Knowledge-Based Systems 224 (2021) 107068

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Predicting loss given default using post-default information

Ke Li [a,c], Fanyin Zhou [a], Zhiyong Li [b,c], Xiao Yao [d,*], Yashu Zhang [a]

[a] School of Statistics and Center of Statistical Research, Southwestern University of Finance and Economics, 555 Liutai Avenue, Chengdu, 611130, China
[b] School of Finance and Fintech Innovation Center, Southwestern University of Finance and Economics, 555 Liutai Avenue, Chengdu, 611130, China
[c] Collaborative Innovation Center of Financial Security, Southwestern University of Finance and Economics, 555 Liutai Avenue, Chengdu, 611130, China
[d] Business School, Central University of Finance and Economics, 39 South College Road, Beijing, 100081, China

ARTICLE INFO

Article history:
Received 13 May 2020
Received in revised form 14 April 2021
Accepted 20 April 2021
Available online 22 April 2021

Keywords:
Credit risk modelling
Loss given default prediction
Post-default information
Online lending

ABSTRACT

The loss given default (LGD) is an essential component for estimating credit risk according to the international regulatory Basel Accord. Traditionally, LGD models are built based on the characteristics of the loan and the borrower prior to default, a practice which fails to consider the post-default information revealed during the repayment process. We start by uncovering a predictive post-default variable (i.e., a flag that indicates whether the defaulted borrower had cooperated with a debt-settlement company) in the defaulted data from the online lending platform. We then propose a stratified modelling framework to incorporate this variable into LGD prediction. The experimental results demonstrate that LGD prediction is significantly improved by the inclusion of post-default information under the proposed stratified modelling framework. We further show that the predictive performance of the proposed models is robust for the choice of training set and input variables. Our results imply the importance of using post-default information in LGD prediction.

1. Introduction

In the context of credit lending, credit loss is a major concern for both investors and financial regulatory authorities. Under the international regulatory Basel Accord [1,2], expected credit loss provisioning mainly depends on three risk factors: *probability of default* (PD), which is defined as the probability that a debt cannot be repaid; and *loss given default* (LGD), which is measured as the fraction of defaulted debt that cannot be recovered, as well as *exposure at default* (EAD). Thus, the calculation of credit loss not only relies on the event of default itself that occurs as a random incident, but also relates to the proportion of the loan amount that can be recovered at default. Compared to the extensive studies on PD modelling, LGD modelling has presented a new, challenging topic in the field of credit risk management for researchers and practitioners alike. Since the LGD distributions tend to be highly irregular, it is difficult to obtain an accurate prediction using existing regression models [3]. Therefore, new predictive LGD determinants need to be continuously searched for and appropriately utilised to try to improve the accuracy of the predictive modelling. For example, when a loan has gone into default, post-default information (such as collection methods) should directly impact the LGD. However, prevailing LGD models are only built based on the determinants observed prior to default, while post-default information is largely neglected. The reason is that in practice, LGD models are developed based on defaulted loans but are applied to performing loans, including newly issued loans, to estimate the potential credit losses of a performing portfolio. Therefore, researchers simply accept the fact that LGD should be estimated only with features prior to default, because the post-default information is available only for non-performing loans. Therefore, it is clear that new methodologies incorporating post-default information are desperately needed.

Online lenders have experienced a period of rapid growth over the past decade due to the boom in information technology. Online lending platforms, such as LendingClub,[1] provide a more convenient channel for personal customers to access credit. This approach is particularly attractive to customers who have a high-risk profile or relatively poor credit record. Therefore, online lending products such as peer-to-peer (P2P) lending are in general exposed to higher default risk than traditional bank loans. Because the transaction volume of online lending has increased significantly, the desire for effective risk management is rising rapidly. Moreover, online lending platforms follow different collection practices than traditional banks do. For

* Corresponding author.
E-mail addresses: liz@swufe.edu.cn (Z. Li), yaoxiao18@cufe.edu.cn (X. Yao).

[1] LendingClub is the largest peer-to-peer online lending platform founded in the United States. The link to their site is https://www.lendingclub.com/.

https://doi.org/10.1016/j.knosys.2021.107068
0950-7051/

example, the internal service team of LendingClub proactively collaborates with external agencies for debt collection,[2] while the primary choice of banks is to conduct the collection process with their internal collection departments, unless in-house collection becomes uneconomical. In that case, banks might consider alternative options, such as collaborating with external agencies or simply selling off the debts, which again suggests that the risk profile and collection strategies of online lending platforms are notably different from traditional commercial banks. Since the post-default information in the collection process is expected to impose a significant impact on the final recovery rate, we argue that LGD prediction for online lending products should be tailored with their unique features in mind, and post-default variables should be included in the predictive models. A variety of statistical models and machine learning techniques have been proposed to establish the relationship between the LGD and the covariates observed prior to default for various credit products, such as corporate bonds [4,5], corporate loans [6,7] and personal loans [8–10]. However, it is more challenging to include post-default information in LGD prediction, in that such variables are not observable prior to default, and thus relevant studies are almost absent from the current literature available on the topic.

Previous studies on LGD modelling consider only the direct relationship between the pre-default information and the eventual LGD. However, there exists a time lag between the determination of the final LGD and the loan application. During this period, the borrower's behaviours (including post-default behaviours) can influence the final LGD. Incorporating the post-default variables provides the opportunity to better utilise the information of the borrower's behaviours, and reveal the relationships between the pre-default information, the post-default behaviours and the LGD. Therefore, we propose to introduce new modelling frameworks to make use of the post-default variables in LGD prediction. In particular, the post-default variable is regarded as a latent variable in the mixture distribution of the LGD conditional on the pre-default variables. Recent studies have found that LGD is significantly associated with post-default behaviours [11,12]. We thus expect that the post-default information brings incremental improvement in terms of predictive performance.

Motivated by the above analysis, our study aims to 'fill in the gap' by incorporating post-default information in LGD prediction, so that online lending products can improve their predictive accuracy. LendingClub in the United States, the world's largest online lending platform offering P2P lending products, is taken as a representative example of this industry. We extracted the data of defaulted P2P loans from LendingClub for the period between January 2013 and June 2015. The data sample includes variables that record the final LGD, the borrower characteristics at the loan initiation date and the behavioural covariates during the loan repayment period. An attractive feature of LendingClub's data is that it shares post-default information and charged-off losses along with classical loan features and borrower characteristics. Our sample includes post-default variables, such as a 'debt-settlement flag' (0 or 1 valued), which indicates whether a defaulted borrower is cooperating with a debt-settlement company when the collection process is initiated.[3] We find that the LGD distributions for the two subsamples, divided by the debt-settlement flags, present significantly different patterns. In particular, the customers who cooperate with the debt-settlement company have a much lower LGD. These findings imply that this flag, as a kind of post-default information, can give predictive power to the LGD.

From the perspective of credit loss evaluation and regulatory capital allocation, the predictive LGD models need to be built using the information prior to default, and those post-default variables, such as the debt-settlement flag, cannot be directly used as input features, although they are observable in the historical data of the recovery process. Therefore, we propose two modelling frameworks, including stratified and mixture architectures, to apply the debt-settlement flag in LGD prediction. The stratified modelling framework generates the final prediction directly from the outputs of the regression model built on one of the two subsamples, by comparing the predicted values of the classification model with the threshold value. In contrast, the mixture modelling framework is analogous to the two-stage modelling framework proposed by [10], which aggregates the regression outputs of each subsample split by a given binary flag, with respect to the predictive probabilities derived from a classification model. In the experimental analysis, a debt settlement flag is adopted to define subsamples in the stratified and mixture modelling frameworks. For the purpose of benchmarking, we also use a binary flag, which indicates whether LGD equals 0 or not for subsample division in both modelling frameworks.

Our study makes contributions to LGD modelling from a new perspective. In previous studies, the post-default information has not been incorporated into the predictive LGD models due to the paradox of LGD modelling. Although the impact of post-default variables on the LGD has been investigated in the literature [11, 13], the incorporation of post-default variables in LGD prediction is still absent. Our study proposes stratified and mixture modelling frameworks to improve the predictive performance of LGD models using post-default variables. The experimental results show that the model's predictive performance is significantly improved with the inclusion of post-default information under the stratified modelling framework. We further test the sensitivity of the experimental results by shrinking the training set and applying an alternative set of input variables, and we find that the predictive performance of the stratified and mixture models is robust, and the advantage of using post-default information in stratified models remains significant.

Our study shows manifold implications for stakeholders in the online lending business. First, the expected loss is dependent on PD, LGD and EAD. An accurate prediction of LGD and expected loss is beneficial for investors to manage portfolio risk more effectively, and to increase their investment returns. Next, the online lending platforms have a strong incentive to improve the risk evaluation of borrowers, by accounting for both PD and LGD in their loan grading to ultimately facilitate more transactions. Therefore, it is also important for online lending platforms to improve the risk evaluation accuracy of the borrower's credit profile to generate more reasonable loan ratings and interest rates for investors. Borrowers and investors also benefit from this kind of environment of transparent and efficient information exchange. Lastly, the exploration of post-default variables allows the debt-settlement company to optimise their strategies in order to communicate with the borrower more efficiently and to negotiate with the platform, which would potentially improve the collection efficiency in support of the investors and defaulted customers.

The remainder of this paper is organised as follows. Section 2 introduces a literature survey of previous studies on LGD modelling and online lending. Section 3 describes the investigated data and proposes the methodologies that incorporate the

[2] https://help.lendingclub.com/hc/en-us/articles/215483768-What-tools-does-LendingClub-have-to-deal-with-delinquent-borrowers-.

[3] A definition of debt-settlement flags is provided in the link: https://help.lendingclub.com/hc/en-us/articles/115004323368-Recent-and-upcoming-changes-to-the-downloadable-data-files-and-API-services.
An introduction to debt management plans and debt-settlement companies is available in
https://www.lendingclub.com/loans/resource-center/debt-management-plan.
https://help.lendingclub.com/hc/en-us/articles/115011819087-debt-settlement-companies.

observable intermediate events into LGD models. The experimental design is introduced in Section 4, and the experimental results and discussion are reported in Sections 5 and 6. Section 7 concludes the paper.

2. Literature review

This study is related to several strands of studies in LGD modelling and prediction. First, our study extends the studies on LGD determinants and predictors by exploring the predictive power of post-default information. For corporate bonds, the characteristics of collateral, firm capital structure and business cycles are all considered to be significant drivers according to [3]. Acharya et al. [14] find that industry distress indicators remarkably influence corporate bond LGD, and similar evidence can be found in [15]. For bank loans, the impact from loan level features has been shown to be more important. For example, LGD is found to be highly dependent on the loan-to-value for both residential [8,9] and corporate loans [7]. Leow et al. [16] find that macroeconomic variables influence the LGD of residential mortgages more significantly than personal unsecured loans. Only a few studies have shown that collection behaviours significantly impact LGD. For example, Thomas et al. [13] discuss the similarities and differences between in-house and third-party collection on LGD, and emphasise that LGD models should be built with respect to the collection method, due to the differences in their LGD distributions. Han and Jang [11] also identify the significance of debt collection actions on LGD, and find that the LGD model which incorporates recovery action covariates is more successfully predictive than the model which uses firm-specific variables. They also report that the proposed model is not applicable for the prediction of LGD at the loan origination point, because recovery actions are only observable after default. A more recent study from Bellotti et al. [12] shows that LGD forecasting performance can be further enhanced by the incorporation of debt collection features related to the bank recovery process. Such features can be derived when the loans are defaulted, which are unavailable for new customers. Our study fills the gap in the literature by proposing a stratified modelling framework that allows post-default information to be applied for newly issued loans.

Next, our study relates to a growing body of literature on credit risk modelling for online lending products. As a novel FinTech innovation in the financial market, online lending products have demonstrated unique risk patterns and driving factors [17–21], and credit risk has only been investigated from the perspective of PD [22–24]. However, studies related to modelling LGD for online lending products are nearly absent. In this regard, our study attempts to explore the predictive power of the post-default variables based on the online lending data from LendingClub.

Our study is also linked with various LGD modelling techniques applied in the literature. Parametric statistical regression models include simple linear regression and fractional response regression [25], wherein a logit link function is imposed to transform the linear combination into a fractional value bounded by 0 and 1. More complicated statistical regression models have also been proposed to improve LGD predictions. For example, beta transformation is considered to be suitable for LGD modelling since it is defined in the interval (0, 1), and it was first proposed by Moody's LossCalc[4] to fit the irregular LGD distributions. This, however, was found to be less effective than expected [4]. In contrast, machine learning techniques such as decision trees [6] and neural networks [4] have emerged as more competitive than traditional statistical regression models. Moreover, support vector regression techniques have also been proven to be competitive alternative candidates for both retail and corporate LGD predictions in recent studies [5,27–30]. Random forest has been found to outperform other techniques for predicting LGD in more recent studies [12,29,31,32]. In recent studies optimisation methods have also been applied to map the credit ratings with LGD [33,34], but these methods are not calibrated for loan-level characteristics. Considering the complexity of LGD distributions, multi-stage modelling frameworks appear to be more attractive. These frameworks attempt to predict LGD by aggregating the predicted values from subsamples, such as [10] and [35]. Mixture regression models have also adopted similar ideas to improve LGD prediction [36,37]. Furthermore, Yao et al. [38] report that the overall predictive accuracy of a two-stage model strongly relies on the classification accuracy of separating the subsamples. The advantage of a two-stage modelling framework is further presented in [39]. Nazemi et al. [40] find that the modelling accuracy can be further improved by applying a fuzzy decision approach to aggregate the predictive outputs of multiple individual algorithms. Papoušková and Hajek [41] propose a two-stage heterogenous ensemble model to predict PD and EAD separately, but LGD is applied as a fixed parameter in the estimation of expected losses due to the data availability.

The pros and cons of the above LGD modelling techniques are summarised in Table 1. Specifically, statistical regression models, such as OLS and fractional response regression, are more suitable to explore the explanatory variables of LGD, because the model estimates are more easily interpreted. However, the predictive performance of statistical regression models is relatively weak compared to machine learning techniques, which are generally recognised to be more powerful in terms of LGD prediction. Due to the 'black-box' structure of techniques such as neural networks and support vector machines, it is more challenging to interpret these models. Although variable importance can be derived from tree-based models such as random forest and XGBoost, it is still difficult to establish clear-cut relationship between the input and outcome variables. Mixture and multi-stage models offer more advantages when the LGD distribution is more skewed to the boundaries of 0 and 1, where LGD is assumed to follow a mixture distribution formulated as a combination of a classification and a regression problem. Both statistical regression and machine learning techniques can be adapted to the multi-stage modelling frameworks. Thus, it is found to be more competitive than the single-stage models when appropriate techniques are selected at the classification and regression stage [38]. However, none of above methods ever incorporate the post-default information in LGD prediction. Our study adds value to prevailing LGD predictive methods by introducing a novel modelling framework to utilise the post-default variable, together with other pre-default features. Similar to multi-stage models, both statistical methods and machine learning algorithms can be flexibly adapted to improve the model's predictive performance.

3. Modelling framework

In this section, the single-stage LGD modelling framework is briefly introduced first, together with its limitations, and then, the focus is shifted to the proposed mixture and stratified modelling frameworks, to show how to incorporate post-default variables.

[4] Moody's latest version of an internally developed LGD model, LossCalc v3.0, moved back to a simple linear regression model rather than applying a beta distribution transformation. Details can be found in Dwyer and Korablev [26].

Table 1
Summary table of literature on LGD modelling.

Methodologies	Examples	Pros and Cons	References
Statistical regression models	OLS, Fractional response regression and beta regression	Statistical regression models such as OLS are widely used to explore the explanatory variables of LGD. OLS is considered to be a benchmark LGD model in the literature.	Acharya et al. [14]; Dwyer and Korablev [26]; Khieu et al. [7];
Machine learning and optimisation algorithms	Neural networks, Support vector machines and random forests	Machine learning techniques have been found to be more powerful than statistical regression methods. However, it is challenging to explain these models, due to their 'black-box' structures.	Bastos [6]; Loterman et al. [27]; Nazemi et al. [40]; Nazemi et al. [30]; Qi and Zhao [4]; Yao et al. [5]; Shi et al. [33,34]
Mixture and multi-stage models	Inflated beta regression, Zero-adjusted gamma and two-stage model	Since the LGD distributions are presented as resembling either a U-shape or a L-shape in the empirical datasets, the response variable (LGD/recovery rate) is assumed to follow a mixture distribution. Mixture and multi-stage models are proposed to improve the LGD prediction by differentiating the zero and full recovery cases from the remaining cases.	Leow and Mues [9]; Bellotti and Crook [42]; Tong et al. [36]; Altman and Kalotay [43]; Calabrese [37]; Hwang et al. [35]; Yao et al. [38]; Papoušková and Hajek [41]; Bellotti et al. [12]

3.1. Classical LGD model

In the context of LGD prediction in credit risk evaluation, a supervised regression model (as follows) is built based on historical data, as:

$$E(Y|\mathbf{X}) = g(\mathbf{X}) \tag{1}$$

where the model input vector $\mathbf{X}$ represents both loan and borrower characteristics observed prior to default, such as the borrower's credit history, and the output Y is the LGD observed at the loan termination. When $g(\cdot)$ is a linear function, Eq. (1) is equivalent to an OLS regression, which has been widely applied in the literature as a benchmark model [7,14]. In recent studies, more complicated functional forms of $g(\cdot)$, such as generalised linear regression and machine learning techniques, have been applied to improve predictive performance based on the loan and borrower characteristics [4,27,36,37].

Post-default information has rarely been studied in LGD modelling, because it cannot be observed at the loan issuance point. In fact, post-default information is expected to be more predictive for LGD since it directly reflects post-default repayment behaviours. It is reasonable to assume that the defaulted borrower's repayment behaviour is not only affected by the borrower's repayment capability but is also driven by his/her willingness to repay defaulted loans. From amongst the defaulted borrowers, a few may stop repaying the defaulted loan due to a temporary lack of repayment capacity, but they may still be willing to repay the defaulted loan at a later date in order to maintain a good credit record.

In the LendingClub data, there is an observable variable called the 'debt-settlement flag'. This variable indicates whether the defaulted borrower is collaborating with a debt-settlement company. By working with a debt-settlement company, borrowers can partially recover their repayment capability and begin repaying the loan. For example, debt-settlement companies can assist defaulters in negotiating with the lending platform to secure a better loan repayment contract. This flag is not available prior to default, but it is observable for all of the charged-off loans. Fig. 1 shows the LGD distributions for the subsamples given by the debt-settlement flag. The shapes of these two distributions differ significantly, which indicates that the variable debt-settlement flag is strongly related to the final LGD, a relationship which should be explored and accounted for in LGD modelling.

A traditional LGD model formed as Eq. (1), however, can only take the features observed prior to default as the model inputs, and thus fails to incorporate post-default variables. The difficulty is how to make use of the post-default information to improve predictive accuracy, and thus innovative modelling architectures are needed to incorporate this intermediate event. Partially inspired by the work in [10,43,44], we propose mixture and stratified modelling frameworks in the following section.

3.2. Proposed modelling frameworks

(a) Mixture Modelling framework

A mixture model is frequently used to account for a latent variable, such as debt-settlement flags. The conditional density function of LGD is given by

$$p(Y|\mathbf{X}) = \pi(\mathbf{X}) \cdot p(Y|Z=1,\mathbf{X}) + [1-\pi(\mathbf{X})] \cdot p(Y|Z=0,\mathbf{X}) \tag{2}$$

where Y is the LGD, $\mathbf{X}$ denotes the vector of observed inputs, Z is the debt-settlement flag, and $\pi(\mathbf{X}) = P(Z=1|\mathbf{X})$ is the probability conditional on $\mathbf{X}$. Eq. (2) implies that

$$E(Y|\mathbf{X}) = \pi(\mathbf{X}) \cdot E(Y|Z=1,\mathbf{X}) + [1-\pi(\mathbf{X})] \cdot E(Y|Z=0,\mathbf{X}) \tag{3}$$

In a classical mixture model, the latent variable Z is unobservable. Therefore, the model parameters of $\pi(\mathbf{X})$, $E(Y|Z=1,\mathbf{X})$ and $E(Y|Z=0,\mathbf{X})$ are usually estimated by iterative algorithms such as the EM algorithm. However, in the context of LGD prediction, the debt-settlement flag Z is observable in the training data. This extra information allows $\pi(\mathbf{X})$, $E(Y|Z=1,\mathbf{X})$ and $E(Y|Z=0,\mathbf{X})$ to be modelled separately. Therefore, the procedure of a mixture model with the debt-settlement flag Z can be given by Fig. 2 and the following procedure description:

Algorithm Flow of the Mixture Modelling Framework

Step 1: Determine a training set $DATA_{train}$. Split $DATA_{train}$ into $DATA^{+}_{train}$ with $Z=1$ and $DATA^{-}_{train}$ otherwise.

Step 2: On $DATA_{train}$, build a model $\hat{\pi}$ to estimate $\pi = P(Z=1\,|\,\mathbf{X})$:

$$\hat{\pi} = \hat{\pi}(\mathbf{X}) \tag{4}$$

Simultaneously, build a model f_{+} to estimate $E(Y|Z=1,\mathbf{X})$ on $DATA^{+}_{train}$ and a model $\hat{f}_{-}$ to estimate $E(Y|Z=0,\mathbf{X})$ on $DATA^{-}_{train}$ by

$$\hat{Y}^{+} = \hat{f}_{+}(\mathbf{X}) \tag{5}$$

$$\hat{Y}^{-} = \hat{f}_{-}(\mathbf{X}) \tag{6}$$

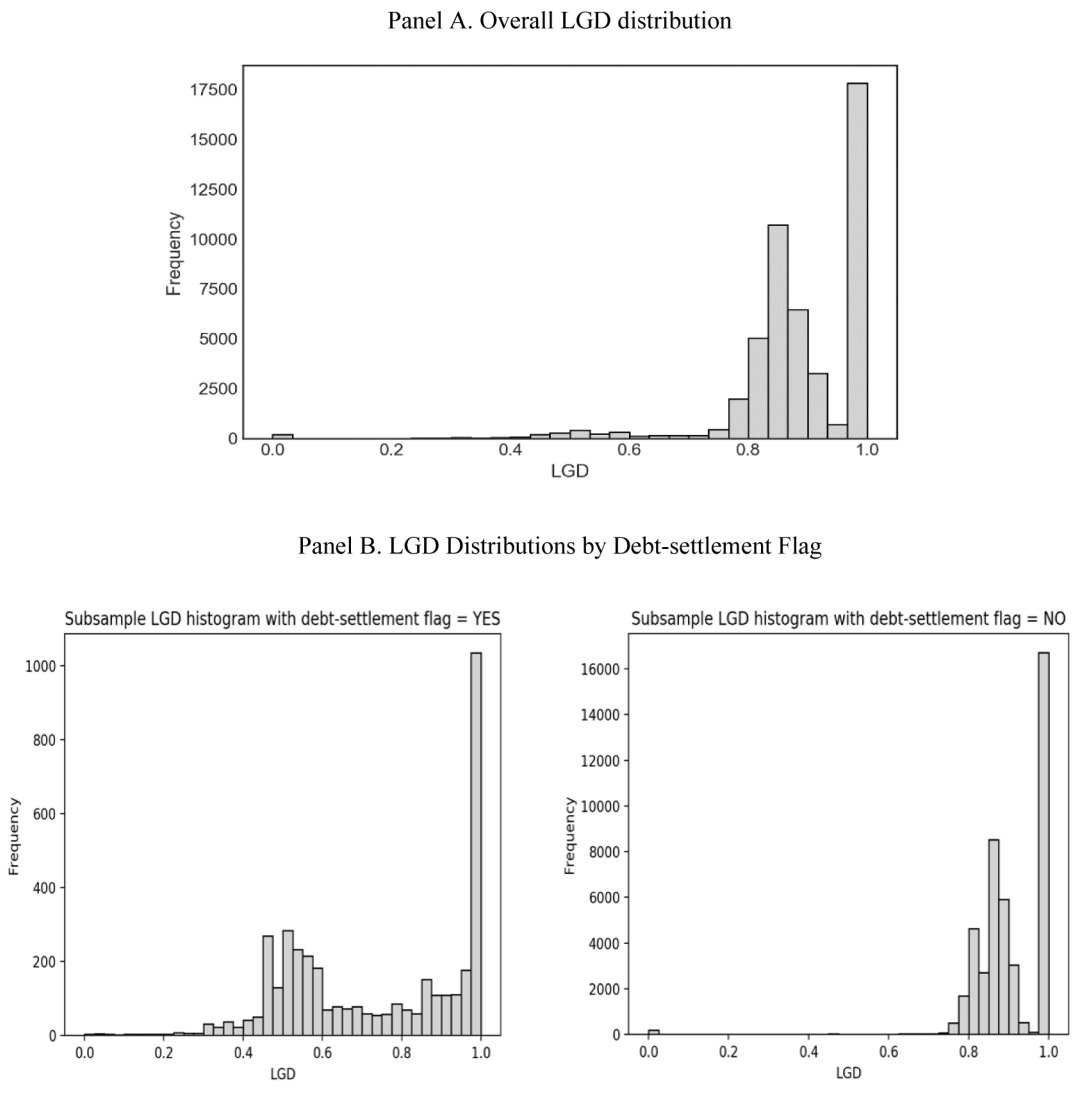

Fig. 1. LGD histogram of defaulted loans.

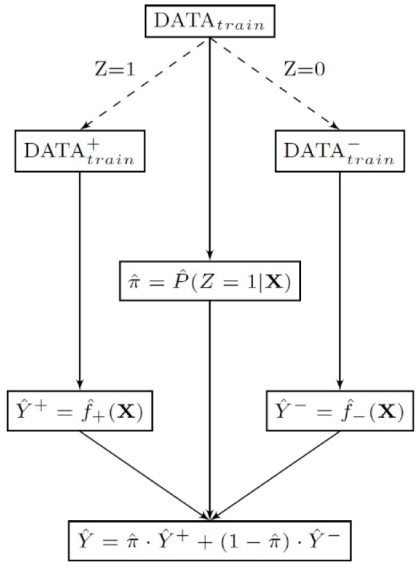

Fig. 2. Flowchart of the mixture modelling framework.

Step 3: Integrate the results to obtain a final model for prediction by

$$\hat{Y} = \hat{\pi} \cdot \hat{Y}^{+} + \left(1 - \hat{\pi}\right) \cdot \hat{Y}^{-} \tag{7}$$

For a new loan with a given X_{new}, calculate and aggregate the results to obtain

$$\hat{Y}_{new} = \hat{\pi}_{new} \cdot \hat{Y}^{+}_{new} + \left(1 - \hat{\pi}_{new}\right) \cdot \hat{Y}^{-}_{new} \tag{8}$$

(b) Stratified Modelling Framework

A classical stratified structure consists of models built separately on subsamples and divided by a given observable categorical feature. With the same notations used in the mixture modelling framework, an adjusted stratified framework with the debt-settlement flag is presented in Fig. 3 and described in the following procedure:

Algorithm Flow of the Stratified Modelling Framework

Steps 1–2: The same as for the mixture model.

Step 3: Integrate the results to obtain a final model for prediction by

$$\hat{Y} = I_{(\hat{\pi} \geq T)} \cdot \hat{Y}^{+} + \left(1 - I_{(\hat{\pi} \geq T)}\right) \cdot \hat{Y}^{-} \tag{9}$$

where T is a threshold hyperparameter that is set manually or by cross-validation.

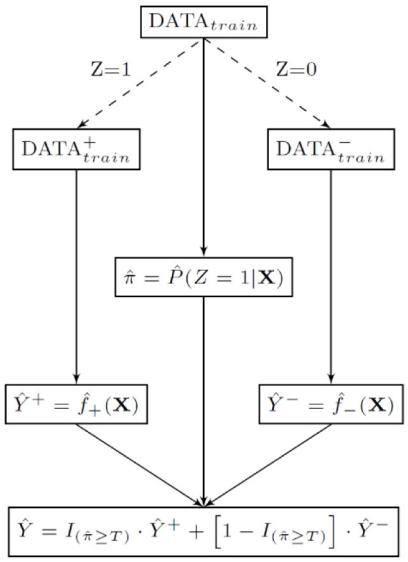

Fig. 3. Flowchart of the stratified modelling framework.

For a new loan with given X_{new}, calculate

$$\hat{Y}_{new} = I_{(\hat{\pi}_{new} \geq T)} \cdot \hat{Y}^{+}_{new} + \left[1 - I_{(\hat{\pi}_{new} \geq T)}\right] \cdot \hat{Y}^{-}_{new} \tag{10}$$

The above two models incorporate post-default information (e.g., debt-settlement flag) from different perspectives. The first model borrows strength from the structure of a mixture distribution via a weighted summation of predictions from subsamples, which can be considered as a model-averaging type of ensemble learning. The second model shares a very similar architecture with the first, but here combine the predictions from subsamples by binary weight (0 or 1), which results in a model which is stratified by the debt-settlement flag. Notably, the stratified model has more flexibility than the mixture model, in the sense that the stratified model has a threshold hyperparameter that can be tuned to achieve different levels of predictive power.

Compared with the existing methods, the proposed approach assumes that the post-default variable is a latent factor in the mixture distribution of LGD conditioning on the application features. Such a hierarchical structure provides a few strengths such as (a) making use of more information drawn from the collection history in LGD prediction, in particular, capturing the variable debt-settlement flag, which cannot be directly applied in traditional predictive models because it is not observable at the application time; (b) applying a combination of a classifier and a regressor that enhances the predictive performance of LGD modelling; and (c) describing the relationship between the application features, post-default behaviours and the final LGD.

On the other hand, a few potential drawbacks should be addressed: (a) The post-default information must be observable so that the proposed approach is applicable. We find the debt-settlement flag to have been a powerful predictor in LGD modelling in the case of the LendingClub dataset, and it is thus necessary to explore other novel powerful post-default variables in other datasets. (b) A more complicated modelling framework might weaken the ease of explanation for the prediction, although such a hierarchical framework is relatively easily explained. (c) While traditional models predict the LGD directly with the inputs of application features, only a single regression technique is needed for prediction. In contrast, both a classifier and a regressor are required in the proposed modelling framework. This approach would potentially increase the complexity of the training process and bring extra computing costs.

3.3. Discussion on the incorporation of post-default variables

Under more general scenarios, an observable variable with post-default information is not limited to a binary flag. Instead, it could be either a multi-class categorical variable or a continuous variable. However, the proposed modelling frameworks are still applicable, as follows:

For a K-class categorical variable Z, the conditional density function of LGD is given by

$$p(Y|\mathbf{X}) = \sum_{i=1}^{K} P(Z = i|\mathbf{X}) \cdot p\left(Y|Z = i, \mathbf{X}\right) \tag{11}$$

where Y is the LGD, $\mathbf{X}$ denotes the vector of observed inputs, Z is the post-default variable, and $P(Z = i|\mathbf{X})$ is the probability conditional on $\mathbf{X}$. This equation implies that

$$E\left(Y|\boldsymbol{X}\right) = \sum_{i=1}^{K} P(Z = i|\boldsymbol{X}) \cdot E\left(Y|Z = i, \boldsymbol{X}\right) \tag{12}$$

Accordingly, to be applicable to the K-class variable, either the mixture modelling framework or the stratified modelling framework can be modified by both adjusting the classification step from binary classification into K-class classification and adjusting the regression step from two regressors into K regressors. For a continuous post-default variable, a simple solution is to bin it as a categorical variable altogether, which can be further processed as discussed above.

If multiple post-default variables do exist, several potential solutions can be considered. For example, multiple single post-default variable models could be established, and the predictions from each individual model aggregated by ensemble methods, such as model averaging. Alternatively, the dimension reduction method, such as principal component analysis or unsupervised autoencoders, could be applied to multiple post-default variables to compress these variables into a single dimensional variable, allowing the modelling framework based on a single post-default variable to become applicable.

4. Experimental design

In this section, we design and implement the experiments on real data from LendingClub to show the effectiveness of the proposed approaches. We first present the data and sample used for the experimental study, and then we briefly introduce the applied regression and classification models. Next, the experimental setup is described with the model performance metrics introduced.

4.1. Sample and variables

P2P lending is a typical form of online lending, and has provided an abundant source of data for research. In this study, we explore LGD prediction based on data publicly released by the world's largest P2P lending platform, LendingClub, in the United States. Our sample includes a total of 41,562 defaulted loans, issued between January 2013 and June 2015. Most of the loan characteristics, along with a few macroeconomic factors, are included to account for their economic impact, except for those with a missing rate greater than 20%. All the modelling candidate variables are shown in Appendix. To validate the model performance, we divide the whole sample into training and testing sets based on the loan issue date. In summary, there are 27,969 loans issued between January 2013 and Dec 2014 in the training set $DATA_{train}$, and the remaining 13,593 loans issued between January 2015 and June 2015 are taken as the testing set $DATA_{test}$.

Table 2
Summarised statistics of LGD.

Panel A. LGD statistics by Loan grade						
Grade	Volume	Mean	Std	Min	Median	Max
A	3422	0.8990	0.1390	0	0.8575	1
B	11 160	0.8921	0.1287	0	0.8532	1
C	13 725	0.8895	0.1220	0	0.8478	1
D	9023	0.8884	0.1233	0	0.8449	1
E	3353	0.8865	0.1233	0	0.8406	1
F	794	0.8792	0.1325	0	0.8380	1
G	85	0.8816	0.1381	0	0.8128	1
Total	41 562	0.8903	0.1230	0	0.8967	1
Panel B. LGD statistics by Home ownership						
Home ownership	Volume	Mean	Std	Min	Median	Max
Mortgage	16 623	0.8848	0.1331	0	0.8448	1
Own	4230	0.8892	0.1259	0	0.8491	1
Rent	20 709	0.8949	0.1198	0	0.8514	1
Panel C. LGD statistics by Debt-settlement flag						
Debt-settlement	Volume	Mean	Std	Min	Median	Max
No	38 199	0.9041	0.1022	0	0.8972	1
Yes	3363	0.7334	0.2253	0	0.7352	1

LendingClub has provided investors with data on numerous loan and borrower characteristics. For example, each loan is assigned a credit grade (from A to G, the safest to the riskiest) as a risk indicator based on an internally developed scoring model. Loans with a safer grade benefit from a lower interest rate. It can be found that there is no significant pattern of the averaged LGD across loan grades according to Panel A of Table 2, and the mean LGD of grades A and B is even higher than the riskier grades, such as F and G. This surprising finding contradicts the intuition that a safer loan grade leads to a lower LGD. Another significant explanatory variable of LGD is home ownership, with its summarised statistics reported in Panel B of Table 2. This aspect shows that the mean LGD of each category remains undifferentiated. Furthermore, Panel C of Table 2 exhibits significant differentiation of the mean LGD with the use of a debt-settlement flag. It can be found that the averaged LGD of the loans for which the borrower has chosen to cooperate with debt collection agencies is notably lower than for the others. This finding confirms the informative value of the debt-settlement flag, as shown in Fig. 1. This evidence suggests that LendingClub does not account for the recovery risk in their internal credit scoring model, and it is therefore necessary to incorporate this post-default information in order to develop a bespoke LGD model which accommodates the features of P2P loans.

The variable importance measure given by the decision-tree-type model is useful for evaluating the strength of the relationship between the input variable and the output variable. A total of 25 variables with variable importance measures above 0.02 are selected as the modelling inputs, given by random forest. Table 3 reports the selected variables and their variable importance values, and Fig. 4 presents the ordered bar chart by variable importance. More details of the variable descriptions can be found in Appendix.

4.2. Model selection

In this section, we briefly introduce the classification and regression models applied in the mixture and stratified modelling framework for the benchmarking study.

4.2.1. Classification model

Logistic regression is a commonly used classification algorithm that maps the input covariates with the binary outputs by a logit function. Logistic regression is quite efficient in terms of computing time and memory requirements, and the model estimates are easily interpreted. Unlike other machine learning techniques, whose performance might be impacted by hyperparameters, no tuning is needed for the logistic regression model. Moreover, it generates posterior predictive outputs that are bounded between 0 and 1, which can be explained as the probability that the model event occurs. In this study, we adopt a penalised logistic regression with L1 regularisation (LASSO logistic) to enforce the penalty on the model parameters. The advantage of LASSO logistic regression is that the input variables can be automatically selected in the model estimation process to avoid overfitting.

4.2.2. Regression models

4.2.2.1. Linear regression. Linear regression provides the simplest form of assessing the quantitative relationship between variables, assuming that all of the input characteristics are linearly related to the dependent variable. Although the predicted outputs of linear regression are not bounded in the interval [0, 1], conflicting with the LGD definition, it is still considered a benchmark model in the context of LGD modelling, and it remains a robust model despite this main drawback [4,27]. The latest version of Moody's internal LGD model [26] has replaced the previous beta distribution transformation method with linear regression. Therefore, linear regression is selected here as the benchmark technique in our experimental study.

4.2.2.2. Neural networks. Neural networks can be regarded as a flexible nonlinear model architecture, inspired by the way the human brain processes information. Neural networks have been proved to be powerful tools for approximating a nonlinear (or linear) loss function with arbitrary accuracy and an appropriate model setup. The structure of neural networks is composed of neurons, weights and activation functions. The neurons represent the input and output values at each step, and the weights are adjusted to approximate the target function with the help of non-linear activation functions. To train a neural network, the various weights of a neural network should first be initialised, and activation functions should then be specified. The output values can be obtained by propagating the input values through the network, and calculating the error based on the cost function. Next, the weights are updated by propagating the error from the output layer back to the input layer to minimise the error. This process repeats until the stopping criteria are met. Neural networks have been proven to yield high prediction accuracy in LGD modelling for corporate bonds [4], and we expect them to still be capable of obtaining competitive predictive performance for online lending products in light of the large volume of data samples.

4.2.2.3. Support vector regression. Support vector regression (SVR) is the application of a support vector machine (SVM) in the field of regression. In contrast to the type of classification problem that seeks to maximise the margin between the hyperplanes, SVR is designed to minimise the "total deviation" of the observations [45,46]. To ensure minimal total errors, the observations are expected to be bounded by the two boundary lines, and the solution of the regression model is equivalent to the maximum margin of hyperplanes, transformed into a classification problem. Slack variables are added into the objective function with a penalty parameter to control the impact of total errors. An attractive feature of using SVM techniques is that the non-linear inseparable classes can be separated into a higher dimensional space by the use of a mapping function. In the modelling process, it is simply necessary to calculate the inner product of these mapping functions. Therefore, kernel functions formed as the

Table 3
Modelling input variables.

Variables	Score	Mean	Std	Min	Median	Max
mo_sin_old_il_acct	0.0435	119.4	54.9	1	124	545
dti	0.0427	19.7	8.2	0	19.49	40
mo_sin_old_rev_tl_op	0.0394	166.8	91.9	4	150	757
annual_inc	0.0375	65 312.4	74 788.8	4000	55 000	8 706 582
loan_amnt	0.0349	12 453.4	7631.2	1000	10 000	35 000
bc_open_to_buy	0.0347	6102.9	10 096.4	0	2599	196 236
avg_cur_bal	0.0327	9709.1	12 148.9	0	4866.5	297 866
bc_util	0.0325	67.0	26.1	0	71.8	255
revol_bal	0.0321	13 860.8	18 495.7	0	10 026.5	1 746 716
revol_util	0.0318	56.9	22.5	0	57.6	180
total_bc_limit	0.0304	15 853.4	16 273.0	0	10 900	560 800
mths_since_recent_bc	0.0287	19.6	25.2	0	11	451
total_rev_hi_lim	0.0281	25 093.6	26 217.8	0	18 900	1 998 700
total_il_high_credit_limit	0.0272	37 336.7	37 947.6	0	28 042.5	597 338
tot_hi_cred_lim	0.0265	128 340.5	137 620.7	300	73 109	3 867 129
total_bal_ex_mort	0.0250	44 352.5	41 872.9	0	33 558	1 896 461
mths_since_recent_inq	0.0246	6.2	5.6	0	5	24
tot_cur_bal	0.0237	104 546.8	124 192.6	0	50 924.5	3 437 283.0
FICO	0.0231	686.5	23.1	662.0	682.0	847.5
num_il_tl	0.0226	8.4	7.3	0	6	89
pct_tl_nvr_dlq	0.0219	94.3	8.2	24.2	97.7	100.0
mo_sin_rcnt_rev_tl_op	0.0219	10.2	12.6	0	7	372
num_rev_accts	0.0216	15.0	8.1	2	14	81
num_bc_tl	0.0209	8.4	4.9	0	8	52
SP500	0.0207	0.0108	0.0263	−0.0362	0.0179	0.0534

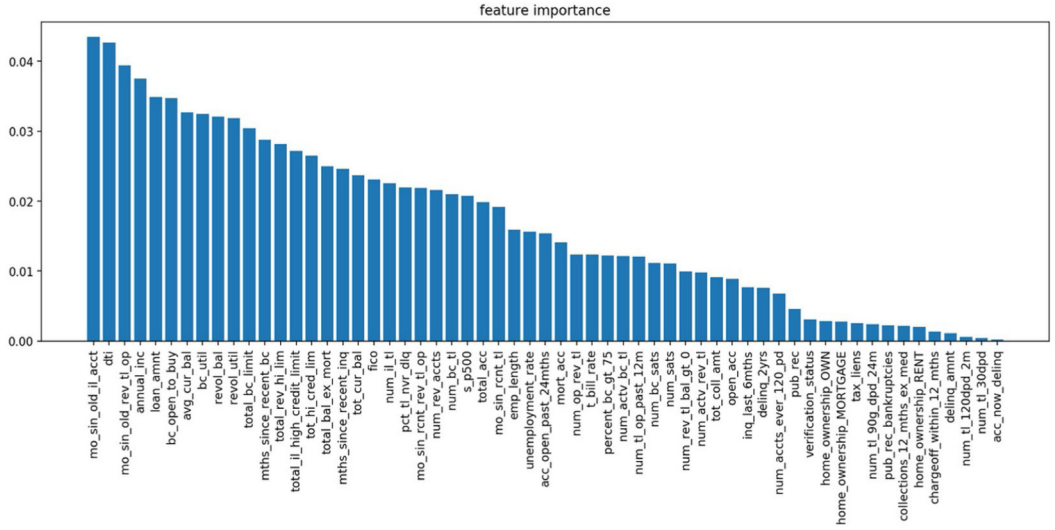

Fig. 4. Variable importance measure by random forest.

inner product of two vectors are introduced. This step significantly simplifies the calculation and is known as the 'kernel trick.' SVR has also been found to be a powerful technique for LGD prediction in the literature [5,28], and thus, it is selected as another benchmark regression technique.

4.2.2.4. Decision tree. The decision tree is a prediction model of the attribute structure, which represents a mapping relationship between the input attributes and output values. A typical regression decision tree, introduced by [47], refers to the CART algorithm. Such a model is obtained by recursively partitioning the feature space and fitting a simple prediction model within each partition. As a result, the partitioning can be represented graphically as a decision tree. Bastos et al. [6] applied the decision tree to the prediction of the default loss rate of bank loans. Qi and Zhao [4] found that the decision tree achieves relatively competitive predictions compared to other non-parametric techniques, and it is not inclined to cause over-fitting issues, which makes this method even more attractive than neural networks.

4.2.2.5. Random forest. Random forest (RF), proposed by [48], combines a collection of individual decision trees and aggregates the prediction by voting or averaging. Specifically, the RF algorithm randomly generates bootstrap samples from the original training set to grow a decision tree, with this step being repeated multiple times. The final predicted value can be averaged for the regression model. Feature selection is conducted in the training process of the RF algorithm, which splits each node randomly and compares the errors generated under different circumstances. The number of selected features and the trees to grow should be tuned to balance both the minimisation of the training error and the maximisation of the generalisation capability. Since RF is effectively an aggregation of the decision trees, it is therefore expected to present a better predictive performance than that offered by decision trees.

4.2.2.6. XGBoost. The extreme gradient boosting decision tree, referred to as XGBoost, has been recognised as one of the most powerful predictive algorithms since it was proposed in 2014 [42].

Gradient boosting was first proposed by [48] as an ensemble learning algorithm that improves predictive accuracy by iteratively reducing training error. To achieve this goal, gradient boosting adjusts each individual classifier or regressor based on the negative gradient of the loss function. Similarly, XGBoost proposes adding a regularisation term in the training objective function to control the model's complexity, and to improve the model's robustness together with the loss term. It uses up to the second order Taylor expansion of the loss function, and allows the user to run a cross-validation at each iteration of the boosting process. Therefore, it is easy to obtain the optimum number of boosting iterations in a single run. Furthermore, XGBoost has several other features that allow it to perform well in many supervised learning tasks.

4.3. Experimental setup

We first construct a classification model to obtain the predicted probabilities that a given instance belongs to a specific subsample, denoted as $\hat{\pi} = \hat{P}(Z = 1|\mathbf{X})$, on the training set $DATA_{train}$. A classical choice is logistic regression, which has been widely applied in classification problems — especially in the two-stage model proposed by [10]. As introduced in Section 4.2, we choose a LASSO logistic model to solve the classification problem and to avoid overfitting. Next, we build regression models to generate the predicted LGD $\hat{Y}_+$ and $\hat{Y}_-$ on the subsamples $DATA^+_{train}$ and $DATA^-_{train}$, respectively. Then, the final predicted LGD $\hat{Y}$ can be derived as in Eq. (7) or (9). Predicted values that are less than 0 or larger than 1 are truncated at 0 and 1, following previous studies. The predicted LGD calculated by the mixture and stratified modelling frameworks are denoted as $\hat{Y}_{mixture}$ and $\hat{Y}_{stratified}$, respectively. Since the performance of stratified models is dependent on the choice of threshold T, the threshold value is fine-tuned, together with other hyperparameters, by five-fold cross-validation on the training set.

To account for the randomness of the model sample, the training and testing sets introduced in Section 4.1 are both randomly partitioned into ten subsets. In summary, there are ten training–testing paired subsets where each candidate algorithm is trained and tested. In the experimental study, categorical features are one-hot coded when necessary, and the model hyperparameters are set as classical values or chosen by cross-validation. Note that the threshold T for the stratified model is set to 0.5. Performance metrics, including the mean absolute error (MAE) and root mean square error (RMSE), are applied for each LGD model evaluation, as shown in Eq. (13).

$$\begin{aligned} \text{RMSE} &= \sqrt{\frac{1}{n}\sum_{\text{i}}(y_i - \widehat{y_i})^2} \\ \text{MAE} &= \frac{1}{n}\sum_{\text{i}}|y_i - \widehat{y_i}| \end{aligned} \tag{13}$$

5. Experiments and discussion

5.1. Experimental results

In this section, the LGD predictive performance of single-stage, mixture and stratified models is presented and discussed.

The single-stage models, including decision trees (DT), linear regression with least absolute shrinkage and selection operator (LASSO), neural networks (NN), ordinary linear regression (OLS), random forest (RF), support vector regression (SVR) and XGBoost, are applied to LGD prediction directly, as introduced in Section 3.1. We also examine the model performance of the mixture and stratified models, which incorporate the debt-settlement flag, based on the discussions in Sections 3.2 and 3.3. In fact, the two-stage model proposed in [10] can be regarded as a type of mixture model that splits the sample based on LGD values. Since a large volume of instances with an LGD of 1 are observed in Fig. 1, we create another binary flag, termed the LGD flag, in the training set to indicate whether LGD equals 1 or not. We employ these two flags in both the mixture and stratified modelling frameworks for more complete comparisons.

In summary, there are five classes of models to be compared: single stage models, mixture models with LGD flag (Mixture_LGD), mixture models with post-default information (Mixture_PDI), stratified models with LGD flag (Stratified_LGD) and stratified models with post-default information (Stratified_PDI). For all the mixture and stratified models, the LASSO logistic regression is chosen to be applied at the classification stage, because the LASSO logistic regression can simultaneously complete the tasks of probability estimation and classification with a tuneable hyperparameter to avoid overfitting. The performance metrics of all five classes of models on the testing set are presented in Tables 4 to 6.

For the single-stage models, Table 4 shows the mean and standard deviations of MAE and RMSE from the ten testing sets. It is noted that the MAE of every model is not significantly different from the MAE of other models, although LASSO and SVR outperform others marginally, and their advantage is more significant in terms of the RMSE. XGBoost and RF are not found to be more advantageous than other learning algorithms. The Mixture_LGD demonstrates slight improvement over the single-stage models, according to Panel A of Table 5. However, the predictive performance of the mixture models with post-default information (Mixture_PDI) shows no significant advantage to the single-stage models, based on Panel B of Table 5. Last, Table 6 shows that Stratified_PDI models outperform other single stage or mixture models in terms of both MAE and RMSE for all learning algorithms, but that the performance of Stratified_LGD models is less competitive than any other class of models. To further validate the superiority of the stratified models, we pick the best model of each class based on MAE and RMSE, and the four best models are compared with each other using the paired t-test. The p values are presented in Panels A and B of Table 7.

Several key pieces of evidence can be summarised according to the outputs reported in Tables 4 to 7. First, the incorporation of post-default information improves the model's overall predictive power. We find that the performance of Mixture_PDI models is in general comparable to any of the single-stage models or the LGD-flag models. Moreover, the Stratified_PDI models are more competitive than all of the other models, and the improvement is statistically significant at either the 1% or 5% level in terms of MAE or RMSE. In summary, the LGD prediction can be significantly improved with the inclusion of debt settlement flags and an appropriate choice of modelling framework.

Next, we find that the performance of single-stage models is not sensitive to the choice of regression algorithms. LASSO and NN appear to show better performance under Mixture_LGD and Stratified_LGD modelling frameworks. In contrast, the predictive accuracy of SVR under the Mixture_PDI and Stratified_PDI frameworks represents a notable improvement when compared with other classification techniques. This finding is consistent with the evidence presented in [38], which shows the advantageous performance of SVR in two-stage models. Surprisingly, XGBoost does not demonstrate any superiority, despite having consistently dominated in many data science contests, such as those provided by Kaggle.

Lastly, although the Stratified_PDI models are more competitive than the others according to Table 7, the Stratified_LGD models present no advantage compared to the single stage or the

Table 4
LGD prediction of single stage models.

Model	MAE		RMSE	
	Mean	Std	Mean	Std
DT	0.0838	0.0035	0.1199	0.0051
Lasso	0.0836	0.0035	0.1189	0.0053
NN	0.0839	0.0034	0.1192	0.0051
OLS	0.0836	0.0048	0.1191	0.0058
RF	0.0850	0.0042	0.1206	0.0050
SVR	**0.0819**	0.0023	**0.1176**	0.0050
XGBoost	0.0864	0.0065	0.1231	0.0091

Table 5
LGD prediction of mixture models.

Panel A. Mixture_LGD				
Model	MAE		RMSE	
	Mean	Std	Mean	Std
DT	0.0835	0.0033	0.1198	0.0049
Lasso	**0.0833**	0.0036	0.1192	0.0049
NN	0.0844	0.0037	0.1198	0.0050
OLS	0.0839	0.0045	0.1191	0.0054
RF	0.0847	0.0047	0.1207	0.0052
SVR	0.0840	0.0058	**0.1189**	0.0069
XGBoost	0.0856	0.0049	0.1215	0.0054
Panel B. Mixture_PDI				
Model	MAE		RMSE	
	Mean	Std	Mean	Std
DT	0.0841	0.0035	0.1197	0.0051
Lasso	0.0837	0.0035	0.1191	0.0056
NN	0.0840	0.0029	0.1191	0.0052
OLS	0.0836	0.0048	0.1193	0.0059
RF	0.0852	0.0046	0.1209	0.0070
SVR	**0.0825**	0.0025	**0.1178**	0.0052
XGBoost	0.0847	0.0040	0.1201	0.0055

mixture models. Both the stratified model and the mixture model rely on the classification stage performed prior to the regression stage, but the roles of the classification stage are notably different. For the stratified models, the prediction of the regression stage is taken as the final predictive output by comparing the classification result and the threshold, while the mixture models apply the weighted average of the outputs from both regression models based on the classification outputs. When the sample is divided by the LGD flag, the final prediction from the Stratified_LGD models is taken as $I_{(\hat{\pi}\geq T)}\cdot 1+\left(1-I_{(\hat{\pi}\geq T)}\right)\cdot\hat{Y}^{-}$, where $\hat{Y}^{-}$ is the prediction from the subsample with LGD less than 1. Thus, for an observation with LGD less than 1, its predicted LGD is be 1 only if it has been misclassified, so the true value and the predicted value tend to be far away from each other. Therefore, the cost of such misclassification can become too high for the Stratified_LGD models. In contrast, the final prediction of the Stratified_PDI models is given by $I_{(\hat{\pi}\geq T)}\cdot\hat{Y}^{+}+\left(1-I_{(\hat{\pi}\geq T)}\right)\cdot\hat{Y}^{-}$, where both $\hat{Y}^{+}$ and $\hat{Y}^{-}$ are bounded in the interval of [0, 1] and are less likely to be affected by the misclassification. It can be noted that the Mixture_LGD models are better than the Stratified_LGD models because the prediction is given by the weighted average of 1 and $\hat{Y}^{-}$ with respect to the estimated probabilities of each class. Taking the above into account, it is not surprising to find that the performance of different modelling frameworks is dependent on the definitions and characteristics of the flag variables.

5.2. Sensitivity analysis

5.2.1. Alternative sample selection

To test the sensitivity of the predictive performance presented in Section 5.1, we replicate our experiments by altering the training sample. Specifically, the original training set defined based on

Table 6
LGD prediction of stratified models.

Panel A. Stratified_LGD				
Model	MAE		RMSE	
	Mean	Std	Mean	Std
DT	0.0897	0.0043	0.1382	0.0085
Lasso	0.0897	0.0042	0.1378	0.0090
NN	0.0888	0.0044	**0.1343**	0.0073
OLS	0.0904	0.0048	0.1384	0.0083
RF	0.0905	0.0046	0.1387	0.0079
SVR	**0.0865**	0.0048	0.1354	0.0120
XGBoost	0.0909	0.0050	0.1390	0.0075
Panel B. Stratified_PDI				
Model	MAE		RMSE	
	Mean	Std	Mean	Std
DT	0.0798	0.0031	0.1169	0.0058
Lasso	0.0792	0.0030	0.1162	0.0061
NN	0.0802	0.0028	0.1165	0.0057
OLS	0.0801	0.0032	0.1168	0.0055
RF	0.0810	0.0044	0.1181	0.0076
SVR	**0.0779**	0.0025	**0.1150**	0.0058
XGBoost	0.0807	0.0033	0.1173	0.0056

loans issued from January 2013 to December 2014 is shortened to two new sets, from January 2014 to December 2014, and from July 2014 to December 2014. The testing set remains. The training and testing sets are divided into ten paired subsets, and the hyperparameters of models are tuned and implemented as above. The modelling outputs are presented in Tables 8 to 10.

In general, the predictive performance of all types of models is almost unchanged when compared to the original results. The slight deterioration of the modelling performance indicates that the LGD models are not sensitive to the selection of a training sample, although more accurate prediction is expected with a longer training set horizon. It can also be noted that Stratified_PDI models outperform others, regardless of how the training sample is reset. LASSO and SVR still excel, while NN also stands out as another strong competitor among the regressors. In summary, the sensitivity analysis provides further evidence that confirms the stability and superiority of the proposed stratified modelling framework, and the importance of incorporating post-default information into LGD prediction.

5.2.2. Alternative input variables

The results of the previous analysis are based on the input variables selected by random forest. To investigate whether the modelling outputs are sensitive to the choice of input variables, we repeat the tests in Section 5.1 by applying the full set of candidate variables as the inputs, and report the modelling outputs in Tables 11 to 13. The paired t-test results between the best classifiers in each modelling class are reported in Table 14. The results, using all candidate variables for prediction, remain consistent with the outputs based on the selected variables. SVR under the Stratified_PDI and Mixture_PDI modelling framework still stands out among all of the models, followed by LASSO under the Mixture_LGD framework, and the Stratified_LGD models are less competitive than the others. This finding suggests that the predictive power of the classification models is stable using selected variables in Section 4.1, which is comparable with the models based on the full set of candidate variables.

In summary, the experimental results present several implications for LGD modelling with the application of post-default information. First, the incorporation of debt-settlement flags based on the proposed stratified modelling framework demonstrates significant improvement in terms of the predictive performance, which implies the positive effect of post-default information in

Table 7
Paired t-test of the best models of each model group.

Panel A. MAE					
Models	Single stage - SVR	Mixture_LGD - Lasso	Mixture_PDI - SVR	Stratified_LGD - SVR	Stratified_PDI - SVR
Single stage - SVR	N.A.				
Mixture_LGD - Lasso	0.0419**	N.A.			
Mixture_PDI - SVR	0.1280	0.1607	N.A.		
Stratified_LGD - SVR	0.0535*	0.2421	0.0912*	N.A.	
Stratified_PDI - SVR	0.0000***	0.0001***	0.0000***	0.0013***	N.A.
Panel B. RMSE					
Models	Single stage - SVR	Mixture_LGD - SVR	Mixture_PDI - SVR	Stratified_LGD - NN	Stratified_PDI - SVR
Single stage - SVR	N.A.				
Mixture_LGD - SVR	0.2972	N.A.			
Mixture_PDI - SVR	0.4352	0.3579	N.A.		
Stratified_LGD - NN	0.0001***	0.0000***	0.0000***	N.A.	
Stratified_PDI - SVR	0.0002***	0.0137**	0.0001***	0.0000***	N.A.

Table 8
LGD prediction under alternative training samples — Single stage models.

Model	Jan2014 to Dec2014				Jul2014 to Dec2014			
	MAE		RMSE		MAE		RMSE	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
DT	0.0856	0.0102	0.1264	0.0224	0.0859	0.0041	0.1302	0.0132
Lasso	0.0836	0.0031	0.1197	0.0066	0.0848	0.0042	0.1206	0.0067
NN	**0.0829**	0.0030	**0.1183**	0.0054	**0.0825**	0.0028	**0.1180**	0.0057
OLS	0.0924	0.0144	0.1273	0.0126	0.1034	0.0569	0.1370	0.0500
RF	0.0850	0.0059	0.1220	0.0093	0.0860	0.0034	0.1222	0.0064
SVR	0.0842	0.0040	0.1190	0.0063	0.0850	0.0055	0.1195	0.0073
XGBoost	0.0860	0.0071	0.1226	0.0103	0.0864	0.0029	0.1223	0.0053

Table 9
LGD prediction under alternative training samples.

Panel A. Mixture_LGD								
Model	Jan2014 to Dec2014				Jul2014 to Dec2014			
	MAE		RMSE		MAE		RMSE	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
DT	0.0818	0.0034	0.1189	0.0065	0.0827	0.0024	0.1193	0.0058
Lasso	**0.0803**	0.0024	**0.1168**	0.0060	**0.0823**	0.0028	**0.1181**	0.0052
NN	0.0816	0.0024	0.1172	0.0055	0.0846	0.0035	0.1196	0.0055
OLS	0.0875	0.0104	0.1235	0.0101	0.0896	0.0187	0.1243	0.0160
RF	0.0817	0.0024	0.1179	0.0055	0.0839	0.0029	0.1200	0.0056
SVR	0.0843	0.0039	0.1193	0.0056	0.0873	0.0046	0.1218	0.0053
XGBoost	0.0817	0.0028	0.1177	0.0057	0.0843	0.0031	0.1198	0.0053
Panel B. Mixture_PDI								
Model	Jan2014 to Dec2014				Jul2014 to Dec2014			
	MAE		RMSE		MAE		RMSE	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
DT	0.0834	0.0039	0.1194	0.0063	0.0834	0.0033	0.1192	0.0058
Lasso	**0.0824**	0.0034	**0.1180**	0.0056	0.0826	0.0033	0.1190	0.0062
NN	0.0830	0.0027	0.1184	0.0051	**0.0825**	0.0027	**0.1179**	0.0057
OLS	0.0902	0.0101	0.1251	0.0091	0.1110	0.0681	0.1458	0.0608
RF	0.0837	0.0034	0.1194	0.0058	0.0841	0.0028	0.1203	0.0056
SVR	0.0834	0.0029	0.1183	0.0055	0.0829	0.0040	0.1181	0.0064
XGBoost	0.0836	0.0034	0.1189	0.0058	0.0844	0.0031	0.1203	0.0057

Table 10
LGD prediction under alternative training samples.

Panel A. Stratified_LGD								
Model	Jan2014 to Dec2014				Jul2014 to Dec2014			
	MAE		RMSE		MAE		RMSE	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
DT	0.0905	0.0046	0.1456	0.0057	0.0924	0.0033	0.1464	0.0048
Lasso	0.0906	0.0046	0.1458	0.0059	**0.0891**	0.0034	**0.1378**	0.0053
NN	**0.0888**	0.0042	**0.1395**	0.0058	0.0914	0.0039	0.1458	0.0052
OLS	0.0911	0.0043	0.1470	0.0063	0.0920	0.0037	0.1466	0.0050
RF	0.0906	0.0046	0.1457	0.0057	0.0919	0.0041	0.1468	0.0058
SVR	0.0907	0.0046	0.1457	0.0057	0.0917	0.0036	0.1466	0.0057
XGBoost	0.0906	0.0046	0.1456	0.0057	0.0918	0.0038	0.1463	0.0051
Panel B. Stratified_PDI								
Model	Jan2014 to Dec2014				Jul2014 to Dec2014			
	MAE		RMSE		MAE		RMSE	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
DT	0.0797	0.0029	0.1169	0.0063	0.0794	0.0027	0.1166	0.0054
Lasso	0.0786	0.0026	0.1157	0.0058	0.0796	0.0024	0.1173	0.0061
NN	0.0800	0.0021	0.1163	0.0055	0.0806	0.0026	0.1168	0.0055
OLS	0.0834	0.0060	0.1203	0.0079	0.1007	0.0528	0.1376	0.0455
RF	0.0803	0.0027	0.1171	0.0060	0.0818	0.0025	0.1189	0.0055
SVR	**0.0780**	0.0023	**0.1152**	0.0057	**0.0792**	0.0025	**0.1159**	0.0057
XGBoost	0.0800	0.0025	0.1165	0.0057	0.0814	0.0025	0.1185	0.0053

Table 11
LGD prediction of single stage models — All candidate variables.

Model	MAE		RMSE	
	Mean	Std	Mean	Std
DT	0.0839	0.0035	0.1204	0.0058
Lasso	**0.0836**	0.0035	**0.1189**	0.0043
NN	0.0840	0.0032	0.1190	0.0049
OLS	0.0881	0.0110	0.1228	0.0088
RF	0.0853	0.0042	0.1212	0.0051
SVR	0.0838	0.0026	0.1189	0.0048
XGBoost	0.0863	0.0060	0.1226	0.0080

LGD prediction. However, the mixture models with post-default variables do not exhibit any advantage over the single-stage models, which suggests that more effort is needed to further explore how to make use of post-default variables more effectively in LGD modelling. Second, the choice of regressors is strongly relevant to the performance of both stratified and mixture models. In particular, the stratified models achieve better predictive accuracy when SVR is selected as the regressor. Although XGBoost does not present competitive performance as expected, it is still worthwhile to investigate the application of machine learning techniques to LGD prediction. Third, the proposed modelling frameworks are robust in terms of both sample and variable selection. The Stratified_PDI models are consistently more competitive than the others when the training sample is shortened. Such conclusions also hold true when a new set of candidate variables is applied as the modelling inputs. Our results suggest that the inclusion of post-default variables brings incremental improvement in LGD prediction, and the evidence is also robust in terms of sensitivity analysis.

Table 12
LGD prediction of mixture models — All candidate variables.

Panel A. Mixture_LGD				
Model	MAE		RMSE	
	Mean	Std	Mean	Std
DT	0.0828	0.0036	0.1195	0.0071
Lasso	**0.0826**	0.0035	**0.1181**	0.0056
NN	0.0831	0.0037	0.1186	0.0059
OLS	0.0829	0.0041	0.1185	0.0057
RF	0.0841	0.0043	0.1200	0.0058
SVR	0.0836	0.0051	0.1186	0.0068
XGBoost	0.0841	0.0043	0.1197	0.0058
Panel B. Mixture_PDI				
Model	MAE		RMSE	
	Mean	Std	Mean	Std
DT	0.0840	0.0033	0.1194	0.0053
Lasso	0.0838	0.0035	0.1192	0.0055
NN	0.0840	0.0031	0.1191	0.0048
OLS	0.0869	0.0078	0.1221	0.0071
RF	0.0852	0.0046	0.1209	0.0064
SVR	**0.0829**	0.0034	**0.1180**	0.0056
XGBoost	0.0849	0.0044	0.1203	0.0057

Table 13
LGD prediction of stratified models — All candidate variables.

Panel A. Stratified_LGD				
Model	MAE		RMSE	
	Mean	Std	Mean	Std
DT	0.0897	0.0047	0.1410	0.0076
Lasso	0.0896	0.0046	0.1400	0.0080
NN	0.0884	0.0041	**0.1357**	0.0066
OLS	0.0904	0.0045	0.1407	0.0068
RF	0.0903	0.0045	0.1409	0.0070
SVR	**0.0876**	0.0053	0.1386	0.0101
XGBoost	0.0903	0.0045	0.1407	0.0070
Panel B. Stratified_PDI				
Model	MAE		RMSE	
	Mean	Std	Mean	Std
DT	0.0797	0.0030	0.1166	0.0059
Lass o	0.0793	0.0030	0.1163	0.0060
NN	0.0802	0.0027	0.1165	0.0054
OLS	0.0824	0.0044	0.1186	0.0061
RF	0.0810	0.0043	0.1182	0.0069
SVR	**0.0782**	0.0029	**0.1152**	0.0058
XGBoost	0.0810	0.0036	0.1177	0.0057

6. Future research directions

In this research, the debt-settlement flag is only taken as a one-dimensional example of post-default variables, to partially incorporate the post-default information. During the loan collection process, borrowers and lenders could take various actions to facilitate the loan recovery, some of which might be predictive for LGD as well. Therefore, we plan to explore such predictive post-default variables for LGD modelling in future.

In addition to the frameworks proposed in this research, we plan to develop other novel frameworks that can provide more effective predictions if more data are available, such as the use of deep learning techniques. Deep learning algorithms have demonstrated superior predictive power in many fields, despite the loss of interpretability. However, recent research has found that deep learning techniques can provide local interpretability in predictive tasks. Therefore, the application of deep learning algorithms can be considered to predict LGD with the incorporation of post-default information.

We propose to explore novel modelling frameworks to predict PD and LGD simultaneously. PD and LGD are known to be correlated with each other, and such correlations should be addressed in the modelling process [33,34]. The portfolio losses might be underestimated if the PD and LGD correlation is not considered. To evaluate the portfolio expected loss more accurately, we aim to develop a PD and LGD joint modelling framework to incorporate the latent factors that drive both PD and LGD, together with the application of variables from the repayment process and any post-default events that may occur.

More generally, post-default information can be regarded as a type of intermediate information that is observable between the time of observing X and the time of observing Y. Therefore, the approach designed in this article should be applicable to the *observable intermediate information* in other predictive tasks. The efficient application of observable intermediate information for supervised learning has not been adequately studied so far. Thus, this research implies a much larger scope in the area of prediction and supervised learning, and provides an exciting future research direction.

7. Conclusions

To date, LGD modelling has started to attract more attention, but relevant studies related to online lending remain extremely limited. In this study, we investigate LGD prediction using post-default information for online lending products. The debt-settlement flag in the repayment process can be regarded as an observable intermediate variable between the loan default and final charge-off, and LGD is a time-lagging event for the defaulted loans. To utilise this post-default variable, we propose that the mixture and stratified modelling frameworks be taken into consideration together with other loan and borrower level characteristics. We find that overall predictive performance is significantly boosted under the stratified framework compared with other models, while mixture models do not represent any significant advantage. The experimental results identify the predictive power of the post-default variable in LGD modelling. Sensitivity analysis further shows that the empirical evidence is robust in terms of the choice of training set and input variables. This result suggests the incomplete utilisation of information in traditional LGD models, where post-default information is neglected due to data availability and model design.

Given the simplicity of the modelling procedure in our experiment, the performance of the models with post-default variables could be further improved with the following suggestions. First, a more comprehensive investigation can be conducted on the classification model in mixture and stratified models. Our study applies a LASSO logistic regression to generate the probabilities at the classification stage. Yao et al. [38] point out that the performance of two-stage models based on flags (LGD $= 0$ and LGD $=$ 1) can be further improved by applying more accurate classifiers. Next, the imbalance issue arises in the classification task. Either a resampling technique or a cost-sensitive algorithm may be a possible solution to improve the accuracy of the proposed models. Moreover, the LGD models are based on an individual learning algorithm, including the mixture and stratified approaches. These ensemble learning algorithms may introduce additional predictive power to the stratified models with post-default information. Lastly, we were only able to investigate how to incorporate a single binary post-default variable into the experimental analysis due to data availability. As discussed in Section 3.3, further studies could be conducted to account for other types of post-default variables in either the mixture or the stratified modelling framework.

In addition, our study provides implications for incorporating intermediate variables in predictive modelling problems. LGD is observed after a recovery process, and is thus a time-lagging outcome, before which a series of post-default actions could be

Table 14
Paired t-test of the best models of each class — All candidate variables.

Panel A. MAE					
Models	Single stage - Lasso	Mixture_LGD - Lasso	Mixture_PDI - SVR	Stratified_LGD - SVR	Stratified_PDI - SVR
Single stage - Lasso	N.A.				
Mixture_LGD - Lasso	0.3012	N.A.			
Mixture_PDI - SVR	0.2789	0.7242	N.A.		
Stratified_LGD - SVR	0.1614	0.0701	0.0967	N.A.	
Stratified_PDI - SVR	0.0001	0.0001	0.0000	0.0017	N.A.
Panel B. RMSE					
Models	Single stage - Lasso	Mixture_LGD - Lasso	Mixture_PDI - SVR	Stratified_LGD - SVR	Stratified_PDI - SVR
Single stage - Lasso	N.A.				
Mixture_LGD - Lasso	0.2679	N.A.			
Mixture_PDI - SVR	0.0705**	0.8158	N.A.		
Stratified_LGD - SVR	0.0001***	0.0001***	0.0000***	N.A.	
Stratified_PDI - SVR	0.0001***	0.0004***	0.0001***	0.0000***	N.A.

taken. These events which occur throughout the process should be considered to be post-default information in the context of LGD prediction. However, beyond the context of LGD modelling, the events happening before the observation of a time-lagging outcome are wide-ranging and numerous, and can be generally termed *intermediate events*. If the intermediate events are observable in retrospective studies, they are *observable intermediate events*. Similar to the paradox of LGD modelling, traditional models for the prediction of time-lagging outcomes fail to capture any observable intermediate events, because these events are unavailable at the time of the model's construction. Our proposed modelling frameworks provide a feasible and effective solution for prediction in more general scenarios where the predicted target is time-lagged and the observable intermediate events occur. Thus, we expect that the potential contributions of this paper can be replicated and validated in other subject areas.

CRediT authorship contribution statement

Ke Li: Conceived of the presented idea, Developed the theory, Wrote the manuscript, Discussed the results, Contributed to the final manuscript. **Fanyin Zhou:** Conceived of the presented idea, Verified the analytical methods, Discussed the results, Contributed to the final manuscript. **Zhiyong Li:** Conceived of the presented idea, Supervised the findings of this work, Discussed the results, Contributed to the final manuscript. **Xiao Yao:** Supervised the project, Wrote the manuscript, Discussed the results, Contributed to the final manuscript. **Yashu Zhang:** Performed the analytic calculations, Discussed the results, Contributed to the final manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors would like to thank for the support from the funding of National Natural Science Foundation in China [No. 71901230] and the Fundamental Research Funds for the Central Universities, China [JBK2003002] [JBK1806002]. This work was also supported by the Joint Lab of Data Science and Business Intelligence at Southwestern University of Finance and Economics, China.

Appendix. Descriptions of modelling variables

Variables	Descriptions
mo_sin_old_il_acct	Months since oldest bank instalment account opened
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
mo_sin_old_rev_tl_op	Months since oldest revolving account opened
annual_inc	The self-reported annual income provided by the borrower during registration.
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
bc_open_to_buy	Total open to buy on revolving bankcards.
avg_cur_bal	Average current balance of all accounts
bc_util	Ratio of total current balance to high credit/credit limit for all bankcard accounts.
revol_bal	Total credit revolving balance
revol_util	Revolving line utilisation rate, or the amount of credit the borrower is using relative to all available revolving credit.
total_bc_limit	Total bankcard high credit/credit limit
mths_since_recent_bc	Months since most recent bankcard account opened.
total_rev_hi_lim	Total revolving high credit/credit limit

Variables	Descriptions
total_il_high_credit_limit	Total instalment high credit/credit limit
tot_hi_cred_lim	Total high credit/credit limit
total_bal_ex_mort	Total credit balance excluding mortgage
mths_since_recent_inq	Months since most recent enquiry.
tot_cur_bal	Total current balance of all accounts
FICO	FICO score of applicants
num_il_tl	Number of instalment accounts
pct_tl_nvr_dlq	Percent of trades never delinquent
mo_sin_rcnt_rev_tl_op	Months since most recent revolving account opened
num_rev_accts	Number of revolving accounts
num_bc_tl	Number of bankcard accounts
SP500	Annual return of S&P index

References

[1] Basel Committee on Banking Supervision, Guidance on paragraph 468 of the framework document, 2005.
[2] Basel Committee on Banking Supervision, An explanatory note on the basel II IRB risk weight functions, 2005.
[3] T. Schuermann, What Do We Know About Loss Given Default?, Working Paper, 2004.
[4] M. Qi, X. Zhao, Comparison of modelling methods for loss given default, J. Bank. Financ. 35 (2011) 2842–2855.
[5] X. Yao, J. Crook, G. Andreeva, Support vector regression for loss given default modelling, European J. Oper. Res. 240 (2) (2015) 528–538.
[6] J.A. Bastos, Forecasting bank loans loss-given-default, J. Bank. Financ. 34 (10) (2010) 2510–2517.
[7] H.D. Khieu, D.J. Mullineaux, H.C. Yi, The determinants of bank loan recovery rates, J. Bank. Financ. 36 (4) (2012) 923–933.
[8] M. Qi, X. Yang, Loss given default of high loan-to value residential mortgages, J. Bank. Financ. 33 (2009) 788–799.
[9] M. Leow, C. Mues, Predicting loss given default (LGD) for residential mortgage loans: A two-stage model and empirical evidence for UK bank data, Int. J. Forecast. 28 (1) (2011) 183–195.
[10] T. Bellotti, J. Crook, Loss given default models incorporating macroeconomic variables for credit cards, Int. J. Forecast. 28 (1) (2012) 171–182.
[11] C. Han, Y. Jang, Effects of debt collection practices on loss given default, J. Bank. Financ. 37 (2013) 21–31.
[12] A. Bellotti, D. Brigo, P. Gambetti, F. Vrins, Forecasting recovery rates on non-performing loans with machine learning, Int. J. Forecast. 37 (1) (2021) 428–444.
[13] L.C. Thomas, A. Matuszyk, A. Moore, Comparing debt characteristics and LGD models for different collections policies, Int. J. Forecast. 28 (2012) 196–203.
[14] V.V. Acharya, S.T. Bharath, A. Srinivasan, Does industry-wide distress affect defaulted firms? Evidences from creditor recoveries, J. Financ. Econ. 85 (2007) 787–821.
[15] N. Mora, Creditor recovery: the macroeconomic dependence of industry equilibrium, J. Financ. Stab. 18 (2015) 172–186.
[16] M. Leow, C. Mues, L. Thomas, The economy and loss given default: Evidence form two UK retail lending data sets, J. Oper. Res. Soc. 65 (3) (2014) 363–375.
[17] O. Rigbi, The effects of usury laws: Evidence from the online loan market, Rev. Econ. Stat. 95 (4) (2012) 1238–1248.
[18] J. Duarte, S. Siegel, L. Young, Trust and credit: The role of appearance in peer-to-peer lending, Rev. Financ. Stud. 25 (8) (2012) 2455–2483.
[19] M.F. Lin, N.R. Prabhala, S. Viswanathan, Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending, Manage. Sci. 59 (1) (2013) 17–35.
[20] R. Emekter, Y. Tu, B. Jirasakuldech, M. Lu, Evaluating credit risk and loan performance in online peer-to-peer (p2p) lending, Appl. Econ. 47 (1) (2015) 54–70.
[21] R. Iyer, A.I. Khwaja, K. Shue, Screening peers softly: Inferring the quality of small borrowers, Manage. Sci. 62 (6) (2016) 1554–1577.
[22] M. Malekipirbazari, V. Aksakalli, Risk assessment in social lending via random forests, Expert Syst. Appl. 42 (10) (2015) 4621–4631.
[23] Y. Guo, W. Zhou, C. Luo, C. Liu, H. Xiong, Instance-based credit risk assessment for investment decisions in P2p lending, European J. Oper. Res. 249 (2) (2016) 417–426.
[24] Z. Li, K. Li, X. Yao, Q. Wen, Predicting prepayment and default risks of unsecured consumer loans in online lending, Emerg. Mark. Financ. Trade 55 (1) (2019) 118–132.
[25] L. Papke, J. Wooldridge, Econometric method for fractional response variables with an application to the 401(k) plan participation rates, J. Appl. Econometrics 11 (1996) 619–632.
[26] D. Dwyer, I. Korablev, Moody's KMV losscalc V3.0. Moody's analytics, 2009.
[27] G. Loterman, I. Brown, D. Martens, C. Mues, B. Baesens, Benchmarking regression algorithms for loss given default modelling, Int. J. Forecast. 28 (1) (2012) 161–170.
[28] E. Tobback, David Martens, T. Gestel, B. Baesens, Forecasting loss given default models: Impact of account characteristics and the macroeconomic state, J. Oper. Res. Soc. 65 (3) (2014) 376–392.
[29] J. Hurlin, A. Patin, Loss functions for loss given default model comparison, European J. Oper. Res. 268 (1) (2018) 348–360.
[30] A. Nazemi, K. Heidenreich, F.J. Fabozzi, Improving corporate bond recovery rate prediction using multi-factor support vector regressions, European J. Oper. Res. 271 (2) (2018) 664–675.
[31] P. Miller, E. Töws, Loss given default adjusted workout processes for leases, J. Bank. Financ. 91 (2018) 189–201.
[32] F. Kaposty, J. Kriebel, M. Löderbusch, Predicting loss given default in leasing: a closer look at models and variable selection, Int. J. Forecast. 36 (2) (2020) 248–266.
[33] B. Shi, X. Zhao, B. Wu, Y. Dong, Credit rating and microfinance lending decisions based on loss given default (LGD), Finance Res. Lett. 30 (2019) 124–129.
[34] B. Shi, G. Chi, W. Li, Exploring the mismatch between credit ratings and loss-given-default: a credit risk approach, Econ. Model. 85 (2020) 420–428.
[35] R.C. Hwang, H. Chung, C.K. Chu, A two-stage probit model for predicting recovery rates, J. Financ. Serv. Res. 50 (3) (2016) 311–339.
[36] E.N.C. Tong, C. Mues, L. Thomas, A zero-adjusted gamma model for mortgage loan loss given default, Int. J. Forecast. 29 (4) (2013) 548–562.
[37] R. Calabrese, Predicting bank loan recovery rates in a mixed continuous-discrete model, Appl. Stoch. Models Bus. Ind. 30 (2) (2014) 99–114.
[38] X. Yao, J. Crook, G. Andreeva, Enhancing two-stage modelling methodology for loss given default with support vector machines, European J. Oper. Res. 263 (2) (2017) 679–689.
[39] Y. Tanoue, A. Kawada, S. Yamashita, Forecasting loss given default of bank loans with multi-stage model, Int. J. Forecast. 33 (2017) 513–522.
[40] A. Nazemi, F.F. Pour, K. Heidenreich, F.J. Fabozzi, Fuzzy decision fusion approach for loss-given-default modelling, European J. Oper. Res. 262 (2017) 780–791.
[41] M. Papoušková, P. Hajek, Two-stage consumer credit risk modelling using heterogeneous ensemble learning, Decis. Support Syst. 118 (2019) 33–45.
[42] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785-794.
[43] E. Altman, E. Kalotay, Ultimate recovery mixtures, J. Bank. Financ. 40 (2014) 116–129.
[44] L. Breiman, Stacked regressions, Mach. Learn. 24 (1) (1996) 49–64.
[45] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.
[46] V. Vapnik, Statistical Learning Theory, John Wiley, New York, 1998.
[47] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Chapman and Hall, Wadsworth, New York, 1984.
[48] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.