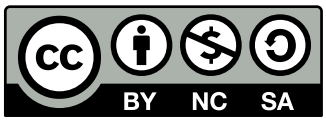
Aalto University
School of Science

Master's programme in Mathematics and Operations Research

Simulation-based Evaluation of Loss Given Default Model Structures

Vesa Ranta-aho

Master's Thesis
2025



Aalto University
School of Science

Author Vesa Ranta-aho

Title Simulation-based Evaluation of Loss Given Default Model Structures

Degree programme Mathematics and Operations Research

Major Systems and Operations Research

Supervisor Prof. Ahti Salo

Advisor D.Sc. (Econ.) Jaakko Sääskilahti

Collaborative partner OP Financial Group

Date 19 September 2025 **Number of pages** 52+20 **Language** English

Abstract

The accurate estimation of Loss Given Default (LGD), or the share of exposure that will be lost in case of a default, is crucial for the profitability and financial stability of banks and other financial institutions. Modelling LGD is challenging due to the complex nature of LGD as a phenomenon, as well as its unusually shaped bimodal distribution. LGD modelling is thus an active field of research and development.

In this thesis, we study how the performance of three different multi-stage LGD model structures changes compared to that of a simple OLS model depending on the shape of the LGD distribution, the proportion of cure, partial recovery and write-off cases, and the predictiveness of explanatory variables for the different components of the multi-stage models. To generate data for the study, we devise a simulation approach that generates LGD data suitable for different types of multi-stage models by combining existing LGD simulation approaches found in the literature.

We show that a multi-stage model performs the best compared to OLS or a simpler multi-stage model when there are variables available that can accurately predict the probabilities related to the component splits of the specific multi-stage structure, when there are enough cases in each component to justify the additional complexity of each split, and when the loss distributions are heterogenous between the components and homogenous within the components.

However, we also show that the average performance differences between the models are small compared to between-dataset variation even within similar data sets, and that the studied multi-stage models do not produce a predicted LGD distribution in the characteristic bimodal shape even when their discriminatory power between the components is strong.

This thesis and its results provide a starting point for LGD model structure choice for modellers, as well as for future research on the behavior of multi-stage models.

Keywords Loss given default, credit risk modelling, multi-stage models, regression, simulation, risk management

Aalto-yliopisto
Perustieteiden
korkeakoulu

Tekijä Vesa Ranta-aho

Työn nimi Simulaatioperusteinen tappio-osuusmallirakenteiden arviointi

Koulutusohjelma Mathematics and Operations Research

Pääaine Systems and Operations Research

Työn valvoja Prof. Ahti Salo

Työn ohjaaja KTT Jaakko Sääskilahti

Yhteistyötaho OP Ryhmä

Päivämäärä 19.9.2025 **Sivumäärä** 52+20 **Kieli** englanti

Tiivistelmä

Tarkka tappio-osuuden (LGD), eli maksukyvyttömyyden sattuessa menetetyn vastuun osuuden maksukyvyttömyyshetken kokonaisvastuusta, ennustaminen on elintärkeää pankkien ja muiden rahoituslaitosten tuottavuuden ja vakauden kannalta. Tappio-osuuden ennustaminen on haastavaa sen monimutkaisen luonteen sekä epätavallisen kaksihuippuisen jakauman vuoksi. Tappio-osuusmallit ovat siksi jatkuva tutkimuksen ja kehityksen kohde.

Tässä diplomityössä tutkitaan, miten kolmen eri monivaiheisen mallirakenteen ennustuskyky muuttuu verrattuna yksinkertaiseen lineaariseen regressiomalliin riippuen tappio-osuusjakauman muodosta, maksukyvyttömyydestä parantuvien tapausten, osittaisten palautusten ja luottotappiokirjaus-tapausten osuuksista, sekä ennustamiseen käytettävien muuttujien ennustavuudesta eri mallikomponenteille. Datan tuottamiseksi työssä kehitetään erilaisille monivaiheisille tappio-osuusmalleille soveltuvaa dataa tuottava simulointimenetelmä yhdistelemällä tappio-osuusmallikirjallisuudesta löytyviä simulointimenetelmiä.

Työn tulokset osoittavat, että monivaiheinen tappio-osuusmalli toimii parhaiten verrattuna lineaariseen regressioon silloin kun käytössä olevien muuttujien avulla voidaan ennustaa tarkasti mallien komponenttijakoon liityviä todennäköisyyksiä; kun jokaiseen mallikomponenttiin kuuluu tarpeeksi tapauksia, jotta komponenttijakojen lisäämä kompleksisuus on perusteltua; ja kun tappio-osuusjakaumat ovat heterogeenisiä mallikomponenttien välillä sekä homogeenisiä niiden sisällä. Tulokset osoittavat kuitenkin myös, että keskimääräiset erot mallien suorituskyvyssä ovat pieniä verrattuna vaihteluun datasettien välillä jopa silloin, kun datasettien erot ovat pieniä; ja että tutkitut monivaiheiset mallit eivät tuota ennusteille tappio-osuusjakaumalle tyypillistä kaksihuippuista muotoa edes silloin, kun mallien erottelukyky komponenttien välillä on korkea.

Tämä diplomityö ja sen tulokset antavat lähtökohdan mallintajille tappio-osuusmallirakenteen valintaan, sekä monivaiheisten tappio-osuusmallien käyttäytymisen jatkotutkimuksiin.

Avainsanat Tappio-osuus, luottoriskimallinnus, monivaiheiset mallit, regressio, simulaatio, riskienhallinta

Contents

Symbols and Abbreviations

A-IRB	advanced internal ratings-based approach
AUC	area under the receiver operating characteristic curve
BCBS	Basel Committee on Banking Supervision
C	cure
CPW	Cure-Partial-recovery-Write-off multi-stage model
Δ	difference
EAD	exposure at default
EC	economic capital
ECL	expected credit loss
EL	expected losses
F-IRB	foundation internal ratings-based approach
gAUC	generalised area under the receiver operating characteristic curve
I_C	cure indicator
I_P	partial recovery indicator
I_W	write-off indicator
IRB	internal ratings-based approach
LGC	loss given cure
LGD	loss given default
LGD_C	loss given default within cure cases
LGD_P	loss given default within partial recovery cases
LGD_W	loss given default within write-off cases
LGD0	zero LGD
LGD1	one LGD
LGNW	loss given no write-off
LGP	loss given partial recovery
LGW	loss given write-off
OLS	ordinary least squares regression
P	partial recovery
PD	probability of default
R^2	coefficient of determination
ROC	receiver operating characteristic curve
RWA	risk-weighted assets
s	default end status
UL	unexpected losses
W	write-off
WNW	Write-off-Non-Write-off multi-stage model
ZFO	Zero-Fractional-One multi-stage model

1 Introduction

Loss given default (LGD) is a critical parameter used in credit risk modelling and risk management. It describes the proportion of exposure at the moment of default that is lost when a borrower defaults. Accurate modelling of LGD and credit risk in general is crucial for the profitability and financial stability of banks and other financial institutions (McNeil et al., 2015).

Predicting LGD accurately is difficult. LGD is the outcome of a highly variable default and workout process, and the typical LGD distribution of a portfolio is bimodal or even U-shaped, with lots of very low and very high losses (Schuermann, 2004). Due to the complex nature of LGD and its unusual distribution, there is no single correct model for LGD, and instead, LGD modelling techniques vary greatly.

Most commonly used LGD models can be categorised roughly into two types: single-stage models and multi-stage models. Single-stage models predict LGD in one step using techniques varying from simple linear models to complex machine learning techniques. Multi-stage models divide the LGD prediction into multiple steps or components, that typically utilise relatively simple modelling techniques, such as linear regression, within the components. The benefit of multi-stage models is that they allow for more detailed modelling of LGD than simple single-stage models, while simultaneously keeping the models easier to interpret than advanced single-stage models. However, developing a multi-stage model often requires considerably more work than developing a single-stage model, since a multi-stage model is essentially a combination of multiple single-stage models. Therefore, a more complex multi-stage model structure is justified only if it significantly improves the performance over a simpler model structure.

Various LGD model comparison studies have been performed. However, most studies focus only on single-stage models or include just one type of multi-stage model in the comparison. Furthermore, we found no studies on how the properties of the LGD and explanatory variable data affect the relative performance of different models. The best study we found in this regard is by Loterman et al. (2012), who include in their comparison two multi-stage model structures and perform the comparison using six different data sets. However, they do not analyse the factors that make one model or model structure perform better for one data set and worse for another.

The aim of this thesis is to fill this gap in the literature for multi-stage models by answering the following two related research questions:

1. What factors affect the performance of different multi-stage model structures relative to a simple single-stage model?

2. For what kind of data can one expect better performance from different multi-stage models than from a simple single-stage model?

To answer the questions, we first simulate a collection of varying LGD data sets following the simulation practices in the LGD literature. Then, using the simulated data, we analyse the relationships between data describing summary statistics and the performance differences of three distinct multi-stage model structures compared to a

simple single-stage model, as well as the performance levels of the different model structures over different value ranges of the summary statistics.

The structure of this thesis is as follows. In section 2, we give background on credit risk, risk management, LGD, LGD modelling and LGD data set simulation based on literature. In section 3, we describe the methodology used in simulating our data sets and in the model analysis, including the models themselves. In section 4, we present and analyse the results. Finally, section 5 concludes the thesis.

2 Background

2.1 Credit Risk and Risk Management

By its broadest definition, credit risk is the financial risk associated with any kind of credit-linked event (Bielecki and Rutkowski, 2013). It includes default risk, which is the risk arising from a borrower's failure to meet their contractual obligations such as loan repayments or bond obligations, as well as the risk of reductions in market value caused by changes in credit quality or ratings and variations in credit spreads (Bielecki and Rutkowski, 2013; Duffie and Singleton, 2003). In this thesis, we focus on the default risk in bank loans.

Credit risk quantification is based on three main parameters: probability of default (PD), exposure at default (EAD) and loss given default (LGD). PD, measured as a decimal between 0 and 1, is the probability that the counterparty will default within a specified time period. EAD, measured as currency, is the outstanding exposure at the time of the default event. Loss given default, measured as a decimal, is the proportion of the exposure at default which the bank is unable to recover after the default and will be lost. Expected credit loss (ECL) can be calculated as the product of these three parameters:

$$ECL = PD \cdot EAD \cdot LGD. \tag{1}$$

Quantifying credit risk serves multiple purposes. It is a critical concern for financial institutions, as it directly affects their profitability and financial stability. By measuring and predicting credit risk accurately, banks are able to tune their loan granting and pricing processes to gain a competitive edge and increase profits while maintaining risk levels accepted by stakeholders, investors and regulation (McNeil et al., 2015).

Effective credit risk management is also important for the society at large, as it relies on functioning and stable banking systems (McNeil et al., 2015). As such, credit risk management is regulated not only by local regulations and laws, but also by international regulation frameworks, such as the Basel Accords.

2.1.1 The Basel Accords

The Basel Accords are an evolving set of standards meant to ensure the capital adequacy of banks, set by the Basel Committee on Banking Supervision (BCBS). BCBS has no official juristical power, but instead it expects that members implement and apply the standards in their own jurisdictions (McNeil et al., 2015).

The first Basel Capital Accord (Basel I, 1988) was motivated by the Latin American debt crisis of the early 1980s and rising concerns of inadequate capital reserves of major international banks (Bank for International Settlements, 2018). It set out to strengthen and unify the international banking system by introducing a minimum standard for capital adequacy with a minimum required ratio of capital to risk-weighted assets (Bank for International Settlements, 2018). However, risk measurement was crude, with claims being divided into only three categories based on the type of counterparty; governments, regulated banks and others, with no differentiation in

risk between different corporate borrowers based on, for example, their credit rating (McNeil et al., 2015).

To address the shortcomings of Basel I, the second Basel Accord (Basel II) was released in 2004. It is a framework of three pillars. The first pillar sets new minimum capital requirements and improved the quantification of required capital, allowing banks to use internal models or external credit-rating systems to achieve better risk differentiation (McNeil et al., 2015). The second pillar requires supervisory review of an institution's capital adequacy and internal assessment process, and the third pillar sets disclosure requirements with the aim of strengthening market discipline and encourage sound banking practices (Bank for International Settlements, 2018).

After the financial crisis of 2007-2009, the third Basel Accord (Basel III) was released in 2010. It revises and extends the three pillars of Basel II by increasing the quality and quantity of required capital, adding countercyclical buffers, leverage and liquidity requirements and stricter requirements for systematically important banks. Since its first release, several revisions to Basel III have been made to further improve capital requirement calculations, with the latest major revision being from 2017, as of May 2025 (Bank for International Settlements, 2018).

2.1.2 Regulatory Capital

The Basel Accords require banks to hold sufficient capital reserves to offset potential losses. This capital is referred to as regulatory capital. The amount of required capital is defined using three quality categories of capital, and risk-weighted assets (RWA). The highest quality capital is Common Equity Tier 1 (CET1) capital, and it is defined as the sum of common shares and stock surplus, retained earnings, other comprehensive income, qualifying minority interest and regulatory adjustments. The second highest quality capital, Additional Tier 1 (AT1) capital, is the sum of capital instruments meeting criteria for Tier 1 and related surplus, additional qualifying minority interest and regulatory adjustments. The lowest quality regulatory capital, Tier 2 capital, is defined similarly as AT1 capital but with the addition of qualifying loan loss provisions, and with lower criteria for qualifying capital instruments. At all times, CET1 capital must be at least 4.5 % of RWA, Tier 1 capital (the sum of CET1 and AT1 capital) must be at least 6 % of RWA, and total capital (the sum of Tier 1 and Tier 2 capital) must be at least 8.0 % of RWA (Basel Committee on Banking Supervision, 2019).

Banks can choose between two approaches to calculate their RWA for bank loans. In the standardized approach, standardized risk weights are assigned to different exposure classes, and RWA is calculated as the product of the standardized risk weights and the exposure amount. With the approval of a bank's supervisor, it can also opt for the internal ratings-based (IRB) approach, where the RWA is calculated based on internal models and assessment (Basel Committee on Banking Supervision, 2019).

In IRB, the RWA-based capital requirements are meant to cover unexpected losses (UL), while expected losses (EL) are to be covered by provisions. EL for an exposure is calculated as the product of PD and LGD, while RWA is calculated using set functions for different asset classes based on PD, EAD and LGD, and sometimes effective maturity of the exposure. Depending on the capabilities of the bank, it may apply for

the foundation IRB (F-IRB) approach, where it can estimate only PD internally, or for the advanced IRB (A-IRB) approach, where all of the parameters can be estimated internally (Basel Committee on Banking Supervision, 2019).

While the standardized approach is more common for small banks, most large banks opt for the IRB-approach, as it provides more flexibility in comparison and they have the resources to satisfy the heavy regulatory and supervisory requirements of the IRB approach. However, even the A-IRB approach does not allow for fully internal RWA quantification, but only the internal estimation of the input parameters PD, EAD and LGD (McNeil et al., 2015).

2.1.3 Economic Capital

In addition to regulatory capital, banks calculate economic capital (EC). It is the bank's own realistic view of the capital required to cover unexpected losses with a high confidence. It is used to assess the bank's own risk position and to allocate capital in the most efficient way according to the bank's strategy by providing a common way to measure risk between different asset classes (Burns, 2004). In contrast to regulatory capital, EC can be quantified using fully internal models.

2.2 Loss Given Default

Loss given default (LGD) is one of the three main parameters used to measure credit risk. It is the proportion of the exposure at default which the bank is unable to recover after the default and will be lost.

The typical LGD distribution of a portfolio is bimodal, with large concentrations near 0 (no loss) and 1 (full loss) (Schuermann, 2004). While LGD is usually between 0 and 1, it can also be greater than 1 in case the whole exposure is lost and there are additional costs related to the recovery process or legal fees. It can also be negative, meaning that a profit was made on the default, due to large collateral recoveries and collected late fees (Salko and D'Ecclesia, 2022). Extreme LGD values (both positive and negative) are more likely to be present when the exposure is small, because it is defined as a ratio of the loss to exposure, and because the relative effect of additional costs and fees is greater than for large exposures.

When a default happens, roughly three scenarios can follow. First, the borrower can resolve the issues that lead to the default, return to the normal payment schedule and cure. This is the best case scenario, where usually only little to no losses incur. Second, the loan can enter collection, where possible collateral is liquidated and other legal measures are taken to recover the exposure. The collection process can lead to a broad range of loss outcomes. Liquidating the collateral and other recoveries might be enough to cover the whole exposure, or, in the third, worst case scenario; a portion of the exposure can not be recovered and has to be written off, leading to high losses.

The most influential factors that affect the loss outcome are the seniority of the loan and the amount of attached collateral Schuermann (2004). Senior debt is repaid first, so a bank is more likely to get good recoveries on a senior loan rather than on a subordinated junior loan. In case the default enters collection, an unsecured loan

is likely to incur high losses, while a fully collateralised loan might get repaid fully. However, even fully collateralised loans may cause losses due to lower-than-expected realisation prices or a prolonged realisation process. The ease and price of realisation are heavily affected by the type of collateral and its location, as well as macroeconomic factors. For example, an apartment in a capital city is easier to liquidate than a house in a demographically regressing rural town, or some highly specialized industrial machinery. Thus, accurate collateral valuation on the bank's part is also a key factor affecting the losses it sees.

Other key factors affecting LGD are the type of customer (e.g., large corporates behave differently than households), nature and severity of the default (e.g., bankruptcy or just late payments), financial situation of the borrower (e.g. employment status, salary, turnover, profit, liquidity) and industry of the borrower (Salko and D'Ecclesia, 2022). In addition to collateral liquidation, macroeconomic factors can also affect different industries differently, and have an overall effect on the borrower's financials and thus LGD.

2.3 Modelling Loss Given Default

The goal in LGD modelling is twofold: risk differentiation and risk quantification (European Central Bank, 2024). Although not completely separate concepts, the difference between the two can be illustrated as follows. Quantifying the level of risk of a portfolio is crucial to ensuring capital adequacy. However, even a simple average model can accurately predict the level of risk at the portfolio level, while providing no insight to where the risk is actually coming from. Only by accurately differentiating high-risk and low-risk customers are banks able to truly manage the risk, through, for example, loan granting decisions, pricing, and targeted preventive action.

To satisfy the two goals, the LGD modelling process is typically divided into two parts:

1. A scoring model, which predicts a raw LGD score with a focus on high differentiation ability or discriminatory power, and does not yet necessarily provide accurate estimates of expected loss levels.

2. A calibration step, where the raw LGD scores are adjusted to match, depending on the use-case of the model, long-run average or downturn loss levels, or to reflect the economical conditions at a certain point in time.

In this thesis, we focus only on the scoring model, which we will refer to as the LGD model.

A major challenge in LGD modelling is the bimodal distribution of LGD values. In theory, models such as linear regression should not be used for this kind of data (Li et al., 2016), because the predictions will be concentrated near the portfolio mean, where in reality the observation concentration is the most scarce. However, as Li et al. (2016) find, even transformations and models designed specifically to fit the bimodal LGD distribution often fail at the task and produce predictions that are more concentrated at the portfolio mean. Thus, in practice, as will be presented in

section 2.3.1, OLS and its variants continue to be used for LGD modelling due to their performance being on a good level despite the theoretical issues, and due to their simplicity and transparency.

Another complicating factor is the complex nature of LGD itself - it is not a single event, but a cumulative final result of a default and a workout process, which are both affected by factors that can not be accurately accounted for on an obligor level. Payment ability, for example, can be measured using financial ratios, employment status, salary or other similar measures, but in reality the individual situation of each obligor can not be fully captured by any number of variables. In addition, the situation of each obligor and the economy as a whole can change unexpectedly before or during the default due to any number of external factors, causing additional uncertainty in the models.

Due to the aforementioned issues, LGD modelling is inherently difficult. The difficulty is reflected in statistical performance, which, for LGD models, is generally quite low (Bellotti and Crook, 2012; Loterman et al., 2012).

Additional biases in LGD models can be caused by the way modelling samples are formed. Realised LGD data is only available from customers or contracts that have defaulted. Yet, most of the bank's portfolio, for which LGD is predicted by the model, will never go into default (e.g., European Investment Bank (2024) report worldwide average annual default rates of 3.56% and 2.59% between 1994 and 2023 for private and public lending, respectively). This causes a representativeness issue, where the historical data does not necessarily represent accurately the overall portfolio. It may miss patterns that are only present in the non-defaulted parts of historical portfolios and that would only materialise, for example, under certain economic conditions. To minimise this bias, banks are required to ensure the representativeness of the modelling data to the overall portfolio in terms of the available risk drivers and other characteristics (Basel Committee on Banking Supervision, 2019).

Gürtler and Hibbeln (2011) also note that a finite observation period of the modelling sample will cause an underestimation of LGDs unless properly accounted for. Modelling data typically consist of only defaults for which the workout process has been fully completed (otherwise some expected recoveries would still be missing, which would in turn overestimate LGDs). This means that long defaults, which have started before the observation period or would end after the observation period, will be excluded from the data, and that the beginning and the end of the observation period will contain only short defaults that fit fully inside the period. Therefore, long defaults, which have on average higher LGDs (Gürtler and Hibbeln, 2011), are underrepresented in the data. Thus, the model will underestimate average LGDs and possibly not capture some patterns present only in the long defaults.

To avoid this issue, Gürtler and Hibbeln (2011) propose restricting the observation period further, such that, based on some observed maximum workout process length T_{max}, after which most of the defaults could already be considered resolved, only the defaults that have started at the latest T_{max} before the end of the observation period, and ended at the earliest T_{max} after the start of the observation period, will be included in the modelling data. This way, all of the defaults in the data set would have the possibility to last for at least the duration of the maximum workout process T_{max}, and

the underrepresentation issue of the long defaults would be solved. The downside of this approach is that some data is excluded, which can be problematic if the data set is small to begin with.

Finally, bias can be caused by the business processes from which the LGD data originates. If, for example, loan granting criteria are stricter for a certain type of business due to its inherent riskiness, then the loan data for these businesses will be comprised of better customers, which may result in lower-than-expected realised LGD levels. This can cause a discrepancy between business expectations and a purely data-based model, especially if the type of business is used as a risk driver in the model. In these situations, the model needs to include relevant risk drivers that capture the fact that these are better customers and explain the lower LGD. Otherwise, the model may work unintuitively by explaining the lower LGD based on the type of business. In general, LGD modellers must be aware of such data issues instead of blindly trusting the data.

To overcome the challenges in LGD modelling, models of varying complexity and structures are used. Section 2.3.1 introduces examples of typical single-stage models and their performance, while section 2.3.2 does the same for multi-stage models, which divide the LGD score prediction into multiple steps or components. Finally, section 2.3.3 discusses different model selection criteria and the model selection process as a whole.

2.3.1 Single-stage Models

The simplest structure for an LGD model is a single-stage structure, where the LGD score is predicted directly in one step. The complexity of the actual model used in single-stage structure can vary greatly, from statistical regression models and decision trees to complex machine learning models (Loterman et al., 2012). In practice, however, financial institutions mainly use statistical models as the final model due to regulation and the need for transparency and business intuition, while machine learning models are for now primarily used for benchmarking and in academic contexts (Bücker et al., 2020).

Loterman et al. (2012) perform a comprehensive benchmark study on the performance of the most common LGD models. For single-stage models, they include ordinary least squares regression (OLS), beta-transformed OLS (B-OLS), beta regression (BR), Box-Cox transformed OLS (BC-OLS), ridge regression (RiR), robust regression (RoR), regression trees (RT), multivariate adaptive regression splines (MARS), least squares support vector machines (LSSVM) and multilayer perceptrons (MLP) in the study. Using six real-world datasets of different loan portfolios and eight diverse performance metrics, they find that the machine learning models LSSVM and MLP outperform the rest of the single-stage models and that also MARS and RT outperform the linear and transformed linear models. Between the purely linear models, OLS, RiR and RoR, they find no consistent performance differences, while they find OLS to outperform the transformed linear models, B-OLS, BR and BC-OLS. They attribute the worse performance of the transformed linear models to the transformations which cannot deal with the high concentration peaks near 0 and

1 in the LGD distribution and the additional inefficiency or bias the transformations introduce. Loterman et al. (2012) note that the better performance of the non-linear models indicates the presence of non-linearity in the relationship between LGD and the independent variables, proving the potential of using non-linear techniques in LGD modelling. However, in the case of LSSVM and MLP, the increased performance comes with the cost of reduced transparency.

Bellotti and Crook (2012) find similar results regarding linear models in their study of LGD models for credit cards - they find that OLS outperforms beta-, fractional logit- and probit-transformed OLS models. Additionally, in their comparison they include a least absolute deviation regression model and a Tobit model, in which a linear regression model is censored between 0 and 1, and is fit using maximum likelihood estimation considering separately the probabilities that LGD is 0, 1, or in between. These models, too, are outperformed by OLS.

Numerous other single-stage model formulations are proposed in various studies. However, the main conclusion is that in terms of statistical performance measures, non-parametric models, such as regression trees or machine learning methods, tend to perform better than parametric models, while for parametric models, such as regression models, more complex methods do not necessarily perform better than simple ones (Li et al., 2016, 2018).

2.3.2 Multi-stage Models

Multi-stage models divide the LGD prediction process into more than one steps or components in order to better capture either the nature of the LGD distribution or the different possible outcome states of a default. The idea is to be able to specify the different model components more accurately for certain types of outcomes without being affected by other outcomes that behave differently, and in order to achieve better overall predictions. In multi-stage models, the final LGD score prediction is obtained as a combination of the predictions of the different components. In each component, the same modelling techniques can be utilised as in the single-stage models.

In their benchmark study, Loterman et al. (2012) include two multi-stage model structures. The first is a combination logistic regression and another regression technique, where in the first stage logistic regression is used to predict the probability that the LGD is exactly 0, and in the second stage a model built using only observations where LGD > 0 is used to predict the LGD in the case it is greater than 0. The final prediction is the probability weighted average of 0 and the second stage prediction. The authors report a slight trend where in the case of linear second stages, the performance is slightly increased compared to the corresponding single-stage linear model, and in the case of non-linear second stages, the performance is slightly decreased.

The authors also note that this model structure could possibly be improved by using an ordinal logistic regression instead of binary logistic regression to also distinguish the LGD = 1 peak from the rest. In fact, Bellotti and Crook (2012) compare a multi-stage model with the same idea, where the probabilities of LGD = 1 and LGD = 0 are determined using two chained logistic regression models. They motivate the model by the large number of 0 and 1 LGD cases which make the division natural, and by the

assumption that there are special conditions which cause the full and no loss outcomes. However, they find that this model performs worse than OLS.

The second multi-stage model structure included in the study by Loterman et al. (2012) is the combination of a first stage OLS model and a non-linear second stage model that is built on the residuals of the first stage model. The final prediction is the sum of the first and second stage predictions. For these models, the performance is reported to be better than for OLS, and very near the performance of the corresponding single-stage non-linear models. Despite the added complexity in model structure, Loterman et al. (2012) state the advantage in this model structure is the comprehensibility of the linear regression component combined with the good performance of the non-linear technique.

In their working paper, Gürtler and Hibbeln (2011) use a similar decision tree -like structure as Loterman et al. (2012) and Bellotti and Crook (2012), but instead of forming the model components based on the LGD outcome (LGD = 0, LGD = 1 and 0 < LGD < 1 cases), they base the components on the end state of the default - write-off or non-write-off. Their model consists of a first stage logistic regression model predicting the probability of write-off, and a second stage of separate linear models for write-off cases and non-write-off cases. The final prediction is the probability-weighted average of the two second stage predictions. Gürtler and Hibbeln (2011) motivate the usage of default outcome instead of LGD outcome by the different characteristics that affect LGD between write-offs and non-write-offs (e.g., collateral value has a high impact in write-off cases where collateral is liquidated, but if collateral is not liquidated, it has no effect on LGD), the fact that losses from non-write-off cases are not always zero, despite often being low, and by a lack of characteristics that separate LGD = 1 cases from other high-loss cases. Using a simulation study, they find the write-off-non-write-off multi-stage model to clearly outperform single-stage OLS.

Tanoue et al. (2017) present a multi-stage model, which is a combination of the LGD-based and default outcome -based multi-stage models. In the first stage, they use logistic regression to estimate the probability of a cure (they call it recovery, but we rename it for consistency with the other models), which they assume to cause no loss. In the second stage, they use another logistic regression model to estimate the probability that the loss is greater than zero for the non-cured cases. In the third stage, they use a logit-transformed OLS model to estimate the greater-than-zero losses. The final prediction is thus the product of the probability that the default does not cure, the probability that the loss is greater than zero, and the loss estimate of the third stage model. Tanoue et al. (2017) find that the multi-stage model has superior predictive accuracy compared to OLS, Tobit, and an inflated beta regression model, where the continuous beta distribution is supplemented with discrete probability masses at exactly 0 and 1 (Ospina and Ferrari, 2010).

Starosta (2021) further expands the multi-stage structure by considering separately cure, partial recovery, and write-off cases. First, the probability of cure is estimated, followed by the probability of write-off given no cure. Finally, the losses are estimated for each of the three cases. They estimate the full model in two ways - in the first, the probabilities are estimated using logistic regression and the losses using OLS, and in the second, the probabilities are estimated using decision tree classifiers and the losses

using regression trees. Starosta (2021) compares the performance of these models to that of OLS and another multi-stage model with the same structure as used by Bellotti and Crook (2012), which separates 0 and 1 losses from the fractional losses. Starosta (2021) uses least squares support vector classifiers to estimate the probabilities and OLS to estimate the losses in this model. They find that the expanded multi-stage model using classification and regression trees performs the best, followed by the expanded model using logistic regression and OLS. Similar to Bellotti and Crook (2012), they also find that the other multi-stage model performs worse than simply using OLS.

2.3.3 Model Selection

Statistical performance is perhaps the most obvious model selection criterion. Loterman et al. (2012) measure performance using a set of eight common performance metrics, which they divide into two types - root mean squared error (RMSE), mean absolute error (MAE), area above the regression error characteristic curve (AOC) and R^2 measure calibration, while area under the receiver operating characteristic curve (AUC), Pearson's r, Spearman's ρ and Kendall's τ measure discrimination. Most other studies also use (a subset of) these same performance metrics.

However, measuring statistical performance in LGD modelling is complicated by the bimodal nature of the LGD distribution - the mean-focused performance metrics give an incomplete picture of the true performance. Li et al. (2018) note that the shape of the predicted distribution is important, for example, for stress testing and conservative LGD adjustments that are often required by regulation due to data limitations, and should therefore also be investigated as part of the model performance analysis. In addition to visual distribution inspection, they use the Kolmogorov–Smirnov (KS) test to quantify the similarity of the predicted distribution to the true one.

In practice, however, model selection does not depend solely on statistical performance, but instead is a balancing act between statistical performance, interpretability, and intuitiveness. Because LGD models are also used in business decision-making, such as loan granting and pricing, they must be intuitive from the business perspective. This means that the modeller must be able to explain why and how each selected variable affects the LGD predictions. For example, adding collateral to a contract must not increase the predicted LGD.

When the complexity of the model structure grows, it becomes increasingly difficult to ensure that such rules hold in all possible scenarios. Especially if the model structure or loan portfolio is segmented (by, for example, a decision tree -like structure), it becomes imperative to ensure there are no unintuitive discontinuities in the LGD predictions when switching from one segment or branch to another. This also holds true for complex modelling techniques, such as machine learning models, where it might be impossible to interpret the effect of each variable.

A complex model structure consisting of multiple components (such as in multi-stage models) has also practical consequences in terms of model development, maintenance and monitoring costs - they are essentially multiplied by the number of components in the model, since the performance and validity of each component must

be ensured separately.

Consequently, provided that adequate statistical performance is achieved, simple models are preferred. This motivates us to compare different model structures with different LGD distributions in order to gain insight on when more complex model structures are worthwhile.

2.4 Simulating Data Sets for LGD Estimation

A challenge restricting LGD studies is the poor availability of empirical data. Banks and other financial institutions will generally not publish details of their data distributions, let alone complete data sets, due to privacy reasons and to avoid revealing useful information to competitors. In order to compare the performance of the different LGD model structures on diverse data, we turn to simulation to generate the data sets. In this section, we explore LGD simulation approaches found in the literature.

Hlawatsch and Ostrowski (2011) propose a simulation approach that is based on the bimodality of LGD. They use a mixture of a right-skewed and a left-skewed beta distribution to capture the bimodal shape of the LGD distribution. To get a variety of data sets, for each mixture distribution, they randomly draw the expected values and variances, as well as a weight parameter, for the two beta distributions from set intervals that ensure a suitable shape for the mixture distribution. From each mixture distribution, 10000 realisations are drawn, an independent identically normal distributed error term is added to each, and the result is limited between 0 and 1 to form the final LGD data.

As explanatory variables, Hlawatsch and Ostrowski (2011) create a beta-distributed ratio with a positive causal relation with LGD, and a normally distributed ratio, a binary distributed indicator, and a beta-distributed ratio with negative causal relations with LGD. The first is assumed to be independent of the other explanatory variables and is drawn directly from the beta-distribution. The relationship for the other three explanatory variables is assumed to be positive. To achieve this dependency structure, Hlawatsch and Ostrowski (2011) use a Gaussian copula (with a randomly sampled correlation matrix for variety between the data sets) to generate observations from dependent uniform distributions, which are then used to generate the final observations from the normal, binary and beta-distributions. We describe copulas and this variable generation approach in more detail in sections 3.1.2 and 3.1.3, respectively.

To combine the explanatory variables and LGD into a complete data set, Hlawatsch and Ostrowski (2011) first sort the simulated LGD, the first explanatory variable independent from others, and the three dependent explanatory variables (based on their copula values to preserve the dependency structure) into quintiles. Then, they randomly match one of the explanatory variable quintiles to each LGD realisation based on its quintile and a 5×5 matrix, which gives the probabilities of matching each explanatory variable quintile to each LGD quintile. Finally, a random realisation of the explanatory variable from the chosen quintile is matched with the LGD. This matching is done separately for the independent explanatory variable and the three dependent variables as one group. By adjusting the matching probability matrix, Hlawatsch and Ostrowski (2011) are able to determine how predictive the explanatory

variables will be.

Gürtler and Hibbeln (2011) use a slightly simplified simulation approach in their comparison of an OLS model and a two-stage model. As explanatory variables, they use five independent standard normally distributed variables, two of which are assumed unobservable and are reserved only for generating the "true" LGD data. Using a Gaussian copula, they transform the first unobservable variable into a uniformly distributed variable on $(0, 1)$, which is correlated with the first two observable (mutually independent) variables. If the value exceeds 0.8, the observation is classified as a write-off case. Then, using another Gaussian copula, they transform the second unobservable variable similarly, but such that it is correlated with the first and third observable (mutually independent) variables. This value is set as the LGD for the write-off cases. For recoveries, LGD is set to zero. This procedure generates a data set with the LGD, a write-off indicator, and three independent standard normally distributed explanatory variables - the first has a causal relationship with the write-off indicator and LGD, the second has a causal relationship with only the write-off indicator, and the third has a causal relationship only with LGD. The second variable, however, also correlates with the final LGD, because the write-off indicator enables LGD to be greater than zero. This enables single-stage models, such as OLS, to also find predictive power from the second variable, even though they do not model write-off probability explicitly.

Li et al. (2018) use another simulation approach, which is based on the inflated beta distribution (Ospina and Ferrari, 2010). Their set of explanatory variables includes a constant, a macroeconomic factor which is based on real data, and nine normally distributed explanatory variables, which, using a copula, are set to have a positive correlation with the macroeconomic factor. For each LGD observation, they set a separate true zero-and-one inflated beta distribution, the parameters of which are determined by the explanatory variable realisations of the same observation through deterministic equations. From this true LGD distribution, the actual true LGD values are drawn. Li et al. (2018) add additional noise to the simulation either by omitting some of the true explanatory variables from the final data set used for model fitting, or by adding normally distributed error terms to the parameters of the true inflated beta distributions.

3 Methods

3.1 Data Set Simulation

3.1.1 Simulating LGD

To simulate LGDs, we follow a slightly modified version of the approach proposed by Hlawatsch and Ostrowski (2011), because it can be readily extended to consider separately cure, partial recovery, and write-off cases. Knowing the end status of the simulated defaults is required to be able to fit some of the multi-stage models.

We assume that the true distributions for the LGDs of cure, partial recovery, and write-off cases are three distinct beta distributions. Their probability density functions are therefore

$$f_s(LGD|\alpha_s, \beta_s) = \frac{1}{B(\alpha_s, \beta_s)} LGD^{\alpha_s - 1}(1 - LGD)^{\beta_s - 1} \; \forall \; s \in \{C, P, W\}, \tag{2}$$

where

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}, \tag{3}$$

$\Gamma(x)$ is the Gamma function, and the possible default end statuses s are denoted by C for cure, P for partial recovery, and W for write-off. To produce the LGD values for one data set, we draw in total 1000 realisations from these distributions, that is, P_C cured cases from f_C, P_W write-off cases from f_W, and $P_P = 1000 - P_C - P_W$ partial recovery cases from f_P. Here P_C, P_P and P_W can be interpreted as the probabilities of sampling from the respective distributions, and in fact, this approach is statistically equivalent to sampling directly from a mixture of the beta distributions similar to Hlawatsch and Ostrowski (2011), with the added benefit of simultaneously classifying each default by the end status. The classification is included explicitly in the data set using separate binary indicators for cure, partial recovery and write-off, I_C, I_P and I_W, respectively.

In total, we simulate 5000 data sets. For diversity between the data sets, we randomly sample the number of sampled cure and write-off observations, P_C and P_W, as well as the beta distribution parameters α_s and β_s for each data set. The number of cure and write-off observations are sampled from the intervals $P_C \in [100, 500]$ and $P_W \in [100, 250]$, which means that $P_P \in [250, 800]$.

To follow the observation that losses for cures should be mainly very low and that a wide range of loss outcomes should be possible for partial recoveries and write-offs with losses concentrating more at the low end for the former and at the high end for the latter (Salko and D'Ecclesia, 2022), and to generate the characteristic bimodal shape for the combined distributions, we choose the possible parameter values for the beta distributions such that the loss distribution will be right-skewed for cures, symmetrical to right-skewed for partial recoveries, and left-skewed for write-offs. The direction of the skew is enforced by the inequalities $\alpha_C \leq \beta_C$, $\alpha_P \leq \beta_P$ and $\alpha_W \geq \beta_W$. To avoid modes greater than 0 and lower than 1 for the right-skewed and left skewed distributions, respectively, we set $\alpha_C \leq 1$, $\alpha_P \leq 1$ and $\beta_W \leq 1$. Additionally, to

enforce unimodality of the distribution of cures, we also set $\beta_C \geq 1$. However, for partial recoveries and write-offs, we allow also bimodal distributions where $\alpha_P < 1$ and $\beta_P < 1$, and $\alpha_W < 1$ and $\beta_W < 1$.

Moreover, we choose the intervals of possible mean and variance values for our simulated distributions by loosely representing empirical LGD means and variances for cures, partial recoveries and write-offs reported by Salko and D'Ecclesia (2022). Their statistics are based on 8755 European real-estate backed loans. To accommodate more variety in our simulated data, we allow values from wide intervals around the reported values. We assume that the real-estate secured loans are low-risk compared to, for example, any unsecured loans, and thus we place the reported mean values at the lower ends of the allowed intervals. We present the statistics by Salko and D'Ecclesia (2022) and our allowed LGD mean and variance value intervals in table 1.

Table 1: Empirical LGD means and variances in a data set comprised of 8755 European real-estate backed loans by Salko and D'Ecclesia (2022), and allowed mean and variance intervals for the simulated distributions.

Statistic	Mean	Allowed mean interval	Variance	Allowed variance interval
(Overall)	(0.1993)		(0.1069)	
Cures	0.0063	[0.005, 0.05]	0.0003	[0.0001, 0.0025]
Partial Recoveries	0.1472	[0.10, 0.50]	0.0876	[0.04, 0.16]
Write-offs	0.5153	[0.50, 0.995]	0.1180	[0.0001, 0.2025]

For each beta distribution, the α and β parameter values are generated in a two-step process. First, the mean value for the distribution is uniformly sampled from the allowed interval. Then, the variance is sampled uniformly from its allowed interval, taking into account also the possible range of values given the sampled mean value - α and β are always greater than zero, and not all mean-variance combinations are possible.

Given a mean μ, the final allowed sampling interval for the variance σ^2 for cures, partial recoveries and write-offs are

$$\sigma_C^2 \in \left[\max\left(\frac{\mu_C^2(1-\mu_C)}{1+\mu_C}, 0.0001\right),\ \min\left(\frac{\mu_C(1-\mu_C)^2}{2-\mu_C}, 0.0025\right)\right], \tag{4}$$

$$\sigma_P^2 \in \left[\max\left(\frac{\mu_P^2(1-\mu_P)}{1+\mu_P}, 0.04\right),\ \min(\mu_P(1-\mu_P), 0.16)\right) \tag{5}$$

and

$$\sigma_W^2 \in \left[\max\left(\frac{\mu_W(1-\mu_W)^2}{2-\mu_W}, 0.0001\right),\ \min(\mu_W(1-\mu_W), 0.2025)\right), \tag{6}$$

respectively. We derive the mean-dependent parts of the intervals in appendix A following a similar derivation of Hlawatsch and Ostrowski (2011).

Once the mean and variance values are sampled, the parameters α and β are computed using the equations

$$\alpha = \mu\left(\frac{\mu(1-\mu)}{\sigma^2} - 1\right) \tag{7}$$

and

$$\beta = (1-\mu)(\frac{\mu(1-\mu)}{\sigma^2} - 1), \tag{8}$$

which are derived in appendix B.

The LGD values simulated from the beta distributions are in the open interval $(0, 1)$. To achieve another characteristic feature of LGD distributions, the large observation concentrations at exactly 0 and 1, and to accommodate multi-stage models which differentiate the exactly 0 or 1 loss cases from others, we first scale each LGD value to the interval $(-0.01, 1.01)$, and then clip the values back to the closed interval $[0, 1]$. Due to the original LGD values being between 0 and 1, the values can be conveniently scaled to any interval $[a, b]$ by

$$\widetilde{LGD} = a + (b-a) \cdot LGD, \tag{9}$$

giving us the final capped LGD values as

$$\overline{LGD} = \min(\max(\widetilde{LGD}, 0), 1). \tag{10}$$

3.1.2 Copulas

Copulas are functions that can be used to combine univariate distribution functions in order to represent a multivariate joint distribution with some dependency structure. They are widely used in finance and risk management (Haugh, 2016).

Formally, a d-dimensional copula, $C : [0, 1]^d :\rightarrow [0, 1]$ is a cumulative distribution function (CDF) with uniform marginals (Haugh, 2016). Crucially, a foundational result in copula theory, Sklar's theorem, states that for any multivariate joint distribution $F(x_1, x_2, \ldots, x_d)$ with continuous marginal distributions $F_1, F_2, \ldots, F_d$, there exists a copula C such that

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)). \tag{11}$$

In essence, Sklar's theorem states that for any continuous marginal distributions, we can use a copula to introduce a dependence structure to form a joint multivariate distribution. We will use this result when generating the explanatory variables for our data sets in section 3.1.3.

Specifically, we will utilise a Gaussian copula, which is derived from the multivariate normal distribution. For a d-dimensional multivariate normal distribution with a correlation matrix P, the corresponding Gaussian copula is defined as

$$C_P^{Gauss}(u_1, \ldots, u_d) = \Phi_P^d(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)), \tag{12}$$

where Φ_P^d is the joint multivariate CDF and Φ is the standard univariate normal CDF (Haugh, 2016).

Because the normal distribution has light tails, the Gaussian copula can underestimate extreme tail behavior and should therefore be used with caution in risk modelling (Haugh, 2016). However, we deal with mainly normal distributions and strong tail dependence is not needed, so Gaussian copulas are suitable for this thesis.

3.1.3 Simulating Explanatory Variables

As explanatory variables, we use eight artificial standard normally distributed variables, A-H. Standard normal distributions are chosen instead of a variety of other possibly more realistic distributions in order to isolate any differences in the shape of predicted LGD distributions to be caused by the model structures themselves instead of different variable distributions being available in some components.

The number of explanatory variables and their correlation structures with the target variables are chosen so that they enable predictive single-stage and multi-stage models. A and B have positive correlations with the cure indicator (I_C), while C and D are positively correlated with the write-off indicator (I_W). These variables represent features which have a causal relationship with - and are commonly used to predict - the end status of the default, such as financial ratios, payment behavior, or collateral type. E has a positive correlation with LGD regardless of the end status, and it represents a general financial feature.

F, G and H are end status -specific predictors. F has a positive correlation with LGD, but only within cured cases (LGD_C). It represents variables, such as the type of loan, which explain losses incurred despite the cure, due to, for example, delayed payments and administrative costs. G and H are designed to explain the larger losses that occur during a collection process and possible collateral liquidation, such as the loan-to-value ratio, and loan seniority. They are set to have positive correlations with LGD within the partial recovery (LGD_P) and write-off (LGD_W) cases such that the correlation within partial recovery cases is independent of the correlation within write-off cases. The correlation between LGD and F, LGD and G, and LGD and H are set to zero within the other end statuses where they are not meant to be predictive. The correlations between A and B, C and D, and G and H are set to be slightly positive for added realism. The correlations between the other variables are not explicitly set.

The actual correlation values are sampled independently for each data set. The intervals from which the correlation values are sampled for each explicitly set variable pair are presented in tables 2 and 3.

Table 2: Correlation intervals between the cure and write-off indicators and explanatory variables A-D.

Variable	I_C	I_W	A	B	C	D
I_C	1		[0.0, 0.5]	[0.0, 0.5]		
I_W		1			[0.0, 0.5]	[0.0, 0.5]
A	[0.0, 0.5]		1	[0.0, 0.2]		
B	[0.0, 0.5]		[0.0, 0.2]	1		
C		[0.0, 0.5]			1	[0.0, 0.2]
D		[0.0, 0.5]			[0.0, 0.2]	1

Following the example of Hlawatsch and Ostrowski (2011), Gürtler and Hibbeln (2011) and Li et al. (2018), we use Gaussian copulas to introduce the specified dependency structures between the explanatory variables. However, we also use the same copulas to introduce the dependency between the explanatory variables and the target variables.

The variable generation procedure is explained in detail below using variables

Table 3: Correlation intervals between LGD in the overall scope and in end status specific scopes and explanatory variables E-H.

Variable	LGD	LGD_C	LGD_P	LGD_W	E	F	G	H
LGD	1	1	1	1	[0.0, 0.5]			
LGD_C	1	1				[0.0, 0.5]	0	0
LGD_P	1		1			0	[0.0, 0.5]	[0.0, 0.5]
LGD_W	1			1		0	[0.0, 0.5]	[0.0, 0.5]
E	[0.0, 0.5]				1			
F		[0.0, 0.5]	0	0		1		
G		0	[0.0, 0.5]	[0.0, 0.5]			1	[0.0, 0.2]
H		0	[0.0, 0.5]	[0.0, 0.5]			[0.0, 0.2]	1

A and B and the cure indicator I_C as an example, but it is identical for all groups of explanatory variables and target. For variables E and F, which have no set correlation with other explanatory variables, the third variable from the example is simply removed. For variables F, G and H, the generation process is performed separately for each of the end status scopes to set the end status -specific correlations with the target. However, the correlations between the explanatory variable pairs are global within each data set, so the same sampled correlation values are used in each scope.

We start by sampling the correlation values r from uniform distributions within the specified intervals and building the correlation matrix:

$$R_{I_C,A,B} = \begin{bmatrix} 1 & r_{I_C,A} & r_{I_C,B} \\ r_{I_C,A} & 1 & r_{A,B} \\ r_{I_C,B} & r_{A,B} & 1 \end{bmatrix}, \tag{13}$$

where $r_{I_C,A} \in [0.0, 0.5]$, $r_{I_C,B} \in [0.0, 0.5]$ and $r_{A,B} \in [0.0, 0.2]$.

Then, we sample the required number N of observations $z_{ij}, i \in \{1, \ldots, N\}, j \in \{1, 2, 3\}$ from three independent standard normal distributions, and use the Cholesky decomposition of the correlation matrix to transform the independent observations into dependent standard normally distributed observations x_{ij} with our specified correlation structure (Hlawatsch and Ostrowski, 2011):

$$x_{i,1} = z_{i,1} \; \forall \, i \in \{1, \ldots, N\}, \tag{14}$$

$$x_{i,2} = r_{I_C,A} \cdot z_{i,1} + \sqrt{1 - r_{I_C,A}^2} \cdot z_{i,2} \; \forall \, i \in \{1, \ldots, N\}, \tag{15}$$

$$x_{i,3} = r_{I_C,B} \cdot z_{i,1} + \frac{r_{A,B} - r_{I_C,A} \cdot r_{I_C,B}}{\sqrt{1 - r_{I_C,A}^2}} \cdot z_{i,2} \tag{16}$$

$$+ \sqrt{1 - r_{I_C,B}^2 - (r_{A,B} - r_{I_C,A} \cdot r_{I_C,B})^2} \cdot z_{i,3} \; \forall \, i \in \{1, \ldots, N\}. \tag{17}$$

Because standard normal distributions are the final (variable) distributions in the analysis, the variable generation process is complete. To combine the explanatory variables with the previously generated target variables (LGD and end status indicators), we first shuffle the target observations into a random order, then sort the targets by $I_{C,i}$ and the explanatory variables by $x_{i,1}$ in ascending order. The shuffling and sorting is always done so that the different target variables or explanatory variables are treated as one observation and their relative order remains unchanged. Finally, we join the

corresponding target and explanatory variable observations after sorting such that $A_i = x_{i,2}$ and $B_i = x_{i,3}$. The shuffling before sorting prevents accidentally introducing unwanted correlations with a different target variable, i.e. LGD in this example, if the targets are already sorted by the different target variable after generating other explanatory variables.

Should we want to use different distributions than the standard normal for the explanatory variables, we could additionally input the standard normally distributed $x_{i,2}$ and $x_{i,3}$ into the standard normal CDF in order to transform them into correlated uniformly distributed variables, and then input the uniformly distributed variables into the inverse CDFs of the desired marginal distributions to generate the final observations from those marginal distributions Hlawatsch and Ostrowski (2011).

3.2 Models

3.2.1 Ordinary Least Squares Regression

The simplest model in our analysis is the ordinary least squares regression (OLS). It assumes a linear relationship between a target y_i and explanatory variables x_{ij}, such that

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i, \tag{18}$$

where β_j are unknown regression coefficients and ϵ_i is a random error, which is assumed to be normally distributed with a zero mean and an unknown but constant variance (Yan and Su, 2009).

The estimates of a fitted OLS model are given by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik}, \tag{19}$$

where $\hat{\beta}_j$ are the fitted regression coefficients. The coefficients are fitted by minimising the sum of squared errors SSE (giving the model its name), i.e. the sum of squared differences between the observed values y_i and the estimated values $\hat{y}_i$ (Yan and Su, 2009):

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij})^2. \tag{20}$$

The closed-form solution for the regression coefficients that minimise the sum of squared errors is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}, \tag{21}$$

where $\hat{\boldsymbol{\beta}}$ are the estimated regression coefficients in vector form, $\mathbf{X}$ is a matrix containing the explanatory variable observations and $\mathbf{y}$ is a vector containing the observed target values (Yan and Su, 2009). However, the matrix calculations required by the closed-form solution are often computationally expensive, so in practice the regression coefficients are usually computed using other numerical methods (Yan and Su, 2009).

We include OLS in the comparison both on its own as a single-stage model, which directly predicts LGD, as well as a part of the multi-stage models.

3.2.2 Logistic Regression

Logistic regression is the most common regression model to use when the target variable is binary (Hosmer Jr et al., 2013). The predictions given by the model are bounded between zero and one, and represent the probability that the target variable has the value one given the explanatory variable values. The predictions are given by:

$$\pi(x_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})}}. \tag{22}$$

An important feature of the model is that when transformed using the logit transformation to represent log-odds, the model is linear (Hosmer Jr et al., 2013):

$$g(x_i) = \ln\left(\frac{\pi(x_i)}{1 - \pi(x_i)}\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}. \tag{23}$$

The regression coefficients are estimated by maximum likelihood estimation, i.e. by finding the coefficient values that maximise the probability of observing the given the data. The likelihood function to be maximised is (Hosmer Jr et al., 2013):

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} (1 - \pi(x_i))^{1 - y_i}. \tag{24}$$

Unlike for OLS, no closed-form solution exists for this maximisation problem, and the solution must be found using numerical methods (Hosmer Jr et al., 2013).

We use the logistic regression model to estimate probabilities of cure, partial recovery, write-off, zero-loss or full-loss in the multi-stage models.

3.2.3 Zero-Fractional-One Multi-stage Model

The Zero-Fractional-One (ZFO) model, based on Loterman et al. (2012) and Bellotti and Crook (2012), divides the LGD prediction into three stages. In the first stage, it predicts the probability that LGD is exactly zero using a logistic regression model that is trained on all observations regardless of LGD. In the second stage, it predicts the probability that LGD is exactly one given that it is greater than zero, using a logistic regression model that is trained only on observations where LGD is greater than zero. In the final stage, it predicts the fractional LGD given it is neither zero nor one using an OLS model that is trained only on observations where LGD is between zero and one. The final LGD prediction is the probability-weighted average of the three outcomes:

$$\begin{aligned} LGD^{ZFO} &= P_0 \cdot 0 + (1 - P_0) \cdot P_1 \cdot 1 + (1 - P_0) \cdot (1 - P_1) \cdot LGD^{OLS} && (25) \\ &= (1 - P_0)(P_1 + (1 - P_1) \cdot LGD^{OLS}), && (26) \end{aligned}$$

where $P_0 = P(LGD = 0)$ is the prediction of the first logistic regression model, $P_1 = P(LGD = 1 | LGD > 0)$ is the prediction of the second logistic regression model and LGD^{OLS} is the LGD prediction of the third-stage OLS model. The structure of the model is illustrated in figure 1.

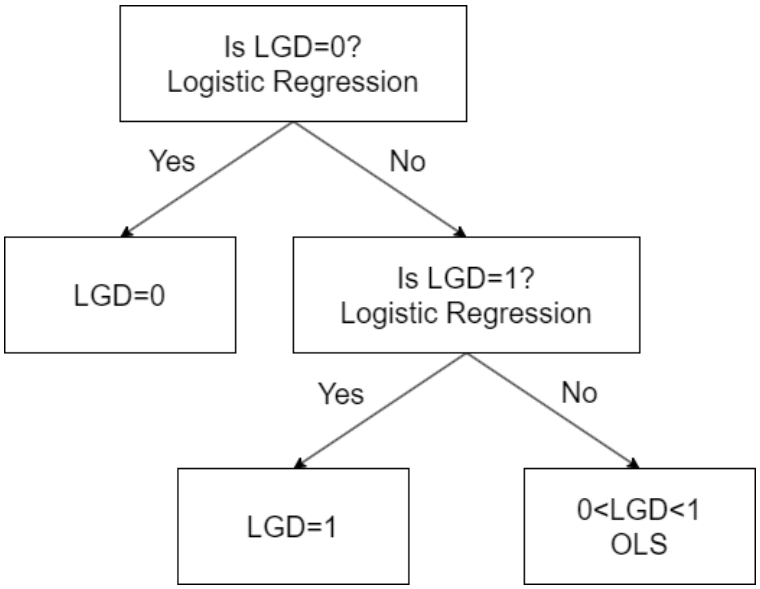


Figure 1: Structure of the Zero-Fractional-One (ZFO) model.

3.2.4 Write-off-Non-Write-off Multi-stage Model

The Write-off-Non-Write-off (WNW) model, proposed by Gürtler and Hibbeln (2011), divides the LGD prediction into two stages. In the first stage, it predicts the probability of Write-off using a logistic regression model, and in the second stage two separate OLS models are used to predict write-off losses and non-write-off losses. The write-off OLS model is trained only on write-off observations (write-off indicator $I_W = 1$) and the non-write-off OLS model is trained only on non-write-off observations ($I_W = 0$). The final LGD prediction is the probability-weighted average of the two outcomes:

$$LGD^{WNW} = P_W \cdot LGW + (1 - P_W) \cdot LGNW, \tag{27}$$

where P_W is the probability of write-off predicted by the logistic regression model, LGW (loss given write-off) is the LGD prediction of the OLS model for the write-off cases and $LGNW$ (loss given no write-off) is the LGD prediction of the OLS model for the non-write-off cases. The structure of the model is illustrated in figure 2.

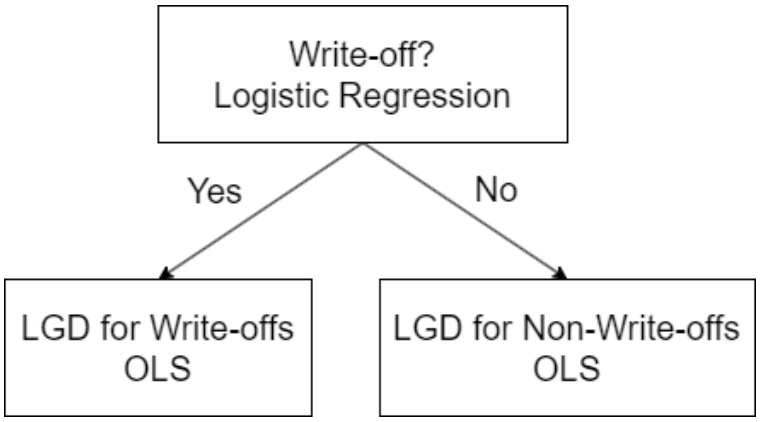


Figure 2: Structure of Write-off-Non-Write-off (WNW) model.

3.2.5 Cure-Partial-recovery-Write-off Multi-stage Model

The Cure-Partial-recovery-Write-off (CPW) model, proposed by Starosta (2021), expands on the WNW model by considering cures, partial recoveries and write-offs separately istead of just write-offs and non-write-offs. First, a logistic regression model is used to predict the probability of cure, after which a second logistic regression model, trained only on non-cured observations, is used to predict the probability of write-off given no cure. Finally, three separate OLS models are used to predict the LGD for the cured, partial recovery, and write-off cases. The OLS models are trained only on observations of the corresponding end status. The final LGD prediction is the probability-weighted average of the three outcomes:

$$LGD^{CPW} = P_C \cdot LGC + (1 - P_C) \cdot (P_W \cdot LGW + (1 - P_W) \cdot LGP), \tag{28}$$

where P_C is the probability of cure predicted by the first logistic regression model, $P_W = P(I_W = 1|I_C = 0)$, is the probability of write-off given no cure predicted by the second logistic regression model, LGC (loss given cure) is the LGD prediction of the OLS model for the cured cases, LGW (loss given write-off) is the LGD prediction of the OLS model for the write-off cases, and LGP (loss given partial recovery) is the LGD prediction of the OLS model for the partial recovery cases. The structure of the model is illustrated in figure 3.

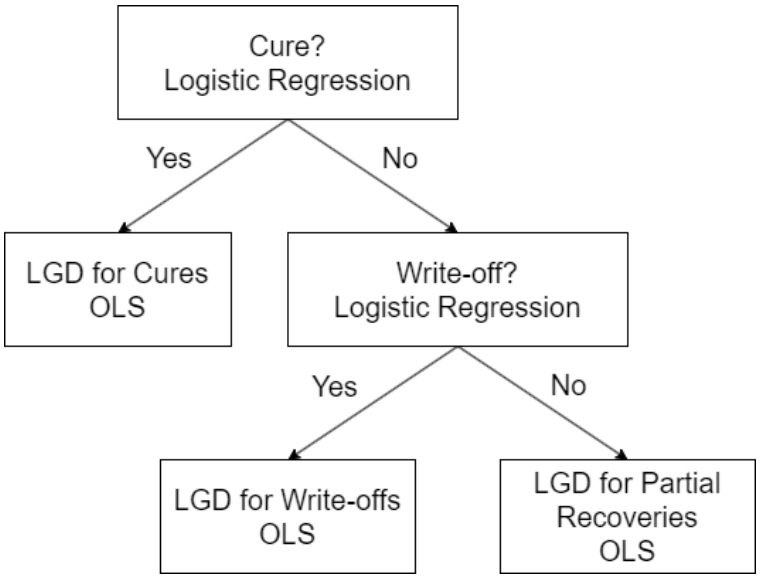


Figure 3: Structure of the Cure-Partial-recovery-Write-off (CPW) model.

3.3 Performance Metrics

3.3.1 Coefficient of Determination R^2

R^2, or the coefficient of determination, is a common measure of model fit and performance. It is defined as (Yan and Su, 2009):

$$R^2 = 1 - \frac{SSE}{SST}, \tag{29}$$

where $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the sum of squared prediction errors, and $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares, i.e. the sum of squared differences between the target values and the target mean. Intuitively, it can be interpreted as the proportion of the total variation in the data that is explained by the model (Yan and Su, 2009). The best possible R^2 value, 1, is obtained when the errors are zero and all of the variation is explained by the model. A model which explains no variation, such as one that always predicts the mean, gets an R^2 value of 0. R^2 can also be negative if the model predicts worse than the mean model.

In the context of LGD, R^2 is used as a measure of calibration, i.e. how close the predicted losses are to the observed losses (Loterman et al., 2012). The advantage of R^2 compared to, for example, raw SSE as a measure of calibration is that it is normalised by the total variation in the data, SST, making R^2 comparable between different data sets (Loterman et al., 2012).

To follow common practice in the LGD literature, we use R^2 to measure the explanatory power and calibration of the models. We also use R^2 to describe the simulated data sets when analysing how the predictiveness of different explanatory variables in terms of LGD affects the performance of the models.

3.3.2 Area Under the Curve

In binary classification problems, prediction performance can be evaluated using, among others, the measures sensitivity (ratio of correct positive predictions to all positive observations), specificity (ratio of correct negative predictions to all negative observations), false positive rate (ratio of false positive predictions to all negative observations, equal to 1 - specificity) and false negative rate (ratio of false negative predictions to all positive observations, equal to 1 - sensitivity) (Nahm, 2022). These measures, however, depend on a classification threshold (e.g. in logistic regression, a threshold for the predicted probability, such that if the predicted probability is higher than the threshold, the observation is classified as 1, and 0 otherwise). Typically, when the threshold is changed such that sensitivity increases, the model also predicts more false positives, and the specificity decreases (Nahm, 2022).

The receiver operating characteristic (ROC) curve plots sensitivity against 1 - specificity, and can be used to assess how this dynamic plays out for a model. For a perfect model, sensitivity is 1 regardless of specificity, and the curve goes from (0,0) to (0,1) and then to (1,1). A random model forms a curve on the diagonal, from (0,0) to (1,1). For a typical model, the curve is somewhere between these two extremes (Nahm, 2022). Figure 4 illustrates ROC curves for a random model, perfect model and an example model.

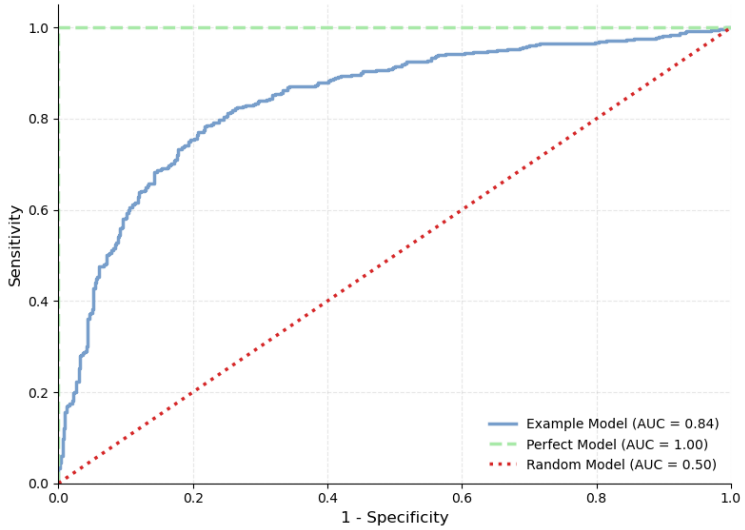


Figure 4: Receiver operating characteristic curves and corresponding AUCs for a random model, perfect model and an example model.

To summarise the performance of a model further, the area under the (receiver operating characteristic) curve (AUC) gives a measure of the model's discriminatory power, which can be interpreted as the probability that the predicted probability of a positive classification is higher for a random positive observation than for a random negative observation (Irwin and Irwin, 2013). The best possible AUC value is 1, while the AUC for a random model is 0.5. Figure 4 includes the AUC values for the

corresponding ROC curves.

AUC is independent of a set classification threshold or the distribution of the classes in the data, which makes model comparisons using the measure meaningful also between different data sets (Irwin and Irwin, 2013). As such, we use AUC in variable selection for the logistic regression models, as well as to describe the simulated data sets in terms of how predictive the different explanatory variables are in terms of the cure and write-off indicators.

3.3.3 Generalised Area Under the Curve

As part of their instructions for reporting the validation results of internal models, European Central Bank (2019) introduced a measure of discriminatory power for continuous and multi-class targets. This measure is the generalised AUC (gAUC), which is based on Somers' D, a common measure of ordinal association:

$$gAUC = \frac{D + 1}{2}. \tag{30}$$

To calculate Somers' D for gAUC, the predicted and observed LGD values are first discretised. If the predictions are continuous (such as in our case) or based on more than 20 unique LGD values, the values are discretised into 12 bins based on the following split points: 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 1 (lower boundary inclusive). Otherwise, if the predicted LGD is already discrete with 20 or less unique values, the unique predicted LGD values are used as split points for the discretisation of the observed LGD values such that the bins are upper boundary inclusive (European Central Bank, 2019).

Then, a contingency table of the frequency of discretised LGD predictions and corresponding discretised observed values is formed. With predicted LGD bins as rows and observed LGD bins as columns in the contingency table, a_{ij} denotes the frequency of cases where the predicted LGD falls into bin i and the observed LGD falls into bin j. From the contingency table, the number of agreements A_{ij} and disagreements D_{ij} are calculated for each table cell (i, j), such that agreements are the total frequency of cases where both the predicted and observed bin indices are greater or smaller than i and j, respectively, and disagreements are the total frequency of cases where the predicted bin index is greater than i but the observed bin index is smaller than j, or vice versa (European Central Bank, 2019):

$$A_{ij} = \sum_{k<i} \sum_{l<j} a_{kl} + \sum_{k>i} \sum_{l>j} a_{kl}, \tag{31}$$

$$D_{ij} = \sum_{k>i} \sum_{l<j} a_{kl} + \sum_{k<i} \sum_{l>j} a_{kl}. \tag{32}$$

The calculation of A_{ij} and D_{ij} is illustrated in figure 5.

Finally, Somers' D is calculated as (European Central Bank, 2019):

$$D = \frac{P - Q}{w_r}, \tag{33}$$

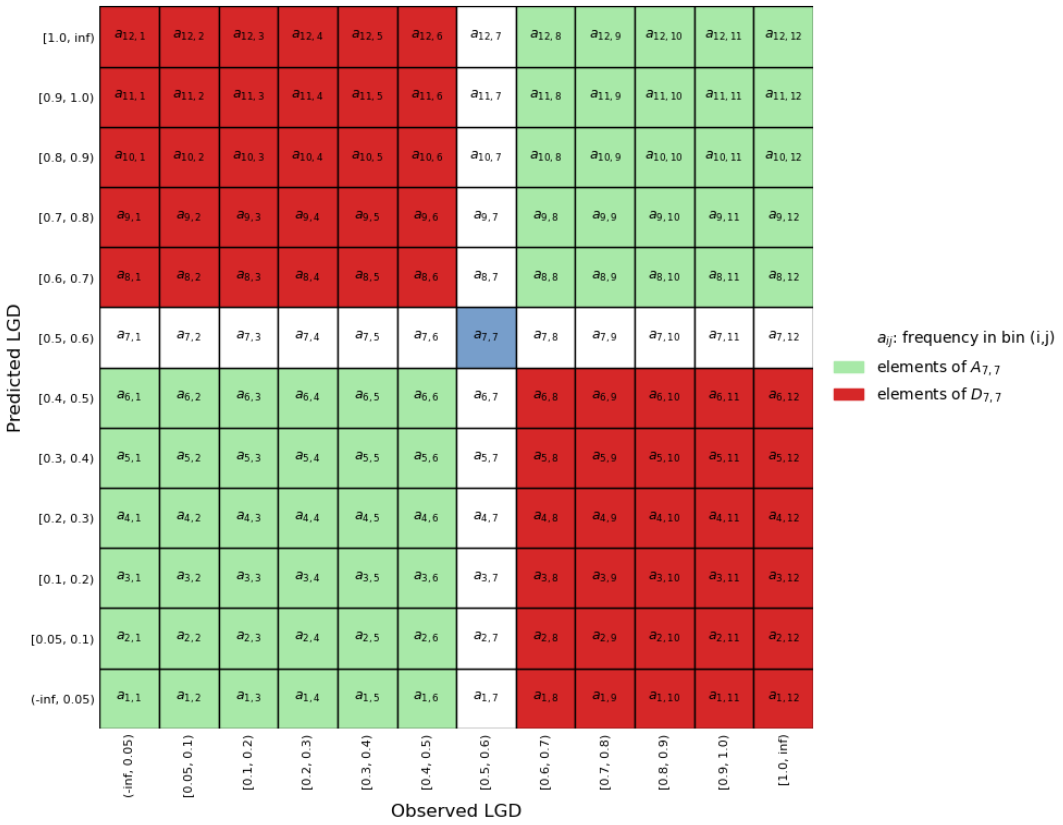


Figure 5: Contingency table for the frequency of binned predicted and observed LGD pairs, which is used to calculate agreements A_{ij} and disagreements D_{ij} for Somers' D and gAUC. $A_{7,7}$ is the sum of the green elements in the table, while $D_{7,7}$ is the sum of the red elements.

where $P = \sum_i \sum_j a_{ij} A_{ij}$, $Q = \sum_i \sum_j a_{ij} D_{ij}$ and $w_r = (\sum_i \sum_j a_{ij})^2 - \sum_i (\sum_j a_{ij})^2$, that is, the total frequency squared minus the sum of the squared row frequencies.

Similar to AUC, the perfect predictor will get a gAUC value of 1, while a random predictor will get a gAUC value of 0.5. If the target variable is binary, gAUC reduces to ordinary AUC (Newson, 2002).

We use gAUC to measure the discriminatory power of the models and in variable selection for the linear regression models.

3.4 Model Fitting

To fit the models to the data, each data set is first randomly divided into a training set and a test set using a 70-30 split. The models are fitted on the training sets, and their performance is measured and reported on the test set.

Variable selection for the single-stage model and each model component in the multi-stage models is performed using the following forward selection algorithm: starting from a model with only a constant, each variable is added to the model one

at a time, and the gAUC of the resulting model is estimated on the training set. The variable which improves the gAUC the most is added to the model, as long as all variables in the model remain significant after the addition (p-value < 0.05). This procedure is repeated until no variable improves the gAUC such that all of the variables in the model remain significant or until all variables are already added to the model.

While crude, the forward selection algorithm is sufficient for this thesis because our simulated data is very clean compared to real world data and does not contain complex dependencies where a variable is only predictive when paired with some other variable. We choose gAUC as the main criterion in the selection algorithm instead of R^2, since, as described in chapter 2.3, high discriminatory power is the main focus of a LGD scoring model.

3.5 Model Analysis Setup and Model Estimation

The goal of the model analysis is to gain insight into how different data factors affect the performance of different multi-stage model structures compared to a simple single-stage model, and to determine what kind of data is beneficial for using more complex multi-stage models instead of single-stage models.

To answer these questions, we assess the relationships between various data summary statistics and the performance difference of the three multi-stage models (ZFO, WNW and CPW) compared to OLS using OLS fits on the gAUC and R^2 difference and summary statistic value pairs of all data sets. Furthermore, we group the data sets based on quantiles of the summary statistics and calculate the average gAUC and R^2 differences compared to OLS and their 95% confidence intervals for each model for a more robust comparison on the performance levels for different summary statistic values. The analysis is combined with a visual assessment on how the shape and composition of the LGD distribution is related to different data set summary statistic values.

In the analysis, we use the three sets of summary statistics in table 4. The first set of statistics describes the shape of the LGD distribution, and the second set describes the end status composition of the data. These two sets of statistics are easily available for any LGD data set, and their effect is studied to get early insights into model structure choice based only on the LGD distribution, without any analysis of the explanatory variables or model fitting. This is useful, because a modeller may not be able to start the modelling process with a complete data set that includes all imaginable explanatory variables, making it necessary to start while data work is still ongoing.

The third set of statistics describes the predictiveness of the available variables. Their purpose is to provide additional information on suitable model structures for a data set based on light analysis of the explanatory variables.

In addition to analysing the model performance by the summary statistics, we analyse the overall performance of the models through gAUC and R^2 statistics over all of the simulated data sets. We also examine visually the shape of the predicted LGD distributions.

Table 4: Summary statistics used to describe the data sets in the model performance analysis.

Statistic	Description
LGD mean	Mean of the LGD distribution.
LGD variance	Variance of the LGD distribution.
Zero rate	The proportion of cases for which LGD is zero.
One rate	The proportion of cases for which LGD is one.
Zero-one rate	The proportion of cases for which LGD is zero or one.
Cure rate	The proportion of cure cases.
Partial recovery rate	The proportion of partial recovery cases.
Write-off rate	The proportion of write-off cases.
AB-cure AUC	AUC of a probability of cure logistic regression model using variables A and B. Describes the predictiveness of the explanatory variables for the probability of cure.
CD-write-off AUC	AUC of a probability of write-off logistic regression model using variables C and D. Describes the predictiveness of the explanatory variables for the probability of write-off.
E-LGD R^2	R^2 of an LGD OLS model using variable E. Describes the predictiveness of the end status -independent explanatory variables for LGD.
F-LGD$_C$ R^2	R^2 of an LGD$_C$ OLS model using variable F. Describes the predictiveness of the explanatory variables for LGD within cure cases.
GH-LGD$_P$ R^2	R^2 of an LGD$_P$ OLS model using variables G and H. Describes the predictiveness of the explanatory variables for LGD within partial recovery cases.
GH-LGD$_W$ R^2	R^2 of an LGD$_W$ OLS model using variables G and H. Describes the predictiveness of the explanatory variables for LGD within write-off cases.

4 Results

4.1 Simulated Data

Figure 6 contains a combined histogram of all of the 5000 simulated LGD distributions and table 5 contains summary statistics for the simulated data sets. The table includes the statistics that are used in model performance analysis as well as the correlations and end status -specific LGD statistics that were used to simulate the data sets. For all correlation pairs, see appendix F, and for a visual representation of how the summary statistics manifest in the data sets, see appendix C, which contains histograms of the LGD data by quantiles of the summary statistics which are used in the model performance analysis.

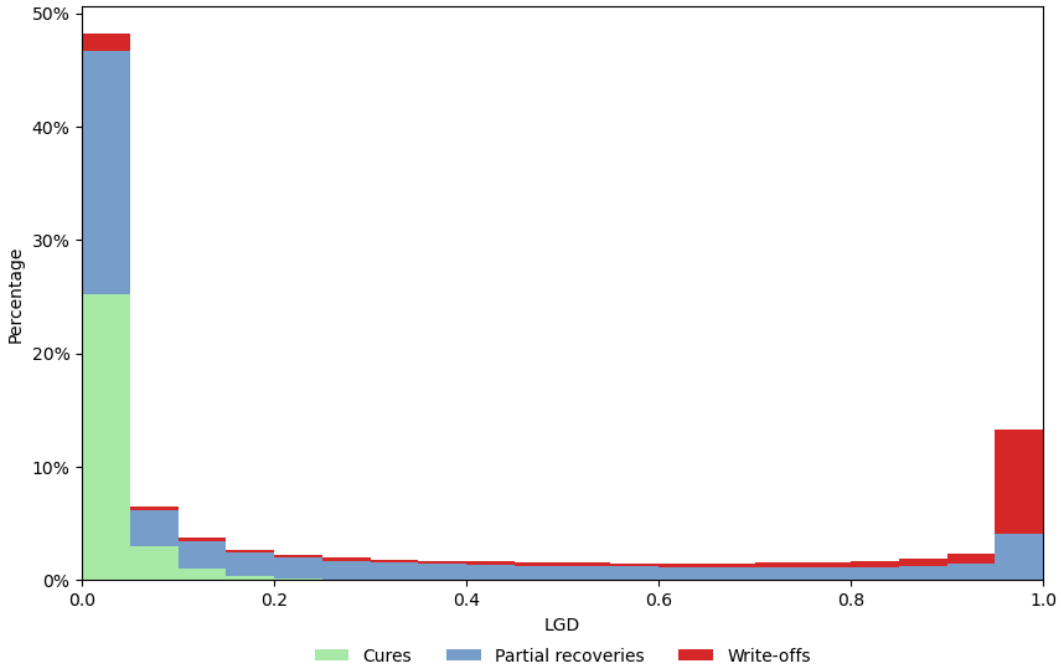


Figure 6: Combined histogram of all 5000 simulated LGD data sets.

The realised LGD mean and variance intervals correspond well to the allowed intervals in table 1. The slight differences are due to randomness in sampling and the scaling and capping that are performed to extend the LGD distributions to exactly 0 and 1. Similarly, the cure rate, partial recovery rate and write-off rate correspond well to the specified sampling intervals.

The distributions for the realised correlation values between the explanatory variables and targets where higher than zero correlation is expected are shifted slightly downward compared to the specified sampling intervals. The difference is caused by the data generation mechanism. Strictly speaking, the specified correlations hold only between the explanatory variables and the latent variable in the copula, which is used to join the explanatory variables to the target values, and not between the explanatory

variables and targets themselves. This is because we join the explanatory variables to the existing target values by rank, instead of freshly sampling the target values from their respective distributions using the latent variable, which would preserve the exact correlation structure. Additionally, because the target distributions are binary in the case of the end status indicators and capped at 0 and 1 in the case of LGD, the relationship of a higher explanatory variable value leading to a higher target value is broken, leading to lower correlation values compared to the continuous latent variable.

Table 5: Summary and correlation statistics for the simulated data sets. The statistics that are used in model performance analysis are bolded.

Statistic	Mean	Min	Max	0.1 quantile	median	0.9 quantile
LGD mean	**0.2961**	**0.1151**	**0.5370**	**0.1976**	**0.2892**	**0.4062**
LGD_C mean	0.0221	0.0002	0.0516	0.0074	0.0215	0.0380
LGD_P mean	0.3002	0.0845	0.5420	0.1416	0.2984	0.4596
LGD_W mean	0.7518	0.4303	1.0000	0.5487	0.7548	0.9529
LGD variance	**0.1370**	**0.0559**	**0.2150**	**0.1008**	**0.1379**	**0.1720**
LGD_C variance	0.0016	0.0000	0.0106	0.0006	0.0016	0.0025
LGD_P variance	0.1062	0.0329	0.1791	0.0602	0.1062	0.1512
LGD_W variance	0.0984	0.0000	0.2238	0.0174	0.0980	0.1804
Zero rate	**0.3114**	**0.0350**	**0.7840**	**0.1260**	**0.2900**	**0.5340**
One rate	**0.0922**	**0.0000**	**0.3150**	**0.0220**	**0.0840**	**0.1760**
Zero-one rate	**0.4035**	**0.0640**	**0.9610**	**0.1949**	**0.3835**	**0.6470**
Cure rate	**0.2985**	**0.1000**	**0.5000**	**0.1380**	**0.2970**	**0.4570**
Partial recovery rate	**0.5265**	**0.2590**	**0.7970**	**0.3610**	**0.5260**	**0.6890**
Write-off rate	**0.1749**	**0.1000**	**0.2500**	**0.1140**	**0.1750**	**0.2350**
A-I_C correlation	0.1826	-0.0693	0.4488	0.0354	0.1809	0.3334
B-I_C correlation	0.1848	-0.0851	0.4481	0.0357	0.1840	0.3334
C-I_W correlation	0.1664	-0.0851	0.4195	0.0314	0.1650	0.3043
D-I_W correlation	0.1678	-0.0655	0.4183	0.0327	0.1665	0.3034
E-LGD correlation	0.2127	-0.0738	0.5336	0.0427	0.2095	0.3858
F-LGD_C correlation	0.1854	-0.2630	0.5435	0.0176	0.1783	0.3636
F-LGD_P correlation	0.0002	-0.1877	0.1939	-0.0557	0.0001	0.0569
F-LGD_W correlation	0.0018	-0.3011	0.3095	-0.1001	0.0014	0.1023
G-LGD_C correlation	-0.0003	-0.2766	0.2783	-0.0793	0.0004	0.0766
G-LGD_P correlation	0.2194	-0.1278	0.5468	0.0372	0.2169	0.4041
G-LGD_W correlation	0.2017	-0.2166	0.6122	0.0190	0.1956	0.3997
H-LGD_C correlation	0.0010	-0.2663	0.2639	-0.0778	0.0003	0.0778
H-LGD_P correlation	0.2183	-0.1274	0.5417	0.0406	0.2170	0.3996
H-LGD_W correlation	0.2014	-0.2120	0.5934	0.0156	0.1974	0.3918
A-B correlation	0.1010	-0.0851	0.2720	0.0137	0.1031	0.1850
C-D correlation	0.0993	-0.1051	0.2862	0.0119	0.0989	0.1865
G-H correlation	0.0988	-0.0932	0.3008	0.0118	0.0989	0.1859
AB-cure AUC	**0.6791**	**0.4985**	**0.8878**	**0.5828**	**0.6858**	**0.7627**
CD-write-off AUC	**0.6886**	**0.5003**	**0.8828**	**0.5875**	**0.6955**	**0.7757**
E-LGD R^2	**0.0617**	**0.0000**	**0.2847**	**0.0019**	**0.0439**	**0.1488**
F-LGD_C R^2	**0.0515**	**0.0000**	**0.2954**	**0.0011**	**0.0319**	**0.1322**
GH-LGD_P R^2	**0.1239**	**0.0000**	**0.4654**	**0.0231**	**0.1121**	**0.2372**
GH-LGD_W R^2	**0.1150**	**0.0000**	**0.4943**	**0.0171**	**0.0981**	**0.2358**

For the explanatory variable and target pairs where the intended correlation is 0, the realised correlation distributions are as expected, with mean and median at 0, but

with some variation due to the randomness in sampling. Similarly, for the correlations between the explanatory variables, the realised intervals align with the specified ones, with the mean and median at the center of the interval and typical variation inside the interval.

4.2 Overall Model Performance

Table 6 shows the overall performance metric statistics for the models, and table 7 presents differences in performance between the models and the corresponding confidence intervals. On average, the Cure-Partial-recovery-Write-off (CPW) model has a higher gAUC and R^2 than the other models, and the differences are statistically significant using a 95% confidence level. The Write-off-Non-write-off (WNW) model performs the second best, with gAUC and R^2 improvements over the Zero-Fractional-One (ZFO) and OLS models also being statistically significant. In terms of R^2, the ZFO model has a statistically significant improvement over OLS, but in terms of gAUC, we do not find a statistically significant difference compared to OLS.

However, it should be noted that the variation in the performance metrics is large between the different data sets, and the relative improvements from one model to the next are low. Especially for gAUC the differences are very low, which is somewhat expected, as the sensitivity of gAUC to small changes in predictions is low due to the binning involved in calculating gAUC.

Table 6: Model gAUC and R^2 statistics over all 5000 simulated data sets.

Model	gAUC mean	gAUC std	gAUC 2.5%	gAUC median	gAUC 97.5%
OLS	0.6423	0.0481	0.5471	0.6432	0.7312
ZFO	0.6423	0.0489	0.5443	0.6429	0.7342
WNW	0.6450	0.0479	0.5513	0.6461	0.7348
CPW	0.6480	0.0476	0.5527	0.6492	0.7385
Model	R^2 mean	R^2 std	R^2 2.5%	R^2 median	R^2 97.5%
OLS	0.1149	0.0700	-0.0013	0.1083	0.2654
ZFO	0.1205	0.0724	0.0015	0.1121	0.2793
WNW	0.1235	0.0719	0.0033	0.1163	0.2789
CPW	0.1289	0.0727	0.0086	0.1218	0.2848

Overall, the results correspond to expectations based on the structure of our simulated data. CPW is the most granular model, and all of the simulated variables serve a purpose within the model components by design, leading to the best performance on average. Similarly, WLW is able to utilise most of the explanatory variables, with C, D, E, G and H aligning directly with the model design. However, for ZFO, none of the explanatory variables are designed to directly explain the probability of zero or one loss, and the induced discriminatory power from the other variables is not enough to justify the more complex model structure compared to OLS on average.

Due to the alignment of the simulated data and the model structures, one should not make strong conclusions about the models in terms of their performance in the general case based on the ranking order shown here. With explanatory variables that support the ZFO model structure, ZFO can perform better than OLS, WNW and CPW,

as shown in section 4.5. This should be kept in mind also when interpreting the results by the shape and end status composition of the LGD distribution in sections 4.3 and 4.4.

Table 7: Mean gAUC and R^2 differences between models over all 5000 simulated data sets.

Comparison	ΔgAUC mean	ΔgAUC mean 95% CI	ΔR^2 mean	ΔR^2 mean 95% CI
CPW-OLS	0.0057	[0.0053, 0.0061]	0.0140	[0.0135, 0.0145]
WNW-OLS	0.0027	[0.0023, 0.0031]	0.0086	[0.0081, 0.009]
ZFO-OLS	-0.0000	[-0.0004, 0.0003]	0.0055	[0.0051, 0.006]
CPW-WNW	0.0030	[0.0027, 0.0033]	0.0054	[0.0051, 0.0058]
CPW-ZFO	0.0057	[0.0054, 0.0061]	0.0085	[0.008, 0.0089]
WNW-ZFO	0.0027	[0.0023, 0.0031]	0.0030	[0.0026, 0.0035]

4.3 Model Performance by LGD Distribution Shape

We analyse the effect of the LGD distribution shape on model performance through LGD mean and variance, and zero rate, one rate and zero-or-one rate. Table 8 contains regression slopes for the effect of the shape statistics on model performance compared to OLS. For illustration, figure 7 shows how mean LGD affects the average shape of the LGD distribution, and figure 8 shows gAUC and R^2 differences compared to OLS for the multi-stage models by mean LGD quantiles and the regression slopes fitted on individual data set performances and mean LGD. The full set of figures for all statistics can be found in appendices C and D.

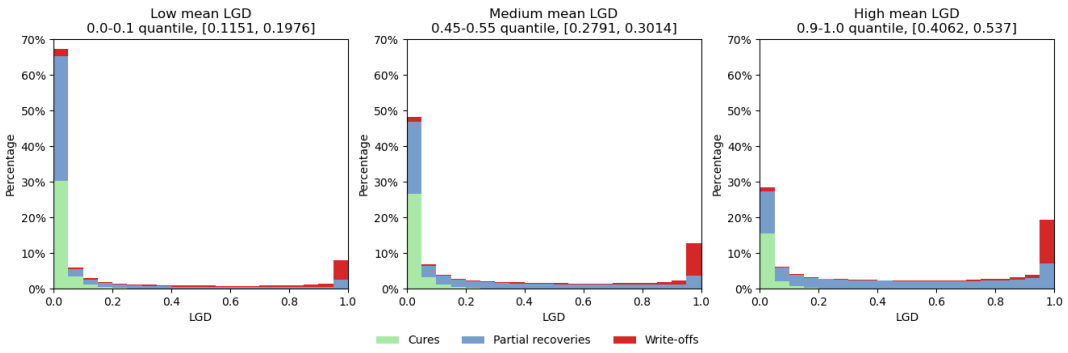


Figure 7: Combined histograms of the LGD data by mean LGD quantiles.

The R^2 difference compared to OLS has statistically significant negative relationships with LGD mean and variance, and statistically significant positive relationships with zero rate and zero-or-one rate for all models. For one rate, the relationship is positive and statistically significant for ZFO and WNW, but for CPW, the relationship is negative but not statistically significant. All of the multi-stage models improve over

Table 8: Regression slopes for gAUC and R^2 differences compared to OLS by statistics that describe the shape of the LGD distribution. Statistically significant slopes are bolded.

Statistic	Model	ΔgAUC slope	ΔgAUC slope p-value	ΔR^2 slope	ΔR^2 slope p-value
LGD mean	ZFO	0.0020	0.4130	**-0.0217**	7.9×10^{-13}
	WNW	-0.0010	0.6963	**-0.0254**	1.4×10^{-15}
	CPW	-0.0036	0.1666	**-0.0276**	2.2×10^{-16}
LGD variance	ZFO	0.0042	0.5570	**-0.0239**	0.0064
	WNW	**0.0211**	0.0030	**-0.0205**	0.0259
	CPW	-0.0009	0.9022	**-0.0474**	1.1×10^{-06}
Zero rate	ZFO	0.0007	0.5799	**0.0111**	5.5×10^{-13}
	WNW	**0.0031**	0.0129	**0.0087**	9.1×10^{-08}
	CPW	0.0015	0.2554	**0.0063**	0.0003
One rate	ZFO	**0.0114**	0.0006	**0.0083**	0.0413
	WNW	**0.0142**	1.8×10^{-05}	**0.0126**	0.0032
	CPW	0.0062	0.0811	-0.0041	0.3617
Zero-or-one rate	ZFO	0.0019	0.0959	**0.0100**	6.3×10^{-13}
	WNW	**0.0042**	0.0002	**0.0085**	5.7×10^{-09}
	CPW	0.0020	0.1051	**0.0046**	0.0029

OLS for the full range of values for all of the statistics, with CPW performing the best, followed by WNW and lastly ZFO.

The common explaining factor for the effect of the shape statistics on R^2 difference compared to OLS is the amount of probability mass in the middle of the LGD distribution, or, conversely, since the distributions are right-skewed, how concentrated the distribution is at zero loss. For highly concentrated distributions, the large mass at zero dominates the fit for OLS, leading to underestimated relationships and low explained variation for the rest of the distribution. The multi-stage models are affected less by the mass concentration, as the ZFO model separates the zero-loss cases by design, and the WNW and CPW models find separation through the end status splits, leading to better performance in relation to OLS.

A high concentration at zero coincides with a low mean LGD, low variance, high zero rate, high zero-or-one rate, and paradoxically, high one rate. Due to the way our data is simulated, when one rate increases, the number of write-off cases increases, and their distribution becomes heavily left-skewed. Simultaneously, the number of partial recovery cases decreases, and their distribution changes from right-skewed to slightly U-shaped for more one-loss cases. As a result, also the number of zero-loss cases increases, and the number of medium losses decreases. This effect is illustrated in figure 9.

This change in the partial recovery distribution also explains why an increasing one rate does not improve CPW R^2 compared to OLS. Specifically, the partial recovery loss distribution starts to resemble the cure loss distribution more closely, so the model loses the benefit of being able to distinguish between cure and partial recovery cases. As a result, the performances of WNW and CPW become more similar as the one rate increases.

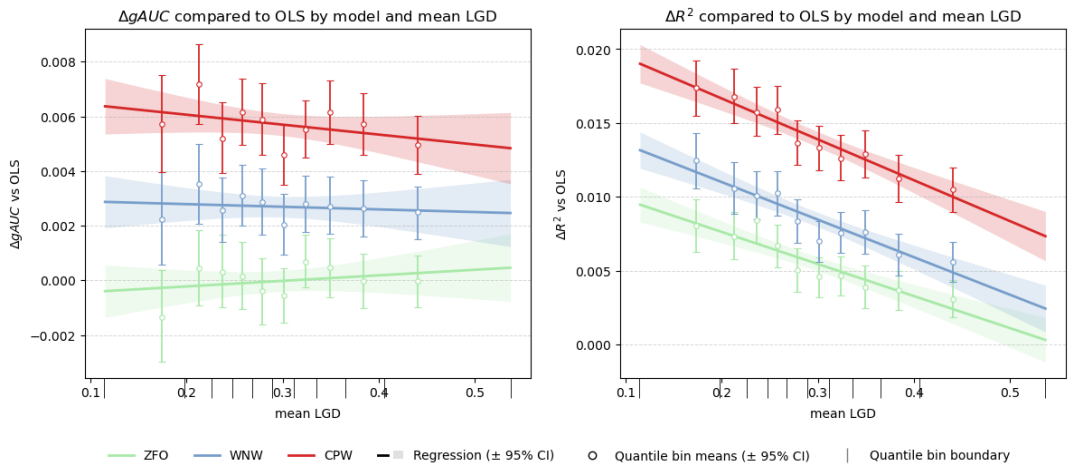


Figure 8: Differences of models' gAUC and R^2 compared to OLS by mean LGD. The lines show regression trends fitted on individual dataset results with a 95% confidence interval. The points show quantile-binned means with a 95% confidence interval. The bin boundaries are indicated by vertical ticks on the x-axis.

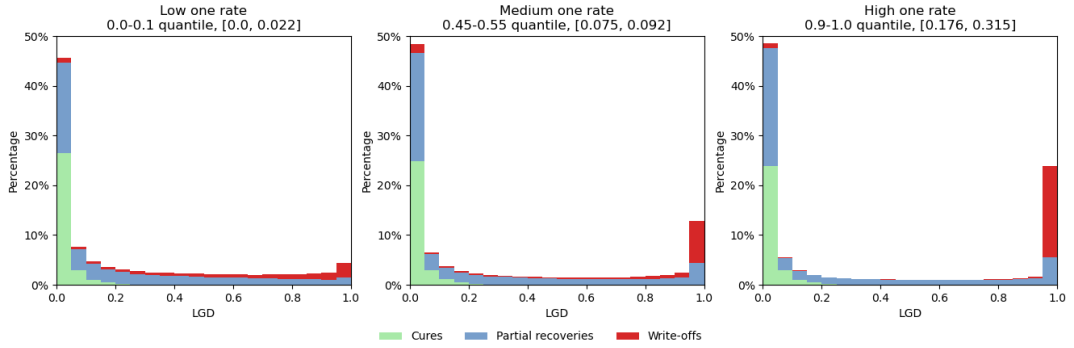


Figure 9: Combined histograms of the LGD data by one rate quantiles.

For gAUC difference compared to OLS, we do not observe a similar significant effect for the distribution shape statistics as for the R^2 difference. This is because small differences in predicted LGD do not affect gAUC due to the binning involved in calculating it, and at the same time additional variation in the predictions does not improve gAUC if binned rank ordering is not affected.

Instead, we observe statistically significant positive relationships for the gAUC difference and LGD variance, zero rate, one rate and zero-or-one rate for WNW, and one-rate for ZFO, that are caused by a greater difference between the loss distribution of write-off cases and others, and the ability of the multi-stage models to separate the different cases. A high variance, zero rate, one rate and zero-or-one rate make the write-off and partial recovery distributions less alike, directly benefitting WNW. Additionally, a higher one rate concentrates the write-off distribution so heavily towards one, that any explanatory variables that are predictive for write-off also become predictive for one loss. This explains the trend for one rate for ZFO.

The trends are, however, not statistically significant for CPW. This is due to the fact that when the partial recovery distribution becomes less similar to the write-off distribution, it becomes more similar to the cure distribution, partially cancelling the effect of improved distinction between partial recovery and write-off cases. Despite this, CPW performs the best out of all of the models in terms of gAUC for the full range of all of the shape statistics, followed by WNW. WNW outperforms OLS for the full range of all of the shape statistics except for one rate, where the improvement is not statistically significant for low one rate values. Unlike in R^2, ZFO does not improve over OLS in gAUC for any shape statistic value.

4.4 Model Performance by End Status Composition

Although cure, partial recovery, or write-off rates do not affect the shape of the end status -specific loss distributions in our setup, they do affect the shape of the combined distribution. Naturally, a high cure rate means a high number of low losses, a high partial recovery rate means a high number of low and medium losses, and a high write-off rate means a high number of medium and high losses. The way we simulate the cure, partial recovery, and write-off rates means that a high cure or write-off rate cause a low partial recovery rate, but the cure rate and write-off rate are not dependent on each other. The effect of cure rate on the distribution is illustrated in figure 10.

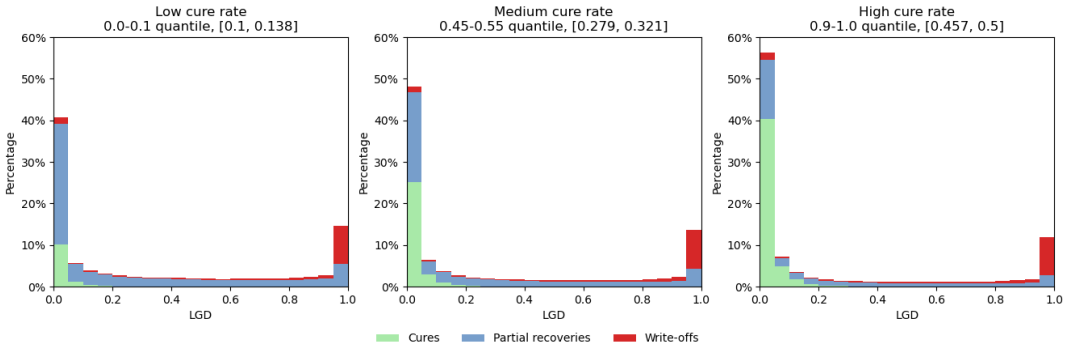


Figure 10: Combined histograms of the LGD data by cure rate quantiles.

Table 9 contains regression slopes for the effect of the end status composition of the LGD data set on model performance compared to OLS. Additionally, figure 8 shows gAUC and R^2 differences compared to OLS for the multi-stage models by cure rate quantiles and the regression slopes fitted on individual data set performances and cure rate.

There are statistically significant positive relationships between R^2 difference to OLS and the cure rate for WNW and CPW, statistically significant negative relationships between the difference and the partial recovery rate for WNW and CPW, and a statistically significant negative relationship and a statistically significant positive relationship between the difference and the write-off rate for ZFO and WNW, respectively. For gAUC difference to OLS, there are statistically significant negative

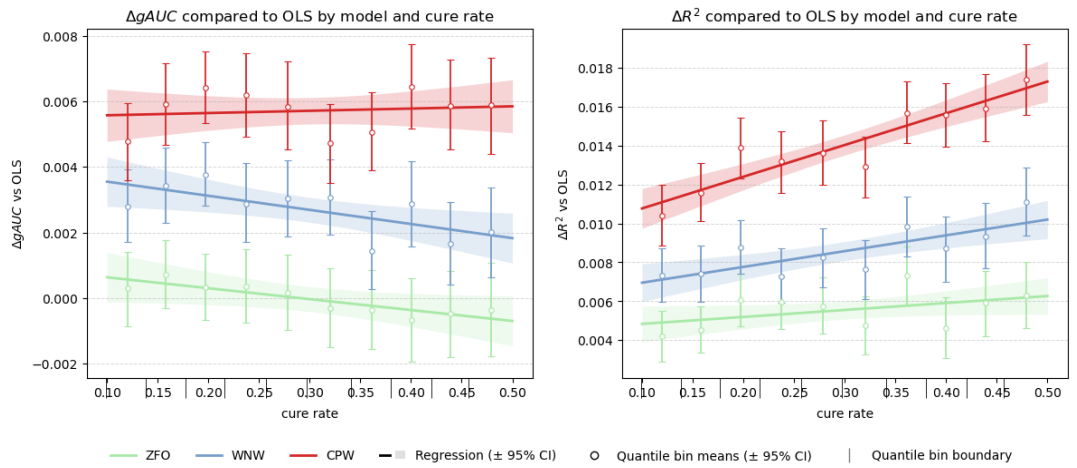


Figure 11: Differences of models' gAUC and R^2 compared to OLS by cure rate. The lines show regression trends fitted on individual dataset results with a 95% confidence interval. The points show quantile-binned means with a 95% confidence interval. The bin boundaries are indicated by vertical ticks on the x-axis.

relationships between the difference and cure rate for ZFO and WNW, and statistically significant positive relationships between the difference and partial recovery rate for ZFO, and write-off rate for WNW.

Table 9: Regression slopes for gAUC and R^2 differences compared to OLS by ends status rates. Statistically significant slopes are bolded.

Statistic	Model	ΔgAUC slope	ΔgAUC slope p-value	ΔR^2 slope	ΔR^2 slope p-value
Cure rate	ZFO	**-0.0033**	0.0463	0.0036	0.0806
	WNW	**-0.0043**	0.0097	**0.0081**	0.0002
	CPW	0.0007	0.6998	**0.0163**	7.3×10^{-13}
Partial recovery rate	ZFO	**0.0040**	0.0118	-0.0003	0.8946
	WNW	0.0027	0.0835	**-0.0086**	2.1×10^{-05}
	CPW	-0.0011	0.5074	**-0.0136**	2.0×10^{-10}
Write-off rate	ZFO	-0.0082	0.0659	**-0.0234**	1.8×10^{-05}
	WNW	**0.0088**	0.0473	**0.0115**	0.0442
	CPW	0.0040	0.3927	-0.0064	0.2943

The gAUC relationships (and lack thereof) are mainly explained by the end status composition within each of the multi-stage model component scopes. Since CPW can distinguish between all three end statuses, we find no statistically significant relationships for gAUC difference for it.

For WNW, a lower cure rate increases the gAUC difference, because the non-write-off component fits better to the stronger relationships of mostly partial recoveries compared to a mix of cures and partial recoveries. Conversely, a higher write-off rate increases the gAUC difference for WNW, because it benefits more from being able to separate the write-off cases from others when there are more of them. Since a higher partial recovery rate lowers, on average, both the cure rate and the write-off rate, we

find no statistically significant relationship between the gAUC difference and partial recovery rate for WNW, although the trend appears to be positive.

For ZFO, the gAUC difference compared to OLS is driven by the relative share of cure, partial recovery and write-off cases within the fractional loss cases. When the partial recovery rate is high, or, conversely, the cure rate and write-off rate is low, the fractional loss OLS component fit is less disturbed by the tails of the cure and write-off distributions, leading to increased performance compared to OLS.

The R^2 relationships, however, are explained both by the end status composition within the model component scopes, as well as the shape of the combined LGD distribution. As described in section 4.3, fewer medium losses and a higher concentration at zero losses improve the R^2 of the multi-stage models compared to OLS. Cure rate and partial recovery rate have a significant impact on the shape of the combined distribution, which explains the relationships for CPW and WNW. However, for ZFO, the combined effect of the shape of the distribution and the end status composition is such that the relationship between R^2 difference and cure rate or partial recovery rate is not statistically significant.

The effect of write-off rate on the combined distribution shape is relatively low, because the medium losses and, to some extent, even the low losses of the partial recovery cases are replaced by write-off losses of similar magnitude, when the write-off rate is increased. The relationships between the R^2 differences and write-off rate are therefore similar and driven by the same factors as for the gAUC differences and write-off rate.

In absolute terms, CPW offers the largest average improvement in gAUC and R^2 over OLS for all cure rate, partial recovery rate and write-off rate values, followed by WNW. ZFO also improves over OLS in terms of R^2 for all end status compositions, but in terms of gAUC, we do not find statistically significant differences to OLS.

4.5 Model Performance by Predictiveness of Explanatory Variables

While the parameters for the correlation structure of the explanatory variables are sampled independently of the LGD distributions, the resulting predictive power of the explanatory variables is not completely independent of the LGD distribution shape. This is because the resulting correlation structure is affected by the capping and flooring of the LGD distributions and the rank joining of the explanatory variables to the targets, as explained in section 4.1. For illustration, figure 12 shows the relationship between E-LGD R^2 and the shape and composition of the LGD distribution.

Table 10 contains regression slopes for the effect of the predictive power of the explanatory variables on model performance compared to OLS. Additionally, figure 13 shows gAUC and R^2 differences compared to OLS for the multi-stage models by AB-cure AUC quantiles and the regression slopes fitted on individual data set performances and AB-cure AUC.

AB-cure AUC has statistically significant positive relationships with gAUC difference compared to OLS for ZFO and CPW, and with R^2 difference compared to OLS

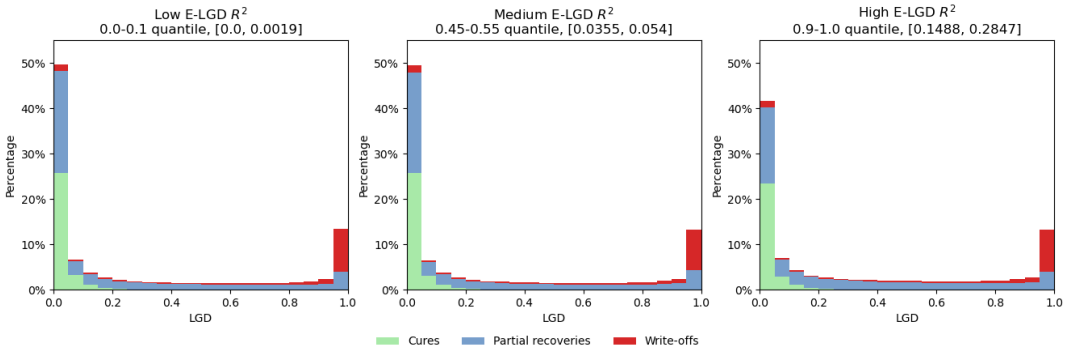


Figure 12: Combined histograms of the LGD data by E-LGD R^2 quantiles.

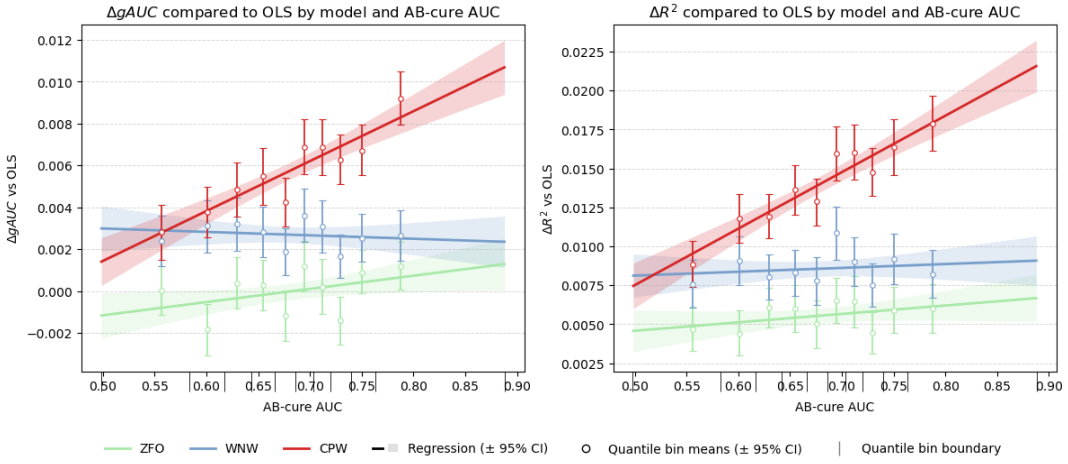


Figure 13: Differences of models' gAUC and R^2 compared to OLS by AB-cure AUC. The lines show regression trends fitted on individual dataset results with a 95% confidence interval. The points show quantile-binned means with a 95% confidence interval. The bin boundaries are indicated by vertical ticks on the x-axis.

for CPW. The relationships for CPW are strong, which can be expected, because a higher AB-cure AUC directly means a better probability of cure component. For ZFO, probability of cure acts as a proxy for the probability of zero loss and one loss, which explains its gAUC relationship and its more gradual slope compared to the slopes for CPW.

CD-write-off AUC has statistically significant positive relationships with gAUC and R^2 difference compared to OLS for all of the models. The relationships are the strongest for WNW, since it relies heavily on the separation of write-off and non-write-off cases. CPW also separates between cured cases and partial recovery cases, so the discriminatory power of the probability of write-off component is not as critical for it. For ZFO, probability of write-off acts as a proxy for the probability of zero loss and one loss, leading to statistically significant positive relationships but the most gradual slopes.

E-LGD R^2 has a statistically significant positive relationship with gAUC difference

Table 10: Regression slopes for gAUC and R^2 differences compared to OLS by explanatory variable predictive power. Statistically significant slopes are bolded.

Metric	Model	ΔgAUC slope	ΔgAUC slope p-value	ΔR^2 slope	ΔR^2 slope p-value
AB-cure AUC	ZFO	**0.0063**	0.0273	0.0054	0.1221
	WNW	-0.0016	0.5653	0.0025	0.5011
	CPW	**0.0238**	2.7×10^{-15}	**0.0362**	7.6×10^{-21}
CD-write-off AUC	ZFO	**0.0061**	0.0244	**0.0292**	1.7×10^{-18}
	WNW	**0.0157**	7.7×10^{-09}	**0.0502**	1.8×10^{-47}
	CPW	**0.0087**	0.0028	**0.0453**	1.1×10^{-34}
E-LGD R^2	ZFO	**0.0160**	2.1×10^{-06}	**0.0543**	5.8×10^{-40}
	WNW	-0.0065	0.0538	**0.0266**	9.4×10^{-10}
	CPW	**-0.0134**	0.0002	**0.0427**	1.4×10^{-20}
F-LGD$_C$ R^2	ZFO	0.0011	0.7675	**-0.0101**	0.0204
	WNW	-0.0033	0.3521	0.0030	0.5152
	CPW	-0.0004	0.9225	0.0014	0.7733
GH-LGD$_P$ R^2	ZFO	0.0002	0.9178	0.0007	0.8065
	WNW	-0.0031	0.1874	**-0.0088**	0.0033
	CPW	-0.0003	0.9045	0.0043	0.1762
GH-LGD$_W$ R^2	ZFO	-0.0038	0.0935	0.0001	0.9840
	WNW	**-0.0055**	0.0142	0.0053	0.0704
	CPW	-0.0038	0.1087	**0.0090**	0.0035

compared to OLS for ZFO, and a statistically significant negative relationship for CPW. For WNW it shows a slightly weaker negative trend, which is not statistically significant. For the R^2 difference compared to OLS, the relationships are positive and statistically significant for all of the multi-stage models.

For CPW and WNW, the gain in gAUC from E-LGD R^2 is smaller than for OLS, because they already discriminate well based on the end status splits. In terms of R^2, CPW and WNW gain more than OLS because of the stronger fits in the more homogenous end status components. The positive relationship of E-LGD R^2 and gAUC difference for ZFO, and the strongest positive relationship for the R^2 difference for ZFO, are due to the fact that in addition to the greater discriminatory and explanatory power within the fractional loss component, a greater E-LGD R^2 also increases the potential of using E to predict the probabilities of zero and one loss.

Interestingly, the end status -specific R^2 values of the explanatory variables have little effect on the relative performance of the models. For F-LGD$_C$ R^2, only the negative relationship with R^2 difference to OLS for ZFO is statistically significant, and it is caused by the slight change in the shape of the LGD distribution of cures, not by the increased R^2 itself. When F-LGD$_C$ R^2 gets higher, the peak at zero LGD shrinks, which means that more cured cases fall into the scope of the fractional loss model component, which in turn complicates the relationships within the component, causing the R^2 difference compared to OLS to decrease.

Similarly, GH-LGD$_P$ has a statistically significant relationship only with the R^2 difference compared to OLS for WNW. An increase in GH-LGD$_P$ R^2 decreases the zero-LGD peak of the partial recovery cases, making the distribution differ more from the cured distribution and increasing the mean LGD of the combined non-write-off

cases slightly closer to that of the write-off cases. This reduces the benefit of the write-off and non-write-off split.

GH-LGD$_W$ R^2 shows negative trends for gAUC difference compared to OLS for all of the models, and positive trends for R^2 difference for WNW and CPW, although only the gAUC trend for WNW and the R^2 trend for CPW are statistically significant. Similarly to how an increase in F-LGD$_C$ R^2 and an increase in GH-LGD$_P$ R^2 decrease the zero-LGD peak of the cure and partial recovery LGD distributions, respectively, an increase in GH-LGD$_W$ R^2 decreases the one-LGD peak of the write-off LGD distribution and makes it flatter. This makes the write-off distribution more similar to the partial recovery distribution, decreasing the possible discriminatory power gains from separating the write-off cases from others, and causing the negative gAUC relationships for CPW and WNW. At the same time, the correlation between write-off and one-LGD decreases, which causes the negative gAUC relationship for ZFO. However, CPW and WNW are able to fully utilise the increase in GH-LGD$_W$ R^2 in their write-off component, which explains the positive relationships between GH-LGD$_W$ R^2 and their R^2 difference compared to OLS.

In general, the predictive power of the explanatory variable does not affect the ranking order of the models in terms of gAUC or R^2 performance within the studied range of values. In almost all cases, CPW improves the most over OLS in terms of gAUC and R^2, followed by WNW. ZFO improves the least over OLS in terms of R^2, and in terms of gAUC there is no statistically significant difference between the performance of ZFO and OLS.

However, for very low AB-cure AUC values, the gAUC and R^2 improvements of CPW fall to the same level and even below those of WNW. Similarly, for very low CD-write-off AUC values, the gAUC and R^2 improvements of WNW over OLS and ZFO fall to zero, and conversely, high E-LGD R^2 values make the gAUC improvement of ZFO over OLS statistically significant and even higher than for WNW and CPW, and the R^2 improvements higher than for WNW. This indicates that the better performance of the multi-stage models is dependent on the presence of variables, which are predictive for the splits into the different components, while the presence of variables which offer additional predictive power within the components is not as important.

4.6 Shape of Predicted LGD Distributions

Figure 14 illustrates the predicted LGD distributions of the models for all simulated data sets combined, and appendix E contains predicted LGD distribution plots by summary statistic quantiles. All of the multi-stage models are somewhat more sensitive to the size of the zero-LGD peak, and are able to produce slightly more right-skewed predicted distributions than OLS. However, overall the predicted distributions are very similar and close to normal for all models, and the summary statistics and explanatory variable predictiveness metrics have little effect on the difference between the shapes of the predicted LGD distributions of the models. Even with highly predictive variables for the component splits, the multi-stage models are unable to produce the characteristic bimodal shape of the LGD distribution from normal explanatory variables.

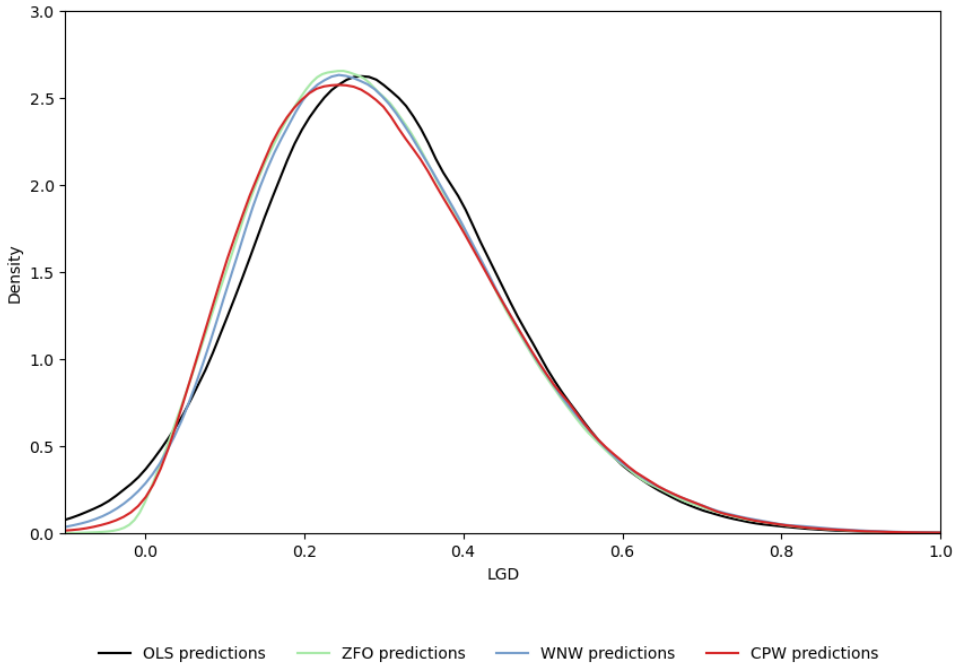


Figure 14: Predicted LGD distributions of all models for all data sets combined.

5 Conclusions

In this thesis, we have devised an approach to simulating LGD data sets that support various types of multi-stage models by combining several existing simulation approaches found in the LGD literature - to generate an LGD data set, we draw realisations from three distinct beta distributions based on three possible end status outcomes of a default: cure, partial recovery, and write-off. Using Gaussian copulas, we then generate normally distributed explanatory variables for the probability of cure, probability of write-off and LGD, with set correlation structures within the cure, partial recovery and write-off cases.

Using this simulation approach, we generated 5000 data sets with varying properties. For the generated data, we studied the relationships between the performance differences of three multi-stage models compared to a single-stage OLS model and the shape of the LGD distribution, the proportion of cure, partial recovery and write-off cases, and the predictiveness of the explanatory variables. We then compared the relative performance of the models for LGD data of different types in these dimensions.

We found that the most critical factor that makes multi-stage models perform better than single-stage OLS is the presence of variables which are predictive for the specific component splits of the model, while the within-component predictive power is not as important. If there are no such variables available, a simpler multi-stage model or OLS will perform similarly and will be preferred due to its simplicity.

In terms of the shape of the LGD distribution, high zero-LGD peaks and a light mass in the center of the distribution, and a larger difference between the shapes of the LGD distributions of the components they separate improve the performance of multi-stage models compared to OLS. Additionally, we found that the end status composition of the LGD distribution affects the performance of the models through the shape of the LGD distribution as well as through the homogeneity of the cases within the components. If the end status composition of the data is such that a multi-stage model component consists mainly of cases of one end status, the model performs better. At the same time, there must be sufficiently many cases in each component to justify each component split.

In terms of the shape of the predicted LGD distribution, we found little difference between the models. We also found that the shape and end status composition of the LGD data and the explanatory variable predictiveness had little effect on the difference between the shapes of the predicted distributions of the different models. While the multi-stage models were able to produce slightly more right-skewed distributions, the studied multi-stage model structures did not transform the predicted distribution into the characteristic bimodal LGD shape.

Despite the model performance relationships found, the average performance improvements between the models were found to be small compared to the variation in performance between the data sets. Thus, a thorough performance analysis of candidate model structures for the specific data set in use remains crucial for the final choice of model structure. Nevertheless, this thesis fills a gap in the LGD literature by formally analysing the relationships between model performance and the nature of the data, and gives LGD modellers a starting point and direction for the choice of model

structure based on the data they have.

The limitations of this thesis and its results relate to data and models alike. To get more comparable results for the different model structures, we only used normally distributed explanatory variables instead of allowing different distributions to be available for some model components and not for others. However, a first extension to this thesis would be to study how the distributions shapes of the explanatory variables affect the performance of the different model structures.

A further improvement to strengthen the validity of the results would be to use real data instead of simulated data. LGD data sets are not widely available, so using simulated data allowed us to perform the analysis in the first place. However, the simulated data does not contain the complex relationships and noise that a real data set would, so a similar analysis to this thesis could be performed with real data to confirm that the relationships that were established are not just a product of the simulation approach and also to possibly find new relationships that are not present in the simple simulated data.

Another direction of further research would be to extend the model comparison to include nonlinear techniques in single-stage models and as components of the multi-stage models. This would be especially important when using real-world data with more nuanced nonlinear relationships.

References

Bank for International Settlements. History of the Basel committee, 2018. URL https://www.bis.org/bcbs/history.htm. [Accessed 29-May-2025].

Basel Committee on Banking Supervision. The Basel framework, 2019. URL https://www.bis.org/baselframework/BaselFramework.pdf. [Accessed 5-January-2025].

Tony Bellotti and Jonathan Crook. Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting*, 28(1): 171–182, 2012.

Tomasz R Bielecki and Marek Rutkowski. *Credit risk: Modeling, valuation and hedging*. Springer Science & Business Media, 2013.

Robert L Burns. Economic capital and the assessment of capital adequacy. *Supervisory Insights*, 2(2):5–11, 2004.

Michael Bücker, Gero Szepannek, Alicja Gosiewska, and Przemyslaw Biecek. Transparency, auditability and explainability of machine learning models in credit scoring, 2020. URL https://arxiv.org/abs/2009.13384. [Accessed 16-June-2025].

Darrell Duffie and Kenneth J Singleton. *Credit risk: Pricing, measurement, and management*. Princeton University Press, 2003.

European Central Bank. Instructions for reporting the validation results of internal models, 2019. URL https://www.bankingsupervision.europa.eu/activities/internal_models/shared/pdf/instructions_validation_reporting_credit_risk.en.pdf. [Accessed 30-July-2025].

European Central Bank. ECB guide to internal models, 2024. URL https://www.bankingsupervision.europa.eu/ecb/pub/pdf/ssm.supervisory_guides202402_internalmodels.en.pdf?2eae47049232d667084335755f1bdef. [Accessed 31-May-2025].

European Investment Bank. Default and recovery statistics: Private and public lending 1994 2023, 2024. URL https://data.europa.eu/doi/10.2867/6241848. [Accessed 18-June-2025].

Marc Gürtler and Martin Hibbeln. Pitfalls in modeling loss given default of bank loans. *Technical Report, Working Paper*, 2011.

Martin Haugh. An introduction to copulas, 2016. URL https://www.columbia.edu/~mh2078/QRM/Copulas.pdf. [Accessed 23-July-2025].

Stefan Hlawatsch and Sebastian Ostrowski. Simulation and estimation of loss given default. *The Journal of Credit Risk*, 7(3):39, 2011.

David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. John Wiley & Sons, 2013.

R John Irwin and Timothy C Irwin. Appraising credit ratings: Does the CAP fit better than the ROC? *International Journal of Finance & Economics*, 18(4):396–408, 2013.

Phillip Li, Min Qi, Xiaofei Zhang, and Xinlei Zhao. Further investigation of parametric loss given default modeling. *Journal of Credit Risk*, 12(4):17–47, 2016.

Phillip Li, Xiaofei Zhang, and Xinlei Shelly Zhao. Modeling loss given default. *FDIC Center for Financial Research Paper*, (2018-03), 2018.

G. Loterman, Iain Brown, David Martens, C. Mues, and Bart Baesens. Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting*, 28, 2012.

Alexander J McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative risk management: Concepts, techniques and tools-revised edition*. Princeton University Press, 2015.

Francis Sahngun Nahm. Receiver operating characteristic curve: Overview and practical use for clinicians. *Korean Journal of Anesthesiology*, 75(1):25–36, 2022.

Roger Newson. Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences. *The Stata Journal*, 2(1):45–64, 2002.

Raydonal Ospina and Silvia LP Ferrari. Inflated beta distributions. *Statistical Papers*, 51:111–126, 2010.

Aida Salko and Rita D'Ecclesia. Decomposing loss given default: A closer look at recovery patterns. 2022. URL https://www.dss.uniroma1.it/en/system/files/pubblicazioni/Salko%2C%20D%27Ecclesia%202%20Decomposing%20Loss%20Given%20Default.pdf. [Accessed 30-May-2025].

Til Schuermann. What do we know about loss given default? *Wharton Financial Institutions Center Working Paper No. 04-01*, 2004.

Wojciech Starosta. Loss given default decomposition using mixture distributions of in-default events. *European Journal of Operational Research*, 292(3):1187–1199, 2021.

Yuta Tanoue, Akihiro Kawada, and Satoshi Yamashita. Forecasting loss given default of bank loans with multi-stage model. *International Journal of Forecasting*, 33(2): 513–522, 2017.

Xin Yan and Xiaogang Su. *Linear regression analysis: Theory and computing*. World Scientific, 2009.

A Derivation of Variance Intervals

The mean μ and variance σ^2 of the beta distribution are

$$\mu = \frac{\alpha}{\alpha + \beta} \tag{A1}$$

and

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}, \tag{A2}$$

respectively. By rearranging A1 for β and α, we get

$$\beta = \frac{\alpha}{\mu} - \alpha \tag{A3}$$

and

$$\alpha = \frac{\beta\mu}{(1-\mu)}. \tag{A4}$$

By substituting A3 into A2, the variance in terms of α and μ is

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{A5}$$

$$= \frac{\alpha(\frac{\alpha}{\mu} - \alpha)}{(\alpha + (\frac{\alpha}{\mu} - \alpha))^2(\alpha + (\frac{\alpha}{\mu} - \alpha) + 1)} \tag{A6}$$

$$= \frac{\alpha^2(\frac{1}{\mu} - 1)}{\frac{\alpha^2}{\mu^2}(\frac{\alpha}{\mu} + 1)} = \frac{\mu^2(\frac{1}{\mu} - 1)}{\frac{\alpha}{\mu} + 1} = \frac{\mu^3(\frac{1}{\mu} - 1)}{\alpha + \mu} \tag{A7}$$

$$= \frac{\mu^2(1-\mu)}{\alpha+\mu} \tag{A8}$$

and by substituting A4 into A8, we can express it in terms of β and μ as

$$\sigma^2 = \frac{\mu^2(1-\mu)}{\alpha+\mu} = \frac{\mu^2(1-\mu)}{\frac{\beta\mu}{(1-\mu)} + \mu} = \frac{\mu(1-\mu)}{\frac{\beta}{(1-\mu)} + 1} \tag{A9}$$

$$= \frac{\mu(1-\mu)^2}{\beta+1-\mu}. \tag{A10}$$

Now, for cures, where $\alpha_C > 0$, $\beta_C \geq 1$ and $0 < \mu_C < 1$, we get an upper bound for the variance by using

$$\alpha_C = \frac{\beta_C\mu_C}{(1-\mu_C)} \geq \frac{\mu_C}{(1-\mu_C)} > 0 \tag{A11}$$

$$\Rightarrow \sigma_C^2 = \frac{\mu_C^2(1-\mu_C)}{\alpha_C + \mu_C} \tag{A12}$$

$$\leq \frac{\mu_C^2(1-\mu_C)}{\frac{\mu_C}{(1-\mu_C)} + \mu_C} \tag{A13}$$

$$= \frac{\mu_C(1-\mu_C)^2}{1+1-\mu_C} = \frac{\mu_C(1-\mu_C)^2}{2-\mu_C}. \tag{A14}$$

Since $\alpha_C \leq 1$, we get the lower bound

$$\sigma_C^2 = \frac{\mu_C^2(1-\mu_C)}{\alpha_C + \mu_C} \geq \frac{\mu_C^2(1-\mu_C)}{1+\mu_C}. \tag{A15}$$

For partial recoveries, with $0 < \alpha_P \leq 1$, we get the upper bound

$$\sigma_P^2 = \frac{\mu_P^2(1-\mu_P)}{\alpha_P + \mu_P} \tag{A16}$$

$$< \frac{\mu_P^2(1-\mu_P)}{\mu_P} = \mu_P(1-\mu_P) \tag{A17}$$

and the lower bound

$$\sigma_P^2 = \frac{\mu_P^2(1-\mu_P)}{\alpha_P + \mu_P} \geq \frac{\mu_P^2(1-\mu_P)}{1+\mu_P}. \tag{A18}$$

Similarly for write-offs, where $0 < \beta_W \leq 1$, we get the upper bound

$$\sigma_W^2 = \frac{\mu_W(1-\mu_W)^2}{\beta_W + 1 - \mu_W} < \frac{\mu_W(1-\mu_W)^2}{1-\mu_W} = \mu_W(1-\mu_W) \tag{A19}$$

and the lower bound

$$\sigma_W^2 = \frac{\mu_W(1-\mu_W)^2}{\beta_W + 1 - \mu_W} \geq \frac{\mu_W(1-\mu_W)^2}{2-\mu_W}. \tag{A20}$$

B Solving Beta Distribution Parameters

From A1-A10 we have

$$\sigma^2 = \frac{\mu^2(1-\mu)}{\alpha+\mu} \tag{B1}$$

$$\Rightarrow \alpha\sigma^2 + \mu\sigma^2 = \mu^2(1-\mu) \tag{B2}$$

$$\Rightarrow \alpha = \frac{\mu^2(1-\mu) - \mu\sigma^2}{\sigma^2} = \mu(\frac{\mu(1-\mu)}{\sigma^2} - 1) \tag{B3}$$

and

$$\sigma^2 = \frac{\mu(1-\mu)^2}{\beta+1-\mu} \tag{B4}$$

$$\Rightarrow \beta\sigma^2 + \sigma^2 - \mu\sigma^2 = \mu(1-\mu)^2 \tag{B5}$$

$$\Rightarrow \beta = \frac{\mu^(1-\mu)^2 - \sigma^2 + \mu\sigma^2}{\sigma^2} = (1-\mu)(\frac{\mu(1-\mu)}{\sigma^2} - 1). \tag{B6}$$

C Simulated LGD Distributions by Summary Statistic Quantiles

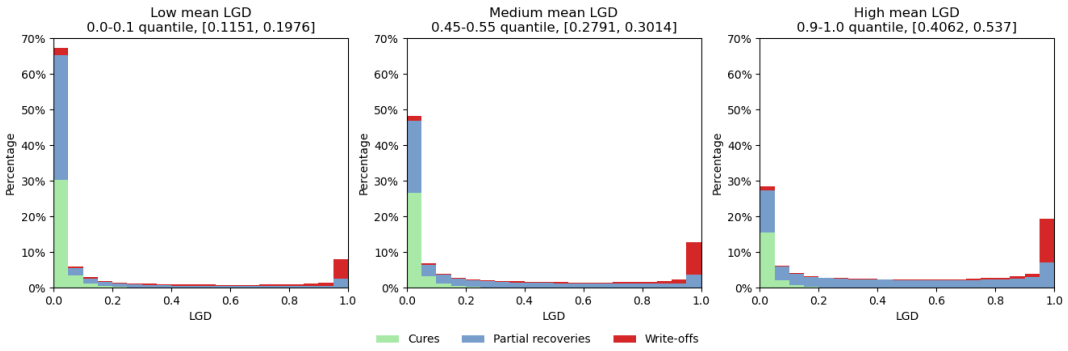


Figure C1: Combined histograms of the LGD data by mean LGD quantiles.

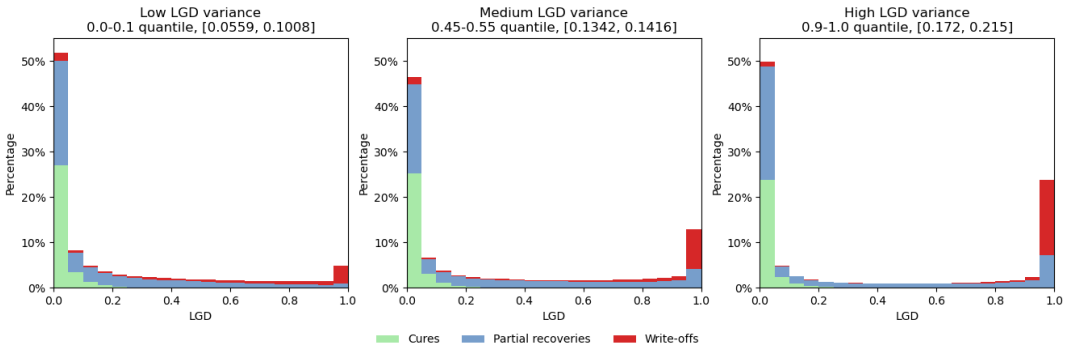


Figure C2: Combined histograms of the LGD data by LGD variance quantiles.

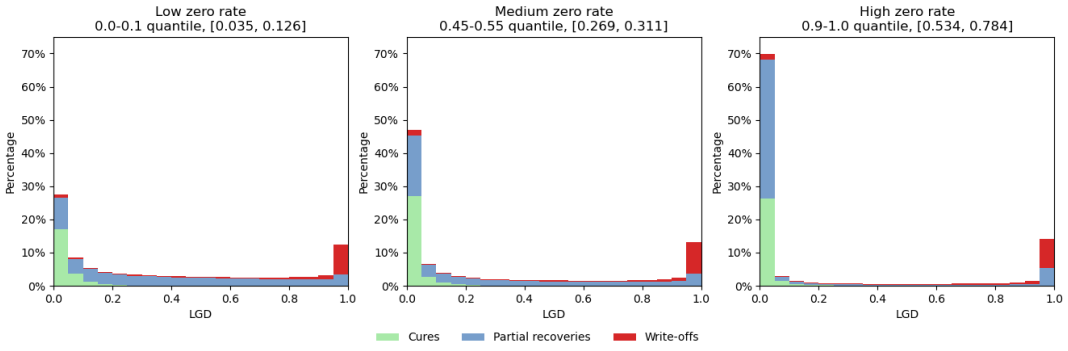


Figure C3: Combined histograms of the LGD data by zero rate quantiles.

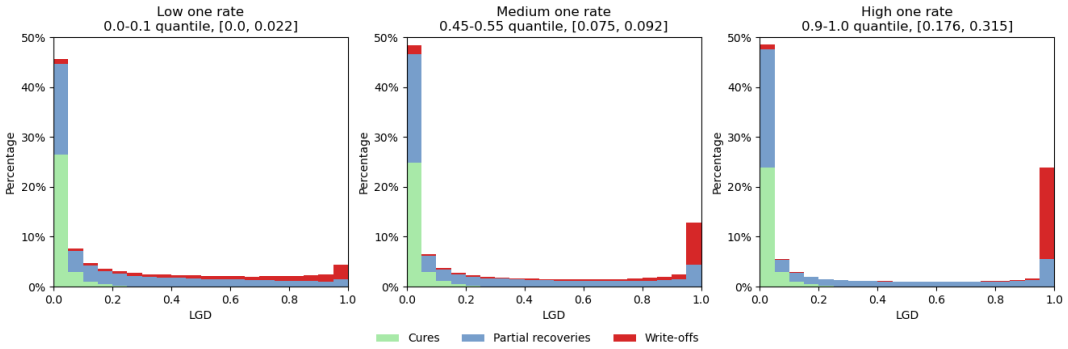


Figure C4: Combined histograms of the LGD data by one rate quantiles.

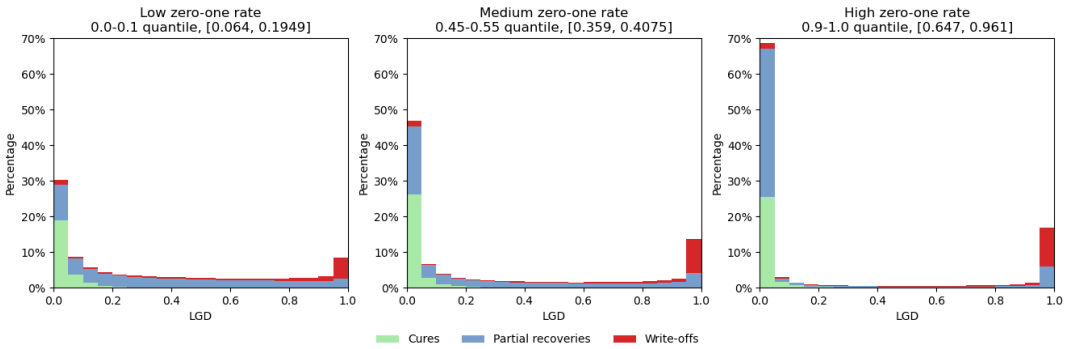


Figure C5: Combined histograms of the LGD data by zero-one rate quantiles.

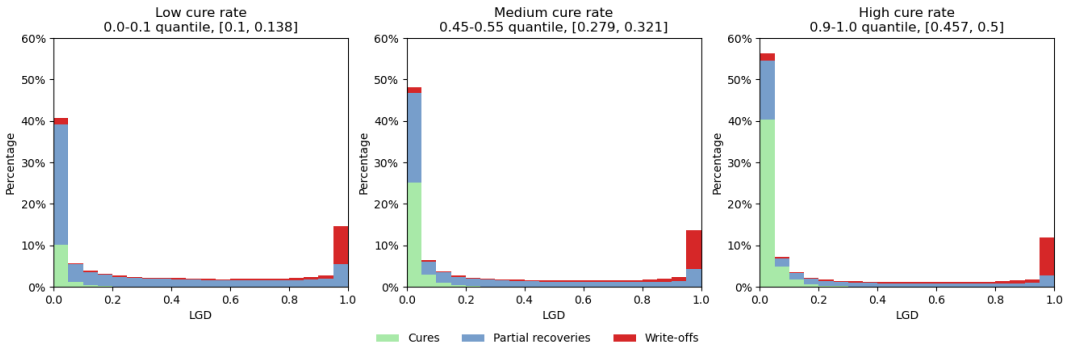


Figure C6: Combined histograms of the LGD data by cure rate quantiles.

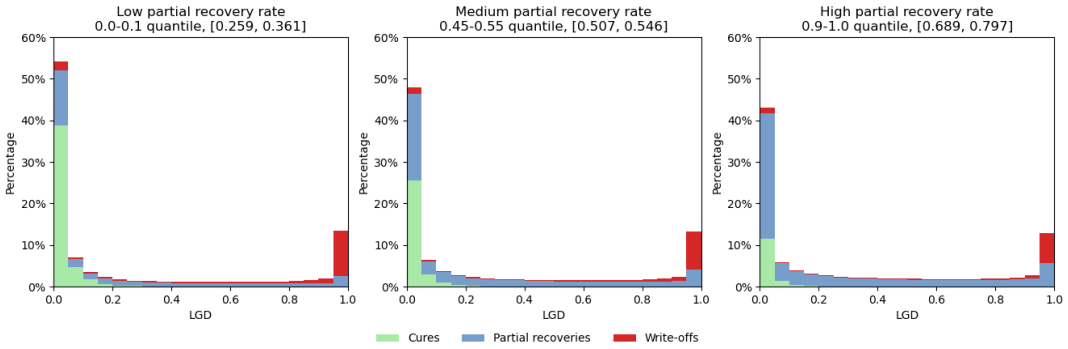


Figure C7: Combined histograms of the LGD data by partial recovery rate quantiles.

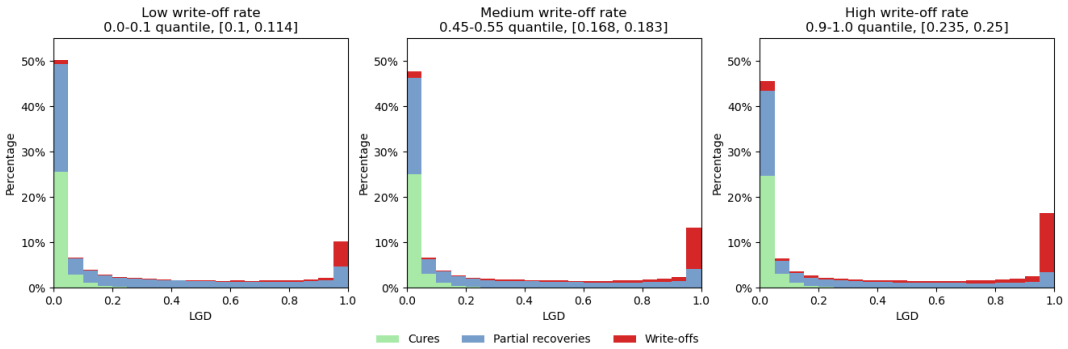


Figure C8: Combined histograms of the LGD data by write-off rate quantiles.

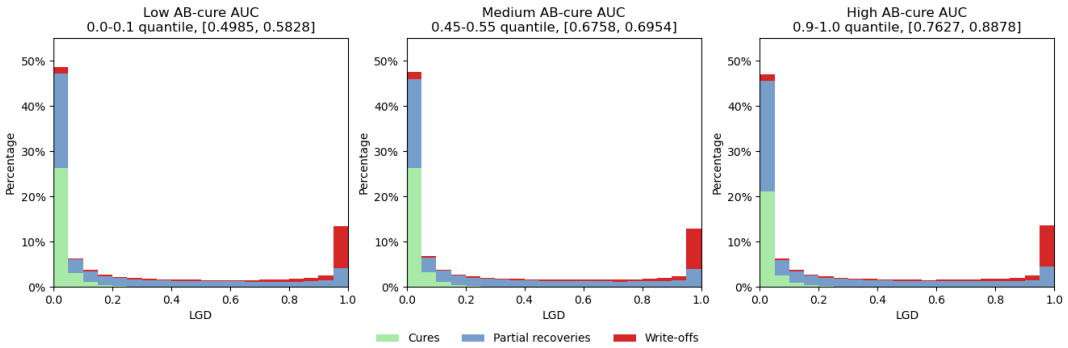


Figure C9: Combined histograms of the LGD data by AB-cure AUC quantiles.

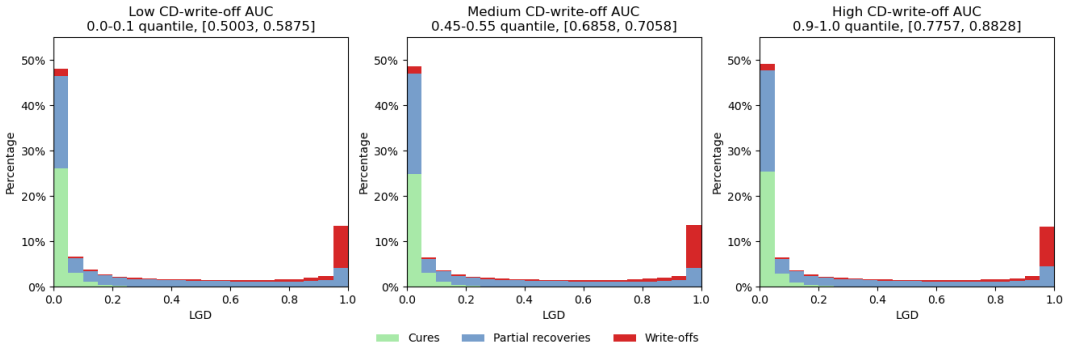


Figure C10: Combined histograms of the LGD data by CD-write-off AUC quantiles.

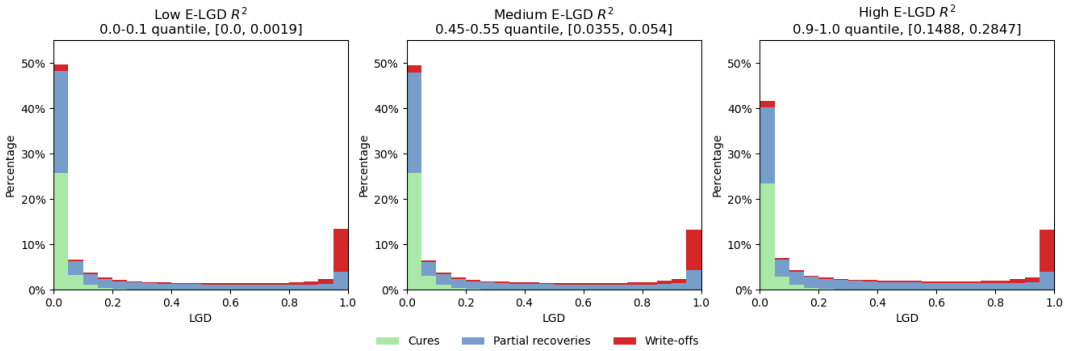


Figure C11: Combined histograms of the LGD data by E-LGD R^2 quantiles.

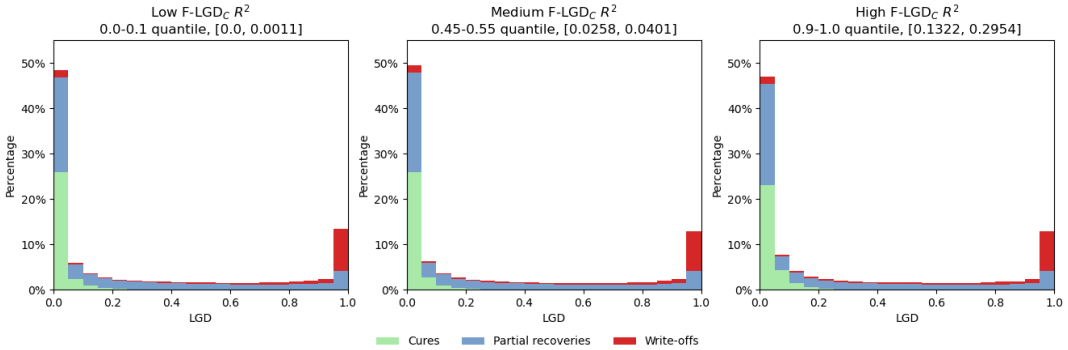


Figure C12: Combined histograms of the LGD data by F-LGD$_C$ R^2 quantiles.

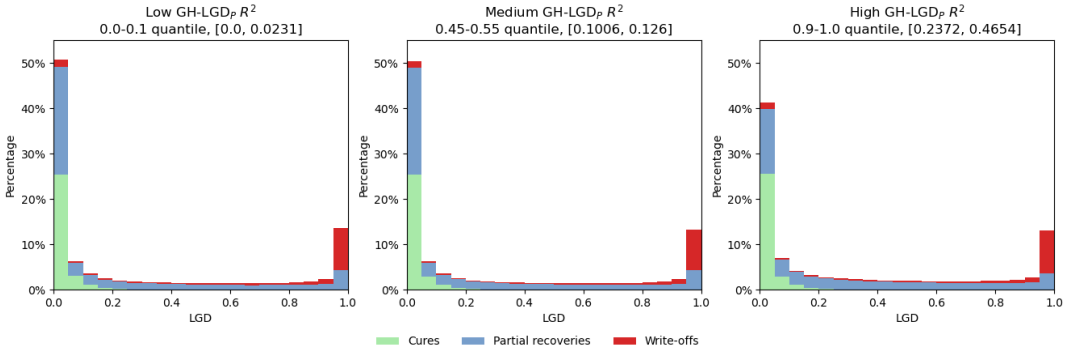


Figure C13: Combined histograms of the LGD data by GH-LGD$_P$ R^2 quantiles.

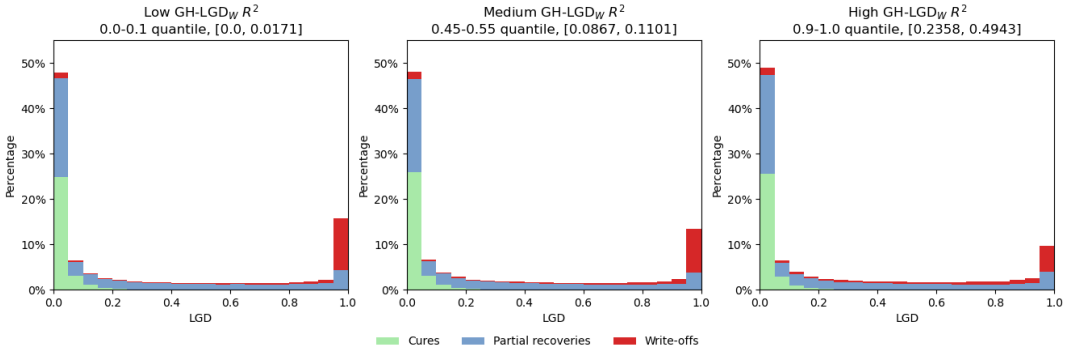
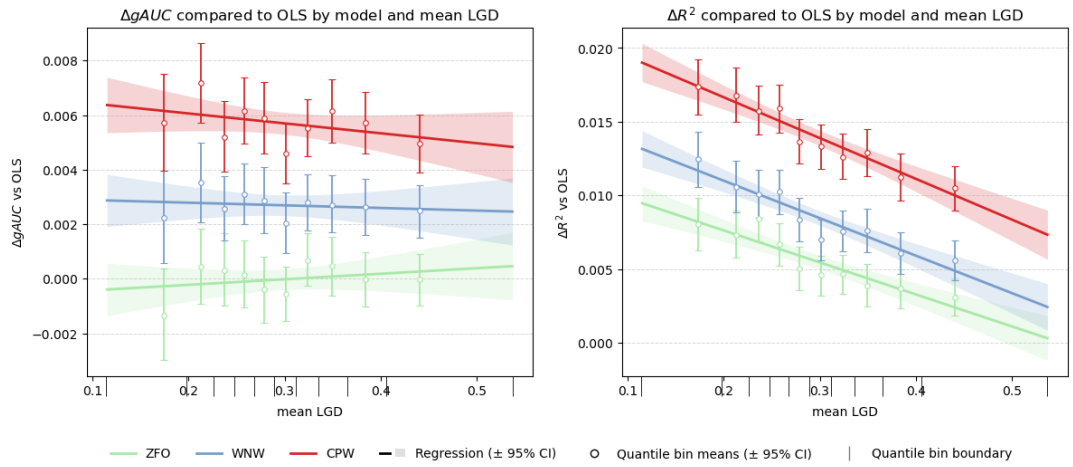
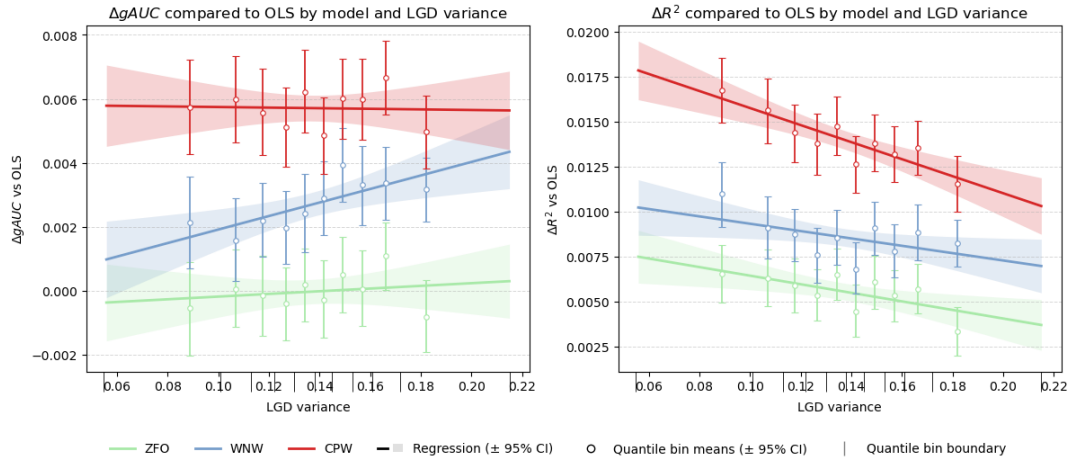


Figure C14: Combined histograms of the LGD data by GH-LGD$_W$ R^2 quantiles.

D Model Performance Figures

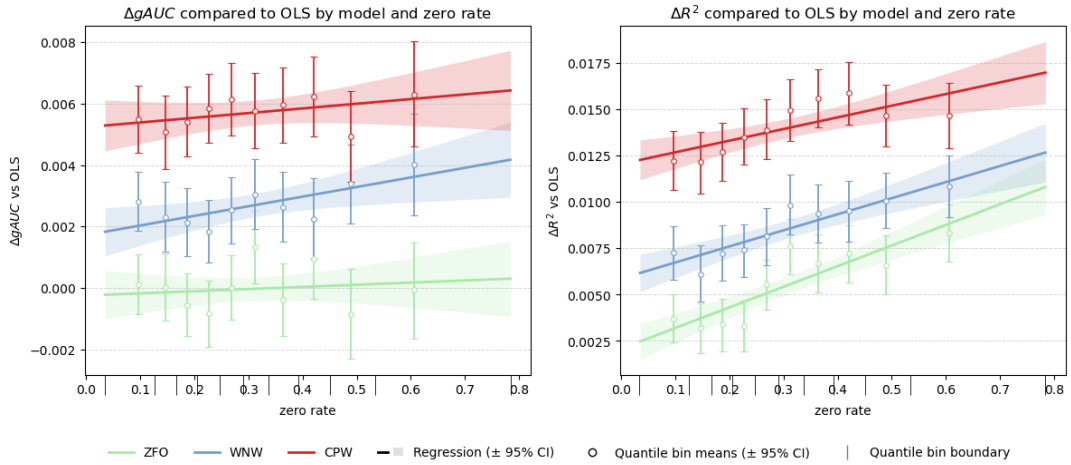


(a) Differences of models' gAUC and R^2 compared to OLS by mean LGD.

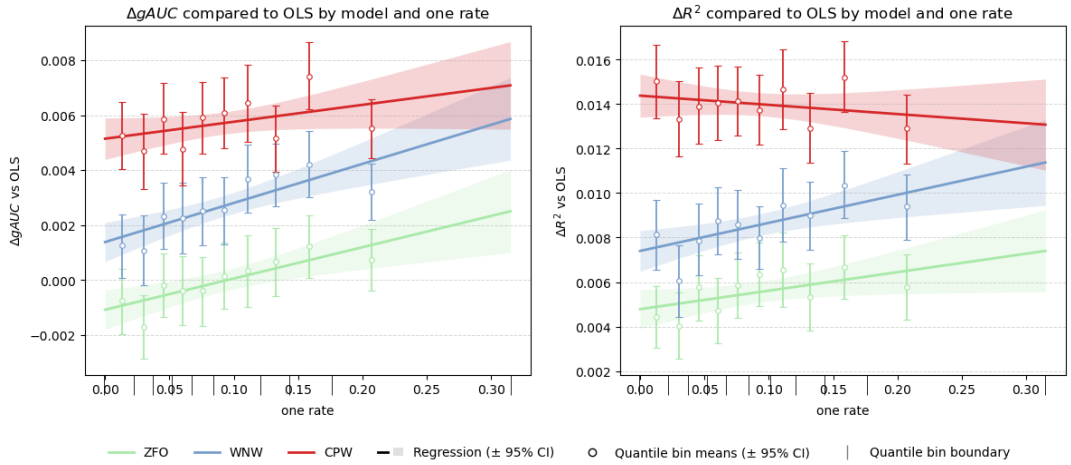


(b) Differences of models' gAUC and R^2 compared to OLS by LGD variance.

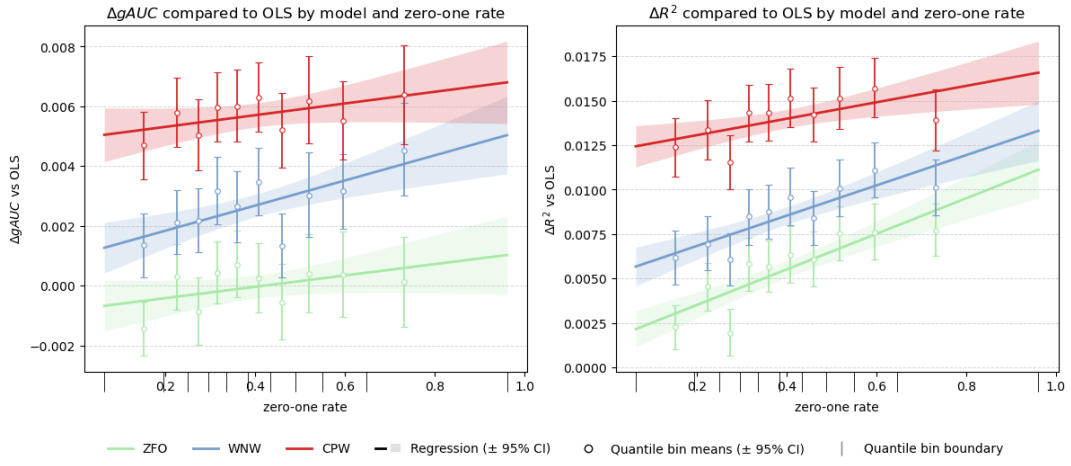
Figure D1: Differences of models' gAUC and R^2 compared to OLS by LGD mean and variance. The lines show regression trends fitted on individual dataset results with a 95% confidence interval. The points show quantile-binned means with a 95% confidence interval. The bin boundaries are indicated by vertical ticks on the x-axis.



(a) Differences of models' gAUC and R^2 compared to OLS by zero rate.

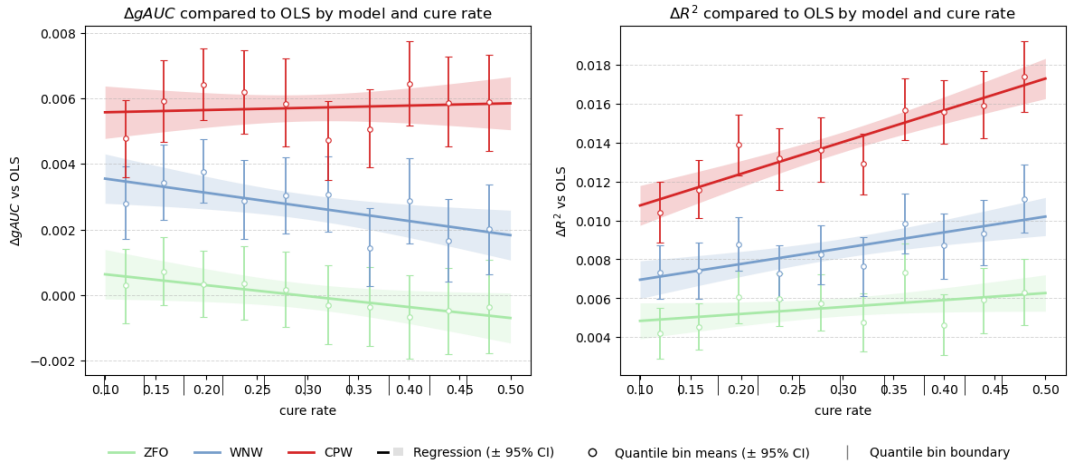


(b) Differences of models' gAUC and R^2 compared to OLS by one rate.

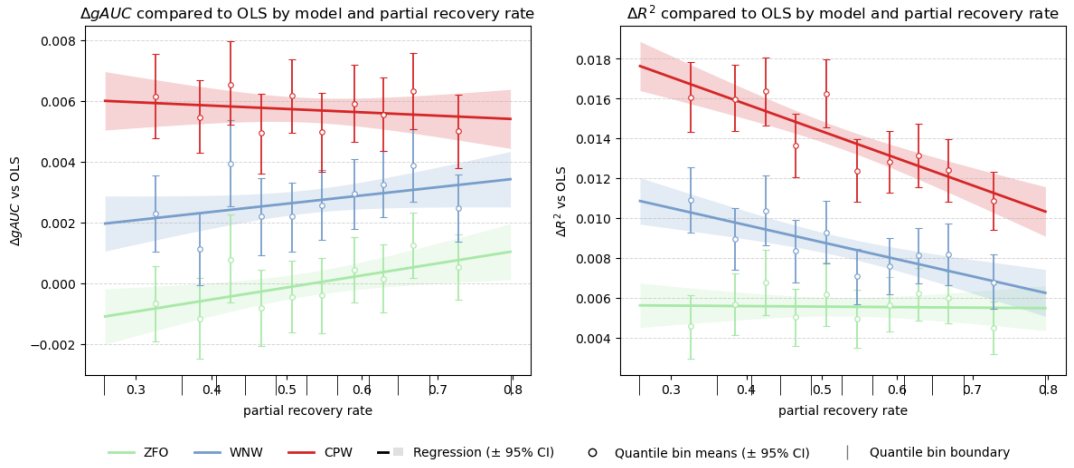


(c) Differences of models' gAUC and R^2 compared to OLS by zero-one rate.

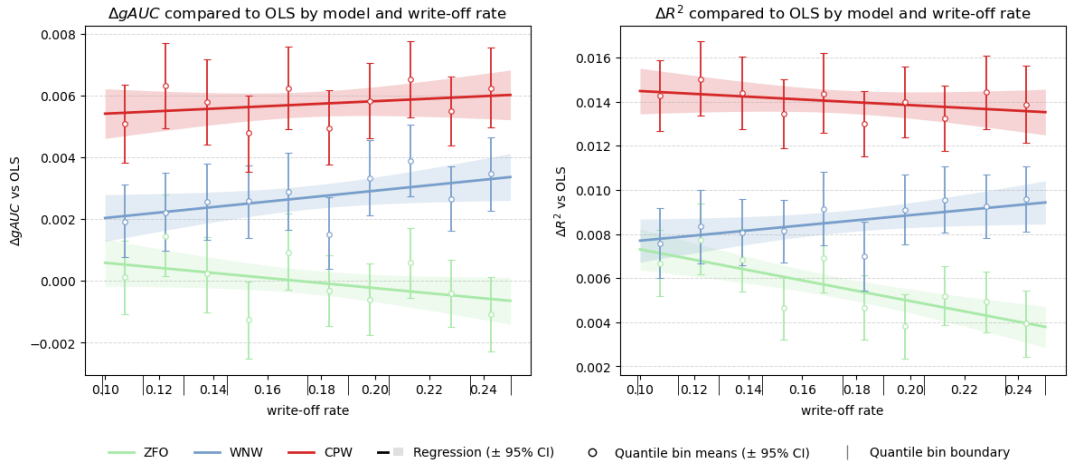
Figure D2: Differences of models' gAUC and R^2 compared to OLS by zero, one and zero-one rates. The lines show regression trends fitted on individual dataset results with a 95% confidence interval. The points show quantile-binned means with a 95% confidence interval. The bin boundaries are indicated by vertical ticks on the x-axis.



(a) Differences of models' gAUC and R^2 compared to OLS by cure rate.

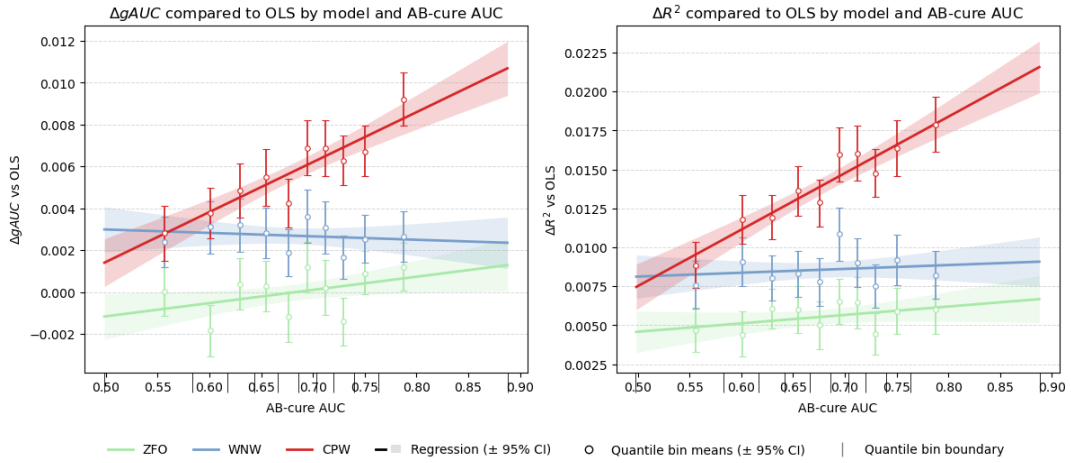


(b) Differences of models' gAUC and R^2 compared to OLS by partial recovery rate.

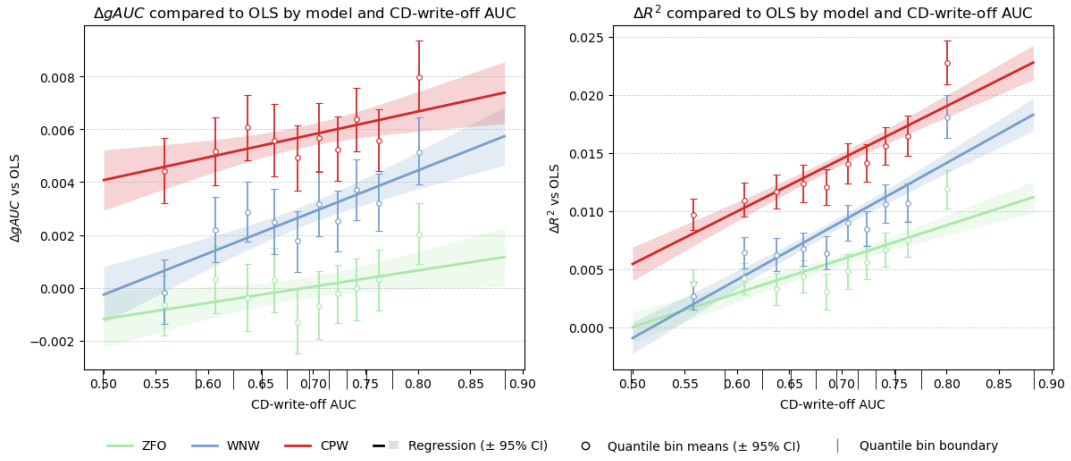


(c) Differences of models' gAUC and R^2 compared to OLS by write-off rate.

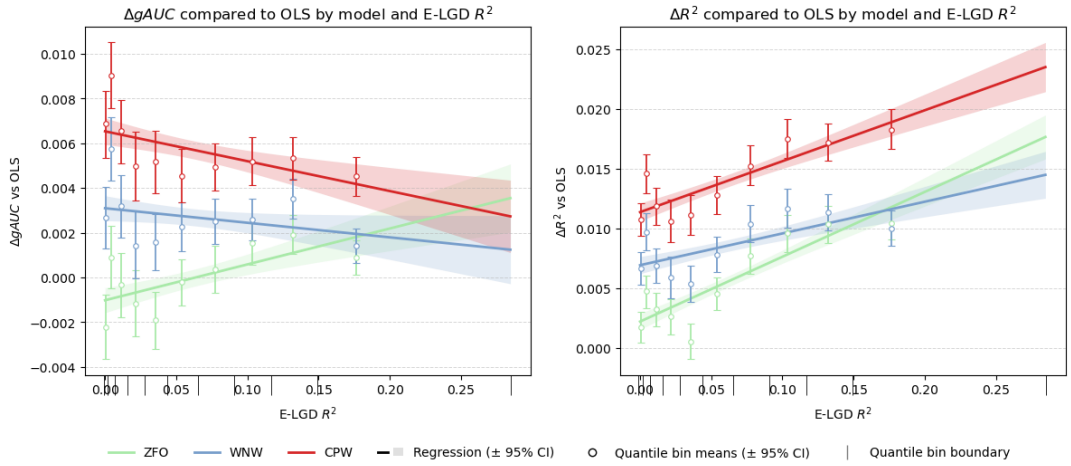
Figure D3: Differences of models' gAUC and R^2 compared to OLS by end status rates. The lines show regression trends fitted on individual dataset results with a 95% confidence interval. The points show quantile-binned means with a 95% confidence interval. The bin boundaries are indicated by vertical ticks on the x-axis.



(a) Differences of models' gAUC and R^2 compared to OLS by AB-cure AUC.

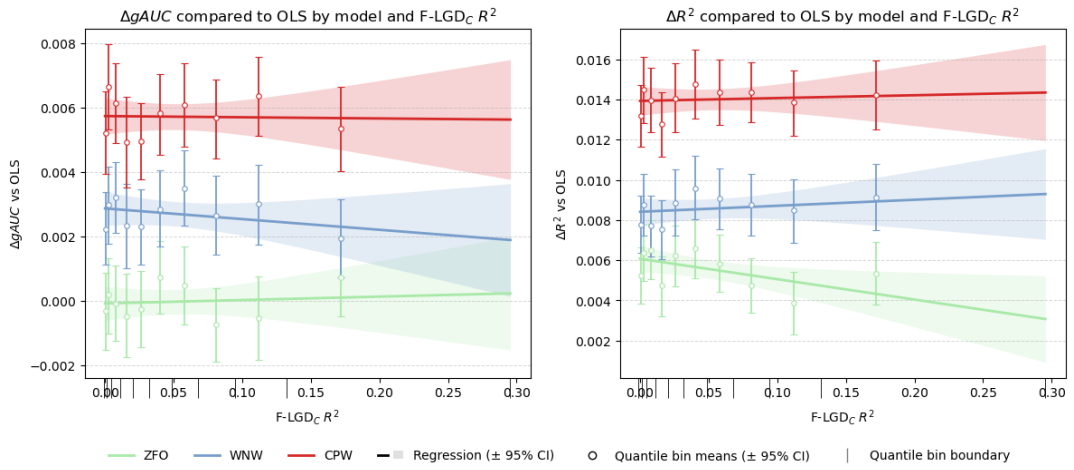


(b) Differences of models' gAUC and R^2 compared to OLS by CD-write-off AUC.

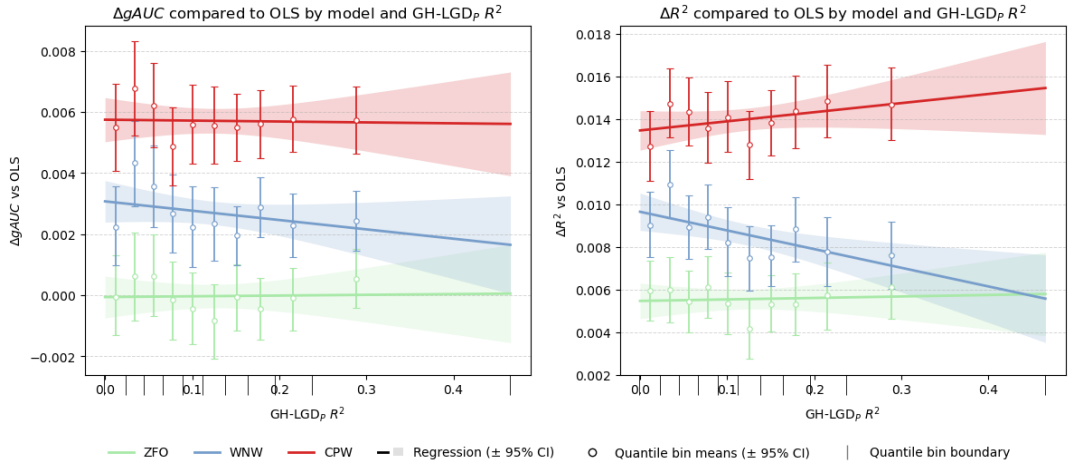


(c) Differences of models' gAUC and R^2 compared to OLS by E-LGD R^2.

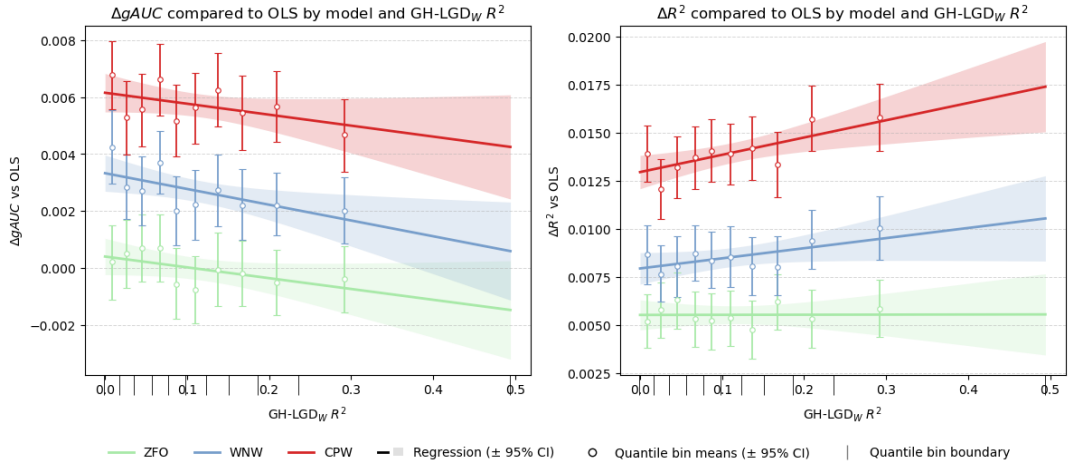
Figure D4: Differences of models' gAUC and R^2 compared to OLS by AB-cure AUC, CD-write-off AUC and E-LGD R^2. The lines show regression trends fitted on individual dataset results with a 95% confidence interval. The points show quantile-binned means with a 95% confidence interval. The bin boundaries are indicated by vertical ticks on the x-axis.



(a) Differences of models' gAUC and R^2 compared to OLS by F-LGD$_C$ R^2.



(b) Differences of models' gAUC and R^2 compared to OLS by GH-LGD$_P$ R^2.



(c) Differences of models' gAUC and R^2 compared to OLS by GH-LGD$_W$ R^2.

Figure D5: Differences of models' gAUC and R^2 compared to OLS by F-LGD$_C$ R^2, GH-LGD$_P$ R^2 and GH-LGD$_W$ R^2. The lines show regression trends fitted on individual dataset results with a 95% confidence interval. The points show quantile-binned means with a 95% confidence interval. The bin boundaries are indicated by vertical ticks on the x-axis.

E Predicted LGD Distributions by Summary Statistic Quantiles

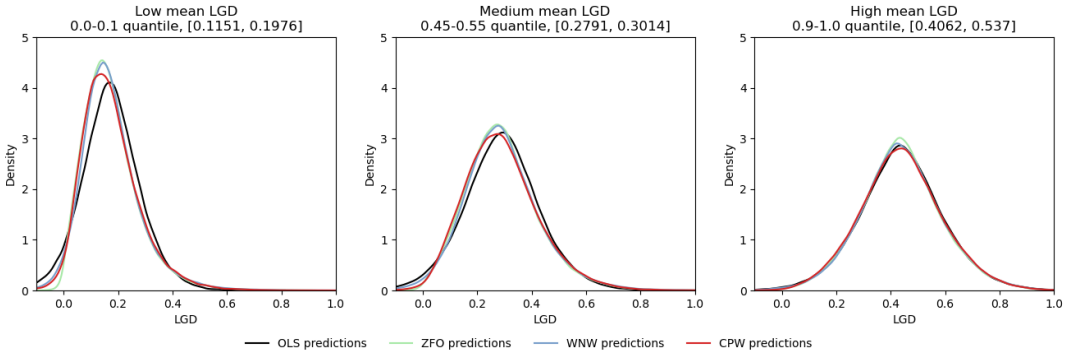


Figure E1: Predicted LGD distributions by mean LGD quantiles.

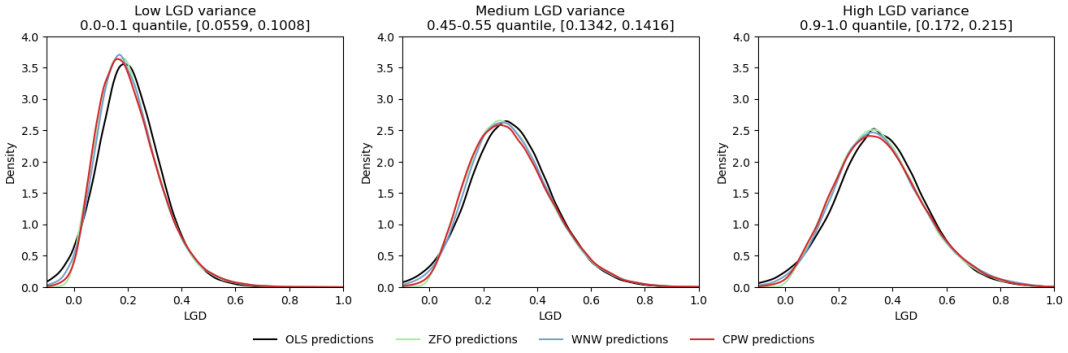


Figure E2: Predicted LGD distributions by LGD variance quantiles.

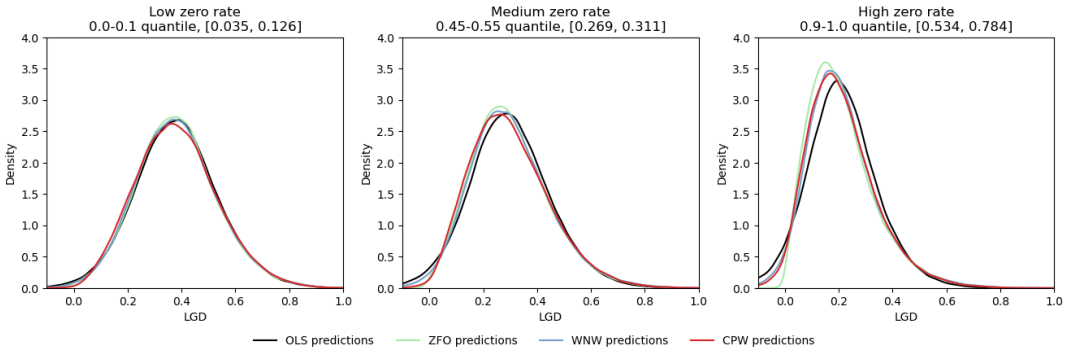


Figure E3: Predicted LGD distributions by zero rate quantiles.

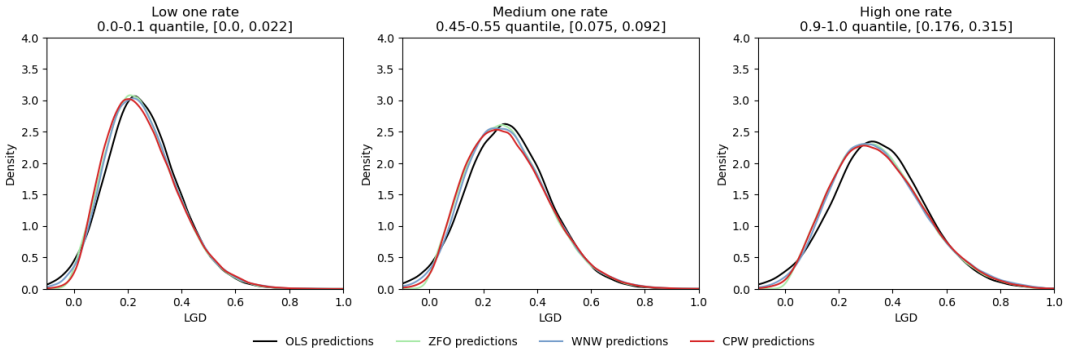


Figure E4: Predicted LGD distributions by one rate quantiles.

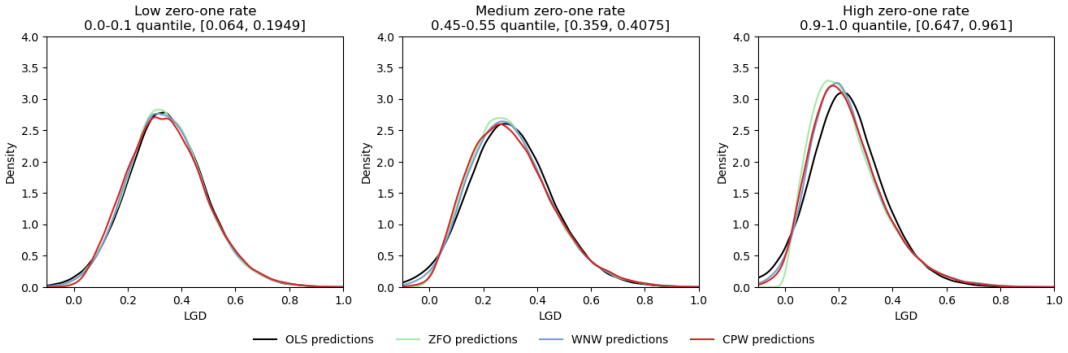


Figure E5: Predicted LGD distributions by zero-one rate quantiles.

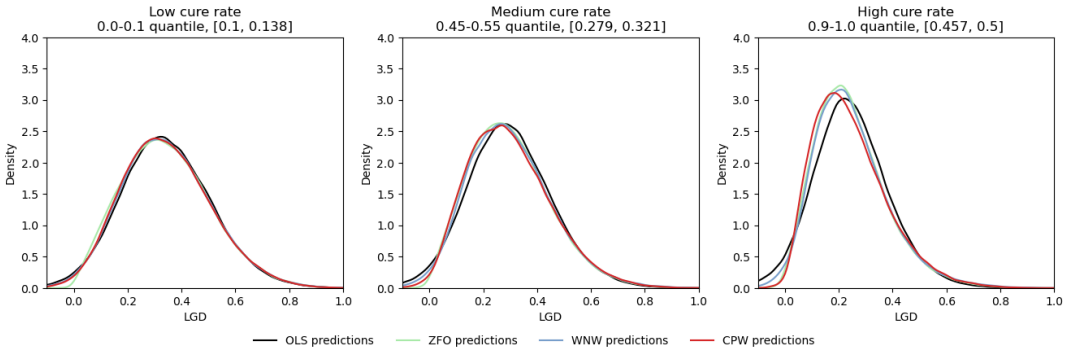


Figure E6: Predicted LGD distributions by cure rate quantiles.

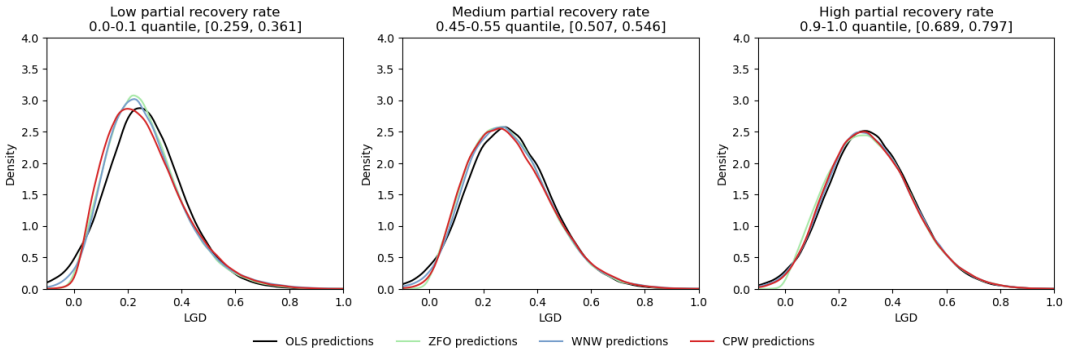


Figure E7: Predicted LGD distributions by partial recovery rate quantiles.

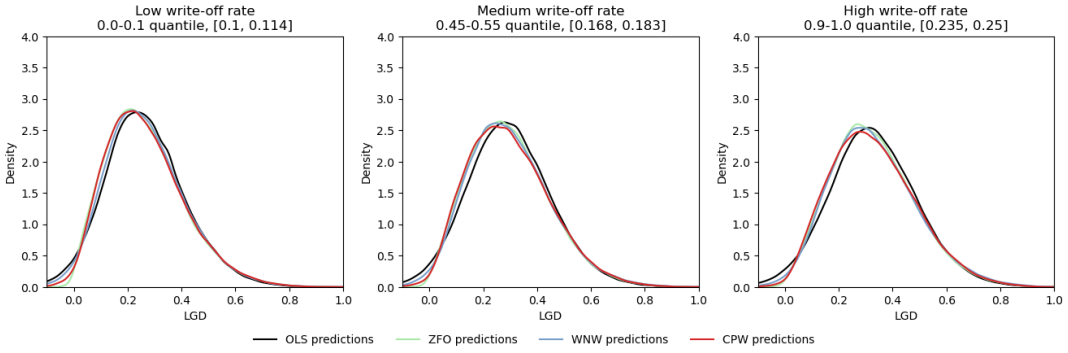


Figure E8: Predicted LGD distributions by write-off rate quantiles.

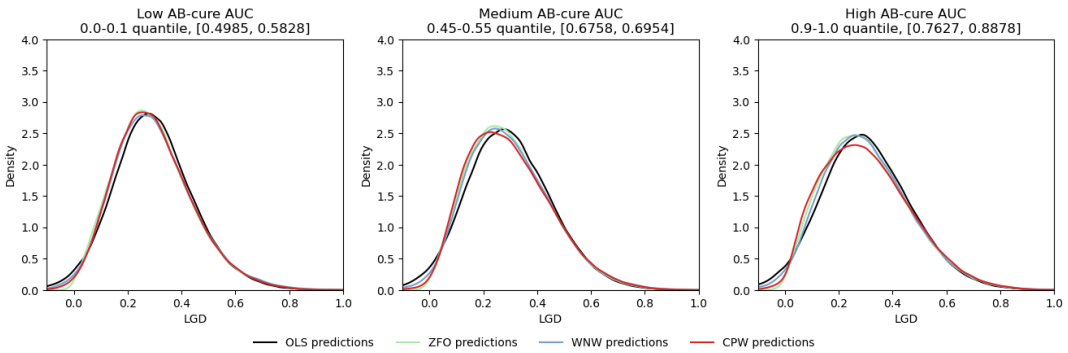


Figure E9: Predicted LGD distributions by AB-cure AUC quantiles.

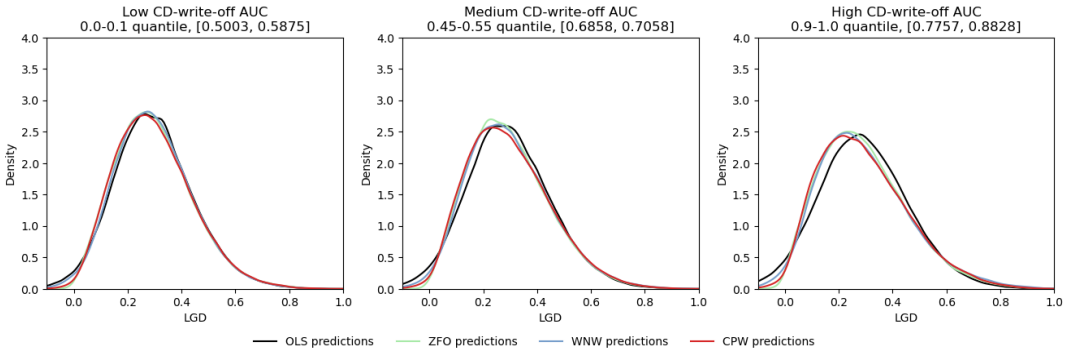


Figure E10: Predicted LGD distributions by CD-write-off AUC quantiles.

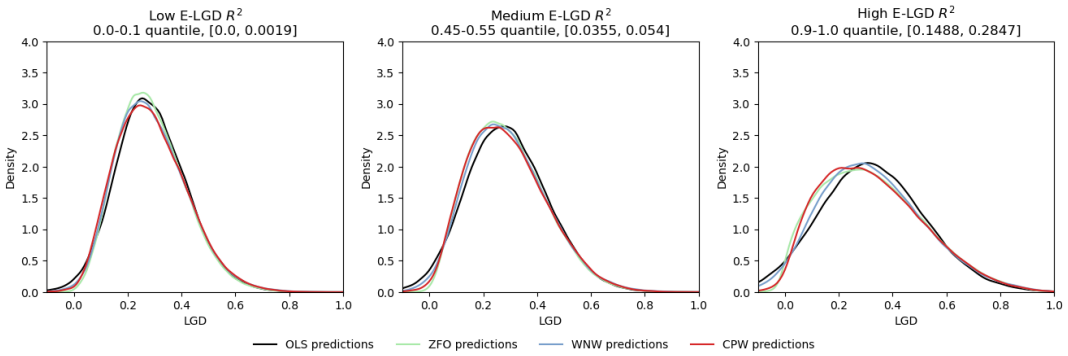


Figure E11: Predicted LGD distributions by E-LGD R^2 quantiles.

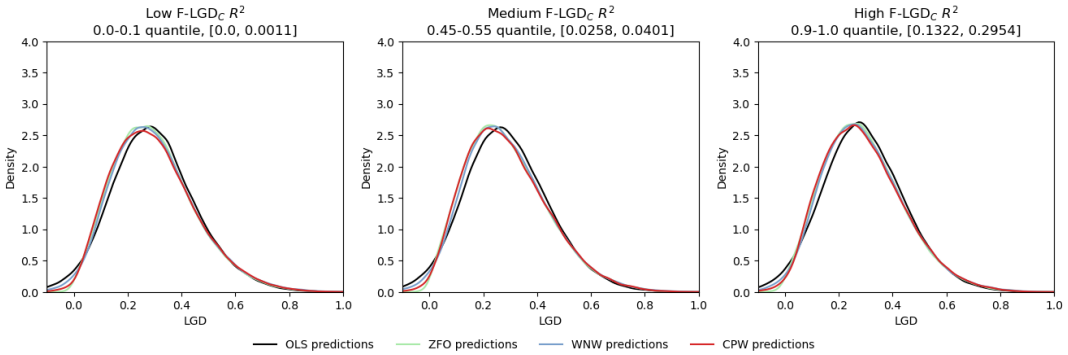


Figure E12: Predicted LGD distributions by F-LGD$_C$ R^2 quantiles.

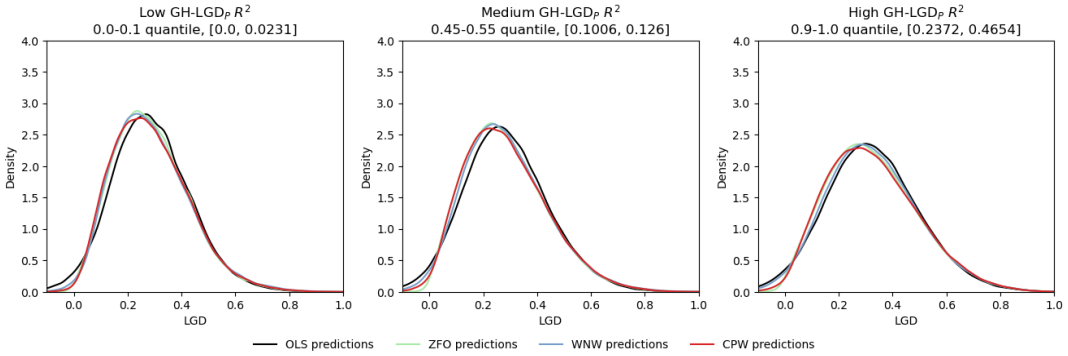


Figure E13: Predicted LGD distributions by GH-LGD$_P$ R^2 quantiles.

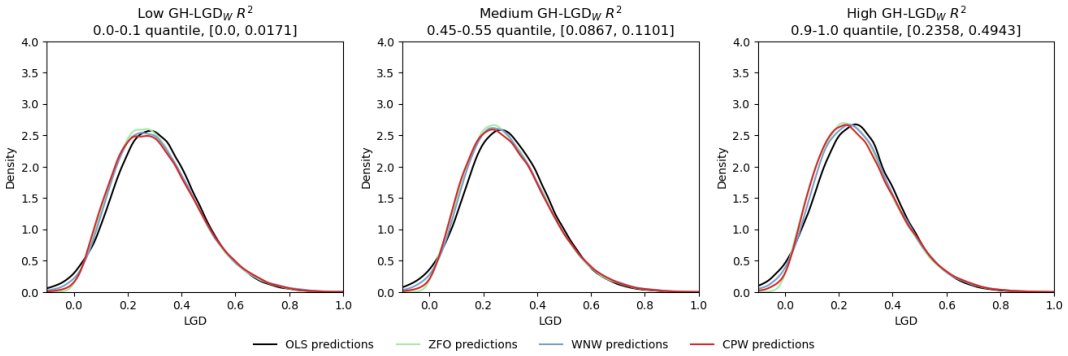


Figure E14: Predicted LGD distributions by GH-LGD$_W$ R^2 quantiles.

F Data Correlations

Table F1: Full correlation statistics for the simulated data sets. The correlation pairs that were explicitly specified in simulation are bolded.

Correlation pair	Mean	Min	Max	0.1 quantile	median	0.9 quantile
A-I_C	**0.1826**	**-0.0693**	**0.4488**	**0.0354**	**0.1809**	**0.3334**
B-I_C	**0.1848**	**-0.0851**	**0.4481**	**0.0357**	**0.1840**	**0.3334**
C-I_C	-0.0503	-0.2679	0.1005	-0.1134	-0.0469	0.0086
D-I_C	-0.0511	-0.2311	0.0961	-0.1151	-0.0481	0.0080
E-I_C	-0.1040	-0.4058	0.1058	-0.2328	-0.0877	0.0002
F-I_C	0.0002	-0.1090	0.1207	-0.0391	0.0001	0.0398
G-I_C	0.0007	-0.1163	0.1567	-0.0407	0.0003	0.0425
H-I_C	0.0007	-0.1112	0.1285	-0.0396	0.0008	0.0419
LGD-I_C	-0.4704	-0.8267	-0.1419	-0.6474	-0.4717	-0.2896
LGD0-I_C	0.2756	-0.5473	0.9581	-0.1218	0.2739	0.6791
LGD1-I_C	-0.1982	-0.5532	-0.0131	-0.3296	-0.1843	-0.0866
A-I_P	-0.1228	-0.3580	0.0827	-0.2310	-0.1207	-0.0188
B-I_P	-0.1244	-0.3599	0.0941	-0.2332	-0.1202	-0.0196
C-I_P	-0.0832	-0.3234	0.0940	-0.1653	-0.0788	-0.0068
D-I_P	-0.0835	-0.3210	0.1144	-0.1683	-0.0788	-0.0059
E-I_P	0.0022	-0.2753	0.3201	-0.0893	0.0033	0.0915
F-I_P	-0.0004	-0.1322	0.1126	-0.0416	-0.0002	0.0409
G-I_P	-0.0001	-0.1183	0.1100	-0.0409	-0.0005	0.0418
H-I_P	-0.0006	-0.1020	0.1141	-0.0411	-0.0009	0.0399
LGD-I_P	-0.0033	-0.6381	0.5505	-0.2654	-0.0011	0.2603
LGD0-I_P	-0.0525	-0.7617	0.7698	-0.4371	-0.0877	0.3941
LGD1-I_P	-0.1865	-0.7679	0.3405	-0.4199	-0.1871	0.0488
A-I_W	-0.0564	-0.2648	0.0985	-0.1246	-0.0514	0.0060
B-I_W	-0.0569	-0.2372	0.0970	-0.1255	-0.0523	0.0061
C-I_W	**0.1664**	**-0.0851**	**0.4195**	**0.0314**	**0.1650**	**0.3043**
D-I_W	**0.1678**	**-0.0655**	**0.4183**	**0.0327**	**0.1665**	**0.3034**
E-I_W	0.1199	-0.1026	0.3861	0.0150	0.1112	0.2407
F-I_W	0.0004	-0.1104	0.1263	-0.0404	0.0007	0.0408
G-I_W	-0.0008	-0.1092	0.1182	-0.0420	-0.0008	0.0407
H-I_W	0.0000	-0.1249	0.1210	-0.0404	-0.0002	0.0405
LGD-I_W	0.5615	0.0741	0.9367	0.3371	0.5734	0.7725
LGD0-I_W	-0.2549	-0.8347	0.3036	-0.4490	-0.2443	-0.0823
LGD1-I_W	0.4774	-0.1414	1.0000	0.1072	0.4824	0.8403
A-LGD	-0.0866	-0.3052	0.1025	-0.1788	-0.0806	-0.0042
B-LGD	-0.0882	-0.3182	0.0848	-0.1833	-0.0814	-0.0060
C-LGD	0.0931	-0.0765	0.3560	0.0063	0.0850	0.1929
D-LGD	0.0942	-0.1052	0.3314	0.0073	0.0878	0.1912
E-LGD	**0.2127**	**-0.0738**	**0.5336**	**0.0427**	**0.2095**	**0.3858**
F-LGD	0.0066	-0.1086	0.1130	-0.0332	0.0062	0.0475
G-LGD	0.1297	-0.0827	0.4152	0.0327	0.1240	0.2355
H-LGD	0.1300	-0.0782	0.3957	0.0361	0.1231	0.2340
LGD0-LGD	-0.5113	-0.9520	-0.2358	-0.6766	-0.5020	-0.3550
LGD1-LGD	0.5639	0.0785	0.9924	0.3198	0.5724	0.7918
A-LGD$_C$	0.0008	-0.2774	0.2824	-0.0788	0.0007	0.0798
B-LGD$_C$	0.0000	-0.2667	0.2683	-0.0824	0.0004	0.0814
C-LGD$_C$	-0.0004	-0.3107	0.3122	-0.0810	-0.0004	0.0808
D-LGD$_C$	0.0009	-0.2942	0.3376	-0.0775	0.0011	0.0785

Table F1: (continued)

Correlation pair	Mean	Min	Max	0.1 quantile	median	0.9 quantile
E-LGD$_C$	0.0933	-0.2735	0.3778	-0.0170	0.0899	0.2115
F-LGD$_C$	**0.1854**	**-0.2630**	**0.5435**	**0.0176**	**0.1783**	**0.3636**
G-LGD$_C$	-0.0003	-0.2766	0.2783	-0.0793	0.0004	0.0766
H-LGD$_C$	0.0010	-0.2663	0.2639	-0.0778	0.0003	0.0778
LGD0-LGD$_C$	-0.5220	-1.0000	-0.3346	-0.6205	-0.5203	-0.4236
LGD1-LGD$_C$	-	-	-	-	-	-
A-LGD$_P$	0.0002	-0.1817	0.1778	-0.0567	0.0006	0.0578
B-LGD$_P$	-0.0012	-0.1867	0.1554	-0.0587	-0.0009	0.0547
C-LGD$_P$	-0.0011	-0.1600	0.1828	-0.0575	-0.0018	0.0555
D-LGD$_P$	-0.0008	-0.1734	0.1812	-0.0599	-0.0009	0.0563
E-LGD$_P$	0.1790	-0.1314	0.5103	0.0268	0.1738	0.3382
F-LGD$_P$	0.0002	-0.1877	0.1939	-0.0557	0.0001	0.0569
G-LGD$_P$	**0.2194**	**-0.1278**	**0.5468**	**0.0372**	**0.2169**	**0.4041**
H-LGD$_P$	**0.2183**	**-0.1274**	**0.5417**	**0.0406**	**0.2170**	**0.3996**
LGD0-LGD$_P$	-0.4954	-1.0000	-0.0913	-0.7473	-0.4871	-0.2608
LGD1-LGD$_P$	0.3880	0.0704	1.0000	0.1501	0.3582	0.6902
A-LGD$_W$	0.0009	-0.3009	0.3443	-0.0996	0.0019	0.1005
B-LGD$_W$	0.0018	-0.2967	0.3061	-0.0977	0.0028	0.1006
C-LGD$_W$	-0.0008	-0.3224	0.3427	-0.1017	-0.0000	0.1001
D-LGD$_W$	-0.0012	-0.2829	0.3252	-0.1014	-0.0016	0.0965
E-LGD$_W$	0.1653	-0.2443	0.7034	-0.0106	0.1489	0.3699
F-LGD$_W$	0.0018	-0.3011	0.3095	-0.1001	0.0014	0.1023
G-LGD$_W$	**0.2017**	**-0.2166**	**0.6122**	**0.0190**	**0.1956**	**0.3997**
H-LGD$_W$	**0.2014**	**-0.2120**	**0.5934**	**0.0156**	**0.1974**	**0.3918**
LGD0-LGD$_W$	-0.5170	-1.0000	-0.1178	-0.8365	-0.5063	-0.2273
LGD1-LGD$_W$	0.5665	0.0891	1.0000	0.3133	0.5683	0.8197
A-LGD0	0.0500	-0.1822	0.3785	-0.0338	0.0403	0.1516
B-LGD0	0.0506	-0.2187	0.4012	-0.0336	0.0384	0.1539
C-LGD0	-0.0419	-0.2752	0.1147	-0.1069	-0.0371	0.0182
D-LGD0	-0.0427	-0.2802	0.1358	-0.1067	-0.0388	0.0167
E-LGD0	-0.1826	-0.4555	0.0711	-0.3345	-0.1803	-0.0343
F-LGD0	-0.0573	-0.2626	0.0971	-0.1306	-0.0512	0.0075
G-LGD0	-0.0842	-0.3513	0.0957	-0.1791	-0.0753	-0.0047
H-LGD0	-0.0855	-0.3542	0.0882	-0.1788	-0.0761	-0.0058
A-LGD1	-0.0371	-0.2071	0.1015	-0.0927	-0.0341	0.0153
B-LGD1	-0.0371	-0.2325	0.0910	-0.0948	-0.0343	0.0160
C-LGD1	0.0785	-0.1092	0.3611	-0.0071	0.0643	0.1900
D-LGD1	0.0800	-0.0984	0.3624	-0.0058	0.0676	0.1905
E-LGD1	0.1334	-0.1092	0.4023	0.0196	0.1252	0.2629
F-LGD1	0.0004	-0.1238	0.1297	-0.0399	0.0003	0.0400
G-LGD1	0.0796	-0.0718	0.2809	0.0085	0.0759	0.1563
H-LGD1	0.0799	-0.1040	0.3139	0.0102	0.0765	0.1557
A-B	**0.1010**	**-0.0851**	**0.2720**	**0.0137**	**0.1031**	**0.1850**
A-C	-0.0099	-0.1191	0.1096	-0.0519	-0.0098	0.0323
A-D	-0.0091	-0.1260	0.1247	-0.0512	-0.0090	0.0323
A-E	-0.0190	-0.1842	0.1185	-0.0679	-0.0179	0.0277
A-F	-0.0003	-0.1058	0.1186	-0.0414	-0.0007	0.0396
A-G	-0.0006	-0.1037	0.1162	-0.0407	-0.0009	0.0398
A-H	-0.0002	-0.1230	0.1540	-0.0399	-0.0005	0.0399
B-C	-0.0099	-0.1377	0.1314	-0.0525	-0.0098	0.0329
B-D	-0.0095	-0.1303	0.1007	-0.0509	-0.0094	0.0325

Table F1: (continued)

Correlation pair	Mean	Min	Max	0.1 quantile	median	0.9 quantile
B-E	-0.0193	-0.1719	0.1106	-0.0696	-0.0172	0.0276
B-F	0.0002	-0.1095	0.1385	-0.0397	-0.0003	0.0400
B-G	0.0004	-0.1021	0.1085	-0.0396	0.0003	0.0408
B-H	-0.0003	-0.1105	0.1065	-0.0407	-0.0001	0.0402
C-D	**0.0993**	**-0.1051**	**0.2862**	**0.0119**	**0.0989**	**0.1865**
C-E	0.0199	-0.1301	0.1669	-0.0269	0.0191	0.0678
C-F	-0.0001	-0.1106	0.1066	-0.0410	-0.0006	0.0413
C-G	-0.0002	-0.1109	0.1009	-0.0410	-0.0001	0.0401
C-H	-0.0003	-0.1184	0.1222	-0.0401	0.0000	0.0407
D-E	0.0209	-0.1114	0.1723	-0.0255	0.0194	0.0689
D-F	0.0004	-0.0993	0.1011	-0.0405	0.0006	0.0405
D-G	-0.0003	-0.1036	0.1101	-0.0410	-0.0005	0.0403
D-H	0.0002	-0.1307	0.1164	-0.0396	0.0007	0.0397
E-F	0.0082	-0.1132	0.1196	-0.0335	0.0080	0.0508
E-G	0.0312	-0.0993	0.2114	-0.0199	0.0283	0.0856
E-H	0.0319	-0.0981	0.2035	-0.0197	0.0296	0.0872
F-G	-0.0002	-0.1179	0.1058	-0.0402	-0.0006	0.0399
F-H	-0.0005	-0.1104	0.1090	-0.0411	-0.0008	0.0401
G-H	**0.0988**	**-0.0932**	**0.3008**	**0.0118**	**0.0989**	**0.1859**