A novel generalised extreme value gradient boosting decision tree for the class imbalanced problem in credit scoring

Journal of the Operational Research Society

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tjor20

A novel generalised extreme value gradient boosting decision tree for the class imbalanced problem in credit scoring

Junfeng Zhang, Raffaella Calabrese & Yizhe Dong

To cite this article: Junfeng Zhang, Raffaella Calabrese & Yizhe Dong (01 Nov 2024): A novel generalised extreme value gradient boosting decision tree for the class imbalanced problem in credit scoring, Journal of the Operational Research Society, DOI: 10.1080/01605682.2024.2418882

To link to this article: https://doi.org/10.1080/01605682.2024.2418882

Published online: 01 Nov 2024.

JOURNAL OF THE OPERATIONAL RESEARCH SOCIETY
https://doi.org/10.1080/01605682.2024.2418882

ORIGINAL ARTICLE

OPEN ACCESS Check for updates

A novel generalised extreme value gradient boosting decision tree for the class imbalanced problem in credit scoring

Junfeng Zhang[a], Raffaella Calabrese[a,b] and Yizhe Dong[a]

[a]University of Edinburgh, Edinburgh, UK; [b]European University Institute Via della Badia dei Roccettini, Fiesole, Italy

ABSTRACT

The performance of credit scoring models can be compromised when dealing with imbalanced datasets, where the number of defaulted borrowers is significantly lower than that of non-defaulters. To address this challenge, we propose a gradient boosting decision tree with the generalised extreme value distribution model (GEV-GBDT). Our approach replaces the conventional symmetric logistic sigmoid function with the asymmetric cumulative distribution function of the GEV distribution as the activation function. We derive a novel loss function based on the maximum likelihood estimation of the GEV distribution within the boosting framework. This modification allows the model to focus more on the minority class by emphasising the tail of the response curve, and the shape parameter of the GEV distribution offers flexibility in controlling the model's emphasis on minority samples. We examine the performance of this approach using four real-life loan datasets. The empirical results show that the GEV-GBDT model achieves superior classification performance compared to other commonly used imbalanced learning methods, including the synthetic minority oversampling technique and the cost-sensitive framework. Furthermore, we conduct performance tests on several datasets with varying imbalance ratios and find that GEV-GBDT performs better on extremely imbalanced datasets.

ARTICLE HISTORY
Received 6 September 2023
Accepted 15 October 2024

KEYWORDS
Credit scoring; gradient boosting decision tree; generalised extreme value distribution; imbalanced sample

1. Introduction

Regulatory reforms and changes have been extensively pursued by regulators since the onset of the global financial crisis in 2008, including the enhancement of the Basel regulatory framework and introduction of the International Financial Reporting Standard 9 (IFRS 9), to strengthen the risk management practices of various kinds of financial institutions. These changes highlight the importance of the computation, prediction, and reporting of credit losses. However, estimating and predicting the probability of default remains a challenging and complex process. Credit scoring serves as a key instrument for financial institutions used to facilitate credit risk assessment and distinguish good from bad loan applicants. Given the significant impact that even a small improvement in default prediction accuracy can have on the profitability and risk level of financial institutions (Hand and Henley, 1997; Mushava and Murray, 2022), many studies have focused on developing novel credit scoring models to improve predictive performance (Chen et al., 2023; Gunnarsson et al., 2021; Jiang et al., 2019; Medina-Olivares et al., 2022; Shi et al., 2024).

In the past, conventional statistical linear regression was extensively used in credit scoring applications. However, in the era of big data and artificial intelligence, financial institutions have increasingly turned to advanced machine learning (ML) techniques such as support vector machines, tree-based learners, clustering algorithms and deep neural networks to assess the probability of loan default among applicants (Lagna and Ravishankar, 2022). Numerous studies have demonstrated that ML-based credit scoring models can leverage the non-linear relationships within the data with great computing efficiency to provide more accurate predictions and better overall performance than standard statistical techniques (Dastile et al., 2020). However, many of them do not adequately consider the impact of data complexities, including the lack of data representativeness (selection bias or survival bias), dataset shifts, noisy data, and other factors (Clarke, 2016). Particularly, in the classification tasks, class overlapping and severe class distribution skews can make the modelling work harder.

In classification problems, the presence of imbalanced datasets is a common scenario, where one class (the majority class) significantly outnumbers the remaining minority classes. This situation is particularly prominent in credit scoring, as the majority of loan applicants have a good credit record with

CONTACT Raffaella Calabrese raffaella.calabrese@ed.ac.uk; raffaella.calabrese@eui.eu Business School, University of Edinburgh, 29 Buccleuch Place, EH8 9JS, Edinburgh, UK; European University Institute, Via della Badia dei Roccettini, Fiesole, Italy.

sufficient debt-paying ability (the majority class), while only a small proportion of applicants are unable (or unwilling) to repay their loans (the minority class). The complex characteristics inherent in imbalanced datasets impose challenges when conducting data analytics in real-world applications. Most standard learning methods designed for classification assume an equal distribution of classes in the training samples. Therefore, the imbalanced classification problem leads to a credit scoring model that produces biased estimates and tends to lean towards the major label, making it difficult to accurately predict observations from the minor class (defaults) correctly (Krawczyk et al., 2014; López et al., 2013).

For lending institutions, the cost incurred from misclassifying bad applicants is significantly higher than the loss of revenue from erroneously rejecting good ones (Bahnsen et al., 2015). As a result, both industry and academia have proposed various approaches over the past few decades to address the imbalanced classification problem, aiming to reduce the risk of misclassification and limit economic losses for banks. Two commonly used approaches are the cost-sensitive framework and the resampling method. The first includes a misclassification cost matrix (or cost ratio) that represents the costs related to the misclassification of the various data classes (Liu et al., 2022b; Ren et al., 2022). The latter alters the class distribution in the training data so that the model is trained with balanced data (Brown and Mues, 2012; Haixiang et al., 2017).

Nevertheless, each approach faces unique challenges when used in real-world applications. For example, the cost-sensitive approach requires expert knowledge and an appropriate method for quantifying costs, which often requires additional data and a more complicated deployment process. On the other hand, the resampling approach does not effectively increase the variety of the representation of default cases and could potentially lead to overfitting or underfitting of classifiers, rather than improving model performance (He and Garcia, 2009).

In the field of credit scoring and credit risk assessment, it is crucial to acknowledge that the available data is often insufficient to accurately describe the minority class. To better address the challenges posed by imbalanced data and to effectively isolate rare class events, this paper introduces a novel gradient-boosting decision tree (GBDT) model, which leverages the distribution of generalised extreme values (GEV), referred to as GEV-GBDT. The common link functions used in the class of generalised linear models are symmetric. They can effectively model observations with balanced binary response outcomes by ensuring that the predicted probabilities mirror the likelihood of an event occurring and not occurring at the same rate. However, these symmetric link functions may be inappropriate for imbalanced datasets as they tend to underestimate the probability of rare events (La Rocca et al., 2023). To address this issue, it is suggested that asymmetric link functions are used to fit the dataset with skewed class proportions (Calabrese and Osmetti, 2013). Therefore, in this study, we adopt an asymmetric or skewed link function based on the GEV distributions. The GEV distribution has been widely applied in various fields such as finance, climatology and computer vision. Using the GEV helps the model focus on the tail of the response curve for positive rare events. Thus, this approach effectively overcomes the limitations associated with the symmetry property of logit and probit link functions and enables more accurate modelling of skewed datasets. A recent study by Mushava and Murray (2022) proposed an extension to the extreme gradient boosting (XGBoost) algorithm for learning class-imbalanced datasets. However, our study differs in its methodology: the aforementioned study solely modified the link function in XGBoost and used focal-loss as their loss function, while we modify both the loss and the link functions to more effectively address imbalanced learning. Specifically, we use the cumulative distribution function of the GEV distribution as the activation function, replacing the logistic Sigmoid function, and correspondingly derive the loss function based on the maximum likelihood estimation of GEV within XGBoost. The incorporation of the GEV distribution encourages the classifier to place greater emphasis on minority samples (Wang and Dey, 2010). By controlling the shape parameter of the GEV distribution, our model offers users the flexibility to determine the extent to which the model leans towards minority samples, thereby enabling improved handling of imbalanced data.

To authenticate the efficacy of the model we propose for predicting imbalanced credit defaults, we apply the model to four unique real loan datasets associated with different levels of imbalanced ratios. The loans are granted by different types of lending institutions for individuals, and micro, small or medium enterprises. The results demonstrate that our proposed GEV-GBDT model exhibits superior predictive performance compared to benchmark models, indicating its effectiveness in handling imbalanced datasets.

The rest of the paper is organised as follows. Section 2 provides the literature review on credit scoring, with a particular focus on the use of tree-based learners in this domain and the issue of imbalanced data. Section 3 presents the methodology employed in the study, offering a step-by-step explanation of the mathematical derivations used in the development of the proposed model. Section 4

describes the data used in the study. Section 5 presents the experimental results, wich are divided into three parts. The first subsection evaluates the classification performance of the proposed GEV-GBDT on all four datasets, comparing it to other commonly used benchmark models. The second subsection examines the performance of GEV-GBDT on datasets with different levels of class imbalance, offering insights into parameter selection and demonstrating how GEV-GBDT outperforms the benchmark model. The third subsection discusses how GEV-GBDT works with imbalanced data. Finally, the conclusion of the study is presented in Section 6

2. Literature review

2.1. Default prediction and GBDT

For many years, statistical methods such as linear discriminant analysis and logistic regression have served as the established benchmarks for default prediction (Crook et al., 2007). However, these methods often fall short due to their reliance on strict statistical assumptions and their limited capacity to model non-linear relationships. Recently, ML algorithms have demonstrated superior accuracy in predictive tasks compared to statistical models. This improvement is largely attributed to their superior capabilities for generalisation and exploiting non-linear relationships among variables (Baesens et al., 2003; Dastile et al., 2020; Lessmann et al., 2015). ML models are now widely used in credit risk assessment (Cao and Zhai, 2022; Dastile et al., 2020). The application of ML models has proved particularly beneficial in areas such as small and medium enterprises (SMEs) and peer-to-peer (P2P) lending markets, due to the unstructured nature and high dimensionality of applicant data (Guo et al., 2016; Jiang et al., 2018; Mezei et al., 2018; Papouskova and Hajek, 2019; Zhu et al., 2019).

The commonly used ML techniques in credit scoring include the K nearest neighbour, decision tree, support vector machine, artificial neural network, random forest and boosting methods (Dastile et al., 2020). Among them, GBDTs and their variants are particularly popular because of their superior classification performance. GBDTs, proposed by Friedman (2001), use multiple decision trees that are sequentially built, with each tree learning from the residual errors of the previous trees and summing up the results of all learners. Recent research has proposed several modifications to GBDTs to address class imbalance issues. GBDTs and their extensions have demonstrated promising performance in identifying rare classes (the defaulters in credit scoring datasets) for credit risk analysis (Dumitrescu et al., 2022; Liu et al., 2022a). One of the most popular techniques is to combine GBDTs with the cost-sensitive framework, where the weights are defined as the real misclassification costs. For example, Xia et al. (2017) developed a cost-sensitive boosted tree model for loan credit scoring by incorporating a cost-sensitive link function with XGBoost to enhance its ability to identify potential default borrowers. The study finds that the combination of direct cost-sensitive methods and XGBoost outperforms existing individual cost-sensitive evaluation models. Li et al. (2021) took the instance-based misclassification cost as the sample weight to force the model to lean towards positive samples. This approach assigns a higher weight (as the coefficient in the regular loss function) to the minority positive class. Liu et al. (2022b) proposed a focal-aware cost-sensitive light gradient boosting machine (LightGBM-focal) for credit scoring. They introduce a customised cost-aware focal loss function, replacing the commonly used binary entropy loss, to mitigate the model's bias towards the majority classes. Moreover, Sun et al. (2022) introduced an asymmetric bagging ensemble strategy, which creates multiple balanced datasets through repeated undersampling of majority class samples. They then integrate the asymmetric bagging algorithm with the LightGBM ensemble classifier for multi-class imbalanced enterprise credit evaluation. These studies have demonstrated that GBDTs have significant predictive capabilities for handling complex credit scoring tasks. In the subsequent section, we will conduct a detailed review of the imbalanced learning problem and highlight the contributions of our study.

2.2. Imbalance learning problem

An imbalanced learning problem is defined as a classification problem where one class has a lower number of observations than the other class. This presents a challenge for predictive modelling since the majority of classification machine learning algorithms were formulated under the assumption of a uniform class distribution. In the context of credit scoring, imbalanced classification is a common issue, with the minority class being the class of interest (e.g. default loans) from a learning perspective. The disparity between default loans and legitimate loans often exhibits a significant imbalance, with the misclassification of default loans incurring substantially higher costs for lending institutions than the misclassification of legitimate loans (Brown and Mues, 2012). As a result, it is crucial to explore and develop effective algorithms for imbalanced classification to accurately identify default loans. Krawczyk (2016) provided a comprehensive overview of common techniques used in imbalanced

learning, encompassing data-level methods such as re-sampling techniques employed during pre-processing, algorithm-level methods that modify objective functions to prioritise the minority class (including cost-sensitive frameworks), and mixed-level methods that combine both approaches.

Although there are various types of approaches for addressing imbalanced learning problems, no single approach has emerged as a clear dominant strategy. Each method possesses distinctive advantages and disadvantages. In case of the re-sampling method, oversampling methods (e.g. SMOTE and ADASYN) have been employed extensively for imbalanced learning tasks since they were invented for their good performance (Fernandez et al., 2018). Recently, more advanced generalisation models have been applied in this area, including the conditional tabular GAN (Xu et al., 2019) and variational autoencoders (Wan et al., 2017). The core of the oversampling method is to synthesise samples of the minority class based on existing data. In this case, oversampling methods are sometimes subjected to criticism due to overfitting problems, given that the algorithm can only get limited information based on real data. Moreover, the question of whether the data synthesised by the algorithm can be applied in real-world contexts remains unresolved. Conversely, while under-sampling methods only use real data, the reduction of majority class data typically results in the loss of information.

In contrast to resampling methods, which can only be used as a pre-processing step at the data level in classification tasks, cost-sensitive learning can be applied at the algorithmic level, forcing the models to focus on the minority samples during the training stage (Kaur et al., 2019). In their study, Ting (2002) introduced a sample weighting algorithm for cost-sensitive decision trees, which was demonstrated to be highly effective in reducing the cost of misclassification. By adjusting the class distribution within the induced tree, with a focus on the higher weight/cost class, the algorithm successfully minimises errors in the high-cost class. Consequently, the total misclassification cost is significantly reduced in practical scenarios, while the model's capacity to classify minority samples is enhanced. Nevertheless, it is worth noting that the cost-sensitive framework does have its drawbacks. In contrast to data-level methods, the cost-sensitive approach is not flexible. Additionally, its implementation necessitates domain-specific expertise to accurately calculate costs or benefits.

Another approach to addressing the imbalanced learning problem is to modify the classification algorithms. In the last decade, researchers have proposed numerous improved models based on modifying mainstream machine learning classifiers (Kaur et al., 2019). These modifications involve alterations to kernel functions, activation functions, or the design of basis/activation/kernel functions to enhance the discriminatory power of the classifiers. Compared with the re-sampling method and cost-sensitive framework, classification algorithm modification does not require the artificial generation of new data or the sacrifice of information, while maintaining a certain degree of generalisation and transferability. This research topic has gained attention from scholars across various fields including biology, healthcare, business management, etc (Gao et al., 2016; Pai et al., 2011; Wu et al., 2016).

The study conducted by Brown and Mues (2012) made a significant step in bringing attention to the issue of imbalanced learning in credit scoring. Their systematic comparison of different credit scoring models across multiple datasets highlighted the importance of addressing the imbalance problem. Since then, many scholars have made efforts to improve the predictive accuracy of credit scoring using various imbalanced learning techniques (Calabrese and Osmetti, 2013; He et al., 2018; Lei et al., 2020; Marqués et al., 2013). Some recent studies attempted to integrate the GEV distribution into boosting algorithms in fields such as weather and catastrophe forecasting (Koh, 2021; Velthoen et al., 2021). These studies leverage the generalised Pareto distribution to improve the tail distribution in regression tasks. In the context of classification tasks, Mushava and Murray (2022) extended the gradient boosting tree with a GEV link and a modified focal loss function. These modifications aim to make the algorithm pay more attention to rare cases in the minority class, thereby achieving superior predictive performance. It's worth noting that although Mushava and Murray (2022) and the our study both incorporate the generalised extreme value distribution into gradient boosting decision trees for imbalanced learning, we take different approaches. Mushava and Murray (2022) used a modified focal loss derived from object detection tasks, while our study employs a GEV loss function directly derived from the maximum likelihood estimation of the GEV distribution. This distinction allows our study to contribute to the ongoing exploration and development of imbalanced learning techniques in the context of credit scoring.

3. Methodology

3.1. GBDT classifier with cross-entropy loss

In supervised classification problems, a binary dependent variable and a vector of independent variables $\mathbf{x}$ with a joint probability distribution $P(\mathbf{x}, y)$ are considered. The objective is to identify the optimal

classifier $F^*(\mathbf{x})$ that minimises the expected value $E_{\mathbf{x},y}$ of the loss function $L(y, F(\mathbf{x}))$, where $F(\mathbf{x})$ is a classifier function.

$$F^*(\mathbf{x}) = \arg\min_F E_{\mathbf{x},y}\left[L(y, F(\mathbf{x}))\right].$$

The number of times the boosting process is repeated, known as the number of boosting rounds, is indicated by M. In gradient boosting method, the ultimate classifier $F_M(\mathbf{x})$ is the estimated optimal final classifier at round M

$$F_M(\mathbf{x}) = \sum_{m=1}^{M} \beta_m f_m(\mathbf{x})$$

where β_m represents the weight of the classifier in round m and f_m is the base classifier generated in round m. $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ is the training set. The initial classifier $F_0(\mathbf{x})$ is defined as follows

$$F_0(\mathbf{x}) = \operatorname{argmin}_{\alpha} \sum_{i=1}^{n} \left[L(y_i, \alpha)\right].$$

The GBDT classifier $F_m(\mathbf{x})$ proceeds to iteratively augment *via* a greedy approximation approach (Friedman, 2001)

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \operatorname{argmin}_{f_m} \sum_{i=1}^{n} \left[L(y_i, F_{m-1}(\mathbf{x}_i) + f_m(\mathbf{x}_i)\right]$$

In some GBDT extensions, such as XGBoost, the procedure approximates the original loss function through second-order Taylor expansions at $f_t = 0$ (Chen and Guestrin, 2016). With $\Omega(f_m)$, the regularisation term that avoids the overfitting problem by punishing the loss function if the trees are excessively deep or there are too many leaves, the objective function can be expressed as

$$F_m(\mathbf{x}) \cong \sum_{i=1}^{n} \left[L(y_i, F_{m-1}(\mathbf{x}_i)) + g_i f_m(\mathbf{x}_i) + \frac{1}{2} h_i f_m^2(\mathbf{x}_i)\right] + \Omega(f_m)$$

where g_i and h_i represent the first- and second-order derivatives of the loss function, respectively.

Within a gradient boosting tree, when considering a specific tree structure denoted as $q(\mathbf{x})$, the representation of f_m can be expressed as $f_m = \omega_{q(\mathbf{x})}$, where $\omega_j \in R^T$ represents the weights assigned to node j within the tree structure $q(\mathbf{x})$. Here, $q : R^d \to 1, 2, 3, \ldots, T$, with T denoting the number of all leaves within the given tree structure. Then $\Omega(f_m)$ is defined as

$$\Omega(f_m) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2$$

Here, γ and λ are two different regularisation parameters to control the magnitude of regularisation. Then, we can get the optimal leaf node weight ω_j^* to minimise the loss function using the gradient such that

$$\omega_j^* = -\frac{\sum_{i\in I} g_i}{\sum_{i\in I} h_i + \lambda}$$

and the value of the function, when ω_j is optimal, is

$$F_m(\mathbf{x}) = -\frac{1}{2}\sum_{j=1}^{T} \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda} + \gamma T.$$

Finally, let I_L and I_R denote the sample sets of left and right nodes after splitting and $I = I_L \cup I_R$, and then the gain after splitting can be represented as

$$F_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right] - \lambda$$

The above equation is usually used in practice for evaluating the split candidates to find the best split nodes.

In binary classification problems, one of the widely applied loss functions is the log-loss, which is a binary cross-entropy loss function (Vovk, 2015)

$$L(y, \hat{y}) = -\frac{1}{n}\sum_{i=1}^{n} y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)$$

where y represents the actual label and $\hat{y} \in (0, 1)$ is the predicted probability

$$\hat{y} = \Theta(t)$$

with $\Theta(t)$ the logistic (sigmoid) link function:

$$\Theta(t) = \frac{1}{1 + e^{-t}}$$

The logistic function is a symmetric S-shaped function that can convert any real number to a value between 0 and 1, which can be regarded as a probability. The symmetric shape property helps any model using the logistic function as the link function behaves well in common binary classification tasks. However, in imbalanced learning problems where the data is highly skewed and the minority class is usually associated with higher importance (e.g. disease diagnosis and loan default) (He and Garcia, 2009), the estimated probabilities for the minority class tend to be underestimated when using a symmetric logistic function (Calabrese et al., 2016; Ogundimu, 2019; Wang and Dey, 2010). To overcome this disadvantage, we propose the GEV-GBDT model to better classify imbalanced datasets.

3.2. GEV – GBDT classifier

In the GEV–GBDT model, the GEV function replaces the logistic function as the link function of the GBDT classifier. The GEV distribution can be skewed and symmetric. It is widely used to model the tail of a

distribution (De Haan and Ferreira, 2007). The cumulative distribution function of the GEV $\pi(t)$ can be represented in the form of $e^{-h(t)}$, with

$$h(t) = \begin{cases} (1+\tau\frac{t-\mu}{\sigma})^{-\frac{1}{\tau}}, & \tau \neq 0 \\ e^{-\frac{t-\mu}{\sigma}}, & \tau = 0 \end{cases}$$

where $\mu \in \mathbb{R}$ is the location parameter, $\sigma > 0$ is the scale parameter and $\tau \in \mathbb{R}$ is the shape parameter. We can control the skewness of GEV distribution by changing τ. We can change τ to handle different imbalance ratios: for a higher imbalance ratio, a highly skewed GEV distribution can achieve higher classification accuracy (Wang and Dey, 2010). The GEV distribution has different requirements for its support, depending on the value of τ :

$$t \in \begin{cases} \left[\mu - \frac{\sigma}{\tau}, +\infty\right), & \tau > 0 \\ (-\infty, +\infty), & \tau = 0 \\ \left(-\infty, \mu - \frac{\sigma}{\tau}\right], & \tau < 0 \end{cases}$$

Outside the support of the GEV distribution, the cumulative distribution function is 0.

Going back to the classification tasks, let y and $\pi(\hat{y})$ represent the actual value of the dependent variable and the predicted probability of being 1, respectively. In this case, the link function is

$$\pi(\hat{y}) = \begin{cases} e^{(1+\tau\hat{y})^{-\frac{1}{\tau}}}, & \tau \neq 0 \\ e^{e^{-\hat{y}}}, & \tau = 0. \end{cases}$$

Also, inspired by the maximum-likelihood estimation of the GEV, we derive the GEV loss function used in our proposed model. The log-likelihood function $l(y, \pi(\hat{y}))$ for training set is:

$$l(y, \pi(\hat{y})) = \sum_{i=1}^{n} \left[y_i \cdot \ln(\pi(\hat{y}_i)) + (1-y_i) \cdot \ln(1-\pi(\hat{y}_i))\right]$$

To formulate the loss function, we multiply the log-likelihood function by $\frac{1}{n}$ and take the negative of the whole function. In this way, the model can be optimised by finding best values of $\hat{y}$ that minimise the loss function. After replacing $\pi(\hat{y})$, the loss function becomes

$$L(y, \pi(\hat{y})) = \begin{cases} -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \cdot \ln\left(e^{(1+\tau\hat{y}_i)^{-\frac{1}{\tau}}}\right) + (1-y_i) \cdot \ln\left(1-e^{(1+\tau\hat{y}_i)^{-\frac{1}{\tau}}}\right)\right], & \tau \neq 0 \\ -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \cdot \ln\left(e^{e^{-\hat{y}_i}}\right) + (1-y_i) \cdot \ln\left(1-e^{e^{-\hat{y}_i}}\right)\right], & \tau = 0 \end{cases}$$

The GBDT model requires the second-order Taylor expansion of the loss function used for optimisation. We compute the first- and second-order derivatives *grad* and *hess* of the loss function w.r.t. $\hat{y}$:

grad

$$= \begin{cases} -\frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-y_i)e^{-(\tau\hat{y}_i+1)^{-\frac{1}{\tau}}}(\tau\hat{y}_i+1)^{-\frac{1}{\tau}-1}}{1-e^{-(\tau\hat{y}_i+1)^{-\frac{1}{\tau}}}} - y_i(\tau\hat{y}_i+1)^{-\frac{1}{\tau}-1}\right], & \tau \neq 0 \\ -\frac{1}{n}\sum_{i=1}^{n}\left[\frac{e^{\hat{y}_i}(e^{e^{\hat{y}_i}}y_i-1)}{e^{e^{\hat{y}_i}}-1}\right], & \tau = 0 \end{cases}$$

hess

$$= \begin{cases} -\frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{1}{\tau}+1\right)\tau y_i(\tau\hat{y}_i+1)^{-\frac{1}{\tau}-2} \right. \\ +(1-y_i)\left[\frac{e^{-(\tau\hat{y}_i+1)^{-\frac{1}{\tau}}}(\tau\hat{y}_i+1)^{-\frac{2}{\tau}-2}}{1-e^{-(\tau\hat{y}_i+1)^{-\frac{1}{\tau}}}} + \frac{e^{-2(\tau\hat{y}_i+1)^{-\frac{1}{\tau}}}(\tau\hat{y}_i+1)^{-\frac{2}{\tau}-2}}{\left(1-e^{-(\tau\hat{y}_i+1)^{-\frac{1}{\tau}}}\right)^2}\right. \\ \left.\left. -\frac{\left(\frac{1}{\tau}+1\right)\tau e^{-(\tau\hat{y}_i+1)^{-\frac{1}{\tau}}}(\tau\hat{y}_i+1)^{-\frac{1}{\tau}-2}}{1-e^{-(\tau\hat{y}_i+1)^{-\frac{1}{\tau}}}}\right]\right], & \tau \neq 0 \\ -\frac{1}{n}\sum_{i=1}^{n}\left[e^{\hat{y}_i}y_i - \left[\frac{e^{\hat{y}_i-e^{\hat{y}_i}}(1-e^{\hat{y}_i})}{1-e^{-e^{\hat{y}_i}}} - \frac{e^{2\hat{y}_i-2e^{\hat{y}_i}}}{(1-e^{-e^{\hat{y}_i}})^2}\right](1-y_i)\right], & \tau = 0 \end{cases}$$

After computing the required derivatives of the original loss function, we can approximate the loss function and fit it using Python's LightGBM library to establish our GEV-GBDT model. The Python code developed for this study is publicly available at https://github.com/EzioClark/GEV-GBDT.

4. Data

For our empirical experiments, we use four datasets of loans granted by a P2P lending platform and traditional financial institutions. Each dataset contains a unique set of features describing a borrower and loan characteristics. The characteristics of these datasets are as follows:

- The first dataset (LC) consists of 1,640,003 loans granted by the U.S. P2P lending platform, Lending Club. Each loan is described by 16 features, including 15 independent variables describing the loan information and borrower credit profile and a class label indicating whether the loan is defaulted. Among the 1,640,003 loans, 1,296,371 are "good" borrowers and 343,632 (20.95%) are bad. The average loan value is 14,194.83 US dollars and the average term is 41.04 months.
- The second dataset (SMF-1) consists of 3,076 loans for middle and small enterprises (SMEs) granted by a Chinese commercial bank. This credit scoring data is very unbalanced and it includes 3009 "good" applications and only 67 (2.18%) "bad" ones. Each loan sample is associated with 80 attributes covering borrowing companies' detailed

financial performance, owners' background information, loan characteristics, payment history, other non-financial information, and local macroeconomic variables. The data set has a mean loan value of 10,387.72 CNY.

- The third dataset (SMF-2) contains 2,157 loans made to micro and small businesses granted by a bank. Among those loans, 1,911 are instances of creditworthy applications and 246 are instances with defaulted payments. Each instance has a class label and 60 features including demographic attributes, financial information, guarantor information, and local macroeconomic indicators.
- The fourth dataset (SMF-3) consists of 4,424 loans made to agricultural entrepreneurs and households granted by a state owned commercial bank in China. Each loan is described by 30 features covering the borrower's demographic and financial information, credit history, loan characteristics, and guarantor information. The proportion of default or positive cases is 11.64% (515 default loans). The mean loan value is 34,472.68 CNY.

Table 1 provides a brief overview of the key attributes of the aforementioned four datasets. These loan samples are collected from different sources and the feature sets capture various aspects of borrowers and loan details. In addition, they have different sample sizes and the imbalance ratios range from 1:4 to 1:45. These heterogeneous characteristics make them suitable for use in conducting empirical credit scoring experiments and evaluating the performance of different models and techniques. Table 2 shows the different default percentages for the four analysed samples. Table A2 in Appendix describes the variables in the LendingClub dataset and Table A3 reports those of the small business finance datasets.

5. Results

5.1. Classification performance of GEV-GBDT

In this section, we compare the classification performance of GEV-GBDT against several benchmark models. We apply the z-score transformation to continuous variables to make different features on the same scale.To evaluate the efficiency of the models, we implement a cross-validation process that avoids the biased selection of the sub-sets and improves the reliability of the estimates. For all three SMF datasets, we conduct a 100 times repeated 5-fold cross-validation. Each of the five random partitions acts as an independent holdout test set, the remaining four partitions are used to train the credit scoring model. The training sets are used to estimate the model's parameters, and the classification performance is assessed on the holdout sets. The overall performance score is an average across all five test set partitions. Compared to the SMF datasets, the Lending Club dataset covers a much longer period and the observations may not be fully independent of each other. Therefore, we perform a 5-fold time series/sequential cross-validation for this dataset, which splits the data into 5 folds by preserving the chronological order of the observations. Similarly, we employ different combinations of hyperparameters for each model and repeat the process 100 times to report the average result.[1] Figure 1 illustrates the two different cross-validation methods used.

Table 1. Characteristics of four credit scoring datasets.

	Input size	Area	Number of variables
LC	1,640,003	North America	15
SMF-1	3,076	China	87
SMF-2	2,157	China	60
SMF-3	4,424	China	29

To gain a comprehensive understanding of how GEV-GBDT performs in the context of imbalanced learning, we select two popular imbalance learning techniques as the benchmark methods: the cost-sensitive framework (CS) for data-level method and SMOTE as an algorithm-level handling. We also use three benchmark machine learning models: Logistic Regression (LR), Random Forest (RF) and the regular GBDT. LR is widely used for its simplicity, interpretability, and the ease with which it handles binary outcomes—a natural fit for credit scoring where the primary objective is to predict whether a borrower will default or not (Hosmer et al., 2013). The output of LR can be directly interpreted in terms of odds ratios, providing clear insights into the influence of various factors on the probability of default. This interpretability is crucial for regulatory compliance and for understanding key drivers of risk. For the LR, we control the regularisation parameters to avoid overfitting. The RF, an ensemble learning method, has gained popularity due to its ability to handle large datasets with numerous predictors and its robustness to overfitting (Breiman, 2001). By aggregating multiple decision trees, the RF can capture complex interactions between variables, which can be especially valuable in credit scoring where relationships between variables are often non-linear and interactive. For simplicity and interpretability, we control the number of leaves and the number of trees to provide a stable performance without overfitting. The regular GBDT, as introduced in the previous section, takes a different approach to ensemble learning by iteratively training decision trees on the residuals of the previous trees, allowing itself to focus on the most challenging examples and gradually improve the overall

Table 2. Distributions of target variables in four datasets.

	LC	SMF-1	SMF-2	SMF-3
Positive samples	343,632(20.95%)	67(2.18%)	246(11.40%)	515(11.64%)
Negative samples	1,296,371(79.05%)	3,009(97.82%)	1,911(88.60%)	3,909(88.36%)
Total	1,640,003	3,076	2,157	4,424

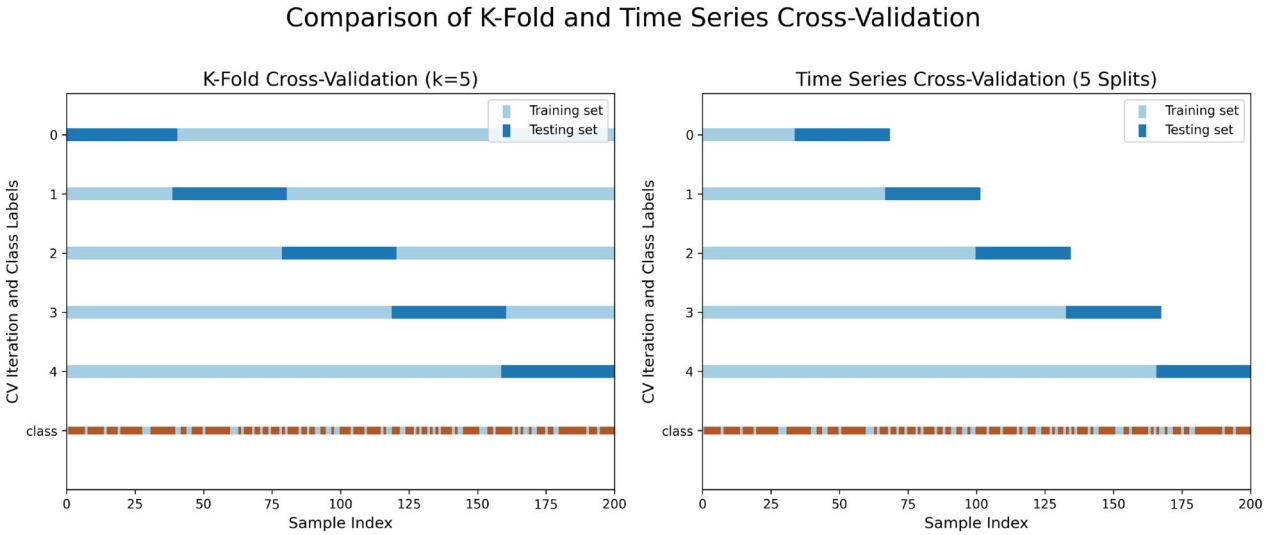

Figure 1. K-fold and time series cross-validation.

performance. It is robust and can handle different types of variables. To be able to compare the two models, the hyperparameter setting of the regular GBDT and the GEV-GBDT are the same. The most important hyperparameters are the number of trees, the learning rate, and the number of leaves.

The cost-sensitive framework can be implemented using the Scikit-learn library. We compute the misclassification cost for each sample and use the normalised cost as the sample weight in the training process (Bahnsen et al., 2014). For the datasets where the principal amount and the interest rate are available, we assign the principal amount to the positive sample and the total amount of the interest earned on the loan to the negative sample. Therefore, we assigned a higher misclassification cost to positive cases (default). When the principal amount and interest rate are unknown (e.g. SMF-2 dataset), we assign misclassification costs that reflect the ratio of positive cases to negative cases.

To avoid the curse of dimensionality, we apply a feature selection to choose the 20 features with the highest feature importance generated by a GBDT.To evaluate the classification performance, we employ three commonly used evaluation metrics: the Area Under the Curve (AUC), the Kolmogorov-Smirnov (KS) statistic, and the H-measure. The area under the Receiver Operating Characteristics (ROC) curve is a widely used metric in statistical analysis and machine learning for evaluating the performance of a classification model (Fawcett, 2006). The ROC curve is a graphical representation that plots the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings. The AUC measures the entire two-dimensional area underneath the entire ROC curve. It provides an aggregated measure of performance across all possible classification thresholds. The value of AUC ranges from 0 to 1. A model with an AUC of 0.5 indicates no discriminative ability, equivalent to random chance, while an AUC of 1.0 indicates perfect discrimination (Hand and Till, 2001).

The KS statistic is a non-parametric test used to evaluate the discriminatory power of classification models. It measures the maximum distance between two cumulative distribution functions (CDFs) of the positive and negative outcomes in a binary classification model. In model evaluation, the KS statistic is particularly useful in identifying the threshold where the separation between the distributions of the true positives (TPR) and false positives (FPR) is the maximised (Hosmer et al., 2013). Unlike the AUC, which considers the entire range of thresholds, the KS statistic focuses specifically on the optimal point at which a model effectively identifies true positives while minimising false positives. Moreover, previous research suggests that the H-measure is a more effective metric for evaluating the predictive performance of models developed for imbalanced datasets (Chen et al., 2024; Hand and Anagnostopoulos, 2014). The H-measure addresses the incoherence of the AUC by introducing costs for the different types of misclassification (Hand, 2009). It ranges from zero for a random classifier to one for a perfect classifier.

The results shown in Tables 3 and 4 demonstrate that the GEV-GBDT achieves the highest predictive accuracy compared to all tested techniques across all

Table 3. The Area Under the Curve (AUC) for different classification models.

	LC	SMF-1	SMF-2	SMF-3
CS-LR	0.6122	0.7584	0.5567	0.5717
CS-RF	0.6746	0.8422	0.6363	0.5992
CS-GBDT	0.6836	0.9089	0.6049	0.6625
SMOTE-LR	0.6125	0.7613	0.5404	0.5528
SMOTE-RF	0.6625	0.8574	0.6508	0.5955
SMOTE-GBDT	0.6619	0.9101	0.6332	0.6532
GEV-GBDT	**0.6966**	**0.9354**	**0.6711**	**0.6839**

The bold values show the highest AUC.

Table 4. The Kolmogorov-Smirnov (KS) statistics for different classification models.

	LC	SMF-1	SMF-2	SMF-3
CS-LR	0.1639	0.6583	0.1639	0.1780
CS-RF	0.2522	0.7527	0.2818	0.1781
CS-GBDT	0.2646	0.8282	0.2337	0.2822
SMOTE-LR	0.1644	0.6658	0.1541	0.1159
SMOTE-RF	0.2318	0.7614	0.3032	0.1653
SMOTE-GBDT	0.2333	0.8301	0.2467	0.2857
GEV-GBDT	**0.2824**	**0.8432**	**0.3217**	**0.3088**

Table 5. Hand (H) measures for different classification models.

	LC	SMF-1	SMF-2	SMF-3
CS-LR	0.0431	0.4786	0.0544	0.0559
CS-RF	0.1019	0.6942	0.1267	0.0666
CS-GBDT	0.1112	0.7841	0.0903	0.1195
SMOTE-LR	0.0433	0.4928	0.0463	0.0196
SMOTE-RF	0.0893	0.7018	0.1488	0.0608
SMOTE-GBDT	0.0870	0.7746	0.1078	0.1061
GEV-GBDT	**0.1219**	**0.7999**	**0.1606**	**0.1260**

four datasets. Table 5 presents the predictive result of the tested models evaluated by H-Measure with our datasets. The results reaffirm the superiority of GEV-GBDT over the other benchmark models.

For each model, we compute the confusion matrix for the optimal cutoff obtained by maximising the difference between the TPR and the FPR for each fold. We then sum the confusion matrices for all folds to form the final confusion matrix for the model. This matrix provides counts of TP, true negatives (TN), FP, and false negatives (FN), providing a comprehensive overview of the model's classification accuracy. From this matrix, we derive several additional metrics to evaluate the performance of models. Accuracy measures the overall correctness of the model, calculated as the proportion of true results (both TP and TN) out of all the cases. Recall, or TPR, measures the model's ability to identify all relevant instances, calculated as TP/(TP+FN), thus reflecting the model's sensitivity to detecting positive cases. Precision, on the other hand, assesses the proportion of true positive predictions among all positive predictions made, calculated as TP/(TP+FP). Lastly, the F1-Score serves as a balanced measure of precision and recall, calculated as $2 * (\text{precision} * \text{recall})/(\text{precision} + \text{recall})$. This metric is particularly useful in scenarios where maintaining equilibrium between the FP and FN is crucial. These metrics

Table 6. Matrix-based metrics for different classification models.

		LC	SMF-1	SMF-2	SMF-3
Accuracy	CS-LR	0.5930	0.8633	0.5336	**0.6702**
	CS-RF	0.6065	0.8863	0.6815	0.4948
	CS-GBDT	0.6236	0.9250	0.6365	0.5335
	SMOTE-LR	0.5880	0.8395	0.4965	0.3515
	SMOTE-RF	0.6000	0.9025	**0.7450**	0.4652
	SMOTE-GBDT	0.6033	0.9168	0.6370	0.5402
	GEV-GBDT	**0.6304**	**0.9413**	0.7070	0.5554
Recall	CS-LR	0.2722	0.0500	0.1435	0.1619
	CS-RF	0.3000	0.0684	0.1967	0.1518
	CS-GBDT	0.3096	0.1048	0.1711	0.1708
	SMOTE-LR	0.2711	0.0471	0.1404	0.1285
	SMOTE-RF	0.2959	0.0720	0.2295	0.1466
	SMOTE-GBDT	0.2944	0.0952	0.1784	0.1673
	GEV-GBDT	**0.3175**	**0.1353**	**0.2342**	**0.1768**
Precision	CS-LR	0.5630	0.6923	0.6220	0.4388
	CS-RF	0.6583	0.8077	0.5813	0.7282
	CS-GBDT	0.6475	0.8462	0.5691	0.7806
	SMOTE-LR	0.5723	0.7692	0.6667	0.7903
	SMOTE-RF	0.6283	0.7917	0.5244	0.7456
	SMOTE-GBDT	0.6397	0.8462	0.6057	0.7417
	GEV-GBDT	**0.6583**	**0.8846**	**0.6911**	**0.7709**
F1-score	CS-LR	0.3670	0.0933	0.2332	0.2365
	CS-RF	0.4121	0.1261	0.2939	0.2513
	CS-GBDT	0.4189	0.1864	0.2632	0.2803
	SMOTE-LR	0.3680	0.0887	0.2320	0.2210
	SMOTE-RF	0.4023	0.1319	0.3193	0.2451
	SMOTE-GBDT	0.4033	0.1712	0.2757	0.2731
	GEV-GBDT	**0.4284**	**0.2347**	**0.3498**	**0.2876**

collectively provide a thorough evaluation of the model's performance, particularly emphasising its ability to accurately identify positive samples.

Table 6 shows the confusion matrix based metrics. The GEV-GBDT achieves the best performance on all four datasets based on recall rate, precision rate, and F1 score. On datasets SMF-2 and SMF-3, the accuracy rates of GEV-GBDT are only in second place, as the best baseline models make more correct predictions in the majority class at the cost of lower recall rate and precision rate on the predictions of positive samples. In general, the GEV-GBDT not only outperforms baseline models from the high-level perspective but also demonstrates superior discriminative capability to the other models, with the highest recall rate and F1-score across the different datasets, proving the success of our model design in terms of forcing GEV-GBDT to focus on the minority positive samples.

Given the large sample size and long period of coverage, we use the LendingClub as a case study to produce calibration plots that compare observed and predicted probabilities for assessment of prediction model performance. Figure 2 displays the calibration plots for all tested models using LendingClub dataset. In a calibration plot, the x-axis represents the predicted probabilities provided by the model, while the y-axis represents the actual probabilities (or frequencies) of the outcomes. A perfectly calibrated model would have all predictions lying on the diagonal line running from the bottom left to the top right of the plot. Figure 2 shows the superiority of the GEV-GBDT over the baseline models, as the other models' predicted values are largely concentrated in the range *[*0.2, 0.8*]*, thereby

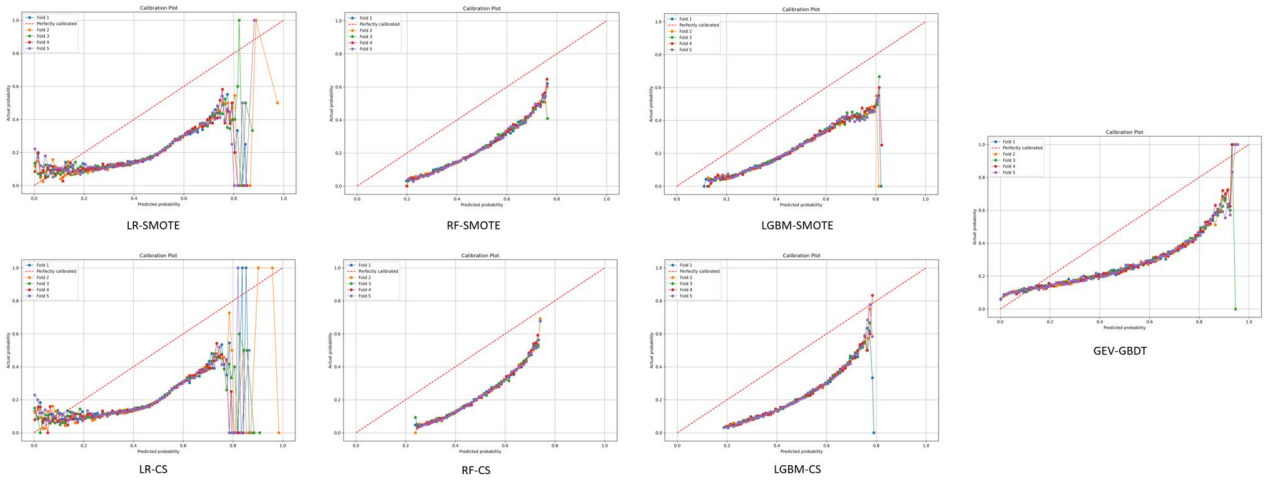

Figure 2. Plots for different classification models on the LendingClub dataset.

underestimating the tail risk even with some corrective measures applied. Although the two LR-based models span all probability bins, their predictions are not stable across different folds. On the contrary, the GEV-GBDT has a stable probability output within the range *[*0, 0.9*]*, closely approaching the diagonal line at both ends. The instability problem only arises beyond 0.9.

To demonstrate the economic benefits of the GEV-GBDT in real credit risk management scenarios, we compare the real loss against the expected loss for all the other models based on the LendingClub dataset. We first calculate the mean actual loss of the test set in each fold. Then, for each model at its optimal cutoff, we compute the value of the predicted loss (i.e. the samples in the TP and the FP groups), the portion correctly predicted (i.e. the samples in the TP group only), and their corresponding predictive error. The actual loss in this test is about 898.48 million US dollars. Figure 3 shows the economic performance of each model. In general, all models tend to overestimate the loss of the portfolio, with all differences between the actual and predicted losses being negative. However, the models have difficulty covering all actual default cases, resulting in a lower value of correctly predicted loss. The GEV-GBDT outperforms the baseline models in this test, with a lower overestimate of predicted loss (52.18 million US dollars less than the best baseline model) and a higher correctly predicted value (24.41 million US dollars higher than the best baseline model). The results of the economic analysis suggest that the GEV-GBDT could help financial institutions to better estimate the expected loss of their loan portfolios. Lending institutions always face a trade-off between pursuing more profit by lending money to more loan applicants and maintaining a lower risk exposure by acting more conservatively. The GEV-GBDT can address this challenge by providing more accurate loss estimation, thus helping financial institutions to pursue more business and higher profit without incurring higher additional risk, as the scoring model is more accurate in identifying bad borrowers. Figures 4 and 5 show that the GEV-GBDT achieves higher AUC and KS for higher imbalanced samples. Figures 6 and 7 plot the optimal tau value for different imbalanced ratios.

For the interpretability of the GEV-GBDT, we can use the built-in feature importance method of LightGBM to identify which features contribute the most information during the tree-splitting process. Some model-agnostic interpretable methods, e.g. SHaP (Lundberg and Lee, 2017), can produce inconsistent results for imbalanced data Chen et al. (2024). Thus, we will further explore the interpretability of the GEV-GBDT in our future works.

5.2. Selection of τ for datasets with various imbalance ratios

In this section, we examine the effect of various τ values on classification performance while keeping other hyperparameters fixed. We also look at how the optimal value of τ changes for datasets with different imbalanced ratios. We randomly sample 109 sub-datasets with varying imbalanced ratios from the LendingClub dataset with an imbalance ratio from 1:1 to 1:2000. For each sub-dataset, we take 201 different τ values ranging from -1 to 1 with an interval of 0.01. e model the regular LightGBM with the hyperparameter 'is_unbalanced' set to 'True' (which can be regarded as a naive approach of cost-sensitive methods by giving positive samples a higher weight according to the imbalanced ratio) and a log-loss function as the benchmark for these sub-datasets. This comparison allows us to analyse the effects of different imbalance ratios and τ values on the classification performance.

The two graphs presented above indicate that as the datasets become more skewed, the GEV-GBDT exhibits a greater advantage over the regular LightGBM model under both evaluation measures. To further strengthen

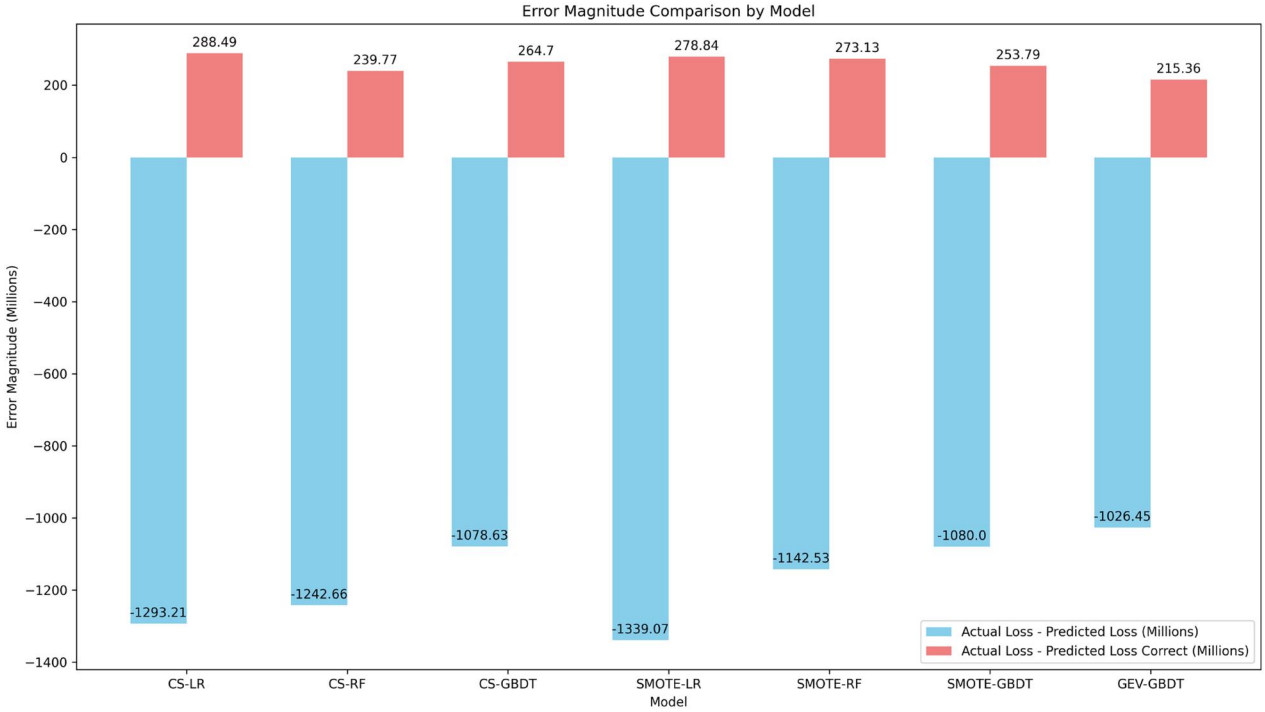

Figure 3. The economic benefit analysis of different classification models.

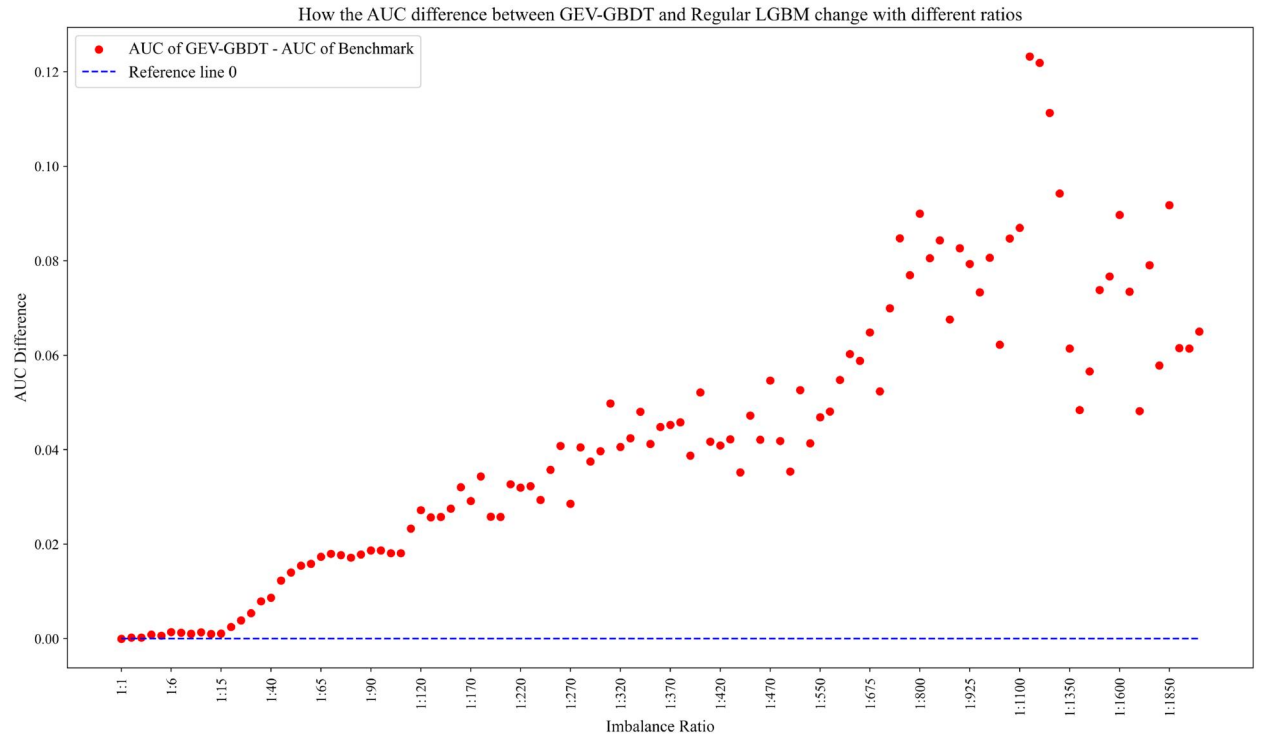

Figure 4. The difference of the AUC between the GEV-GBDT and the Regular GBDT on an Extremely Imbalanced Dataset.

our conclusion, we run a simple linear regression model to test whether the relationship is significant by using the inverse of the imbalance ratio as the independent variable and the performance advantage as the dependent variable. The resulting p-value for the coefficient t-test in the regression model, as shown in Table 7, confirms that this trend is highly statistically significant with respect to both the AUC and KS measures. However, the relationship is not necessarily linear. The purpose of conducting this supplementary t-test of the simple linear regression is to show that, in general, the more skewed the dataset is, the higher is the performance difference between the GEV-GBDT and the regular GBDT.

In addition, the two graphs reveal that as the datasets become more skewed, the optimal value of τ tends to decrease. Similarly, we conduct a t-test to examine the statistical significance of this trend. We consider the

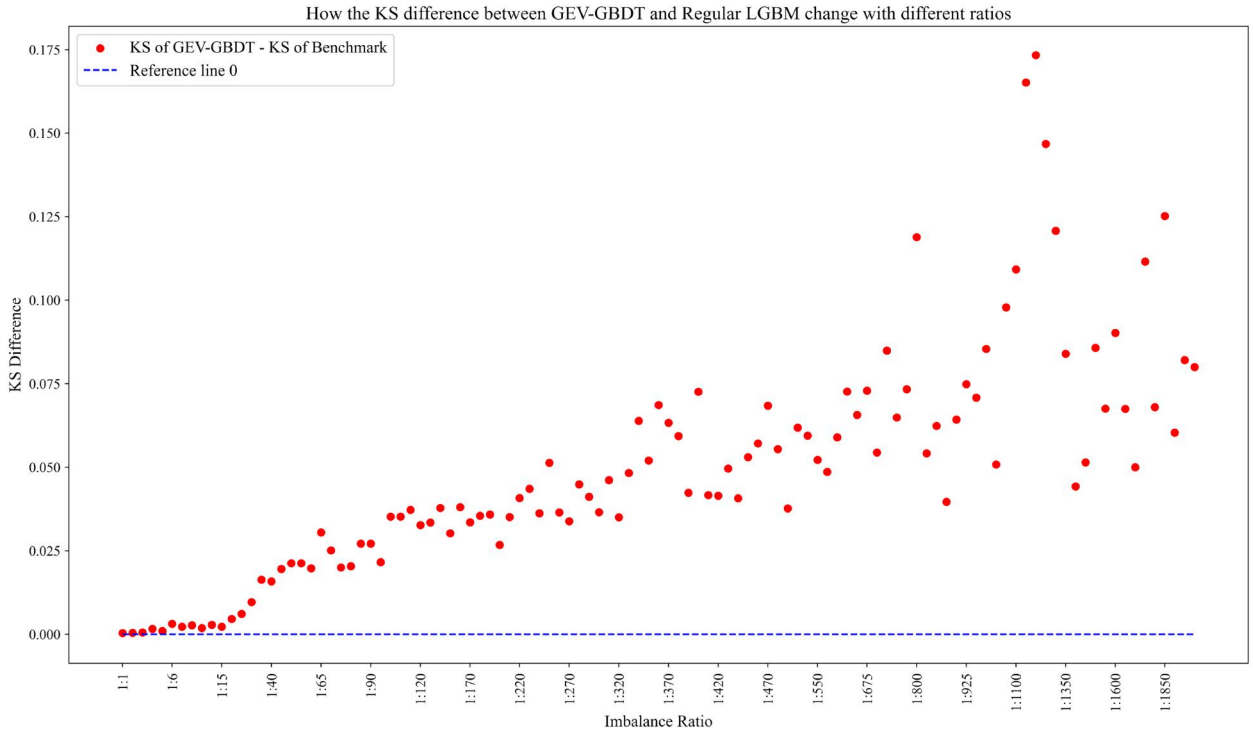

Figure 5. The difference of the KS between the GEV-GBDT and the Regular GBDT on an Extremely Imbalanced Dataset.

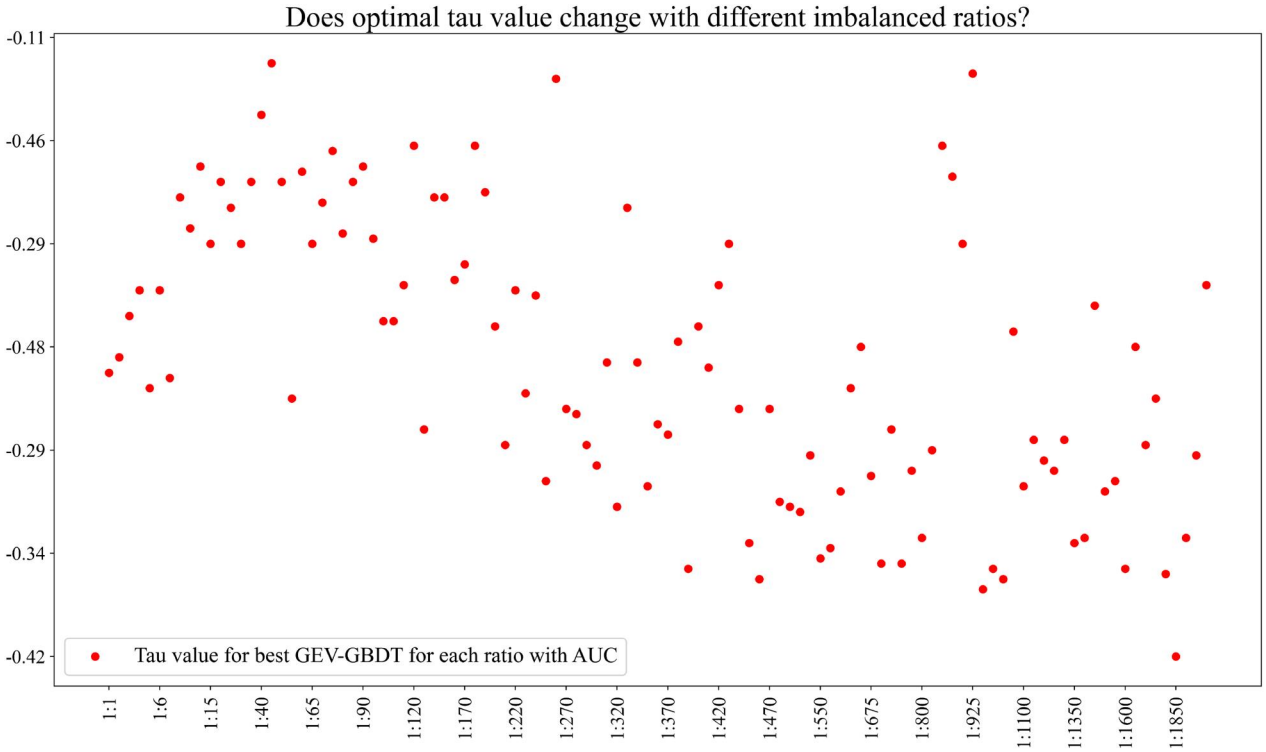

Figure 6. Optimal value of tau for different imbalanced ratios measured by AUC.

inverse of the imbalance ratio as the independent variable and treat the optimal value of τ as the dependent variable. The results presented in Table 8 show this relationship is highly significant if we choose the optimal values of τ based on the AUC but not if it is based on the KS. It is worth pointing out that the graph suggests a random distribution of the τ values, indicating that the correlation between the imbalance ratio and the optimal value of τ may be more complex for the KS measure.

5.3. How GEV-GBDT handles imbalanced data

In the previous sections, we reveal the superior classification performance of the GEV-GBDT on imbalanced datasets compared to the regular LightGBM

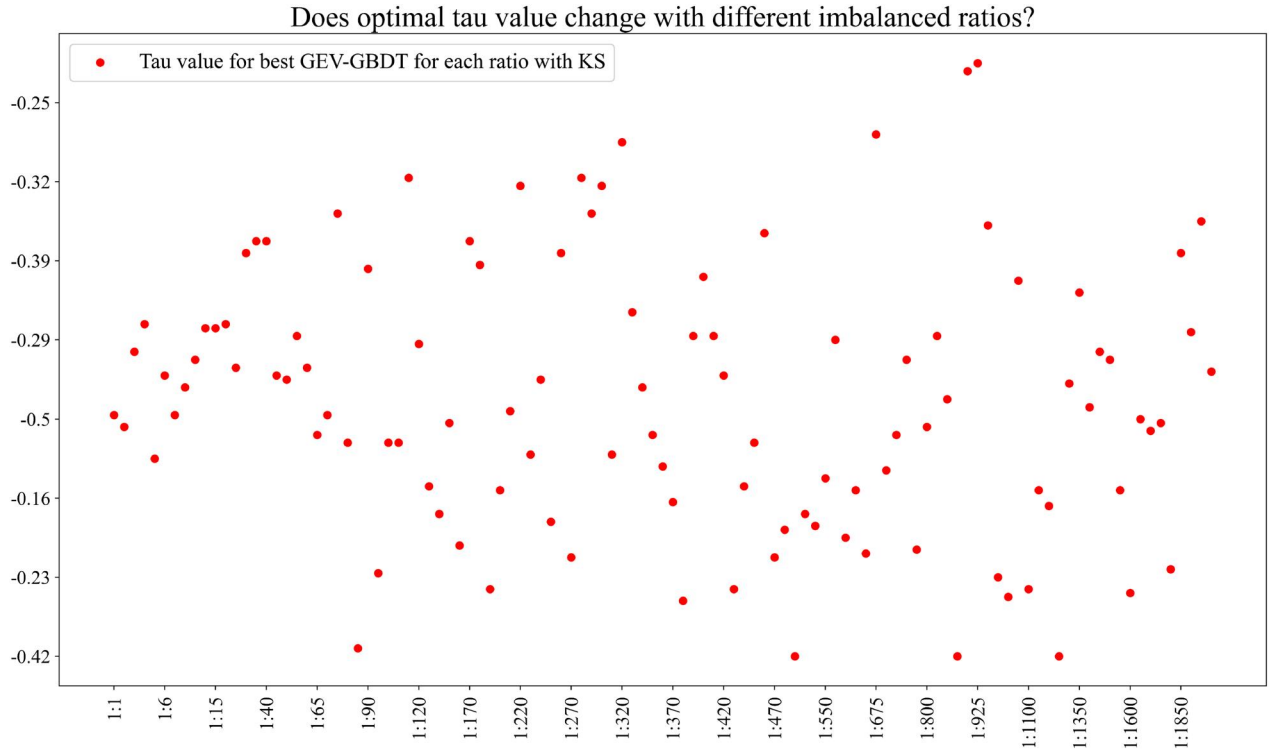

Figure 7. Optimal value of tau for different imbalanced ratios measured by KS.

Table 7. T-test of regression results of imbalance ratios against performance differences.

	Against AUC differences	AGAINST KS differences
test-statistics	4.159e-05	4.525e-05
p-value	2.363e-24***	2.123e-19***

Table 8. T-test on the regression where the dependent variable is the optimal value of τ and the independent variable is the inverse of the imbalance ratio.

	Optimal Tau chosen based on AUC	Optimal Tau chosen based on KS
test-statistics	−0.0002	−5.019e-05
p-value	1.748e-8***	0.395

with the cost-sensitive method. We also discuss the selection of the pre-determined hyperparameter τ. In this section, we further explore the outperformance of GEV-GBDT over the benchmark model by introducing an additional test. We compare the performance of the GEV-GBDT model and the benchmark model on datasets with varying imbalanced ratios based on different evaluation metrics, including the AUC, KS, best cutoff/threshold (defined by the threshold when the difference between TPR and the FPR is maximised), the TPR at best cutoff, and the FPR at the best cutoff of the previous trials.

Table 9 presents the classification performances of both models on the datasets with imbalance ratios ranging from 1:1 to 1:1850. The benchmark model, despite using a cost-sensitive framework, is inadequate in handling highly imbalanced data, as evidenced by a significant decline in the AUC from 0.6888 (1:1) to 0.5308 (1:1850). Moreover, the difference by which the TPR exceeds the FPR at the optimal threshold diminishes. We observe that as the data becomes more skewed, the optimal cutoff point shifts towards larger values, approximately from 0.5 to 1. These results suggest that the benchmark model is very conservative and only classifies those samples predicted with extremely high probabilities as positive. Although this strategy should yield both a low TPR and a low FPR, the benchmark model fails to effectively maintain the false positive rate, resulting in values still reaching around 0.5 for highly skewed datasets. As for GEV-GBDT, it keeps AUC scores higher than 0.6 for all datasets. One interesting finding is that the best cutoff for GEV-GBDT is moving leftward from 0.4743 to 0.0278 instead of getting larger when the data is more imbalanced. By setting the cutoff close to 0 means that both the TPR and the FPR should be high. However, the GEV-GBDT maintains the differences between the TPR and FPR even with an FPR larger than 0.5. Other metrics provide similar conclusions. In other words, the superior classification performance of GEV-GBDT on imbalanced samples stems not only from the higher TPR or lower FPR but from the relative difference between them. The results suggest that the GEV-GBDT focuses on positive samples by assigning near-zero values to negative samples. Thus, using a low threshold can incorporate more potentially positive data points without suffering from high FPR. Importantly, this approach is effective even for highly imbalanced data.

Table 9. Imbalanced classification performance analysis.

	Regular LightGBM (Cost Sensitive)					GEV-GBDT				
Imbalance Ratio	AUC	KS	Best Cutoff	TPR at BC	FPR at BC	AUC	KS	Best Cutoff	TPR at BC	FPR at BC
1:1	0.6888	0.2727	0.4924	0.6589	0.3863	0.6888	0.2727	0.4743	0.6584	0.3857
1:6	0.6882	0.2714	0.5042	0.6352	0.3638	0.6888	0.2724	0.4254	0.6364	0.364
1:15	0.6847	0.2664	0.4947	0.6482	0.3818	0.6856	0.2677	0.3477	0.6362	0.3685
1:40	0.6737	0.25	0.4638	0.6392	0.3892	0.6817	0.2629	0.3033	0.6446	0.3817
1:65	0.6603	0.2253	0.4063	0.687	0.4617	0.6776	0.2568	0.2929	0.657	0.4002
1:90	0.6573	0.2315	0.489	0.5437	0.3122	0.6758	0.2549	0.2371	0.6007	0.3458
1:120	0.6462	0.2176	0.8045	0.6721	0.4545	0.6724	0.2521	0.2426	0.5998	0.3476
1:170	0.6409	0.2099	0.8775	0.7763	0.5664	0.6688	0.2423	0.1709	0.6024	0.3601
1:220	0.6352	0.2062	0.9997	0.4427	0.2365	0.6656	0.2405	0.1058	0.5653	0.3248
1:270	0.6361	0.21	0.9998	0.5001	0.2901	0.6626	0.24	0.1096	0.6421	0.402
1:320	0.63	0.2149	1.0	0.5061	0.2912	0.6663	0.2443	0.1844	0.6561	0.4118
1:370	0.6258	0.1932	1.0	0.4064	0.2132	0.6667	0.2498	0.1194	0.6661	0.4163
1:420	0.625	0.2064	1.0	0.5047	0.2982	0.6662	0.2479	0.0753	0.7005	0.4526
1:470	0.6105	0.181	1.0	0.5654	0.3844	0.6632	0.2463	0.0328	0.7075	0.4613
1:550	0.613	0.1835	1.0	0.508	0.3246	0.6592	0.2376	0.0177	0.6267	0.3891
1:675	0.5915	0.1637	1.0	0.4964	0.3327	0.6533	0.2238	0.0655	0.6497	0.4259
1:800	0.5741	0.1622	1.0	0.5475	0.3853	0.6508	0.2348	0.0384	0.6455	0.4107
1:925	0.5572	0.1249	1.0	0.7024	0.5775	0.6321	0.1854	0.1066	0.5323	0.347
1:1100	0.5406	0.0878	1.0	0.6863	0.5985	0.6222	0.181	0.0174	0.7996	0.6186
1:1350	0.5771	0.1513	1.0	0.6723	0.521	0.6199	0.2152	0.0403	0.5794	0.3642
1:1600	0.5471	0.1145	1.0	0.6525	0.538	0.6258	0.1889	0.0254	0.8209	0.632
1:1850	0.5308	0.0616	1.0	0.5264	0.4648	0.6104	0.1766	0.0278	0.7905	0.6139

6. Conclusion

In this study, we propose a novel ensemble classifier, the GEV-GBDT, based on the gradient-boosting decision tree, to address the challenge of the imbalanced learning problem. Our experimental results demonstrate that the GEV-GBDT outperforms the benchmark models in terms of predictive performance. By integrating the GEV distribution into the loss function of the GBDT, our proposed model extends the use of algorithm-level methods for tackling imbalanced learning problems. It inherits the excellent learning ability of gradient boosting while also allowing for focused learning on minority samples, making it effective in handling imbalanced datasets with varying degrees of class imbalance. Furthermore, our experiments on searching for optimal model parameters reveal that the GEV-GBDT exhibits even better performance when applied to highly imbalanced datasets, further contributing to the literature on extremely imbalanced data classification.

The GEV-GBDT could be adopted by banks and financial institutions as a credit scoring model to identify bad borrowers, even if the number of defaults in the portfolio is very low. Using a more accurate credit scoring model could help financial institutions make more informed lending decisions, reduce the likelihood of granting loans to high-risk borrowers, and mitigate potential financial losses. In particular, this new method is more robust to temporal changes and high-risk profiles. Moreover, the flexibility of the GEV-GBDT model allow risk managers to tailor the model to their specific needs and risk appetite by adjusting the shape parameter τ. This allows lenders to strike a balance between sensitivity to potential defaults and the acceptable level of false positives, depending on the costs associated with each type of misclassification. Implementing the GEV-GBDT model can also streamline the credit approval process by reducing the need for manual intervention and subjective expert judgement. Automating the credit scoring process with a more accurate and robust model could lead to faster and more consistent decisions, improving the overall customer experience and operational efficiency.

Future work can foucs on several directions. First, the applicability of the GEV-GBDT model to classification tasks in other industries can be explored, to assess its robustness and generalizability. By conducting extensive tests on various classification tasks, we can gain a deeper understanding of the model's behaviour and parameter selection. Second, from our experiment, we acknowledge that the computational speed of the GEV-GBDT is slower than the original version of GBDTs since performing the second-order Taylor expansion with respect to the loss function requires a lot of computations. Thus, we can further refine the model to reduce the computational overhead through approximation and improve the efficiency of the model without compromising its classification performance. Lastly, our research can also continue to explore the application of the GEV distribution in other areas, such as its integration into deep learning-based credit scoring models.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Note

1. For the LendingClub dataset, we also conduct a regular 5-fold cross-validation process as a robustness check, and the result remains. Table A1 shows the details of the computing environment.

Funding

Raffaella Calabrese kindly acknowledges the financial support from the Economic and Social Research Council (ESRC) [grant number ES/W010259/1]. Baofeng Shi acknowledges financial support for this research from the Major Project of the National Social Science Foundation of China [Grant No. 23&ZD175].

References

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, *54*(6), 627–635. https://doi.org/10.1057/palgrave.jors.2601545

Bahnsen, A. C., Aouada, D., Ottersten, B. (2014). *Example-dependent cost-sensitive logistic regression for credit scoring* [Paper presentation]. 2014 13th International Conference on Machine Learning and Applications (pp. 263–269). IEEE.

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

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

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

Calabrese, R., Marra, G., & Osmetti, S. A. (2016). Bankruptcy prediction of small and medium enterprises using a flexible binary generalized extreme value model. *Journal of the Operational Research Society*, *67*(4), 604–615. https://doi.org/10.1057/jors.2015.64

Calabrese, R., & Osmetti, S. A. (2013). Modelling small and medium enterprise loan defaults as rare events: The generalized extreme value regression model. *Journal of Applied Statistics*, *40*(6), 1172–1188. https://doi.org/10.1080/02664763.2013.784894

Cao, Y., & Zhai, J. (2022). A survey of ai in finance. *Journal of Chinese Economic and Business Studies*, *20*(2), 125–137. https://doi.org/10.1080/14765284.2022.2077632

Chen, T., Guestrin, C. (2016). *Xgboost: A scalable tree boosting system* [Paper presentation]. Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining (pp. 785–794).

Chen, Y., Calabrese, R., & Martin-Barragan, B. (2024). Interpretable machine learning for imbalanced credit scoring datasets. *European Journal of Operational Research*, *312*(1), 357–372. https://doi.org/10.1016/j.ejor.2023.06.036

Chen, Z.-S., Zhou, J., Zhu, C.-Y., Wang, Z.-J., Xiong, S.-H., Rodríguez, R. M., Martínez, L., & Skibniewski, M. J. (2023). Prioritizing real estate enterprises based on credit risk assessment: An integrated multi-criteria group decision support framework. *Financial Innovation*, *9*(1), 120. https://doi.org/10.1186/s40854-023-00517-y

Clarke, R. (2016). Big data, big risks. *Information Systems Journal*, *26*(1), 77–90. https://doi.org/10.1111/isj.12088

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

De Haan, L., & Ferreira, A. (2007). *Extreme value theory: An introduction*. Springer Science & Business Media.

Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, *297*(3), 1178–1192. https://doi.org/10.1016/j.ejor.2021.06.053

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, *61*, 863–905. https://doi.org/10.1613/jair.1.11192

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232.

Gao, X., Chen, Z., Tang, S., Zhang, Y., & Li, J. (2016). Adaptive weighted imbalance learning with application to abnormal activity recognition. *Neurocomputing*, *173*, 1927–1935. https://doi.org/10.1016/j.neucom.2015.09.064

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

Guo, Y., Zhou, W., Luo, C., Liu, C., & Xiong, H. (2016). Instance-based credit risk assessment for investment decisions in p2p lending. *European Journal of Operational Research*, *249*(2), 417–426. https://doi.org/10.1016/j.ejor.2015.05.050

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73*, 220–239. https://doi.org/10.1016/j.eswa.2016.12.035

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

Hand, D. J., & Anagnostopoulos, C. (2014). A better beta for the h measure of classification performance. *Pattern Recognition Letters*, *40*, 41–46. https://doi.org/10.1016/j.patrec.2013.12.011

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *160*(3), 523–541. https://doi.org/10.1111/j.1467-985X.1997.00078.x

Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine Learning*, *45*(2), 171–186. https://doi.org/10.1023/A:1010920819831

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data*

Engineering, *21*(9):1263–1284. https://doi.org/10.1109/TKDE.2008.239

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

Hosmer, D. W., Jr, Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Jiang, C., Wang, Z., Wang, R., & Ding, Y. (2018). Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. *Annals of Operations Research*, *266*(1-2), 511–529. https://doi.org/10.1007/s10479-017-2668-z

Jiang, C., Wang, Z., & Zhao, H. (2019). A prediction-driven mixture cure model and its application in credit scoring. *European Journal of Operational Research*, *277*(1), 20–31. https://doi.org/10.1016/j.ejor.2019.01.072

Kaur, H., Pannu, H. S., & Malhi, A. K. (2019). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys (CSUR)*, *52*(4), 1–36.

Koh, J. (2021). *Gradient boosting with extreme-value theory for wildfire prediction*. arXiv Preprint arXiv:2110.09497.

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

Krawczyk, B., Woźniak, M., & Schaefer, G. (2014). Cost-sensitive decision tree ensembles for effective imbalanced classification. *Applied Soft Computing*, *14*, 554–562. https://doi.org/10.1016/j.asoc.2013.08.014

La Rocca, M., Niglio, M., & Restaino, M. (2023). Bootstrapping binary gev regressions for imbalanced datasets. *Computational Statistics*, *39*(1), 181–213. https://doi.org/10.1007/s00180-023-01330-y

Lagna, A., & Ravishankar, M. (2022). Making the world a better place with fintech research. *Information Systems Journal*, *32*(1), 61–102. https://doi.org/10.1111/isj.12333

Lei, K., Xie, Y., Zhong, S., Dai, J., Yang, M., & Shen, Y. (2020). Generative adversarial fusion network for class imbalance credit scoring. *Neural Computing and Applications*, *32*(12), 8451–8462. https://doi.org/10.1007/s00521-019-04335-1

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., Zhang, J., Yao, X., & Kou, G. (2021). How to identify early defaults in online lending: A cost-sensitive multi-layer learning framework. *Knowledge-Based Systems*, *221*, 106963. https://doi.org/10.1016/j.knosys.2021.106963

Liu, W., Fan, H., & Xia, M. (2022a). Credit scoring based on tree-enhanced gradient boosting decision trees. *Expert Systems with Applications*, *189*, 116034. https://doi.org/10.1016/j.eswa.2021.116034

Liu, W., Fan, H., Xia, M., & Xia, M. (2022b). A focal-aware cost-sensitive boosted tree for imbalanced credit scoring. *Expert Systems with Applications*, *208*, 118158. https://doi.org/10.1016/j.eswa.2022.118158

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, *250*, 113–141. https://doi.org/10.1016/j.ins.2013.07.007

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4768–4777.

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

Medina-Olivares, V., Calabrese, R., Dong, Y., & Shi, B. (2022). Spatial dependence in microfinance credit default. *International Journal of Forecasting*, *38*(3), 1071–1085. https://doi.org/10.1016/j.ijforecast.2021.05.009

Mezei, J., Byanjankar, A., & Heikkilä, M. (2018). *Credit risk evaluation in peer-to-peer lending with linguistic data transformation and supervised learning* [Paper presentation]. Proceedings of the Annual Hawaii International Conference on System Sciences (pp. 1366–1375). https://doi.org/10.24251/HICSS.2018.169

Mushava, J., & Murray, M. (2022). A novel xgboost extension for credit scoring class-imbalanced data combining a generalized extreme value link and a modified focal loss function. *Expert Systems with Applications*, *202*, 117233. https://doi.org/10.1016/j.eswa.2022.117233

Ogundimu, E. O. (2019). Prediction of default probability by using statistical models for rare events. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *182*(4), 1143–1162. https://doi.org/10.1111/rssa.12467

Pai, P.-F., Hsu, M.-F., & Wang, M.-C. (2011). A support vector machine-based model for detecting top management fraud. *Knowledge-Based Systems*, *24*(2), 314–321. https://doi.org/10.1016/j.knosys.2010.10.003

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

Ren, Z., Zhu, Y., Kang, W., Fu, H., Niu, Q., Gao, D., Yan, K., & Hong, J. (2022). Adaptive cost-sensitive learning: Improving the convergence of intelligent diagnosis models under imbalanced data. *Knowledge-Based Systems*, *241*, 108296. https://doi.org/10.1016/j.knosys.2022.108296

Shi, B., Bai, C., & Dong, Y. (2024). A big data analytics method for assessing creditworthiness of smes: Fuzzy equifinality relationships analysis. *Annals of Operations Research*, 1–31. https://doi.org/10.1007/s10479-024-06054-w

Sun, J., Li, J., & Fujita, H. (2022). Multi-class imbalanced enterprise credit evaluation based on asymmetric bagging combined with light gradient boosting machine. *Applied Soft Computing*, *130*, 109637. https://doi.org/10.1016/j.asoc.2022.109637

Ting, K. M. (2002). An instance-weighting method to induce cost-sensitive trees. *IEEE Transactions on Knowledge and Data Engineering*, *14*(3), 659–665.

Velthoen, J., Dombry, C., Cai, J.-J., & Engelke, S. (2021). *Gradient boosting for extreme quantile regression*. arXiv Preprint arXiv:2103.00808.

Vovk, V. (2015). The fundamental nature of the log loss function. In *Fields of Logic and Computation II* (pp. 307–318). Springer.

Wan, Z., Zhang, Y., & He, H. (2017). *Variational autoencoder based synthetic data generation for imbalanced learning* [Paper presentation]. 2017 IEEE Symposium Series on Computational Intelligence (SSCI). (pp. 1–7). IEEE. https://doi.org/10.1109/SSCI.2017.8285168

Wang, X., & Dey, D. (2010). Generalized extreme value regression for binary response data: An application to b2b electronic payments system adoption. *The Annals*

of Applied Statistics, *4*(4), 2000–2023. https://doi.org/10.1214/10-AOAS354

Wu, D., Wang, Z., Chen, Y., & Zhao, H. (2016). Mixed-kernel based weighted extreme learning machine for inertial sensor based human activity recognition with imbalanced dataset. *Neurocomputing*, *190*, 35–49. https://doi.org/10.1016/j.neucom.2015.11.095

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

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. *Advances in Neural Information Processing Systems*, *32*.

Zhu, Y., Zhou, L., Xie, C., Wang, G.-J., & Nguyen, T. V. (2019). Forecasting smes' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach. *International Journal of Production Economics*, *211*, 22–33. https://doi.org/10.1016/j.ijpe.2019.01.032

Appendix A

Table A1. Computing environment.

Platform	Google cloud platform
OS	Ubuntu 16.04
CPU	Intel Skylake Core 16 vCPU
Memory	104GB
Storage	100GB SSD
Programming Environment	Python 3.8
Key Package	LightGBM, Scikit-Learn

Table A2. Variable list and description of LendingClub dataset.

Variable name	Description
annuity2principal	Generated by feature engineering, it is the ratio of loan annuity to the loan principal
dti_x	The ratio is computed by dividing the borrower's total monthly debt payments, excluding mortgage and the requested LC loan, by the borrower's self-reported monthly income.
total_hi_cred_lim	Total high credit/credit limit
total_bc_limit	Total bankcard high credit/credit limit
annual_inc	Borrower's annual income
revol2inc	Generated by feature engineering, it is the ratio of total credit revolving balance over the annual income
total_il_high_credit_limit	Total installment high credit/credit limit
total_bal_ex_mort	Total credit balance excluding mortgage
revol_util	Revolving line utilisation rate.
CreditHistoryLength	Generated by feature engineering, it is the time length from the borrower has the credit for the first time to now
MonthlyContractAMT	Monthly payment amount
income2principal	Generated by feature engineering, it is the ratio of annual income to the loan principal
revol_bal	Total credit revolving balance
total_rev_hi_lim	Total revolving high credit/credit limit
principal	Principal amount of the loan

Table A3. Variable list of three small business finance datasets.

Dataset SBF-1	Dataset SBF-2	Dataset SBF-3
Financial indicators	**Basic information**	Sex
Total asset	Education	Age
Total liability	Maritial status	Education background
TL/TA	Gender	Marital status
OCF/CL	Age	Residential status
Quick ratio	Residential status	Number of family members
Current ratio	Employment status	Ratio of the number of labours over the number of family members
Cash/RFO	Position Level	Personal monthly income
EBIT/CF	Local registered residence	Family monthly income
NCL/(NCL+E)	ID verification	Total asset
OCF/TA	Client status	Total liability
Equity ratio	Business license	TL/TA
Acid-test Ratio	Time length of business license	Ratio of monthly payment over monthly disposable income
OCF/net profit	Ownership of business premise	Number of previous loans
E/(Clending + NCLending)	Industry	Source of repayment
NFA/E	Number of family members	Principal
Cash ratio	Number of labors	Verification of the asset
(E + NCL)/(FA+Investment)	Number of dependent family members	Has guarantors or not
Total unpaid lending/equity	Number of burden population (Dependent/Labor)	Relationship to joint borrower
Total unpaid lending/TA	Family expenses	Age of joint borrower
OCF/TL	Loan purposes	Sex of joint borrower
EBITDA/TL	**Collateral information**	Joint borrower's location
ROE	Guarantor's sex	Joint borrower's monthly income
OCF/revenue	Guarantor's age	Joint borrower's education background
Net profit margin on sales	Guarantor's marital status	Guarantor's sex
ROA	Guarantor's education background	Guarantor's age
Operating profit ratio	Guarantor's Monthly Income	Guarantor's marital status

(*continued*)

Table A3. Continued.

Dataset SBF-1	Dataset SBF-2	Dataset SBF-3
Net profit/total expense and cost	Relationship to the borrower	Guarantor's monthly income
Gross profit margin	Quality of the relationship to the borrower	Guarantor's education background
Net profit/ cost and expense	Guarantor's credit quality	
EBITDA	Guarantor's industry	
EBITDA/Revenue	**Repayment ability**	
Net profit	Current Ratio	
OCF	TL/TA	
CF from operation activities	TL/E	
Receivables turnover ratio	Equity	
Inventory turnover ratio	Monthly instalment to other banks	
Total asset turnover	Amount of personal borrowing	
Current asset turnover	Monthly payment over net income	
Fixed asset turnover	**Profitability**	
Equity turnover	Net ROA	
Working capital/CA	Net profit	
Rate of return on investment	Sales	
Payable turnover ratio	Operating income	
Cash conversion cycle	Total asset	
Revenue growth	Total ROA	
Net profit growth	Monthly income	
TA growth	Monthly tax	
Rate of capital accumulation	**Operating ability**	
R/E Growth	Accounts receivable turnover rate	
Company and legal representative information	Inventory turnover	
Years of employment in the relevant industry	Turnover tate of total assets	
Audited or Not	Average Fixed Assets	
Recognized level of new product	Turnover rate of fixed assets	
Patent Status	Operating expenses	
Date of Establishment	Operating periods	
Level of famous products	Operating area	
The proportion of the total amount of money collected by enterprises through this bank lending default record of legal representative	Number of employees	
Credit card record of legal representative	**Macroeconomic information**	
Marital Status	Per capita savings balance	
Residential Status	Regional GDP growth rate	
Year of residence of legal representative	CPI	
Gender	GDP per capita	
Location	Per capita disposable income	
Industry	Engel coefficient	
Educational background	Industry climate index	
Age		
The value of vehicle and real estates of legal representative		
Year of employment in this position		
Category of registered capital		
Enterprise credit status in the past three years		
Corporate tax record		
The status of bank accounts		
Sales range of products		
Legal issue		
Compliance status		
Number of defaults		
Macroeconomic indicators		
Industry Index		
Year-end balance of per capita savings of urban and rural residents (yuan/person)		
GDP Growth Rate		
CPI		
Per capita disposable income of urban residents (yuan) / person)		
Engel coefficient		
Loan-level Information		
Principal		
Amount of Interests and Principals		
Score of pledge/collateral		
Loan type		
Currency type		