommendations on selecting the specific sampling techniques: when predictability of risky applications becomes the main issue, random under- and over-sampling may be effective. However, when capital charge errors are the focus and sampling techniques must be employed, SMOTE, OSS and ADASYN algorithms are reasonable choices.

4.3.2. The effects of cut-off value modification

Once the PD of each loan application is estimated, financial institutions must determine the cut-off value. Application with predicted PD above the cut-off value is regarded as risky loans and will not receive credit. To examine the effects of cut-off value on model performance, we follow Bequé and Lessmann (2017) to determine the cut-off value as the fraction of non-default and default applications in the training set. The results of competing models on Lending club and Prosper datasets after cut-off value modification are shown in Table 7. It is worth mentioning that a modification on cut-off value will only affect accuracy, type I and II error rates, misclassification cost and expected return. Thus, we only report these performance

Table 7. Results of out-of-sample validation for Lending club and Prosper dataset (cut-off value modification)

Model	Accuracy	Type I error rate	Type II error rate	Misclassification cost	Expected return (%)
Panel (a) Lending club dataset					
LR	0.78554	0.67351	0.12668	58764.36	0.83653
RF	0.78112	0.66581	0.13015	59707.02	**0.88515**
GBDT	0.78813	0.68473	0.12515	58052.76	0.87890
XGBoost	0.78820	0.68498	0.12511	58035.38	0.87121
LightGBM	**0.78834**	0.68542	0.12503	**57996.60**	0.87538
CatBoost	0.78821	0.68499	0.12510	58031.76	0.86469
DMLP3	0.77605	0.69619	**0.08343**	59502.60	0.71538
DMLP5	0.65872	**0.64134**	0.30161	62583.00	0.31988
Panel (b) Prosper dataset					
LR	0.73798	0.17307	0.53911	8827.44	1.15389
RF	0.74008	0.17168	0.53480	8756.76	1.23911
GBDT	0.74033	0.17151	0.53428	8748.24	**1.25587**
XGBoost	0.72840	0.17940	0.55883	9150.24	0.90120
LightGBM	0.73038	0.17928	0.55103	9042.88	1.05638
CatBoost	0.74035	0.17150	**0.53424**	**8747.64**	1.24063
DMLP3	**0.75095**	0.02970	0.93312	12984.90	0.68578
DMLP5	0.75076	**0.01763**	0.96995	13384.84	0.42873

Note: the best-performing model for each performance measure is highlighted in bold.

measures in Table 7. The empirical results demonstrate the profound effects of cut-off value modification on model performance. First, although accuracy decreases for all the models when compared with default settings. The type II error rate also decreases dramatically and thus benefits misclassification cost. Considering the easy-implementation and compatibility of cut-off value modification, financial institutions may integrate it with original PD models. Specifically, it is possible to make label prediction using models with cut-off value modification and provide PD prediction with the original model. Second, we can observe that the expected returns of models for Prosper dataset are relatively higher than those revealed in Table 4, whereas the results for Lending club dataset exhibits different pattern which can be explained by the example-dependence. Finally, the superiority of GBDT-based method is highlighted again. GBDT-based models become the best-performing model in most cases. Although DMLPs provide good performance on accuracy and Type I error rate for Prosper dataset, they provide comparatively worse misclassification cost and expected return.

4.3.3. The effects of probability calibration

Probability calibration is important at both micro- and macro-level. In Bequé et al. (2017), the researchers claimed that poor probability calibration led to higher regulatory capital charge but only Brier score is employed to evaluate the effectiveness of the different cali-

bration techniques. The proposed regulatory scoring framework provides an opportunity to examine the conclusion empirically. We consider three probability calibration methods, namely rescaling algorithm (Saerens et al., 2002), Platt scaling (Platt, 1999), and isotonic regression. The results of out-of-sample validation for the two datasets are shown in Tables 8 and 9 for Lending club and Prosper dataset, respectively. The results partially support the findings of Bequé et al. (2017): though probability calibration techniques can hardly improve performance on Brier score, we can observe a dramatic decrease on capital charge MAE and MSE after applying rescaling algorithm. However, similar phenomenon does not appear for Platt scaling and isotonic regression. Such a conclusion further encourages the financial institutions to put more attention on the post-modelling and consider the usage of probability calibration.

Table 8. Results of out-of-sample validation for Lending club dataset (with probability calibration)

Model	Accuracy	Brier score	AUC	Type I error rate	Type II error rate	H measure	Misclassification cost	Expected return (%)	Capital charge MAE	Capital charge MSE	AC of capital charge ($\alpha = 5$)
Panel (a) Rescaling algorithm											
LR	0.84654	0.14127	0.67351	0.00013	0.99887	0.09678	70040.72	0.91107	0.00950	30.55354	3014.20753
RF	0.84557	0.13986	0.66579	0.00307	0.98897	0.08664	69574.50	0.90958	0.00967	30.42315	3018.45326
GBDT	0.84629	0.14070	0.68473	0.00148	0.99306	0.10829	69737.30	0.91135	0.00944	30.38520	2997.82847
XGBoost	0.84643	0.14085	0.68498	0.00037	0.99830	0.10875	70019.50	0.91086	0.00944	30.42442	3002.73211
LightGBM	0.84638	0.14078	0.68542	0.00071	0.99676	0.10907	69937.54	0.91089	0.00944	30.40394	2999.94925
CatBoost	0.84643	0.14094	0.68500	0.00042	0.99795	0.10874	69999.06	0.91104	0.00945	30.46261	3005.51008
DMLP3	0.83776	0.12995	0.59751	0.02235	0.93356	0.00014	67179.94	0.34152	0.01612	37.01539	3128.15293
DMLP5	0.83955	0.12995	0.59786	0.01759	0.94816	0.00015	67835.30	0.21554	0.01612	37.01539	3128.15293
Panel (b) Platt scaling											
LR	0.84648	0.12418	0.67271	0.00000	1.00000	0.09607	70110.00	0.91073	0.01563	36.25147	2951.57440
RF	0.84648	0.12630	0.65545	0.00000	1.00000	0.08128	70110.00	0.91073	0.01638	37.77691	2955.33769
GBDT	0.84648	0.12596	0.67160	0.00000	1.00000	0.09516	70110.00	0.91073	0.01635	37.54886	3023.57015
XGBoost	0.84648	0.12536	0.66361	0.00000	1.00000	0.08694	70110.00	0.91073	0.01647	38.53351	2977.99303
LightGBM	0.84648	0.12596	0.66171	0.00000	1.00000	0.08641	70110.00	0.91073	0.01692	39.77797	2990.18325
CatBoost	0.84648	0.12513	0.66999	0.00000	1.00000	0.09497	70110.00	0.91073	0.01651	38.42644	2982.56385
DMLP3	0.84648	0.12895	0.58963	0.00000	1.00000	0.03729	70110.00	0.91073	0.01707	39.89636	3159.73029
DMLP5	0.84648	0.12966	0.57631	0.00000	1.00000	0.02993	70110.00	0.91073	0.01709	39.91439	3178.69865
Panel (c) Isotonic regression											
LR	0.84639	0.12348	0.67214	0.00059	0.99731	0.09540	69966.74	0.91089	0.01570	38.22815	2861.31100
RF	0.84647	0.12486	0.66698	0.00003	0.99989	0.08793	70104.88	0.91074	0.01603	37.37074	2861.31100
GBDT	0.84648	0.12413	0.67757	0.00001	0.99996	0.09960	70107.98	0.91074	0.01613	37.60759	2952.05645
XGBoost	0.84646	0.12416	0.67379	0.00004	0.99988	0.09536	70104.62	0.91070	0.01611	38.41297	2911.73153
LightGBM	0.84647	0.12429	0.67404	0.00002	0.99994	0.09517	70107.24	0.91071	0.01628	38.44589	2935.55935
CatBoost	0.84646	0.12365	0.67962	0.00003	0.99992	0.10179	70106.56	0.91070	0.01592	37.34962	2924.73405
DMLP3	0.84648	0.12890	0.59638	0.00000	1.00000	0.03427	70110.00	0.91073	0.01698	39.96128	3160.42866
DMLP5	0.84648	0.12971	0.56237	0.00000	1.00000	0.03001	70110.00	0.91073	0.01701	39.99521	3175.16890

Table 9. Results of out-of-sample validation for Prosper dataset (with probability calibration)

Model	Accuracy	Brier score	AUC	Type I error rate	Type II error rate	H measure	Misclassification cost	Expected return (%)	Capital charge MAE	Capital charge MSE	AC of capital charge ($\alpha = 5$)
Panel (a) Rescaling algorithm											
LR	0.76153	0.18790	0.72630	0.03383	0.87593	0.16926	12239.68	–0.01687	0.05502	78.89706	1876.65208
RF	0.76904	0.18479	0.72797	0.03653	0.83661	0.17231	11726.04	0.21190	0.05475	77.76578	1876.65208
GBDT	0.77100	0.18207	0.74538	0.05135	0.78237	0.19684	11112.04	0.45110	0.05251	76.06836	1802.12176
XGBoost	0.77200	0.18296	0.74700	0.04290	0.80461	0.19907	11343.58	0.37221	0.05300	76.41118	1819.39928
LightGBM	0.77152	0.18259	0.74560	0.04555	0.79833	0.19739	11280.44	0.38999	0.05265	76.27768	1810.61248
CatBoost	0.76891	0.18602	0.74386	0.02632	0.86893	0.19406	12080.38	0.08984	0.05385	77.02242	1860.73354
DMLP3	0.75219	0.17790	0.63701	0.02535	0.94075	0.06733	13052.14	–0.29912	0.07721	105.35975	1842.40238
DMLP5	0.75310	0.17995	0.61213	0.01484	0.96981	0.04830	13359.12	–0.35451	0.07856	108.86077	1855.46944
Panel (b) Platt scaling											
LR	0.76133	0.16280	0.72602	0.03274	0.88016	0.16889	12288.16	–0.03172	0.07027	99.73941	1616.49522
RF	0.77039	0.16141	0.73076	0.03910	0.82306	0.17732	11563.00	0.26773	0.06966	94.76807	1616.49522
GBDT	0.77219	0.15777	0.74727	0.03719	0.82160	0.20009	11526.86	0.31720	0.06795	93.95840	1562.00377
XGBoost	0.77275	0.15849	0.74798	0.04839	0.78440	0.20061	11114.46	0.45403	0.06870	94.35359	1580.46417
LightGBM	0.77240	0.15836	0.74762	0.04493	0.79663	0.20045	11251.98	0.40550	0.06906	94.70389	1585.42100
CatBoost	0.77052	0.15909	0.74495	0.04895	0.79184	0.19560	11220.82	0.35785	0.06815	92.06064	1587.34018
DMLP3	0.74853	0.17797	0.63399	0.03200	0.93514	0.06350	13032.00	–0.30606	0.07615	102.50805	1823.61049
DMLP5	0.75726	0.18003	0.61195	0.00306	0.98937	0.04841	13526.12	–0.38269	0.07796	107.10701	1847.70186
Panel (c) Isotonic regression											
LR	0.76184	0.16291	0.72509	0.03772	0.86254	0.16787	12090.02	0.03407	0.06946	99.69042	1602.38034
RF	0.76934	0.16125	0.72968	0.03304	0.84626	0.17575	11828.10	0.18119	0.06788	93.92102	1602.38034
GBDT	0.77218	0.15780	0.74660	0.03787	0.81954	0.19899	11504.54	0.32514	0.06676	93.93477	1543.14608
XGBoost	0.77234	0.15780	0.74714	0.04114	0.80867	0.19938	11384.06	0.36324	0.06630	93.66287	1537.71779
LightGBM	0.77209	0.15775	0.74696	0.03827	0.81866	0.19945	11495.96	0.32127	0.06670	94.26824	1543.87741
CatBoost	0.77050	0.15862	0.74418	0.04340	0.80919	0.19444	11410.36	0.29971	0.06613	91.49968	1550.52186
DMLP3	0.75183	0.17848	0.62301	0.02423	0.94577	0.05546	13111.00	–0.30935	0.07641	104.58151	1848.93525
DMLP5	0.75441	0.17965	0.61963	0.01659	0.95896	0.05184	13226.00	–0.32576	0.08009	112.25530	1842.02424

4.4. Analysis on the interpretability

Interpretability becomes an important issue when persuading senior managers in financial institutions to employ a credit scoring model based on a certain type of black-box model, such as SVM, RF, and GBDT. When we focus on the interpretability of black-box models, one may be curious on how the prediction is derived and which variables are crucial that lead a potential customer to be defaulted or not. Luckily, the SHAP approach proposed by Lundberg and Lee (2017) creates a unified framework to interpreted model prediction. SHAP combines definition of Shapley values in coalitional game theory and local surrogate model. The fundamental goal of SHAP is to calculate Shapley value of a feature by quantifying the average marginal contribution of it across the possible combinations of features. Since GBDT-based methods provide superior performance as described in the previous subsections, we mainly focus on the interpretability of GBDT-based models in this subsection. TreeSHAP algorithm (Lundberg et al., 2018), a variant of SHAP designed for tree-based ensemble model, is employed in this paper due to its low computational complexity. TreeSHAP can provide a Shapley value for each feature and features with large absolute Shapley values are regarded as important predictors. Following this idea, we calculate the mean value of the absolute Shapley

values for each feature across the dataset and visualize all the average absolute Shapley values in SHAP feature importance figure. Take XGBoost as an example. Figures 4 and 5 illustrate the 5-fold cross-validation SHAP feature importance of XGBoost under Lending club and Prosper datasets, respectively. From the two figures we can summarize the following findings.

First, predictors concerning loan characteristics (e.g., interest rate, loan amount, and term), borrowers' solvency (income and its verification status) and creditworthiness (e.g., internal or external credit rating) play an important role in making predictions. These predictors have been extensively applied in industry and academia and again highlights the usefulness of traditional data source for credit scoring modeling. Second, we clearly observe some

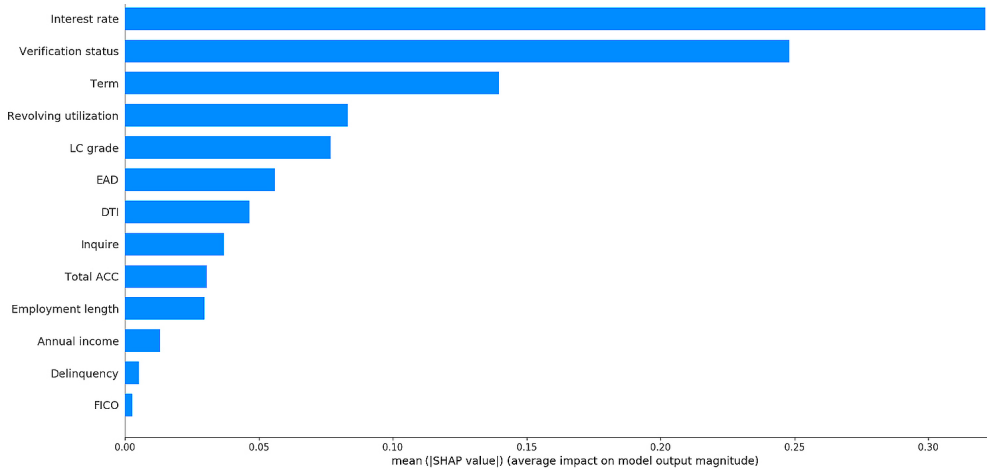

Figure 4. 5-fold cross-validation SHAP feature importance of XGBoost under Lending club dataset. The SHAP feature importance is measured by average absolute Shapley values and visualized in descending order

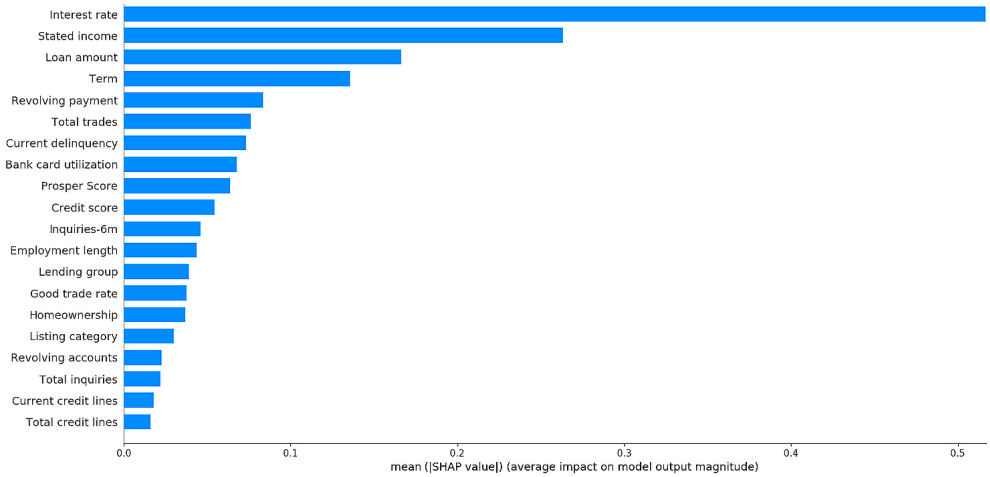

Figure 5. 5-fold cross-validation SHAP feature importance of XGBoost under Prosper dataset. The SHAP feature importance is measured by average absolute Shapley values and visualized in descending order

potential redundant variables, such as FICO and Delinquency in Lending club dataset and credit-lines-related variables in Prosper datasets since they score extremely low SHAP feature importance. However, the result is a bit counter-intuitive since FICO score is a well-reputed external credit scores and should be powerful in discriminating risky loan applications. The reasons may require further investigation. Finally, we assume that it is the collinearity that leads to the low SHAP feature importance score for some features (e.g., FICO), which is partially supported by the high correlation between interest rate and FICO score. The correlation may be due to the fact that interest rates in Lending club highly depend on the credit scores of borrowers.

5. Managerial implications

The regulatory scoring framework proposed in this paper provides a feasible way to quantify the necessary regulatory capital charge and compare the corresponding performance of different models. Furthermore, the proposed comparison framework allows the financial institutions to determine the optimal level of regulatory capital and gives the regulators an opportunity to evaluate internal credit scoring models from a regulatory perspective. Consequently, the managerial implications are organized from two-fold, namely the financial institutions and the regulatory agencies.

Due to the large amount of daily transactions, a minor improvement in credit scoring benefits the financial institutions considerably. As a result, predictive accuracy and profitability of credit scoring models have been emphasized for a long time. The discussion parts of several academic studies (Lessmann et al., 2015; Xia et al., 2018a) mainly debated on the paradox of complex but accurate models and managers' resistance. In this paper, we raise another problem: does accurate credit scoring models really contribute to financial stability? As we all know, financial institutions must meet the requirement of capital charge otherwise they will be penalized. The minimum requirement of capital charge, however, seems to impose a great burden on financial institutions and limits their business volume. Such an opinion is obviously short-termism since capital charge ensures the stability of an individual financial institution or even the whole financial system, which entails much more economic benefit than losing some business. Moreover, the loss incurred by capital charge can be alleviated if the financial institutions maintain an appropriate level of capital charge. The conventional comparison framework and performance measures, however, focus only on the predictive accuracy and profitability and do not reflect the corresponding performance on regulatory capital charge. Our empirical results even show a trade-off among predictive accuracy, profitability and regulatory capital charge. The proposed regulatory scoring provides a tool to quantify and compare the corresponding capital charge error of different credit scoring models. Using such a framework, the financial institutions can choose an ideal credit scoring system that achieves a good trade-off among different performance measures.

The proposed regulatory scoring can potentially benefit the regulators. The global credit market is booming, whereas there is a growing trend of tightening regulations on financial industry since the 2008 financial crisis. The ascent of credit may hinder asset quality and lead to financial fragility, which finally results in systemic risk when exogeneous shocks emerge

(Gorton & Ordonez, 2014). The proposed regulatory framework, however, bridge the gap between financial stability and credit scoring. The experimental results clearly show that accurate or profitable credit scoring models do not necessarily perform well within regulatory scoring framework. The comprehensive comparison of model performance supports the adoption of GBDT-based models when financial stability (represented by capital charge error measures) is the main focus of the regulators. Given that the credit scoring models are required to be transparent (Dastile et al., 2020), the concern of interpretability of GBDT-based models can be further alleviated by SHAP algorithms. In the RegTech era, regulators can even establish an early-warning model themselves by collecting data from multiple sources to handle the information asymmetry between regulators and financial institutions.

Conclusions

Credit scoring has become a research hotspot due to its considerable economic benefits. However, conventional comparison method mainly evaluates models in terms of predictive accuracy or profitability, which only reflects the pursuit of financial institutions and is not suitable for regulatory usage. In this paper, we develop regulatory scoring, a novel comparison framework that quantifies and compares the capital charge errors of different credit scoring models. The proposed regulatory scoring is validated on two real-world datasets and several popular competing models are introduced into the experiment. The empirical results demonstrate that models with good predictive accuracy or profitability do not guarantee a sound performance on capital charge requirement error. Such a finding sheds light on the proposed regulatory scoring framework. GBDT-based models significantly outperform industry benchmarks (i.e., LR and RF) and DMLPs when taking predictive accuracy, profitability and regulatory charge measures into consideration. Sensitivity analysis not only shows that superiority of GBDT-based models remains robust, but also present several interesting findings: though sampling techniques achieve a balanced distribution on two types of errors, they contribute little to profitability or capital charge errors. Cut-off value modification can better off label prediction and thus benefit misclassification cost and expected return. Regarding probability calibration, a simple rescaling algorithm leads to better capital charge errors and demonstrates the potential of probability calibration in regulatory scoring framework. The analysis on the interpretability via TreeSHAP alleviates the concerns on transparency of GBDT-based models, confirms the important roles of loan characteristics, borrowers' solvency and creditworthiness as predictors, and reveal some potential redundant variables in establishing credit scoring models.

This paper remains some limitations. First, the proposed capital charge errors were not considered as fitness function during building credit scoring model. Second, the other risk parameters, such as EAD and LGD were assumed to be fixed. Third, only a few common sampling techniques are considered in this paper and some complex ones lack further investigation. Finally, due to data limitation, the experiment is performed on two mainstream online lending platforms. The efficiency of our proposal may require further confirmation under other types of financial institutions.

Under the framework of regulatory scoring, future research can move in the following directions: first, seven types of simple but popular competing models are considered in this paper. It is interesting to examine the performance of other complex models on regulatory capital charge measures. Second, since only indirect CSL algorithm (i.e., cut-off value modification) is employed in this paper, one may introduce some direct CSL techniques and compare their performance with the prototype of this paper. Third, to simulate the practical modelling process, it would be interesting to estimate other risk parameters in future research within the framework of regulatory scoring. Finally, since the interpretability of credit scoring models is a critical consideration of regulators, comprehensive models accomplished either by self-explanation or rule extraction should be further emphasized in future research.

Acknowledgements

We are grateful for the National Natural Science Foundation of China (71874185; 72103082), the Project of Philosophy and Social Science Research in Colleges and Universities in Jiangsu Province (2020SJA1018), National Social Science Foundation of China (15BTJ033), and National Training Program of Innovation and Entrepreneurship for Undergraduates (202110320012). We wish to thank Prof. Saparauskas and anonymous reviewers for their constructive comments, which helped us to improve the manuscript.

Funding

This work was supported by the National Natural Science Foundation of China under Grant [71874185, 72103082]; the Project of Philosophy and Social Science Research in Colleges and Universities in Jiangsu Province under Grant [2020SJA1018]; National Social Science Foundation of China under Grant [15BTJ033]; and National Training Program of Innovation and Entrepreneurship for Undergraduates [202110320012].

Author contributions

Yufei Xia and Jun Xu conceived the study and were responsible for the design and development of the data analysis. Zijun Liao and Yinguo Li were responsible for data collection and analysis. Yufei Xia and Yinguo Li were responsible for data interpretation. Yufei Xia and Zijun Liao wrote the first draft of this article. Yufei Xia was responsible for the revision of this article.

Disclosure statement

The authors declare no competing financial, professional, or personal interests from other parties.

References

Ala'raj, M., & Abbod, M. F. (2016a). Classifiers consensus system approach for credit scoring. *Knowledge-Based Systems*, *104*, 89–105. https://doi.org/10.1016/j.knosys.2016.04.013

Ala'raj, M., & Abbod, M. F. (2016b). A new hybrid ensemble credit scoring model based on classifiers consensus system approach. *Expert Systems with Applications*, *64*, 36–55. https://doi.org/10.1016/j.eswa.2016.07.017

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23*(4), 589–609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

Anagnostopoulos, I. (2018). Fintech and regtech: Impact on regulators and banks. *Journal of Economics and Business*, *100*, 7–25. https://doi.org/10.1016/j.jeconbus.2018.07.003

Arner, D. W., Barberis, J., & Buckey, R. P. (2016). FinTech, RegTech, and the reconceptualization of financial regulation. *Northwestern Journal of International Law & Business*, *37*, 371. https://scholarlycommons.law.northwestern.edu/njilb/vol37/iss3/2

Bahnsen, A. C., Aouada, D., & Ottersten, B. (2014, December). Example-dependent cost-sensitive logistic regression for credit scoring. *Proceedings of 13th International Conference on Machine Learning and Applications* (*ICMLA*) (pp. 263–269). Detroit, MI, USA. IEEE. https://doi.org/10.1109/ICMLA.2014.48

Bahnsen, A. C., Aouada, D., & Ottersten, B. (2015). Example-dependent cost-sensitive decision trees. *Expert Systems with Applications*, *42*(19), 6609–6619. https://doi.org/10.1016/j.eswa.2015.04.042

Basel Committee on Banking Supervision. (2005). *An explanatory note on the Basel II IRB risk weight functions*. Bank for International Settlements.

Baxter, L. G. (2016). Adaptive financial regulation and RegTech: A concept article on realistic protection for victims of bank failures. *Duke Law Journal*, *66*(3), 567–604. https://scholarship.law.duke.edu/dlj/vol66/iss3/5

Bellotti, T., & Crook, J. (2009). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society*, *60*(12), 1699–1707. https://doi.org/10.1057/jors.2008.130

Bensic, M., Sarlija, N., & Zekic-Susac, M. (2005). Modelling small-business credit scoring by using logistic regression, neural networks and decision trees. *Intelligent Systems in Accounting, Finance & Management*, *13*(3), 133–150. https://doi.org/10.1002/isaf.261

Bequé, A., Coussement, K., Gayler, R., & Lessmann, S. (2017). Approaches for credit scorecard calibration: An empirical analysis. *Knowledge-Based Systems*, *134*, 213–227. https://doi.org/10.1016/j.knosys.2017.07.034

Bequé, A., & Lessmann, S. (2017). Extreme learning machines for credit scoring: An empirical evaluation. *Expert Systems with Applications*, *86*, 42–53. https://doi.org/10.1016/j.eswa.2017.05.050

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. https://doi.org/10.1016/j.eswa.2011.09.033

Chen, N., Ribeiro, B., & Chen, A. (2016). Financial credit risk assessment: A recent review. *Artificial Intelligence Review*, *45*(1), 1–23. https://doi.org/10.1007/s10462-015-9434-x

Chi, B.-W., & Hsu, C.-C. (2012). A hybrid approach to integrate genetic algorithm into dual scoring model in enhancing the performance of credit scoring model. *Expert Systems with Applications*, *39*(3), 2650–2661. https://doi.org/10.1016/j.eswa.2011.08.120

Crone, S. F., & Finlay, S. (2012). Instance sampling in credit scoring: An empirical study of sample size and balancing. *International Journal of Forecasting*, *28*(1), 224–238. https://doi.org/10.1016/j.ijforecast.2011.07.006

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. https://doi.org/10.1016/j.ejor.2006.09.100

Dastile, X., Celik, T., & Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, *91*, 106263. https://doi.org/10.1016/j.asoc.2020.106263

Demma, C. (2017). Credit scoring and the quality of business credit during the crisis. *Economic Notes: Review of Banking, Finance and Monetary Economics*, *46*(2), 269–306. https://doi.org/10.1111/ecno.12080

Duarte, J., Han, X., Harford, J., & Young, L. (2008). Information asymmetry, information dissemination and the effect of regulation FD on the cost of capital. *Journal of Financial Economics*, *87*(1), 24–44. https://doi.org/10.1016/j.jfineco.2006.12.005

Eisenbeis, R. A. (1977). Pitfalls in the application of discriminant analysis in business, finance, and economics. *The Journal of Finance*, *32*(3), 875–900. https://doi.org/10.2307/2326320

Feng, X., Xiao, Z., Zhong, B., Dong, Y., & Qiu, J. (2019). Dynamic weighted ensemble classification for credit scoring using Markov Chain. *Applied Intelligence*, *49*(2), 555–568. https://doi.org/10.1007/s10489-018-1253-8

Finlay, S. (2010). Credit scoring for profitability objectives. *European Journal of Operational Research*, *202*(2), 528–537. https://doi.org/10.1016/j.ejor.2009.05.025

Florez-Lopez, R., & Ramon-Jeronimo, J. M. (2015). Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. A correlated-adjusted decision forest proposal. *Expert Systems with Applications*, *42*(13), 5737–5753. https://doi.org/10.1016/j.eswa.2015.02.042

Gordy, M. B. (2003). A risk-factor model foundation for ratings-based bank capital rules. *Journal of Financial Intermediation*, *12*(3), 199–232. https://doi.org/10.1016/S1042-9573(03)00040-8

Gorton, G., & Ordonez, G. (2014). Collateral crises. *American Economic Review*, *104*(2), 343–378. https://doi.org/10.1257/aer.104.2.343

Gunnarsson, B. R., Vanden Broucke, S., Baesens, B., Óskarsdóttir, M., & Lemahieu, W. (2021). Deep learning for credit scoring: Do or don't? *European Journal of Operational Research*, *295*(1), 292–305. https://doi.org/10.1016/j.ejor.2021.03.006

Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning*, *77*(1), 103–123. https://doi.org/10.1007/s10994-009-5119-5

Hanson, S. G., Kashyap, A. K., & Stein, J. C. (2011). A macroprudential approach to financial regulation. *Journal of Economic Perspectives*, *25*(1), 3–28. https://doi.org/10.1257/jep.25.1.3

He, H., Zhang, W., & Zhang, S. (2018). A novel ensemble method for credit scoring: Adaption of different imbalance ratios. *Expert Systems with Applications*, *98*, 105–117. https://doi.org/10.1016/j.eswa.2018.01.012

Herasymovych, M., Märka, K., & Lukason, O. (2019). Using reinforcement learning to optimize the acceptance threshold of a credit scoring model. *Applied Soft Computing*, *84*, 105697. https://doi.org/10.1016/j.asoc.2019.105697

Hoese, S., & Huschens, S. (2013). Stochastic orders and non-Gaussian risk factor models. *Review of Managerial Science*, *7*(2), 99–140. https://doi.org/10.1007/s11846-011-0071-8

Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *17*(3), 299–310. https://doi.org/10.1109/TKDE.2005.50

Hurlin, C., Leymarie, J., & Patin, A. (2018). Loss functions for Loss Given Default model comparison. *European Journal of Operational Research*, *268*(1), 348–360. https://doi.org/10.1016/j.ejor.2018.01.020

Kadan, O., Madureira, L., Wang, R., & Zach, T. (2009). Conflicts of interest and stock recommendations: The effects of the global settlement and related regulations. *The Review of Financial Studies*,

22(10), 4189–4217. https://doi.org/10.1093/rfs/hhn109

Kavassalis, P., Stieber, H., Breymann, W., Saxton, K., & Gross, F. J. (2018). An innovative RegTech approach to financial risk monitoring and supervisory reporting. *The Journal of Risk Finance*, *19*(1), 39–55. https://doi.org/10.1108/JRF-07-2017-0111

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. https://doi.org/10.1016/j.ejor.2015.05.030

Li, Z., Tian, Y., Li, K., Zhou, F., & Yang, W. (2017). Reject inference in credit scoring using Semi-supervised Support Vector Machines. *Expert Systems with Applications*, *74*, 105–114. https://doi.org/10.1016/j.eswa.2017.01.011

Ling, C. X., & Sheng, V. S. (2011). Cost-sensitive learning. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 231–235): Springer. https://doi.org/10.1007/978-0-387-30164-8_181

Lohmann, C., & Ohliger, T. (2019). The total cost of misclassification in credit scoring: A comparison of generalized linear models and generalized additive models. *Journal of Forecasting*, *38*(5), 375–389. https://doi.org/10.1002/for.2545

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, *30*, 4765–4774.

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*

Ma, L., Zhao, X., Zhou, Z., & Liu, Y. (2018). A new aspect on P2P online lending default prediction using meta-level phone usage data in China. *Decision Support Systems*, *111*, 60–71. https://doi.org/10.1016/j.dss.2018.05.001

Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., & Niu, X. (2018). Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research and Applications*, *31*, 24–39. https://doi.org/10.1016/j.elerap.2018.08.002

Maldonado, S., Peters, G., & Weber, R. (2020). Credit scoring using three-way decisions with probabilistic rough sets. *Information Sciences*, *507*, 700–714. https://doi.org/10.1016/j.ins.2018.08.001

Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems with Applications*, *42*(10), 4621–4631. https://doi.org/10.1016/j.eswa.2015.02.001

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. https://doi.org/10.1057/jors.2012.120

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance*, *29*(2), 449–470. https://doi.org/10.2307/2978814

Moosa, I. A. (2010). Basel II as a casualty of the global financial crisis. *Journal of Banking Regulation*, *11*(2), 95–114. https://doi.org/10.1057/jbr.2010.2

Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications*, *165*, 113986. https://doi.org/10.1016/j.eswa.2020.113986

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. https://doi.org/10.1016/j.asoc.2018.10.004

Papouskova, M., & Hajek, P. (2019). Two-stage consumer credit risk modelling using heterogeneous ensemble learning. *Decision Support Systems*, *118*, 33–45. https://doi.org/10.1016/j.dss.2019.01.002

People's Bank of China. (2019). *China financial stability report 2019*. http://www.pbc.gov.cn/en/3688235/3688414/3710021/index.html

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. J. Smola, P. Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), *Advances in large-margin classifiers* (Vol. 10, pp. 61–74). MIT Press.

Pławiak, P., Abdar, M., & Acharya, U. R. (2019). Application of new deep genetic cascade ensemble of SVM classifiers to predict the Australian credit scoring. *Applied Soft Computing*, *84*, 105740. https://doi.org/10.1016/j.asoc.2019.105740

Saerens, M., Latinne, P., & Decaestecker, C. (2002). Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, *14*(1), 21–41. https://doi.org/10.1162/089976602753284446

Schotten, P. C., & Morais, D. C. (2019). A group decision model for credit granting in the financial market. *Financial Innovation*, *5*(1), 1–19. https://doi.org/10.1186/s40854-019-0126-4

Serrano-Cinca, C., & Gutiérrez-Nieto, B. (2016). The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. *Decision Support Systems*, *89*(2), 113–122. https://doi.org/10.1016/j.dss.2016.06.014

Shen, F., Wang, R., & Shen, Y. (2019). A cost-sensitive logistic regression credit scoring model based on multi-objective optimization approach. *Technological and Economic Development of Economy*, 1–25. https://doi.org/10.3846/tede.2019.11337

Sun, J., Lang, J., Fujita, H., & Li, H. (2018). Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. *Information Sciences*, *425*, 76–91. https://doi.org/10.1016/j.ins.2017.10.017

Tang, L., Cai, F., & Ouyang, Y. (2019). Applying a nonparametric random forest algorithm to assess the credit risk of the energy industry in China. *Technological Forecasting and Social Change*, *144*, 563–572. https://doi.org/10.1016/j.techfore.2018.03.007

Thomas, L. C. (2000). A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting*, *16*(2), 149–172. https://doi.org/10.1016/S0169-2070(00)00034-0

Tsai, C.-F., & Wu, J.-W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, *34*(4), 2639–2649. https://doi.org/10.1016/j.eswa.2007.05.019

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. https://doi.org/10.1016/j.ejor.2014.04.001

Wiginton, J. C. (1980). A note on the comparison of logit and discriminant models of consumer credit behavior. *Journal of Financial and Quantitative Analysis*, *15*(3), 757–770. https://doi.org/10.2307/2330408

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82. https://doi.org/10.1109/4235.585893

Xia, Y. (2019). A novel reject inference model using outlier detection and gradient boosting technique in peer-to-peer lending. *IEEE Access*, *7*, 92893–92907. https://doi.org/10.1109/ACCESS.2019.2927602

Xia, Y., He, L., Li, Y., Fu, Y., & Xu, Y. (2021a). A dynamic credit scoring model based on survival gradient boosting decision tree approach. *Technological and Economic Development of Economy*, *27*(1), 96–119. https://doi.org/10.3846/tede.2020.13997

Xia, Y., He, L., Li, Y., Liu, N., & Ding, Y. (2020a). Predicting loan default in peer-to-peer lending using narrative data. *Journal of Forecasting*, *39*(2), 260–280. https://doi.org/10.1002/for.2625

Xia, Y., Li, Y., He, L., Xu, Y., & Meng, Y. (2021b). Incorporating multilevel macroeconomic variables into credit scoring for online consumer lending. *Electronic Commerce Research and Applications*, *49*, 101095. https://doi.org/10.1016/j.elerap.2021.101095

Xia, Y., Liu, C., Da, B., & Xie, F. (2018a). A novel heterogeneous ensemble credit scoring model based on bstacking approach. *Expert Systems with Applications*, *93*, 182–199. https://doi.org/10.1016/j.eswa.2017.10.022

Xia, Y., Liu, C., & Liu, N. (2017a). Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending. *Electronic Commerce Research and Applications*, *24*, 30–49. https://doi.org/10.1016/j.elerap.2017.06.004

Xia, Y., Liu, C., Li, Y., & Liu, N. (2017b). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, *78*, 225–241. https://doi.org/10.1016/j.eswa.2017.02.017

Xia, Y., Yang, X., & Zhang, Y. (2018b). A rejection inference technique based on contrastive pessimistic likelihood estimation for P2P lending. *Electronic Commerce Research and Applications*, *30*, 111–124. https://doi.org/10.1016/j.elerap.2018.05.011

Xia, Y., Zhao, J., He, L., Li, Y., & Niu, M. (2020b). A novel tree-based dynamic heterogeneous ensemble method for credit scoring. *Expert Systems with Applications*, *159*, 113615. https://doi.org/10.1016/j.eswa.2020.113615

Xiao, J., Wang, Y., Chen, J., Xie, L., & Huang, J. (2021). Impact of resampling methods and classification models on the imbalanced credit scoring problems. *Information Sciences*, *569*, 508–526. https://doi.org/10.1016/j.ins.2021.05.029

Xiao, J., Zhou, X., Zhong, Y., Xie, L., Gu, X., & Liu, D. (2020). Cost-sensitive semi-supervised selective ensemble model for customer credit scoring. *Knowledge-Based Systems*, *189*, 105118. https://doi.org/10.1016/j.knosys.2019.105118

Xu, D., Zhang, X., & Feng, H. (2019). Generalized fuzzy soft sets theory-based novel hybrid ensemble credit scoring model. *International Journal of Finance & Economics*, *24*(2), 903–921. https://doi.org/10.1002/ijfe.1698

Yu, L., Li, X., Tang, L., Zhang, Z., & Kou, G. (2015). Social credit: a comprehensive literature review. *Financial Innovation*, *1*(1), 1–18. https://doi.org/10.1186/s40854-015-0005-6

Yu, L., Wang, S., & Lai, K. K. (2008). Credit risk assessment with a multistage neural network ensemble learning approach. *Expert Systems with Applications*, *34*(2), 1434–1444. https://doi.org/10.1016/j.eswa.2007.01.009

Yu, L., Yue, W., Wang, S., & Lai, K. K. (2010). Support vector machine based multiagent ensemble learning for credit risk evaluation. *Expert Systems with Applications*, *37*(2), 1351–1360. https://doi.org/10.1016/j.eswa.2009.06.083

APPENDIX

The lists of abbreviation and mathematical symbols can be found on supplementary online material.

Table A1. List of abbreviations

Abbreviation	Meaning
AC	Asymmetric cost
ADASYN	Adaptive synthetic
ANN	Artificial neural network
ARR	Annualized rate of return
AUC	Area under the receiver operating characteristic curve
ASRF	Asymptotic single risk factor
CSL	Cost-sensitive learning
DMLP	Deep multilayer perceptron network
DT	Decision tree
EAD	Exposure at default
EL	Expected loss
GBDT	Gradient boosting decision tree
IRR	Internal rate of return
LGD	Loss given default
LR	Logistic regression
MAE	Mean absolute error
MSE	Mean squared error
OSS	One-side selection
PD	Probability of default
P2P	Peer-to-peer
RF	Random forests
SHAP	Shapley additive explanations
SMOTE	Synthetic minority oversampling technique
SVM	Support vector machine
UCI	University of California Irvine
UL	Unexpected loss
VaR	Value at Risk

Table A2. List of mathematical symbols

Symbol	Description
C	Cost parameter for the calculation of misclassification cost
EAD_j	Exposure at default of the j-th observation in the test set
LGD_j	True loss given default of the j-th observation in the test set
$\widehat{LGD}_j$	Predicted loss given default of the j-th observation in the test set
M	Maturity duration in years
$\widehat{PD}_{i,j}$	Predicted probability of default on the j-th observation in test set (portfolio) provided by the i-th model
PD_j	True probability of default on the j-th observation in test set (portfolio)
R_j	Value of the j-th credit in the portfolio
S	Enterprise's annual sales turnover in millions
Z	Single systematic risk factor
$\hat{c}_{i,j}$	Predicted label on the j-th observation in test set provided by the i-th model
c_j	True label on the j-th observation in test set
$\widehat{cc}_{i,j}$	Predicted capital charge on the j-th observation in test set provided by the i-th model
cc_j	True capital charge on the j-th observation in test set
e	Natural logarithm
n_e	Number of observations in test set
n_{e0}	Number of non-risky observations in the test set
n_{e1}	Number of risky observations in the test set
n_t	Number of observations in training set
r_j	ARR of the j-th observation
α	Confidence level for the calculation of Value at Risk
γ	Maturity adjustment function
ϵ	Idiosyncratic risk factor
θ	Asymmetric cost parameter for the calculation of asymmetric cost of regulatory charge
π	Decision threshold for label prediction
ρ	Correlation coefficient
max $(\cdot,\cdot)$	Maximizing function
$I(\cdot)$	Indicator function
$L(\cdot,\cdot)$	Loss function
$\Phi(\cdot)$	Probability density function of the standard normal distribution