Essays on LGD Models for Residential Mortgage Loan

A thesis presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy in Finance

at Massey University, Albany, New Zealand.

Justin Rylie Tang
2026

Abstract

This thesis introduces innovative models for predicting loss (recovery) rates of defaulted and uncured residential mortgages through three research topics: (1) decomposing loss (recovery) rates into three stages: prior to collateral disposition (Stage 1), collateral disposition (Stage 2), and post-collateral disposition (Stage 3); (2) further breaking down loss (recovery) rates by resolution types; and (3) examining loss (recovery) rates across periods that include the 2008 global financial crisis (GFC). All analyses utilize account-level US single-family prime mortgage data from Freddie Mac spanning 1999-2019, extended to 2021 for the third topic.

For the first research topic, we demonstrate that splitting uncured recovery rates into three meaningful stages and improving accuracy for each stage enhances overall accuracy in specific cases. Each recovery stage exhibits distinct loss (recovery) rate distributions, historical trends, and driving factors with varying magnitudes and directions. This challenges the traditional approach of predicting loss (recovery) as a single proportion of exposure at default (EAD), which overlooks valuable information. This contribution enables recovery rates to be modelled in three distinct, banker-relevant modules that can be independently improved. Several models for each stage outperformed nominated benchmarks in RMSE and r-square metrics when applied to the Freddie Mac dataset.

The second research topic extends the first by developing separate three-stage models for each resolution type, as evidence suggests different outcomes yield varying recovery levels and distributions across stages. These models are combined with estimated probabilities of occurrence for each resolution type.

The third research topic investigates new external predictors relevant to periods of economic stress, incorporating them into models from the second topic to enhance prediction accuracy.

Collectively, these research topics contribute to more accurate mortgage recovery rate predictions, thereby improving collection strategy effectiveness and the accuracy of capital and accounting provisions. This thesis advances the literature by introducing new perspectives and uncovering valuable information in loss given default (LGD) modelling, ultimately enhancing credit risk quantification accuracy.

Acknowledgements

Justin would like to extend his sincere gratitude to all the organisations that have generously supported his PhD journey. In particular, he wishes to thank Kiwibank Ltd. and DebtManagers Ltd. for their financial contributions, which played a vital role in making this research possible.

He is especially grateful to Dr David Woods, who volunteered his time and expertise to co-supervise the PhD throughout its entirety. His unwavering support and guidance have been invaluable. Justin also thanks Dr Jens Nielsen for his early insights and thoughtful feedback on the initial proposal, which helped shape the direction of the research from the very beginning. Beyond their academic contributions, both Dr Woods and Dr Nielsen have provided mentorship that has been deeply meaningful on both a professional and personal level.

Justin also wishes to acknowledge the following individuals and committees for their constructive feedback, which significantly contributed to the development of this work:

- Vietnam Symposium in Banking and Finance (VSBF) 2021 (Essay 1)
- Noah Urban, a co-participant in VSBF who provided helpful feedback for Essay 1
- The 2022 New Zealand Finance Colloquium PhD Symposium, where the Essay 1 was awarded with the *Best PhD Paper Award* and has since been accepted for publication in the Annals of Operations Research
- Dr Narongdech Thakerngkiat (aka Kenny) for his feedback on the draft Confirmation Presentation
- Dr Martin Berka & Dr George Wu, for feedback during Justin's Confirmation Report in 2021
- Dr Alan Pope, for feedback on Essay 3

Finally, Justin expresses his heartfelt thanks to his supervisory team — Dr Hung Do, Dr David Tripe, and Dr David Woods — for their steadfast guidance, encouragement, and mentorship over the six years of this programme. Their support has been instrumental in bringing this research to fruition.

Table of contents

List of Tables

List of figures

Chapter 1- Introduction

This thesis is about improving loss given default (LGD) models for residential mortgages. It does so by looking at potentially meaningful modelling structures and including driving factors that the literature may have missed due to opportunities present at the time of research and writing.

In the context of loans, credit risk "is the potential that a bank borrower or a group of borrowers will fail to meet its contractual obligations and the future loss associated with that" (Bluhm et al. (2016). In quantifying credit risk, it is common practice to break it down to probability of default (PD), exposure, default (EAD) and loss given default (LGD) (Bastos, 2010; Hoskin & Irvine, 2009; Leow & Mues, 2012; Qi, 2013; Schuermann, 2004). Forecasting expected losses in credit risk means predicting each component and assembling them together in product form: PD x EAD x LGD.

1.1. Credit risk components and LGD

PD is the probability of a loan to default over a specific time horizon, and PD models have already been well explored. Some of them are hazard models (Bajari et al., 2008; Foote et al., 2008; Francesca, 2012), decision tree models (Feldman & Gross, 2005), maximum entropy models (Stokes & Gloy, 2007), generalised linear regression models (Demyanyk et al., 2012; Do et al., 2018; Elul et al., 2010; Papouskova & Hajek, 2019; Tanoue et al., 2017; Tong, 2015), use of dynamic parameters (Breeden, 2016; Djeundje & Crook, 2019) and non-parametric models (Li et al., 2012), each highlighting different drivers of default and how they are related to default events.

EAD is the total amount owed when an account defaults. EAD models are relatively simpler and mostly revolves around probabilities of drawdown and limit changes. For accounts with payment plans like term loans, amortisation is also considered.

Lastly, LGD is the unrecoverable portion of EAD. Recovery rate is 1-LGD, so predicting recovery rate is also equivalent to predicting LGD and these two terms are used interchangeably all throughout this paper. The usual range of recovery rate is between 0 and 1 where 0 means the entire EAD is lost and a value of 1 means the entire EAD is recovered. Although LGD is just as important as the first two, it is still not well explored (Calabrese, 2014b; Khieu et al., 2012; Renault & Scaillet, 2004; Seidler & Jakubík, 2009) mainly due to

scarcity of data (Bade et al., 2011; Tanoue et al., 2017) and thus still has room for accuracy improvement in predictions for some types of lending.

For this thesis, LGD estimates will be defined from the perspective of banks, which means insurance proceeds are treated as recovery.

1.2. LGD for retail mortgages

Banks offer loans to different segments of customers, including business loans, unsecured retail (credit cards, overdrafts and personal loans) and residential mortgages. The latter is the focus of this research.

Residential mortgage lending is "by far the most important asset class" in a bank's book (Do et al., 2020). Federal Reserve data indicates that between 2008-2018, the total lending portfolio of US commercial banks ranged from 3.7-4.3 trillion USD (Economic Research, 2021), with residential mortgages accounting for 2 to 2.2 trillion USD of this amount (Economic Research, 2021). While not considered to be as risky, due to collateral prior to the GFC, the GFC has shown that even collateral values can drop dramatically (Do et al., 2020), which resulted to higher losses for banks (Andersson & Mayock 2014). As such, predicting recovery rates is indeed important for credit risk modelling (Höcht & Zagst, 2014).

When a mortgage account defaults, it may return to performing or be prepaid through liquidation of other assets or borrowing from friends and relatives. The focus of this research is forecasting the recovery rate of defaulted mortgage accounts that did not return to performing or were prepaid after defaulting, referred to as uncured defaults.

Given the overall development of LGD models in the literature, the objective is then to explore new perspectives and develop a new generation of LGD models that outperforms typical models in the literature. This is done by nominating a naïve model and another which is a potential best model from the literature.

1.3. LGD in the context of regulatory and accounting standards

Some of the applications for LGD are support for optimal collection strategy, calculation of regulatory capital figures and calculation of accounting provisions (Bellotti & Crook, 2012).

In the presence of credit risk, regulators require banks to hold a safety cushion to account for unexpected credit losses. This quantity, called regulatory capital, is usually an imposed fixed proportion of a bank's risk-weighted-assets (RWA). Rules on RWA calculation are prescribed

in a document called Basel II (2004): "International Convergence of Capital Measurement and Capital Standards: A Revised Framework". Among several rules, this document states that a bank can be accredited as Internal Ratings Based (IRB) if it has proven to regulators that it has enough resource and capability to build credit risk (and other risks) models for predicting PD, EAD and LGD. Others who have not achieved this accreditation use a standardised calculation as a way of regulators making banks compensate for uncertainty by being more conservative (Bank for International Settlements, 2006). Consequently, capital generally decreases as a model becomes more accurate. In fact, some banks were motivated to acquire an advanced IRB rating to maximise return on capital (Yao et al., 2017). However, while most banks want to increase the level of accuracy for their calculation, prediction models can be built to include only one recession: the 2008 GFC, which is not sufficient for modelling. Data for recessions prior to this are non-existent or are kept in a way that cannot be analysed, and this is especially true for low default lending such as residential mortgages. Because of this, LGD modelling for residential mortgages turned out to be not as developed compared to that of unsecured retail and corporate lending.

In 2014, a new accounting standard (IFRS 9) was created to recognise credit risk in a forward-looking manner (FSI Connect, 2018). This appears to have based its earlier recognition of deteriorating credit risk quality from Spain's prudential regulator's style of dynamic provisioning (Roldán & Saurina, 2012; Saurina, 2009), through a concept called "significant increase in credit risk" (SICR) for calculating expected credit losses in IFRS 9 (Deloitte, 2016). In addition, IFRS 9 requires models to be unbiased, among other requirements like being based on reasonable and supportable information and without undue cost or effort (Deloitte, 2016; FSI Connect, 2018; KPMG, 2014; PWC, 2016; XRB, 2014a). There are two ways of estimating expected loss for a performing account under IFRS 9: 12 month expected credit loss (ECL) which is calculated when an account is performing without any signs of deteriorating credit quality, and lifetime ECL which is calculated when there is SICR (Deloitte, 2016; FSI Connect, 2018; KPMG, 2014; PWC, 2016; XRB, 2014a). As this is a significant change from its predecessor, IAS39, which requires banks to calculate provisions for accounts only as they become non-performing - with allowance for delayed recognition of non-performing loans (emergence period), and even from regulatory capital calculations, especially for standardised banks, there has been a struggle within the industry in implementing IFRS 9 compliant models (S&P Global, 2017).

In the US Generally Accepted Accounting Principles (GAAP), their new version of accounting rules called Current Expected Credit Loss (CECL) is generally like IFRS 9, except they only estimate losses using a lifetime horizon (Guégan et al., 2018). Because of this, CECL provisions usually end up with higher numerical values.

1.4. Gaps on mortgage LGD

Among several unique perspectives in modelling LGD, three are explored in this research work.

First (gap 1) is the source of cash flow for recovery from defaulted loans and its corresponding recovery stage in the process of collection. These are broken down into three stages:

- Stage 1: repayments prior to collateral disposition, reflecting borrower willingness and/ or payment capacity, or a lender's collection effort.
- Stage 2: proceeds from collateral disposition. Typically, the largest recovery amount, driven by property value (Biswas et al., 2020).
- Stage 3: post disposition recoveries, such as mortgage insurance payouts or shortfall repayments.

Each stage is influenced by distinct drivers and may vary over time, suggesting that stage-specific modelling can improve LGD forecasts. To the best of our knowledge, no prior studies have addressed this decomposition.

The second (gap 2) perspective is the type of resolution. Different resolutions—such as short sale, foreclosure, or deed-in-lieu—lead to different LGD outcomes due to varying collection processes. Building on Clauretie & Daneshvary (2011) and Pennington-Cross, (2010), this research integrates resolution type into LGD modelling. Recent empirical evidence (Gabriel et al., 2020) confirms that foreclosure results in significantly higher credit losses than alternatives like short sales, reinforcing the importance of resolution-sensitive LGD estimation.

Lastly (gap 3), modelling LGD during major economic downturns, such as the 2008 Global Financial Crisis, poses unique challenges. In such periods, market dynamics shift rapidly, and traditional valuation and recovery assumptions may break down. To address this, the research introduces a median-based house price index and a Distressed House Price Discount Index (DHPDI), which together allow for more granular tracking of price discounts across crisis scenarios. These indices help capture how external drivers of LGD behave differently under stress, enabling more accurate and resilient LGD models during economic contractions.

The existing literature on mortgage default and recovery has mainly focused on the United States, where non-recourse provisions in many states limit lender recovery to the collateral value. Ghent and Kudlyak (2011) demonstrate that recourse provisions materially affect borrower default behaviour. In full-recourse jurisdictions such as New Zealand and Australia, lenders retain the right to pursue borrowers for any deficiency following property sale, which fundamentally alters both borrower incentives and expected recovery rates. While the present study uses US data from Freddie Mac, the modular modelling framework is designed to be jurisdiction-agnostic: the structure of the model transfers directly, though the specific coefficients and resolution categories would require recalibration to reflect local assumptions.

1.5. Data Description and Empirical Evidence

1.5.1. Overview of Freddie Mac dataset

Our study employs the single-family fixed rate prime loan-level dataset provided by Freddie Mac.[1] The data consists of loans originated from 1999 to 2019 with monthly performance observed from 1999 to 2020. This data set is fit for LGD modelling purpose because it not only contains observations over the 2008/2009 GFC, but also represents the entire country's loss experience given it contains loans from most, if not all, states.

Several prior studies have used this data set. An earlier version of this data set was used by Lin et al. (2009), who studied spillover effects of foreclosure, and Mei et al. (2019) who developed a better repayment quantity design to minimise foreclosure chances by adjusting mortgage repayments based on current property prices. Others included Lekkas et al. (1993) which proposed a put option valuation to calculate LGD, and Pennington-Cross (2003) and Calem & LaCour-Little (2004) who contributed to the "debate of reform" of Basel Accord by incorporating economic drivers and PD distribution, showing a divergence between standardised and economic capital. It is important to note, however, that while the Freddie Mac dataset spans a wide variation of mortgage loans across the USA, all observations are prime loans, which may be less sensitive to economic movements compared to less credit-worthy borrowers under subprime lending. As a result, it provides a better basis for international comparisons of mortgage lending, particularly with countries where sub-prime lending is less prevalent.

[1] Available at: http://www.freddiemac.com/research/datasets/sf_loanlevel_dataset.page

The data set comes in two parts: performance and origination. The performance portion contains quarterly updates of monthly records for each outstanding loan, including delinquency status, outstanding principal balance, current fixed mortgage interest rates, and latest maturity date when an account is performing, while fields such as net property sale proceeds, foreclosure expenses, reason for having zero balance as an account's last performance record, MI recovery and non-MI recovery are populated when a defaulted account is resolved or disappears from the dataset. The origination portion contains all information taken as at origination like FICO score, LTV for the specific loan, combined LTV (CLTV) of the customer including secondary mortgages disclosed by the mortgage seller, starting balance, interest rate, maturity, origination date, state where property is located, number of units for the property, occupancy status, mortgage insurance coverage, debt-to-income (DTI) ratio, property type for the main property under mortgage, loan purpose, number of borrowers, name of entity that sold the mortgage to Freddie Mac, and name of loan servicer.

1.5.2. Necessary filters in coming up with modelling dataset

To build a reliable model, it is essential to exclude some observations. For example, accounts that are still performing after a default event are excluded from this exercise. **Table 1** presents the reasons for exclusions and the number of observations excluded for each category. Starting with 1.69M defaulted accounts, fewer than 310k loans are left for modelling.

Table 1. Table of exclusions

Exclusion category	Description	Accounts
Starting accounts	Starting accounts	1,688,854
Invalid resolutions	It is composed of accounts that have defaulted but have not cured and are still not matured.	380,675
Missing CLTV	CLTV is the basis for one of the most important drivers of R2; R2 is usually the highest driver of R'. Given this, we exclude accounts with missing CLTV.	58
Unknown Net Sale Amount	R2 is usually the highest contributor to R', and unknown net sale means R2 is indeterminate mainly due to awaiting resolution type. Given this, EAD can be determined but LGD cannot be.	191,678
Unknown property type	Property type specifically tagged as unknown (99) are to be excluded	40
Missing DTIO	DTI was also expected to be important. There are cases when values are assigned to be 999 or blank. These accounts to be excluded	86,802
Unknown MIP	When MI is present, MIP needs to be available (i.e. value should not be 999), otherwise the account can't be used for LGD calculation	4,944

Missing FICO	FICO scores that are missing can't be replaced by a proxy so should be excluded as well	2,762
Observations with > 1% change in EAD and R	Current release where EAD and R are within 1% of that in the next release.	15,129
Defaulted accounts that are back to performing	There are accounts that have gone back to performing after going into default. While they should theoretically be contributing to the probability of cure, the reality is that nobody knows if they will go back to default at any point before maturity. Because of this, these observations are excluded	238,887
Cured accounts and have left	There are accounts that have gone back to performing and have prepaid after going into default. This is the population where cure = 1 when cure is modelled. However, that is not part of this research topic.	271,667
MSA is missing in Freddie Mac or HPI does not exist for MSA	Because DLTV is obtained by merging with an external data set, it is not surprising to not get a successful merge. These are cases when MSA as a joining field is missing or the MSA simply does not have any HPI from FHFA. This includes US territories like Puerto Rico.	95,477
DTI index is not available for specific MSA for specific dates to be joined	After excluding for unmatched FHFA HPI, another external data set that represents MSA level DTI indices also has a significant number of unmatched observations primarily for MSAs that are smaller. These are excluded for the preliminary analysis but may be re-included if DDTI or DICR are shown to be insignificant in modelling LGD	91,414
Undefined S2	S2 is undefined because A1 = EAD, but property liquidation still proceeded.	1
Missing UMP	Unmatched UMP	234
Outlier INT	INT > 50%	1
Remaining accounts	Remaining accounts	309,085

This table exhaustively reports the data clean-up process taken prior to modelling

Note that the remaining sample is still subject to sampling depending on the type of test being performed, which means actual data used for modelling is lower than 310k. To ensure robustness, we undertake random sample selections (with replacement) and out of time predictions over several rolling sample windows which are repeated 10 times.

1.5.3. Variables derived from additional datasets

To add information value, four external data sources are used: Federal Housing Finance Agency (FHFA) house price index (HPI), DTI index, foreclosure-related state laws, and state-level monthly seasonally adjusted unemployment rate.

To upgrade both origination LTV and combined origination LTV to default period, FHFA HPI was matched by MSA and date to scale origination LTV from origination to default period and origination loan amount was replaced with default unpaid principal balance (UPB). We illustrate with DLTV = (CLTV x origination HPI/default HPI x default UPB). This is common industry practice previously used by Do et al. (2020); Leow & Mues (2012); Qi & Yang (2009); Tong et al. (2013). To account for MSA definition changes and to ensure most loans have matching HPI, a few MSA codes were replaced, mostly for old records.

DTI was also scaled by MSA level to the period when the default occurred to ensure origination DTI is still relevant at the time of default. To do this, a DTI index provided by the Federal Reserve was used. We define a new variable $DDTI = DTI \times \frac{Index\ at\ Default}{Index\ at\ origination}$. While less common, this follows a similar logic as LTV.

In most US states such as Arizona, Connecticut, Delaware, Hawaii, Illinois, Iowa, Louisiana, Maryland, Massachusetts, Mississippi, Montana, New Hampshire, New York, Oklahoma, Pennsylvania, South Carolina, Texas, Washington D.C., and West Virginia (*United States Foreclosure Laws, 2020)*, the statutory right of redemption allows defaulted borrowers to reclaim collateral by paying everything that is due. If negotiations have failed, this could be the last hope for defaulted borrowers to reclaim their property. Otherwise, this becomes another cost to be deducted from the total recovered amount. While it may be expected that states where statutory right of redemption is prohibited will result in lower pre-collateral recoveries, Do et al. (2018) found the opposite; they highlighted that a statutory right of redemption also means more court rulings, which take time and money. This lowers the chance of zero loss and increases non-zero LGD.

There are also states with judicial foreclosures such as Connecticut, Delaware, Florida, Illinois, Indiana, Iowa, Kansas, Kentucky, Louisiana, Maine, Nebraska, New Jersey, New Mexico, North Dakota, Ohio, Pennsylvania, and South Carolina (*United States Foreclosure Laws, 2020)*, which means lenders need a court ruling to foreclose on a defaulted borrower. This adds to cost of recovery.

Recovery on shortfalls may only happen when allowed by law in the US given prohibition of deficiency judgment in a few states such as Delaware, Iowa, Massachusetts, Mississippi, Missouri, Nebraska, West Virginia (*United States Foreclosure Laws, 2020)*. This is consistent with findings of Clauretie & Herzog (1990); Do et al. (2018) where LGD is higher when lenders

are left to write-off a shortfall after realising that foreclosure sale is unable to cover the total outstanding amount.

While all three types fall under post collateral disposition recoveries, they can be thought of as two components: MI and non-MI. Unfortunately, there is no way to distinguish between deficiency judgement and other more technical recoveries under non-MI recovery. It also becomes challenging to determine the correct order between MI and non-MI. Given these constraints in the Freddie Mac dataset, all post collateral recoveries are simply grouped under Stage 3.

Unemployment rate is joined by state at default time. It is expected to drive recovery rates down.

These quantities are listed in detail in **Table 2** for variations of LTV and DTI, while the three definitions of state laws and unemployment rate are defined in **Table 3**.

1.5.4. Variable transformation using all datasets

A few transformations were performed to derive key variables in modelling recovery rate R. The first one is default definition; because the dataset keeps a monthly record of delinquency status, default is simply defined as the first time an account becomes at least 90 days past due or goes into REO sale, whichever is earlier. REO sale is a type of resolution following default where the lender, through court proceedings, gains ownership of the collateral/s allowing them to sell it in the market to cover what was owed. The amount owed usually includes allowable expenses and fees. The definition of default is chosen to be practical yet aligned with regulatory standards (Bank for International Settlements, 2006) and common banking practice.

Because Freddie Mac contains real-life accounts, some accounts may default, cure and re-default. To avoid confusing the models, an account is defined as having cured if there is no record of 30+ delinquency for 12 consecutive months. To prevent confusing models with information from subsequent defaults, only first defaults are used in this study.

EAD is defined as the total amount outstanding at default after future modifications from debt consolidation are added. For modelling, modifications are defined as any instance when outstanding principal balance increases over time. To control for the possibility of bias from modifications, a dummy independent variable is added.

While account performances are populated monthly under normal circumstances, there are times when a record skips due to data quality issues. Because of this, the unpaid principal balance right before collateral disposition also needs to contain the modification amount. In the case of Freddie Mac dataset, EAD is defined as the higher value of default or pre-collateral liquidation balance.

A1 for the Freddie Mac dataset is defined as the difference between EAD and the unpaid principal balance right before collateral disposition. This is the case when pre-collateral liquidation balance is higher, A1 = 0.

The dataset does not have gross property sale proceeds, but only sales proceed net of collateral related expenses. Foreclosure expense is defined separately and is not related to the maintenance and disposition of collateral. This is considered under A2. Following Goodman & Zhu (2015), we do not consider interest accrued during delinquency.

We summarize the definition of these quantities in **Table 2** below.

Table 2. *Freddie Mac concepts, given and derived quantities*

Concept	**definition**
Default	90 days past due or any loss experienced, in alignment with both IFRS9 and Basel II regulations
Date of default (t)	The date when an account first hits 90 days past due, if sudden loss without delinquency, then the loss date will be recognised as date of default
Cure	A once defaulted account which has never been 30+ days past due for 12 consecutive months.
Scheduled balance	Scheduled balance if no modification or advance payment has been made since origination
Balance at default	Total unpaid balance as at date of default (90 days past due or earlier if went to REO earlier)
Total Modification	Pre-collateral balance – balance at default
Pre-collateral balance	Unpaid principal balance before zero balance code is assigned. Collateral recovery and Stage 3 recovery happen in zero balance code assignment. This is zero if an account cures.
EAD	Exposure at default; max (balance at default, Pre-collateral balance)
Accrued interest	(First delinquency unpaid balance – non-interest-bearing unpaid balance) * (Current Interest rate – 0.35) * (Months between Last Principal & Interest paid to date and zero balance date) * 30/360/100. This formula is defined in the Freddie Mac User Guide in Freddie Mac (2019). This quantity will not be used as prescribed by Goodman & Zhu (2015)
Origination UPB	Origination unpaid principal balance
MI recovery	Dollar amount recovered from MI
Non-MI recovery	Dollar amount recovered from non-collateral and non-MI source post collateral disposition
MV	Net proceeds from property sale. The amount credited to the lender when collateral has been sold; net of allowable expenses related to selling.
C0	Origination collateral value extracted from origination LTV
DDTI	Dynamic debt-to-income ratio. DTI at origination x DTIt/DTI0. DTIt is DTI index at time of default, DTI0 is DTI index at time of origination. DTI index matched by MSA.
LDTI	Liquidation Dynamic debt-to-income ratio. Defined as DTI at origination*DTIL/DTI0. DTIL is DTI index at time of property liquidation, DTI0 is DTI index at time of origination. DTI index matched by MSA.
DLTV	Default UPB / (C0 / HPI0 * HPIt), HPIt is from FHFA at default and matched by MSA
LLTV	Liquidation UPB / (origination security value / HPI0 x HPIL), HPIL is from FHFA at liquidation and matched by MSA.
MI	Indicator whether an account has mortgage insurance or not. 1 if MI present, 0 otherwise.

Concept	definition
A	A1 + A2 + A3
A1	(Balance at default – pre-collateral balance + total modification)
A2	MV – selling expenses – other foreclosure expenses
A3	MI recovery + non-Mi recovery

This table defines core modelling quantities, concepts, and their formula, if applicable, as they are used in Freddie Mac dataset. Some quantities like DDTI and DLTV are not solely obtained from the Freddie Mac dataset. DDTI is DTI at origination scaled by a DTI index obtained from the Federal Reserve. DLTV on the other hand is CLTV where the collateral value is scaled using FHFA HPI and loan amount is the updated unpaid principal balance at default.

1.5.5. Definition of dependent variables and outlier handling[2]

Outliers were capped and floored at values presented in **Table 4** for modelling but preserved for out of sample calculations of expected vs actual recovery, RMSE and R-square. The rationale for this is that models should not be confused by outliers. But since outliers are real-world observations, they should be included in assessing model performance, regardless of how infrequent they are. These bounds are shown in the third and fourth columns of **Table 4** and **Table 3**. These outliers are mostly loans with small EAD where expenses dominate most of the net recovery.

[2]We refer the definitions of continuous and categorial independent variables as well as the outliers handling process to **Table 3**.

Table 3. *Variable definition for Freddie Mac*

Variable	Definition	Floor	Cap
CLTV	Cross collateralised loan to value ratio at origination	0	1
DICR	$DDTI/DTI_0$	0	2
DLTVCR	DLTV/CLTV	0	2.5
DTI_0	Origination Debt-to-Income ratio. Valid values range from 0 to 65%. Calculated using sum of borrower's monthly debt payment & housing expense at time of delivery to Freddie Mac / total monthly income at origination	-	-
FHB	First home buyer flag. 1 if loan is a mortgage for borrower's first home. 0 otherwise.	N/A	N/A
FICO	FICO score at origination	-	-
LC	Liquidity Constraint. Balance at default/scheduled balance - 1	-0.13	0.03
LICR	LDTI/DDTI	0	2
LLTVCR	LLTV/DLTV	0	2
LOB	log of origination UPB	-	-
LP	Purpose for the mortgage loan: · Purchase – used to purchase a property · Refinance with cash out – no specific purpose to the loaned amount. Was not used to secure property. · Refinance with no cash out - limited to paying off first mortgage, or loans for other properties used to secure current mortgage, and cash out of min (2% of refinance amount, $2000) · Refinance but not specified · Unknown	N/A	N/A
MF	Flag for modifications that resulted in unpaid principal balance increase. If total modification is positive, 1, else 0	N/A	N/A
MIP	MI percentage. If MI = 1, MI is the percentage of the shortfall covered by insurance provider on the event of default after collateral disposition	-	-
MOB	Months on book. The age of the loan when it defaulted	-	-
NB	Number of borrowers: 1 2+ blank	N/A	N/A
NJF	States with non-judicial foreclosure: Connecticut, Delaware, Florida, Illinois, Indiana, Iowa, Kansas, Kentucky, Louisiana, Maine, Nebraska, New Jersey, New Mexico, North Dakota, Ohio, Pennsylvania, South Carolina (United States Foreclosure Laws (2020)	N/A	N/A
NSRR	States with Statutory right of redemption: Arizona, Connecticut, Delaware, Hawaii, Illinois, Iowa, Louisiana, Maryland, Massachusetts, Mississippi, Montana, New Hampshire, New York, Oklahoma, Pennsylvania, South Carolina, Texas, Washington D.C., and West Virginia (United States Foreclosure Laws (2020)	N/A	N/A

Variable	Definition	Floor	Cap
OO	1 if property under mortgage is owner occupied, 0 if not.	N/A	N/A
PDJ	Prohibited deficiency judgement. The following states prohibits deficiency judgement: Delaware, Iowa, Massachusetts, Mississippi, Missouri, Nebraska, West Virginia (United States Foreclosure Laws (2020)	N/A	N/A
PT	Property type · CO = Condo, Co-op or manufactured housing · PU = PUD · SF = Single Family homes	N/A	N/A
SNS	Seller (originator) name is not the same as servicer name (1) else 0	N/A	N/A
TID	Months between first 30 DPD and default date	0	100
TTR	Time to resolution. Months between default date to property disposition, which is also when A2 and A3 are realised	0	150
INT	Mortgage rate at the time of default	N/A	N/A

Table 4. *Derived dependent variables definition for Freddie Mac*

Variable	Definition	Winsorise floor	Winsorise cap
R1	$\frac{A1}{EAD}$	0	1
R1'	R1 if R1 > 0 and 0 otherwise	-	-
R2	$\frac{A2}{EAD}$	-0.7	1.5
R2'	R2 if A2 > 0, else 0.	-	-
R3	$\frac{A3}{EAD}$	-0.4	1
R3'	R3 if A3 > 0 and 0 otherwise	-	-
R	$R1 + R2 + R3$	-1.11	2.04
R'	R1 if A > 0 and 0 otherwise	-	-

This table defines dependent variables for modelling and their formula as they are used in Freddie Mac dataset. Some quantities like R, R1, R2 and R3 need winsorisation for the training dataset to ensure treatment for outliers.

1.6. Thesis Outline

This thesis addresses the identified gaps through three interconnected essays that progressively build a comprehensive residential mortgage LGD model.

The first essay decomposes uncured recovery rates into three collection stages, exploring optimal ways to express and predict recovery at each stage. By combining stage-specific predictions, this approach is tested against benchmark models including simple linear regression and the two-part model by Do et al. (2020), using the Freddie Mac single family loan dataset.

The second essay extends this framework by further breaking down each recovery stage into specific resolution types (prepayment, repurchase, REO disposition, charge-off, note sale, third party sale, and short sale). Building on the multi-stage structure from Essay 1, it models both resolution-specific recovery parameters and the probability of each resolution occurring, providing granular insights into recovery outcomes.

The third essay enhances the overall model by incorporating a distressed house price index as an additional predictor, demonstrating that distinguishing between standard and distressed property sales significantly improves LGD forecast accuracy—a refinement of the comprehensive approaches developed in the previous essays.

Together, these essays create a progressively sophisticated LGD modelling framework: from stage-based decomposition (Essay 1) to resolution-specific modelling (Essay 2), to market-condition enhancement (Essay 3). Each essay validates improvements using RMSE, R-square, and expected versus actual recovery comparisons, ensuring both statistical rigor and practical applicability for the banking industry.

2. Chapter 2- Literature review

2.1. Structure categories for LGD models under all types of lending

There are two classes of LGD models: structural, which follows a more theoretical framework based on the underlying principle that assets = liability + equity and is used to price bad debt if a market price is not readily available, and workout LGD, which focuses on the recovery cash flow and uses this information to predict LGD.

2.1.1. LGD models based on asset pricing

Carty & Lieberman (1996) of Moody's came up with data sets that contain secondary market prices of default bonds which could be used to extract the premium for risky performing corporate bonds as a proxy for recovery rates. Gupton et al. (2000) later continued the idea before Düllmann & Trapp (2004) emphasised that the use of LGD specific systematic factors is also important. Assuming this information is available, it wouldn't matter how much cost is recoverable from each individual bad debt as it will be floored at a value close to the market price. Unfortunately, this kind of information is only available for large and liquid markets like defaulted corporate bonds. And even then, there are still huge uncertainties hiding behind irrationalities in markets (Tomarchio & Punzo (2019) which imply that a uniform expectation on market value through time is not always a safe bet.

While one could argue that pricing for bad debt sold to 3rd party debt collection agencies could be used under this category for non-tradeable assets, this system is still too crude, and one would expect pricing to be more of a one-size-fits-all rather than risk-based. Given this, other loan classes would need an alternative approach.

Fortunately, in the absence of market price, there is a category of models built from asset pricing. For this generation of models, actual recovery information is not necessary except for validation of results. This generation of models is more helpful for low default portfolios and/or data with low loss experiences. The logic behind this model class is based on several key assumptions. First, that this is a risk neutral world where there is no arbitrage present; this means "expected return + capital gains = risk-free return + risk adjustment" (Lekkas et al., 1993). In plain English, it says that the value of a loan is essentially the same as the present value of all future cash flows using a discount rate that presumably already factors in credit risk. Second, that the market is efficient (i.e. logical) which implies the existence of an optimal point when defaulting would benefit a borrower. From this, the value of a put option should be

the same as the premium of a risky performing debt on top of an already defaulted one after adjusting for time value using some risk-free rate. According to Altman (2003), this category of credit risk models can be widely categorised in 3 generations.

Generation 1. In 1974, Merton (1974) started pricing models for risky debt: a European put option (underlying asset is a zero coupon bond) on the borrowing firm's asset as underlying against a strike price equal to the borrowing company's total liability. PD was then implied to be the probability that liabilities will be greater than the assets while recovery amount is the residual from the asset after removing the price of the option for the reason being a bank would need to buy a put option of the same type to protect itself from losses/ as an insurance equivalent. Building on a set of simplifying, and at times, unrealistic, assumptions, several foundational papers sought to enhance the robustness of this concept by changing it from an assumption of single creditor to accommodate seniorities in a multiple lender environment (Black & Cox, 1976), allowing coupon payments (Geske, 1977), which resembles a business term loan with principal and interest payments, and lastly, relaxation of the assumption that a borrower only defaults at maturity (generation 2).

Generation 2. Structural pricing models, default probability is implied from American put options (default can happen at any point during the life of a loan) and the pay-out to the lender is scaled by a constant representing a "recovery factor" as a proportion of outstanding amount in place of looking at total assets minus put option price in the first generation. This "recovery factor" was derived from historical observation on similar loan profiles and so it is independent from firm's asset value by construction (Longstaff & Schwartz, 1995).

Generation 3. While the first two generations obtained LGD implied from put options, the next generation, reduced form models, are simply based on the premium between an already defaulted asset and a similar one that is still performing. The idea of this is that risk premium can be treated as a rate of return. That said, a discount rate of R = risk-free + PD * LGD was used to equate a performing corporate bond's face value to an already defaulted one that shares the same risk profile as the former. Some assumptions were made like fixed values of risk-free rate for time value and PD from survival models and LGD is obtained by solving back. This has been shown by Duffie & Singleton (1999); Jarrow et al. (1995) for corporate bonds and business loans, and Lekkas et al. (1993) for mortgages, which was llaterimproved in Crawford & Rosenblatt (1995) by incorporating costs & fees. Despite, once again, being logically sound, this generation of models had a lot of weaknesses: availability of market prices to solve the

return R from, assumption that the market price of both performing and defaulted assets are fair and fully reflective of risks, and the fact that the return R might also contain other risks like liquidity risk.

While the academic world has been exploring theoretical models, the banking industry has quietly been implemented widely-used models with details not fully disclosed due to proprietary reasons. Most of these are commercial models like Gupton et al. (2002); Gupton & Stein (2005) who built simple models that assumed LGD follows a beta distribution.

Given this, there was nowhere else to go but to consider historically obtained cash flows from recovery.

2.1.2. Workout LGD models

A coin always has two sides: instead of looking at the price of a bad debt to use as floor to credit risk losses, another option is to predict the net cash flow following default during workout processes by mostly looking at historical experience. This category of models can be grouped into 3 structure classes which roughly aligns with the categorisation introduced by Lyn C. Thomas et al. (2016): historical averages, regression techniques and hybrids between structural and regression techniques.

2.2. Review of LGD modelling techniques for all types of lending

Taking historical averages is perhaps the simplest approach in modelling LGD. Workout LGDs started with very basic historical averages estimated for recovery rates (Hamilton et al., 1999) broken down by lending type (Basel Committee on Banking Supervision, 2001, 2021), before being split by cash flow component where loss is the net of write-off, interest lost, interest payments, unanticipated recovery and miscellaneous repayments. Asarnow & Edwards (1995) and Hurt (1998) did their study using Citibank's corporate loans secured in different levels located in North and South America. This eventually led to look-up tables, usually as a function of security coverage (Leow & Mues, 2012), when implemented, and something that Gupton & Stein (2005) criticised due to lack of robustness. While this simple approach looks promising, not all types of costs can be factored in considering a perpetually changing environment in banking. Lastly, historical averages work only for unimodal and symmetric distributions, which LGD does not satisfy given the recent works of Chen & Wang (2013); Hlawatsch & Ostrowski (2016); Hurlin et al. (2018); Krüger & Rösch (2017); Tong (2015); Witzany et al.

(2012) which, in different ways, have shown that LGD is either multimodal and/ or asymmetric.

Given historical records, a statistical regression technique is perhaps the most popular methodology in building LGD models.

For a typical parametric regression model, generalised linear models included, LGD is calculated as a single quantity (Altman et al., 2005; Chava et al., 2011; Gupton et al., 2000b; Siddiqi & Zhang, 2004) or estimated as separate groups of constant cash flows (Asarnow & Edwards, 1995; Hurt, 1998).

The simplest form is ordinary least squares (OLS) or linear regression (Bellotti & Crook, 2012; DeFranco, 2001); however, OLS models were only shown to be predictive with enough predictors and observations-a situation not many banks are in. From this point, model improvements came in two forms: improvement of model structure and consideration of additional predictors.

In the face of limited predictors, structures were used to "assist" models in forecasting. This is where the linear regression models started to become more generalised with different link functions like logit, probit, tobit, etc. (Chalupka & Kopecsni, 2009; Chava et al., 2011; Pennington-Cross, 2010; Rösch & Scheule, 2014).

Given some claims that LGD should be bound within 0 and 1 (Tomarchio & Punzo, 2019), a number of papers have dealt with this boundary problem using logit link function (Chava et al., 2011; Hao & Ala, 2010; Karwański et al., 2015; Pennington-Cross, 2010) which refers to a link function that is the log of odds ratio between an event happening and the same event not happening, fractional logit (Bellotti & Crook, 2007, 2009, 2012) which is essentially logit with a different likelihood function to optimise, probit (Chava et al., 2011; Hwang et al., 2016) that uses a standard normal CDF as a link function, Tobit models (Bellotti & Crook, 2012; Do et al., 2018; Sigrist & Stahel, 2010; Yashkir & Yashkir, 2013) which is, once again, a logit function with censoring on both ends of 0 and 1, and other censored distributions (Sigrist & Stahel, 2010; Yashkir & Yashkir, 2013).

Some even tried accounting for the distribution of LGD like Krüger & Rösch (2017) who pointed out that using simple averages in estimating LGD is not appropriate given that it is not unimodal in nature. Following this, several literature attempted to address this issue by using quantile regressions Krüger & Rösch, 2017; Siao et al., 2016; Somers & Whittaker, 2007),

modified gamma distributions (Tong et al., 2013; Yashkir & Yashkir, 2013), and beta based transformations and regressions (Bellotti & Crook, 2012; Calabrese, 2014a, 2014b; Hlawatsch & Ostrowski, 2016; Loterman et al., 2012; Qi & Zhao, 2011).

There were also studies that split LGD into zero and non-zero values like Bijak & Thomas (2015); Leow & Mues (2012); Somers & Whittaker (2007); Tong et al. (2013); Wood & Powell (2017); Yao et al. (2017) and adding another split on predicting the chance that LGD = 1 through multinomial models like Tomarchio & Punzo (2019). Do et al. (2020) pointed out that a non-zero LGD is not entirely independent of the zero LGD so they started with the assumption that a non-zero loss is conditional on a zero LGD not happening, which they then connected to the bigger picture by conditioning both cases after PD in Do et al. (2018) because, after all, LGD is a conditional probability—conditional on default happening.

Thomas et al. (2012) LGD from different ends depending on collections policy where in-house collection would be split into zero LGD vs LGD > 0 and those that went to 3rd party collections would be split by LGD = 1 and LGD < 1 while Chen (2018), on the other hand, has treated time to recovery as a random variable.

After this, there is also the use of cross-sectional effects where some studies have pointed out that LGD drivers can have varying influence on LGD through another domain. With this, some proposed to allow parameters to have cross sectional effects through time (Bade et al., 2011) using cross sectional dummies representing different points of loan age, origination date and observation time.

Non-parametric models, on the other hand, would then allow data to define its own distribution like kernelling which was done by Calabrese & Zenga (2010); Chen et al. (2019); Chen & Wang (2013); Hurlin et al. (2018); Nazemi et al. (2018); Renault & Scaillet (2004) where distributions are fitted from data rather than imposed, or the urn-based approach used by Cheng & Cirillo (2018), (2019) that incorporates time to resolution using combinatorics technique. On the hybrid side, machine learning was used by Bellotti et al. (2019) to model recovery for non-performing mortgage loans.

Then there were those which employed the use of machine learning techniques to train models based on data like vector support regression/machine done by Chen et al. (2009); Hurlin et al. (2018); Loterman et al. (2012); Nazemi et al. (2017), (2018); Tobback et al. (2014); Yao et al. (2015), (2017); Zhang et al. (2014), local logit regression by Sopitpongstorn et al. (2017) and other machine learning techniques like neural networks by Hurlin et al. (2018); Loterman et al.

(2012); Nazemi & Fabozzi (2018); Qi & Zhao (2011). While these models have produced really accurate in-sample predictions, possibly because non-linear relationships are also captured (Qi & Zhao, 2011), they have been revealed to be overfit when tested out of sample and out of time. Furthermore, this generation of models are also difficult to explain to key bank stakeholders, let alone be maintained and monitored, which is a requirement in both regulatory and accounting standards.

Finally, regression techniques need not be independent from structural models. In fact, bringing some steps in collection processes into the model might help make models more transparent and therefore, understandable. By merging banking intuition with data-driven techniques, we achieve both transparency and predictive accuracy.

Several researchers have attempted different modelling approaches: some have employed survival models such as Markov chains (Crook & Bellotti, 2010); some have deployed proxies for certain nodes using regression trees like Bastos (2010); Hurlin et al. (2018); Loterman et al. (2012); Nazemi et al. (2017); Nazemi & Fabozzi (2018); Qi & Zhao (2011), and others have implemented complex decision trees, as demonstrated by Bellotti & Crook (2012); Bonini & Caivano (2016); Thomas et al. (2010).

As a variation of Merton (1974), a new workout LGD model was designed by Frontczak & Rostek (2015); Pelizza & Schenk-Hoppé (2019) which treats the collateral as the new underlying asset in pricing put options where other parameters like haircut mean and volatility were estimated from real data. This has enabled application for secured retail loans like mortgages.

Because LGD highly depends on the type of recovery cash flow, it is only appropriate to distinguish LGD between different types of lending. While the previous sections have covered all types of lending including business, unsecured and secured retail lending, the focus of this study will be on predicting residential mortgage LGD as it is still relatively unexplored and most models built are usually one of the following: (i) structural which has sound business interpretation but has poor predictive performance like Ambrose et al. (1997); Crawford & Rosenblatt E (1995); Lekkas et al. (1993), (ii) only based on predicting house prices by defining LGD to be the proportion of EAD that is not covered by asset sale and considering foreclosure expenses like Andersson & Mayock (2014); Biswas et al. (2020); Chen & Chen (2010); Clauretie & Herzog (1990); Crawford & Rosenblatt (1995); Harrison & Mathew (2008); Lekkas et al. (1993); Leow et al. (2014); Leow & Mues (2012); Pelizza & Schenk-Hoppé

(2019); Pennington-Cross (2003), which is self-explanatory given collaterals usually take up majority of the recovery portion but fails to consider that other cash flows could also be significant, and (iii) statistical models where a list of drivers are picked intuitively and some numerical relationship, including imposition of what a loss distribution should look like, for some, is enforced against LGD. This is shown in the works of Biswas et al. (2020); Calem & LaCour-Little (2004); Chen (2018); Clauretie & Herzog (1990); DeFranco (2001); Do et al. (2018), (2020); Leow et al. (2014); Leow & Mues (2012); Pennington-Cross (2003); Qi & Yang (2009); Tong et al. (2013).

Mathematical structures are not yet that developed for mortgages. Among the many ways to mathematically link the relationship between losses (or recoveries) to different drivers, the simplest form, linear regression, was done by Chen (2018); Clauretie & Herzog (1990); DeFranco (2001); Pennington-Cross (2003), followed by generalised linear regression which is essentially linear regression with a link function (Leow & Mues, 2012; Qi & Yang, 2009; Tong et al., 2013), and finally, other mathematical structures like multi-step conditional approaches were also explored to accommodate the multimodality of LGD. Leow et al. (2014) built separate models for probability of repossession and a 'haircut model' for repossessed accounts while Do et al. (2018), (2020) built conditionally related models through probability of zero loss and expected value of non-zero loss. Generally, the predictions turned out be more accurate in-sample with more explanatory variables.

On the other hand, having too many predictors come at a risk of over specification and multicollinearity. For mortgages, two main predictors were found to explain LGD and its components like probability of repossession and non-zero LGD: collateral coverage, and affordability. Chen & Zhang (2011) of Moody's illustrated this using loan to value ratio (LTV) and debt service coverage ratio (DSCR) and named this the "double trigger framework". Do et al. (2020) demonstrated that LC and positive equity (1-DLTV, when positive) explained zero-loss for foreclosed mortgages and negative equity (1-DLTV, when negative) drove non-zero loss. Similarly, Leow & Mues (2012) found that the probability of repossession is driven by LTV at default and any history of being previously in default, and the realisable proportion of valuation of security at default (i.e. Haircut) is driven negatively by LTV, ratio of valuation of security at default to average property valuation in that region, history of previous default, time on book, security value, and property age.

2.3. Detailed discussion of Benchmark Model Inferences

A strong negative relationship between R and a loan-to-value-ratio variable in general is trivial and was highlighted by Chava et al. (2011); Do et al. (2020); Goodman & Zhu (2015); Greve & Hahnenstein (2016); Somers & Whittaker (2007). While this way of defining the LTV related variables is uncommon, it helps avoid obvious multicollinearity while the coefficients make intuitive sense and align with the different transformations of LTV used in the literature (Chava et al., 2011; Do et al., 2020; Goodman & Zhu, 2015; Greve & Hahnenstein, 2016; Park & Bang, 2014; Somers & Whittaker, 2007).

LOB has one of the highest influences on uncured recovery R. This is likely due to prioritisation to sell assets secured by loans with higher outstanding balances. This may be because lenders prioritise recovery from bigger loan sizes at origination, consistent with the findings of Calem & LaCour-Little (2004); Do et al. (2020); Pennington-Cross (2003) that the relationship of loan size and non-zero LGD generally has a convex parabola (concave for recovery).

Positive coefficients of MIP point to the existence of insurance. Naturally, the higher the insurance coverage, the higher chance that more of EAD may be recovered.

TTR likely represents the amount of fees and interest charged from default until liquidation of collateral. A net negative relationship may have resulted from increasing expenses with higher time to resolution, thereby lowering recovery.

While DRD is a field not often used in estimating LGD, this attempts to link current portfolio risk level to recovery levels as the former is expected to influence day-to-day operations, including collection effort and quality. High DRD for a specific date may mean longer queues for collections and thus might push for delayed or lowered recoveries. As such, it is not surprising to see this variable in the top 5 list of predictors for R.

Aside from those highlighted, there is no material difference between benchmark model B1 and B2 for R. Additionally, material drivers for P(A>0) of B2 are MOB and OO in place of MIP and DRD. Because P(A>0) can be interpreted as the probability of non-charge-off, a high MOB points to more chances of being charged-off while owner occupied collaterals tend to reduce this possibility.

While MOB (loan age at default) was highlighted by Do et al. (2018); Qi (2013) to have a positive relationship with their probability of cure, which was defined as zero loss, although this came about after controlling for the parabolic effect. For the Freddie Mac dataset, this

parabolic relationship ends up with a net negative coefficient using the Freddie Mac dataset, likely due to more concentration of younger accounts (<36 months old) at default.

Owner occupied (OO) properties generally yield to less zero-losses (i.e. charge-offs) and are also generally sold at a higher price given owner occupiers usually take more care of properties; Clauretie & Daneshvary (2011); Do et al. (2020); Qi (2013); Qi & Yang (2009) have also observed this.

2.4. Gaps identified and intended to be filled in this research

There are two ways to contribute to the literature for mortgage LGD modelling. First, is to incorporate additional drivers contributing materially to information value. Second, is to make the models more robust by deriving a modelling framework that is aligned with business interpretation and is accurate not only for in-sample tests but also in out of sample and out of time. As outlined in Section 1.4, Essay 1 and Essay 2 contribute through the second point (addressing gaps 1 and 2 respectively), while Essay 3 satisfies the first (addressing gap 3).

2.5. Discounting of recovery cashflows for LGD modelling

There are times when the time interval between recovery cash flows is significant. In which case, time value plays a significant part in estimating LGD. Several papers in the literature such as Brady et al. (2011); Do et al. (2018), (2020); Duffie & Singleton (1999); Rapisarda & Echeverry (2016) have accounted for this. However, the way that they have accounted for it differs significantly.

When discounting cash flows, there are two components: the discounting factor which is determined by a discount rate, and the time interval to bring the value of the cash flow back to. While the period for discounting is simple and is usually just calculated in hindsight, the rate choice differs for different papers. For what intuitively makes sense, Scheule & Jortzik (2020) stated that "discount rates for LGD should be based on opportunity cost of comparable financial instruments and include the risk-free rate and a premium for non-diversifiable risk". As such, identified weighted average cost of capital (WACC) (i.e. capital ratio and debt ratio weighted cost of funding) and market equilibrium methods (i.e. a base rate + premium for systematic risk) are better than the use of contract rate. Using US data, Do et al. (2018), (2020) have discounted losses using LIBOR rate matched at default date with a fixed premium of 3%. Using NZ data, Harrison & Mathew (2008) defined discount rate differently depending on the source of cash flow. If it is from foreclosure of a loan, the rate is risk-free rate + 6% premium for

being foreclosed and the lending rate to the insurance company if it is from insurance receivables.

This is not to be confused with the discounting rate used to bring all future expected credit losses (ECL) back to a specific time point when ECL is being reported; for the latter purpose, discounting rate is determined as the rate that brings all future contractual cash flows to origination value (a.k.a. effective interest rate). For application in IFRS 9, XRB (2014) states that the appropriate rate should be higher than the risk-free rate and lower than or equal to the effective interest rate. Further, they have recommended the use of effective interest rate.

In the Freddie Mac dataset, the performance data already incorporates the effects of accrued interest during the delinquency period through the relationship between EAD and the unpaid principal balance. The time-to-resolution variable (TTR) used in our models further captures time-dependent effects on recovery, including the accumulation of carrying costs, property depreciation during vacancy, and market timing effects that are not fully reflected in accrued interest alone. While alternative discounting conventions, such as using market-implied discount rates or modelling discount rates as endogenous to market conditions, could in principle yield different LGD estimates, the approach here follows standard industry practice and is consistent with the treatment used in the benchmark models against which our framework is compared.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Justin Rylie Tang
Name and title of main supervisor:	Dr. David Tripe, Adjunct Professor of Banking
In which chapter is the manuscript/published work?	Yes

Describe the contribution that the student and members of the supervisory team have made to the manuscript/published work:[1]

This study introduces a three-stage LGD decomposition framework for residential mortgages—pre-disposition, disposition, and post-disposition recovery—aligned with collection processes. Using Freddie Mac loan-level data (1999–2020), the approach outperforms traditional single-component models in out-of-time predictions and uncovers stage-specific drivers. The contribution lies in advancing theoretical understanding of recovery dynamics, which enables improved practical risk management, capital allocation, and loss provisioning.

Please select one of the following three options:

- (●) **The manuscript/published work is published or in press**
  Please provide the full reference of the research output:
  https://link.springer.com/article/10.1007/s10479-025-06860-w
- ( ) **The manuscript is currently under review for publication**
  Please provide the name of the journal:
- ( ) **It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:	Justin Rylie Tang — Digitally signed by Justin Rylie Tang Date: 2025.12.06 06:00:36 +13'00'	Main supervisor's signature:	

This form should be placed at the beginning of each relevant thesis chapter.

[1] Refer to the Massey University Publishing and Authorship guidelines (OneMassey for staff, Stream for students) and/ or Contributor Roles Taxonomy (CRediT) guidelines for guidance.

3. Chapter 3 - Essay 1: Predicting loss severities for residential mortgage loans: A decomposition approach

This essay has been accepted for publication in the Annals of Operations Research

3.1. Introduction

The optimisation of credit risk models remains a critical operational challenge in financial institutions, particularly in the domain of Loss Given Default (LGD) estimation. As established in the main introduction, the optimisation of LGD estimation presents unique computational and methodological challenges.

Current approaches to LGD modelling exhibit several operational limitations. Traditional Ordinary Least Square (OLS) methods treat LGD as a unified quantity (Clauretie & Herzog, 1990), ignoring the multi-stage nature of the recovery process. While PD modelling has evolved to incorporate sophisticated techniques such as advanced scorecard modelling (Qian et al., 2024), LGD modelling often remains confined to simplified unidimensional approaches folded within modelling of overall credit losses (Cui et al., 2024). Existing decomposition approaches focus on value ranges rather than operational stages (Crook & Bellotti, 2010; Yao et al., 2017), limiting their practical application. introduced a binary decomposition between cure and non-cure components, while Starosta (2021) proposed a three-way split considering cures, partial recoveries, and write-offs.

Despite these advances, current approaches fail to capture the fundamental nature of mortgage recovery processes. Recovery cash flows in mortgage lending originate from distinct sources: (i) direct borrower payments, (ii) collateral disposition or debt transfer, and (iii) insurance proceeds and residual collections. These sources exhibit different historical patterns, respond to different economic drivers, and require different modelling approaches. Traditional single-component models may produce inconsistent predictions when sub-components move in opposing directions over time, particularly during economic downturns.

Loss-Given-Default (LGD) modelling is fundamentally an Operations Research problem, because it may involve:

- decomposing a complex, stochastic recovery process into manageable sub-problems,
- forecasting cash flows under uncertainty, and
- optimisation of collection related processes.

By decomposing LGD into three components, we create a modular framework that readily supports optimisation for both modelling and operations. This mirrors the demonstration by Demyanyk & Hasan (2010), where Operations Research methods are used to explain systemic banking distress, predict default dynamics and design optimal intervention strategies for financial crises. In the same spirit, our three-Stage LGD model not only sharpens predictive accuracy but also yields actionable insights on when to liquidate collateral, how to prioritise collection efforts and how to structure collection policies - thereby enhancing the resilience of credit risk management.

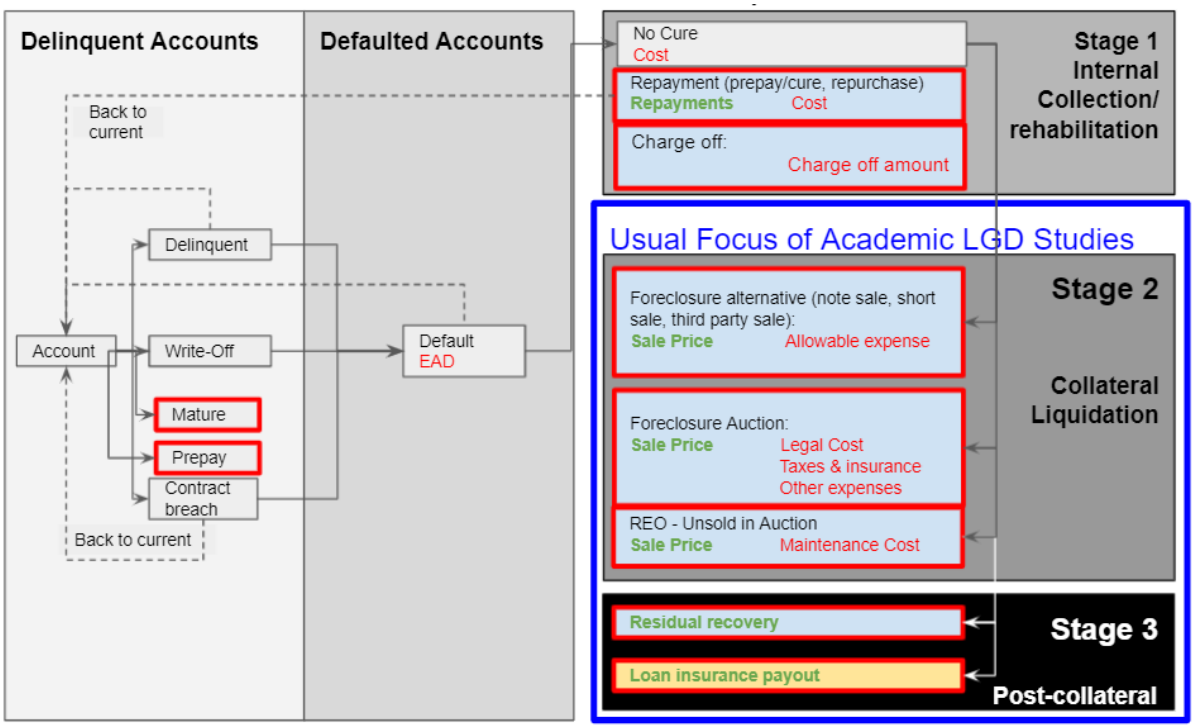

Figure 1. *Mortgage Loan Recovery Process and Stage Decomposition.*

The lifecycle of a delinquent mortgage loan is grouped into three vertical columns: delinquency, default, and collection. The focus of this study is on collection, which is broken further down into 3 stages of recovery.

This paper introduces a novel three-stage decomposition framework aligned with typical bank collection processes, as illustrated in **Figure 1**. The diagram captures the complete loan lifecycle, from initial delinquency through to final resolution, demonstrating how a performing account may transition to default through various paths including write-off, maturity, *or* prepayment. Once an account defaults, the recovery process unfolds across three *categorical* stages, each representing different recovery channels and associated costs.

Building on the three stages identified in Section 1.4 (gap 1), we separate loss realisation into: Stage 1 (pre-collateral disposition recovery), encompassing internal collection and rehabilitation efforts including potential cure through repayment or repurchase; Stage 2 (collateral disposal), covering various foreclosure alternatives and sale processes; and Stage 3 (post-collateral disposition recovery), including residual recoveries and loan insurance payouts. While previous academic studies have primarily focused on Stage 2 (collateral liquidation), our framework captures the full spectrum of recovery channels and their associated costs.

Using the dataset described in section 1.5, we demonstrate that these stages exhibit distinct statistical properties and economic relationships. Stage 1 recoveries primarily reflect borrower capacity and bank collection efforts, Stage 2 responds to property market conditions, and Stage 3 correlates with insurance industry dynamics. Our analysis reveals several key findings supporting the proposed decomposition. Analysing the Freddie Mac dataset, Goodman & Zhu (2015) found that uninsured mortgages with high equity paradoxically showed higher losses than insured mortgages with low equity, suggesting the importance of separating insurance proceeds from other recovery sources. Recovery patterns across stages show distinct distributional characteristics and temporal trends, particularly during economic stress periods. Each stage demonstrates unique relationships with economic and operational drivers, supporting the need for stage-specific modelling approaches. While Cui et al. (2024) establish techniques to model aggregate portfolio losses, and Qian et al. (2024) explore borrower-level PD scoring, this paper focusses on improving model interpretability and accuracy of the LGD component, consequently improving modelling for predictions of losses.

The proposed three-stage decomposition offers several advantages: enhanced transparency in bank operations, improved model accuracy through stage-specific prediction, better-informed loss minimisation strategies, and more precise risk assessment capabilities. Our empirical results demonstrate that a combined modelling approach - using OLS for Stages 1 and 3 with a two-step model for Stage 2 - consistently outperforms traditional single-component models across both out-of-time tests and random sample validations.

The remainder of this essay is structured as follows. Section 3.2 develops the theoretical framework and empirical strategy for our three-stage decomposition approach. Section 3.3 delivers the empirical results, Section 3.4 presents empirical results and discussion, and Section 3.5 concludes.

3.2. Modelling concepts

As a simplification of **Figure 1**, **Figure 2** displays a sequence of recoveries matched with the three stages in a collection process post a default event.

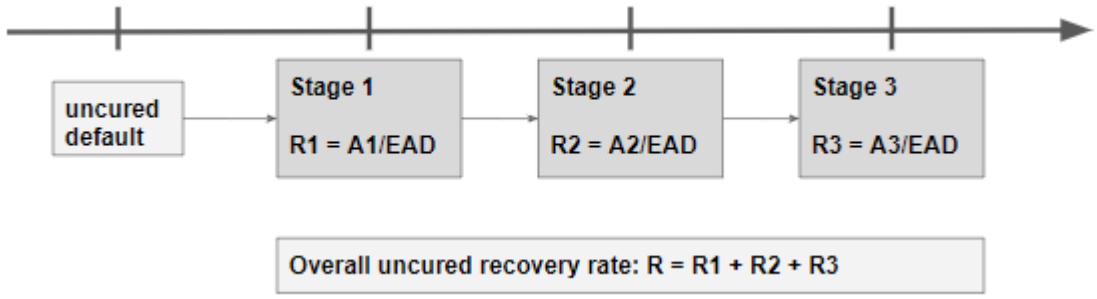

Figure 2: *Three-stage diagram for uncured recovery rate.*
A defaulted account that does not cure (prepay or repurchase) goes through internal in-house collection (Stage1) where A1 is recovered, expressed as % of total amount owed EAD, R1. In the presence of shortfall, collection process proceeds to Stage 2 where collateral is liquidated, or contract is transferred (case of note/ reperforming sale) after deducting for allowable property/contract disposition expenses, A2, expressed once again as a proportion of EAD, R2. If shortfall is still present, process goes to Stage 3 where insurance proceeds and other miscellaneous cash flows are involved. These quantities add up to A3. Expressed as a proportion of EAD, we have R3. Together, the total amount recovered is R.

If a defaulted loan does not cure, the servicer will first attempt to negotiate repayment plans before considering more drastic options. This may involve interest rate modifications or extension of loan life, which may occur as debt consolidation. When the outlook from assessment points to a negative net cash flow, the lender might be better off charging/writing a loan off and saving costs. Any repayments over this period are aggregated after adjusting for time value and defined as quantity A1. From the perspective of the lender, they may have a higher chance of collecting more under R1 if the time to resolution is longer. Because this sequence excludes accounts that cured, it will only lead to immediate charge-off or to Stage 2.

When an account gets to Stage 2, it usually indicates a significant deficiency in loan repayments that the lender no longer believes it can recover much without more severe measures. The collateral or the entire loan contract can be disposed of in several different ways. At this point, the borrower, with the lender's permission, may choose to sell the property at a distressed price (short sale) or sell it in a foreclosure auction (third party sale). If the lender did not approve or the property is unsold through short sale or third-party sale, the lender may take ownership of the property through court rulings and take over property disposal to cover the outstanding debt

and some or all its costs (real-estate-owned, REO). As an alternative to selling the property, the lender can sell the mortgage loan to another party for them to continue the loan with the property under new ownership (reperforming sale) or sell the mortgage loan to another lender who believes the uncured defaulted loan could be worth more than what they are paying (note sale). With these, we end up with quantity A2, which includes expenses related to foreclosure and property/ loan contract disposition.

In some cases, even loan/property disposition is not enough to cover the outstanding debt and other costs. In these cases, lenders can claim insurance proceeds for loans with mortgage insurance (typically those with origination LTV > 80%). Otherwise, the lender can pursue the defaulted borrower to claim deficiencies on shortfalls, when permissible by law. The net amount collected in this stage is defined as A3.

Together, A1, A2 and A3, each valued at default date, sum up to total recovery A so uncured recovery rate, $R = R1 + R2 + R3$, where $R1 = \frac{A1}{EAD}$, $R2 = \frac{A2}{EAD}$ and $R3 = \frac{A3}{EAD}$. Once again, we note that this recovery definition comes from the perspective of a lender, which means proceeds from mortgage insurance are treated as recovery.

As there is no way to determine whether a defaulted account eventually cures other than in hindsight, predicting LGD would mean simply multiplying the expected uncured recovery rate $R, E(R)$, by the proportion of accounts that did not cure. Probability of cure may be a separate model which is not covered in this research.

3.2.1. Recovery Rate Components

Building on the three-stage framework introduced in Section 1.4 (gap 1) and the workout LGD models discussed in the main literature review, we define five key elements of recovery rates:

Exposure at Default (EAD): The unpaid principal balance at default, potentially increased due to loan modifications from debt consolidation and delinquency interest charges.

Stage 1 Recovery ($R1 = A1/EAD$):

- A1 represents the difference between EAD and unpaid principal balance prior to collateral disposition
- Captures early collection efforts and borrower payments

Stage 2 Recovery ($R2 = A2/EAD$):

- A2 represents net sale proceeds from collateral disposition
- Includes various foreclosure alternatives (note sale, short sale, third-party sale)
- Accounts for property disposition expenses and foreclosure costs

Stage 3 Recovery ($R3 = A3/EAD$):

- A3 comprises mortgage insurance (MI) recovery and non-MI recoveries
- Non-MI recoveries include:
    - Repurchase/make-whole proceeds
    - Tax/insurance refunds
    - Hazard insurance proceeds
    - Rental receipts
    - Positive escrow
    - Other miscellaneous credits

Total Recovery[3] ($R = R1 + R2 + R3$):

- Aggregate recovery rate across all stages
- Represents total recovery as a proportion of EAD

3.2.2. Definitions

We adopt four core definitions – Default, Cure, LGD and EAD:

First, default happens when either the bank deems the obligor unable to pay back a loan or the obligor goes more than 90 days past due for a material amount owed (Bank for International Settlements, 2006). When a defaulted account is non-delinquent for 12 consecutive months, it is considered cured. This study ignores subsequent defaults. This definition is aligned with Dermine & de Carvalho (2006) who caution against biases when considering each default case for multiple default customers.

Second, cure happens when a retail mortgage account defaults but eventually prepays or gets repurchased due to adverse selection. When this happens, there is no loss recorded.

[3] Cash flow discounting is explained in detail in Section 2.5.

Third, most mortgage LGD studies like Leow et al. (2014); Leow & Mues (2012); Tong et al. (2013) assume foreclosure or repossession as the point of default. While this enables them to find stronger predicting power, it misses accounts which have defaulted but have not been repossessed. This study explores a more comprehensive sample, and predictions may thus not be directly comparable with prior literature. The models proposed predict LGD of impaired accounts which did not end up cured.

Lastly, EAD is an important core quantity given that LGD is a proportion of this. When a defaulted account eventually prepays or matures, and the lending relationship ends without any write-offs, it is considered to have been cured. When an account does not cure, LGD is $1 - R$, where R is the total amount recovered as a proportion of EAD.

Table 5 enumerates all core quantities important for modelling. When considering the source for R, cash derived during Stage 1 is defined as A1. The proportion of EAD recovered under Stage 1 is R1. Similarly, for A2 recovered under Stage 2, $\frac{A2}{EAD}$ is defined as R2. A3 represents the quantity recovered during Stage 3 and R3 is defined as $\frac{A3}{EAD}$.

Table 5. *Definitions of core quantities and ratios*

Variable	Definition	Formulae
R	The proportion of EAD recovered, for non-cured accounts.	$R = R1 + R2 + R3$
A1	The total dollar amount collected during Stage 1. Each cashflow discounted to the default date if there is a discount rate.	
A2	The total discounted dollar amount collected during Stage 2.	
A3	The total discounted dollar amount collected during Stage 3.	
R1	The recovery rate for Stage 1.	$R1 = \frac{A1}{EAD}$
R2	The recovery rate for Stage 2.	$R2 = \frac{A2}{EAD}$
S2	The conditional recovery rate for Stage 2 (the amount recovered over the amount still unrecovered at the start of Stage 2).	$S2 = \frac{A2}{EAD - A1} = \frac{R2}{1 - R1}$
R3	The recovery rate for Stage 3.	$R3 = \frac{A3}{EAD}$

This table defines core modelling quantities and their formula. In general, R is used for recovery rate (e.g., R, R1, R2, R3), and A is used to define dollar amounts (e.g., A, A1, A2, A3).

Sequentially following Stage 1, another way to think of recovery under Stage 2 is to express it as a proportion of the uncollected amount in EAD, which we define as S2. That is, $S2 = \frac{A2}{EAD - A1}$. Given all quantities in this formula can be expressed as a proportion of EAD, we can simplify this term as $S2 = \frac{R2}{1 - R1}$.

3.2.3. Modelling approaches

Our primary hypothesis is that the decomposition of total recovery into three distinct stages, each modelled separately, enhances the prediction accuracy of the overall recovery rate R compared to modelling R directly. This section provides an overview of the model structure with succeeding sections detailing each component.

3.2.4. Benchmark models

To establish a baseline, we define the **Naïve model** as an Ordinary Least Squares (OLS) regression of R. OLS is widely used in industry for its simplicity and robust performance. As a second benchmark, we employ a **two-step selection model**, as proposed by Do et al. (2020), which has demonstrated superior performance over OLS. Additionally, we incorporate a machine learning benchmark using a random forest model with 100 trees, leveraging 60% of

the data for training. This choice is informed by A. Bellotti et al. (2019), who identified random forests as effective for LGD modelling for unsecured lending.

3.2.5. Basic decomposition approach

We decompose recovery rates into three stages: $R = R1 + R2 + R3$. This decomposition allows for independent modelling of each stage, expressed as $R = f(R1) + g(R2) + h(R3)$. This assumes independence among stages, enabling the use of separate models for R1, R2 and R3. If all three were modelled using OLS, the combined result will be the **Naïve model**.

3.2.6. Proof that decomposition works for OLS

To illustrate that this is true: given an unknown vector R, a matrix of independent variables X and their corresponding unknown coefficients, b, we have a simple system of linear equations:

$R = bX$ 3.1

The solution for this system under OLS is $b = (X'X)^{-1}X'R$ where X' is the transpose of matrix X and $(X'X)^{-1}$ is the matrix inverse of $X'X$. Given this, the error term ε is given as:

$\varepsilon = R - E(R)$, where $E(R) = (X'X)^{-1}X'RX$ 3.2

Similarly, the solution for each of $R1 = b_1X$, $R2 = b_2X$ and $R3 = b_3X$ is $(X'X)^{-1}X'(R1)X$, $(X'X)^{-1}X'(R2)X$ and $(X'X)^{-1}X'(R3)X$, respectively. Since by definition, $R = R1 + R2 + R3$, we have $E(R) = E(R1) + E(R2) + E(R3)$ and so ε would be:

$$\varepsilon = R - E(R1) - E(R2) - E(R3)$$

$$= R - (X'X)^{-1}X'(R1)X - (X'X)^{-1}X'(R2)X - (X'X)^{-1}X'(R3)X \quad \text{..... 3.3}$$

which is equivalent to:

$E = R - (X'X)^{-1}X'[R1 + R2 + R3]X$ 3.4

Since $R1 + R2 + R3$ can be simplified to R, we have shown that they share the same error term ε. This is true because $(X'X)^{-1}X'(R1 + R2 + R3)X = (X'X)^{-1}X'(R1)X + (X'X)^{-1}X'(R2)X + (X'X)^{-1}X'(R3)X$, which also means the coefficients from R1, R2 and R3, when added, is equal to the coefficient obtained for R.

We have now established a way to split R into three stages without compromising its accuracy.

3.2.7. Advanced decomposition approach

Recognising the potential for non-linear relationships, we propose advanced models that enhance the basic decomposition approach. For Stage 2, which often dominates over R and represents net collateral recovery, we introduce conditional modelling: R2 is expressed as a proportion of the remaining recovery, $S2 = \frac{R2}{1-R1}$. This approach assumes independence between R1 and S2, allowing for the prediction $E(R2) = E(S2)[1 - E(R1)]$.

We explore several modelling strategies, such as:

Modelling R2 directly. This was introduced as part of the basic decomposition model.

R2 as a conditional recovery rate. Another option is to express R2 as a proportion of what is left to be collected, $1 - R1$. We define $S2 = \frac{R2}{1-R1}$. To predict R2, we have $E(R2) = E(S2)[1 - E(R1)]$ under the assumption that R1 and S2 are independent.

Modelling recovery rate in 2-steps. One option is to look at the possibility of not being charged off (A1, A2 and A3 = 0) and only predict R1, R2, S2, and R3 when A1, A2, A2 and A3 are non-zero, respectively. We define non-zero counterparts for the dependent variables by adding an apostrophe (') to the end of the variable name (R1' for R1, S2' for S2, etc), so that, as an example, we have $E(R2) = P(A2 > 0)E(S2')[1 - E(R1)]$ for Stage 2 recovery rates using 2-step conditional recovery rate.

Modelling recovery rate using a simple machine learning model. Given that a known weakness of machine learning models is overfitting for LGD, we test whether there is improvement by modelling more elementary components through decomposition. Due to the complexity of a simple machine learning algorithm, convergence is only obtained with less sparse datasets. To ensure convergence, sampling had to be performed with replacement.

The machine learning model used is a random forest model with 100 trees, with 60% of the training data used in training the trees, inspired from the work of Bellotti et al. (2019). For notation purposes, we define recovery rates obtained using random forest as $E(R1_{RF})$, $E(R2_{RF})$, $E(R3_{RF})$ and $E(S2_{RF})$.

Exhausting alternatives, we have:

- $E(R1) = P(A1 > 0)E(R1')$ where R1' is R1 for non-zero A1,
- $E(R1) = E(R1_{RF})$,
- $E(R2) = P(A2 > 0)E(R2')$ where R2' is R2 for non-zero A2,

- $E(R2) = P(A2 > 0)E(S2')[1 - E(R1)]$ where S2' is S2 for non-zero A2,
- $E(R2) = E(R2_{RF})$
- $E(R2) = E(S2_{RF})[1 - E(R1)]$
- $E(R3) = P(A3 > 0)E(R3')$ where R3' is R3 for non-zero A3
- $E(R3) = E(R3_{RF})$

To predict R, one can mix and match any of these options. As an example, if one chooses the first option for modelling recovery rate for each stage, we end up with the basic decomposition model. If one chooses the first option for Stage 1 and Stage 3, and the 5th option for Stage 2, then we end up with $E(R) = E(R1) + E(S2_{RF})[1 - E(R1)] + E(R3)$.

We summarize the models and their corresponding names/notations in **Table 6**.

Table 6. *Summary of the model specifications*

Recovery stage	Model name/notations	Model description/specifications
Stage 1	#1	OLS R1
	#1.1	$P(A1 > 0)E(R1\|A1 > 0)$
	#1.2	$E(R1_{RF})$
Stage 2	#2	OLS R2
	#2.1	$P(A2 > 0)E(R2\|A2 > 0)$
	#2.2	$E(R2_{RF})$
	#2.3	$E(S2_{RF})[1 - \#1.1], S2 = \frac{A2}{1 - A1}$
	#2.4	$E(S2_{RF})[1 - \#1.2], S2 = \frac{A2}{1 - A1}$
Stage 3	#3	OLS R3
	#3.1	$P(A3 > 0)E(R3\|A3 > 0)$
	#3.2	$E(R3_{RF})$
Combined models	M1	OLS benchmark
	M2	2-step benchmark
	M3	Random forest benchmark
	M4	#1 + #2 + #3
	M5	#1.1 + #2.3 + #3.2
	M6	#1.2 + #2.4 + #3.2

This table summarizes the model specifications and their associated names or notations. OLS denotes the Ordinary Least Square approach, 2-step benchmark denotes the 2-step selection model proposed by Do et al. (2020), and RF denotes the machine learning random forest techniques.

One of the models explored above introduces a critical methodological assumption of independence between R1 and S2. While mathematically elegant, this assumption warrants careful scrutiny within the complex landscape of credit risk management. The independence premise allows for a simplified computational approach, enabling E(R2) to be calculated as $E(S2)[1 - E(R1)]$, but may potentially obscure interdependencies inherent in these variables especially during periods of market stress.

During economic downturns, the correlation between R1 and S2 may become more pronounced, potentially leading to systematic underestimation or overestimation of expected recovery rates. This suggests a need for more sophisticated modelling approaches that can capture the complex interdependencies of recovery mechanisms across different economic regimes and sector-specific contexts.

Empirical validation remains crucial. Researchers and risk managers should approach this assumption with caution, implementing rigorous sensitivity analyses and exploring alternative modelling techniques such as copula methods or conditional probability frameworks. While the proposed model provides a valuable analytical tool, it should be understood as a sophisticated approximation rather than an absolute representation of recovery dynamics.

Potential mitigation strategies include comprehensive correlation analysis, development of more sophisticated dependency modelling techniques, and extensive sensitivity testing across different economic regimes.

Ultimately, this modelling approach serves as a proof of concept, demonstrating the potential for recovery rate decomposition into R1, R2, and R3 components. The independence assumption, while acknowledged as a simplification, provides a foundational framework for understanding the complex mechanics of credit recovery. The primary value lies in the decomposition methodology, which offers a novel approach to breaking down and analysing recovery rates, despite the recognised limitations of the current assumptions.

3.3. Empirical results

In this section, we compare predictive performance among alternative models introduced in section 3.2 and with the benchmark models.

3.3.1. Benchmark models 1 and 2: naïve and 2 step model

In predicting uncured recovery rate, one of the simplest model structures is a linear regression model or ordinary least squares (OLS). Here, the expected value of uncured recovery rate R is modelled as:

$$E(R) = f(X) \quad \text{..... 3.5}$$

where f is a generic linear combination of independent variables that explains R. Theoretically, the quality of prediction from a linear regression model increases with more independent nominated drivers, but too many drivers come at a risk of overfitting/over specification. On the other hand, a new generation of models with mathematical structures like multi-step conditional approaches came about after exploration of new variables in predicting LGD have slowed down. Several iterations were explored to accommodate the multimodality of LGD. Leow et al. (2014) built separate models for probability of repossession and a 'haircut model' for repossessed accounts and this multi-step model has become standard.

Following similar spirit, we consider the benchmark model 2 by following Do et al. (2020) to build a two-step conditional framework consisting of probability of positive (i.e., greater than 0) recovery and expected value of positive recovery. The expected recovery is, therefore, can be constructed as:

$$E(R) = P(A > 0)E(R|A > 0) \quad \text{..... 3.6}$$

where,

$A = R \times EAD$, and P(A>0) is the probability that A>0.

Table 7 shows the standardised in-time coefficient estimates for R using a simple linear regression. We find five main factors for benchmark model 1: CLTV, DLTVCR, LLTVCR, LOB and MIP. The effects of five continuous variables align with Table 1 presented in the main manuscript except for CLTV, which now shows a negative relationship. This may be MIP controlling for MI related effects on R especially for high CLTV cases, leaving CLTV to explain what is left. Besides, CLTV has weaker influence on uncured recoveries relative to how much CLTV has changed from origination to default, and from default to collateral liquidation.

Table 7. *Standardised regression coefficients for R using Benchmark Models 1 and 2*

Parameter	M1: E(R)	M2: P(A>0)	M2: E(R')
Intercept	0.6361***	2.9051***	0.6541***
	(0.0007)	(0.0247)	(0.0007)
CLTV	-0.0714***	-0.1906***	-0.0702***
	(0.0012)	(0.0142)	(0.0009)
DLTVCR	-0.1195***	-0.2012***	-0.1171***
	(0.0009)	(0.0196)	(0.0009)
LLTVCR	-0.0979***	-0.2101***	-0.0938***
	(0.0008)	(0.0121)	(0.0007)
DTIO	-0.0027***	0.002	-0.0029***
	(0.0007)	(0.0104)	(0.0007)
DICR	0.0164***	0.134***	0.014***
	(0.0008)	(0.0134)	(0.0007)
LICR	0.0176***	0.1547***	0.0144***
	(0.0007)	(0.0149)	(0.0007)
FICO	0.0058***	0.034***	0.0059***
	(0.0008)	(0.0116)	(0.0007)
LC	0.0096***	0.0349*	0.0069***
	(0.002)	(0.0199)	(0.0015)
LOB	0.1373***	0.9504***	0.1107***
	(0.001)	(0.0167)	(0.0009)
MOB	-0.0335***	-0.1688***	-0.0275***
	(0.0011)	(0.0126)	(0.0009)
MIP	0.0919***	0.2588***	0.087***
	(0.0009)	(0.0162)	(0.0009)
TID	0.0023**	-0.0009	0.002**
	(0.0011)	(0.0121)	(0.0009)
TTR	-0.0506***	-0.2862***	-0.042***
	(0.001)	(0.011)	(0.0008)
LP_C	-0.0161***	-0.0598***	-0.0141***
	(0.001)	(0.0131)	(0.0008)
LP_P	0.0394***	0.1553***	0.0379***
	(0.001)	(0.0171)	(0.0009)
NB_01	-0.0127***	-0.0654***	-0.0123***
	(0.0007)	(0.0122)	(0.0007)
PT_CO	0.0023***	0.1532***	-0.0034***
	(0.0007)	(0.0132)	(0.0007)
PT_PU	0.026***	0.2692***	0.025***
	(0.0006)	(0.031)	(0.0007)
FHB	0.0006	-0.0186	0.001
	(0.0008)	(0.0149)	(0.0007)
SNS	-0.0004	-0.0264**	0.0001
	(0.0007)	(0.0119)	(0.0007)
OO	0.035***	0.1412***	0.0271***
	(0.0008)	(0.0086)	(0.0007)
MF	0.0061***	0.0556***	0.0051***
	(0.0008)	(0.0125)	(0.0007)

Parameter	M1: E(R)	M2: P(A>0)	M2: E(R')
NJF	0.0185***	0.0864***	0.0168***
	(0.0008)	(0.0107)	(0.0007)
NSRR	-0.0098***	0.0005	-0.0091***
	(0.0007)	(0.0114)	(0.0007)
PDJ	-0.0064***	-0.0097	-0.0068***
	(0.0008)	(0.0099)	(0.0007)
INT	-0.0106***	-0.091***	-0.0068***
	(0.001)	(0.0111)	(0.0008)
UMP	-0.0284***	-0.0906***	-0.0263***
	(0.0009)	(0.0131)	(0.0008)

*This table reports the average coefficients obtained for models under the two benchmark models based on 10 rounds of random sampling without replacement. Level of coefficient significant as follows: *** for p-value < 0.01, ** for p-value < 0.05 and * for p-value < 0.1. Values come in the following order: coefficient, degree of significance/ standard error. The estimates are based on in-time training sample. M1 is chosen as a benchmark model due to its popularity in both the literature and industry. M2 is the second benchmark model which is a slightly modified version from Do et al. (2020). It has been chosen for being relatively recent, simple and for claiming outperformance against OLS.*

Figure 3 shows that R decreases with CLTV until CLTV reaches 70-80%, at which point R starts to increase again up to a certain point. The reason behind the change in direction for R is the presence of MI.

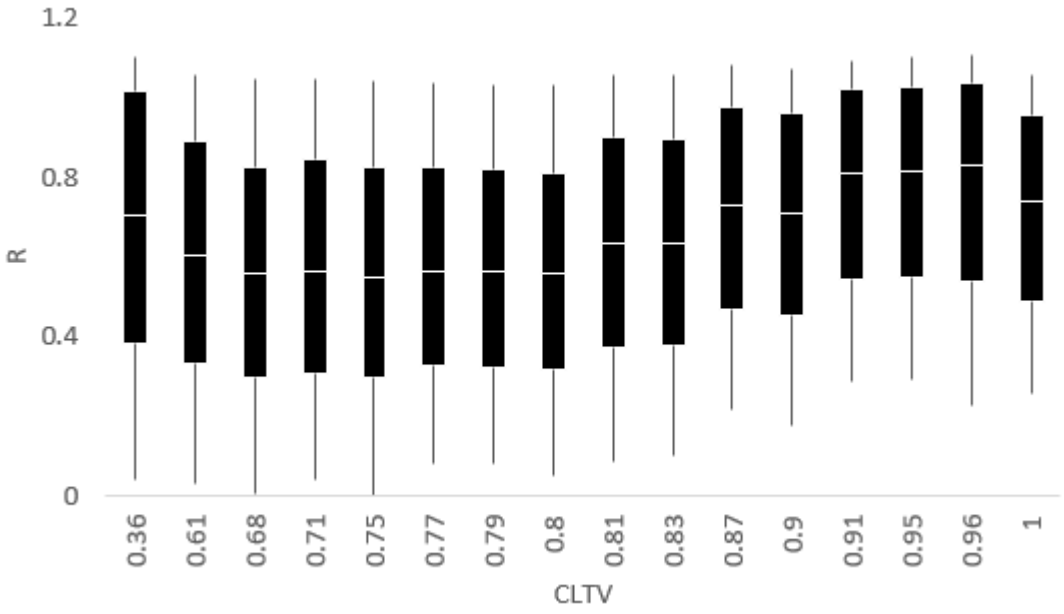

Figure 3. *R distribution through CLTV domain*

This is an illustration using the Freddie Mac dataset. Vertical axis denotes the domain of the distribution of R, and horizontal axis denotes values of CLTV. Each box and whiskers plot represents a distribution of R where the top and bottom lines are the 5% and 95% percentile, the box contains values within the 20% and 80% percentile, and the middle line (median) is the 50% percentile.

All DTI-related variables have very weak influence on uncured recovery. While intuition suggests that DTI at origination may affect uncured recovery, this is most likely due to the correlation of debt with EAD. Having LOB as control diminishes the predictive power of DTI variables. DICR and LICR, on the other hand yield immaterial and/or positive estimates. This is likely because actual DTI (if known) would have had a negative coefficient, and default would have been caused by an income shock. Unfortunately, this information is not in the data set.

FICO score at origination returns a small positive relationship with uncured recovery because borrowers with low DTI at origination are likely to have high FICO scores. This is demonstrated in **Figure 4**. While statistically significant, the immateriality of this effect may be explained by the irrelevance of FICO score recorded at origination. Also, the FICO score is known to be a determinant of PD and not LGD. Although they may at times be correlated, other more relevant factors like DLTVCR and LLTVCR carry most influence on R.

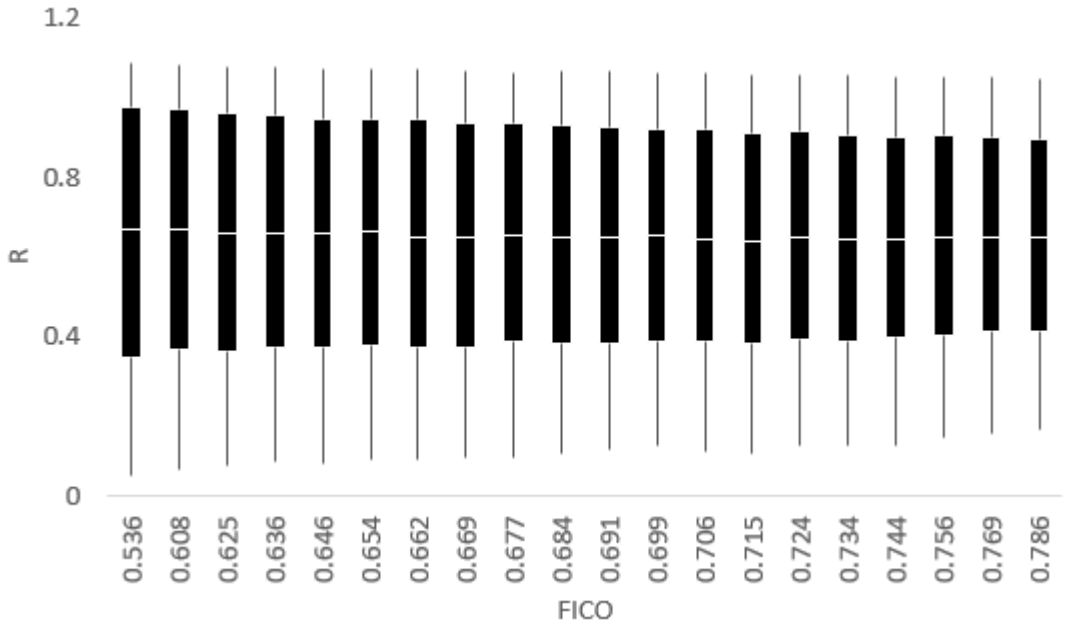

Figure 4. *R distribution through FICO domain*

This is an illustration using the Freddie Mac dataset. Vertical axis denotes the domain of the distribution of R, and horizontal axis denotes values of CLTV. Each box and whiskers plot represents a distribution of R where the top and bottom lines are the 5% and 95% percentile, the box contains values within the 20% and 80% percentile, and the middle line (median) is the 50% percentile.

LC yields a weak but counterintuitive estimate by having positive relationship with recovery. This is likely due to the presence of non-linearity in their relationship. This was highlighted by Do et al. (2020) when they used a spline-knot technique to address non-linearity.

LOB has one of the highest influences on uncured recovery R. As mentioned above, this is likely due to prioritisation to sell assets securing loans with higher outstanding balances. **Figure 5** shows average R follows a slight parabolic pattern of increasing initially as origination balance increases until around $100k (around before LOB of 11.5), followed by a slight decrease as uncertainty further decreases. This shows that lenders prioritise recovery from bigger loan sizes at origination, consistent with the findings of Calem & LaCour-Little (2004); Do et al. (2020); Pennington-Cross (2003), that the relationship of loan size and non-zero LGD generally has a convex parabola (concave for recovery).

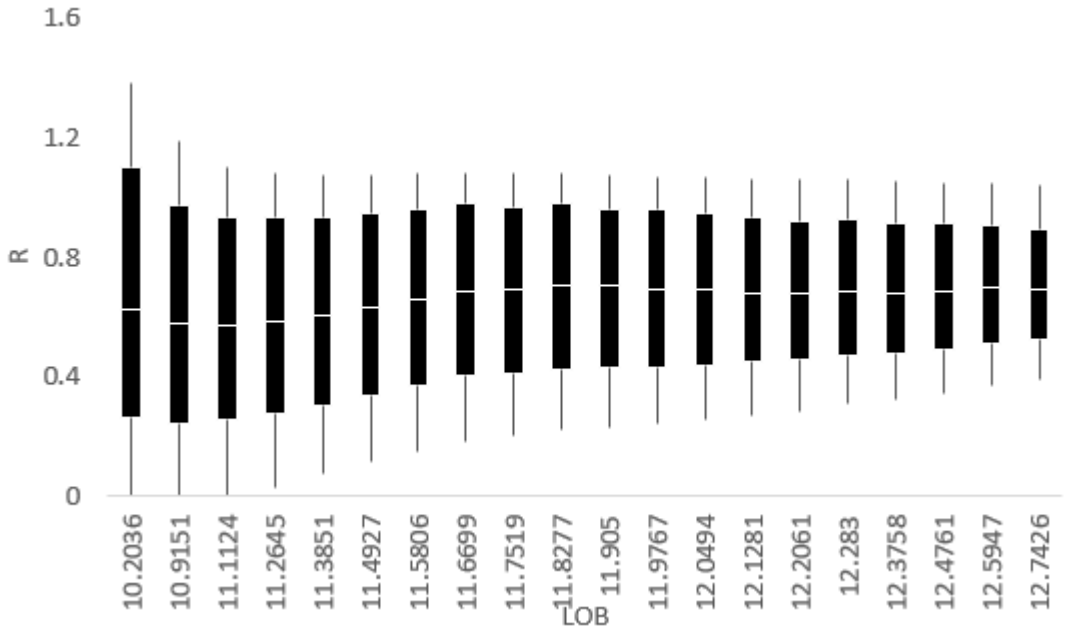

Figure 5. *R distribution through LOB domain*

This is an illustration using the Freddie Mac dataset. Vertical axis denotes the domain of the distribution of R, and horizontal axis denotes values of CLTV. Each box and whiskers plot represents a distribution of R where the top and bottom lines are the 5% and 95% percentile, the box contains values within the 20% and 80% percentile, and the middle line (median) is the 50% percentile.

While MOB (loan age at default) was highlighted by Do et al. (2018); Qi (2013) as having a positive relationship with recovery rates, **Figure 6** shows that mean R gradually decreases until

3 years (MOB = 36 months) before increasing as an account is older at the time of default. This parabolic relationship ends up with a net negative coefficient using the Freddie Mac dataset, likely due to a greater concentration of younger accounts (<36 months old) at default.

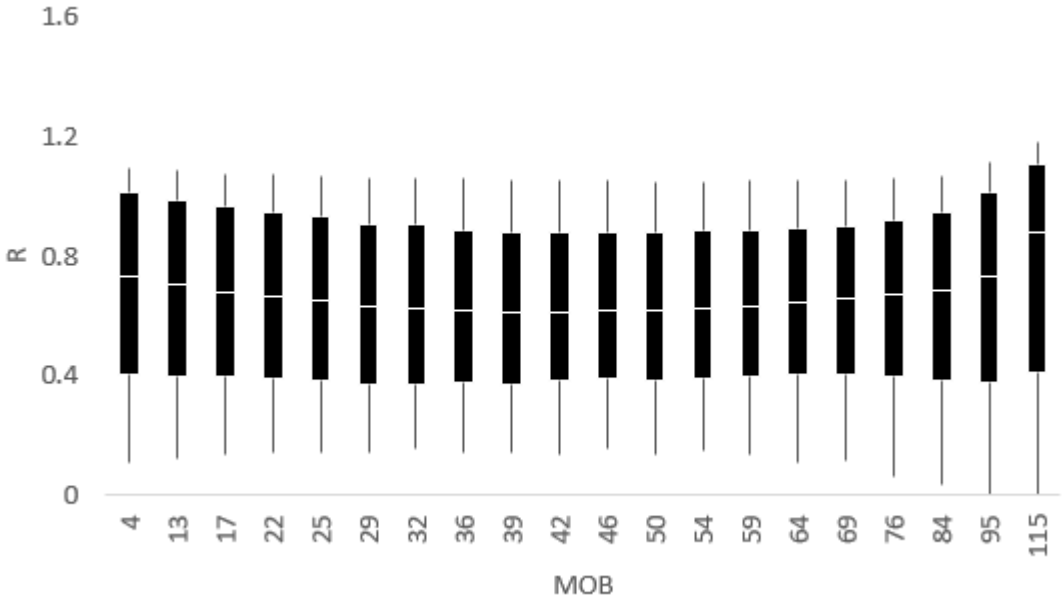

Figure 6. *R distribution through MOB domain*

This is an illustration using the Freddie Mac dataset. Vertical axis denotes the domain of the distribution of R, and horizontal axis denotes values of CLTV. Each box and whiskers plot represents a distribution of R where the top and bottom lines are the 5% and 95% percentile, the box contains values within the 20% and 80% percentile, and the middle line (median) is the 50% percentile.

Positive coefficients of MIP point to the existence of insurance. The higher the insurance coverage, the higher chance that more of EAD may be recovered.

TID shows a weak positive relationship with R for M1 and E(R') for M2 while it is negative for P(A>0). The interpretation for this is the chance of getting anything back is lower for accounts that have delinquency but for cases when recoveries are obtained, it is often higher as TID becomes higher. In **Figure 7**, recovery is significantly lower for accounts that were 2 months delinquent at default. After this, the relationship is relatively flat. This continues until TID gets to 5 years where both mean R and the level of uncertainty start to increase, which may reflect the impact of outliers given significantly lower volume of accounts with high TID. Given the immaterial relationship, this will not be explored further in this paper.

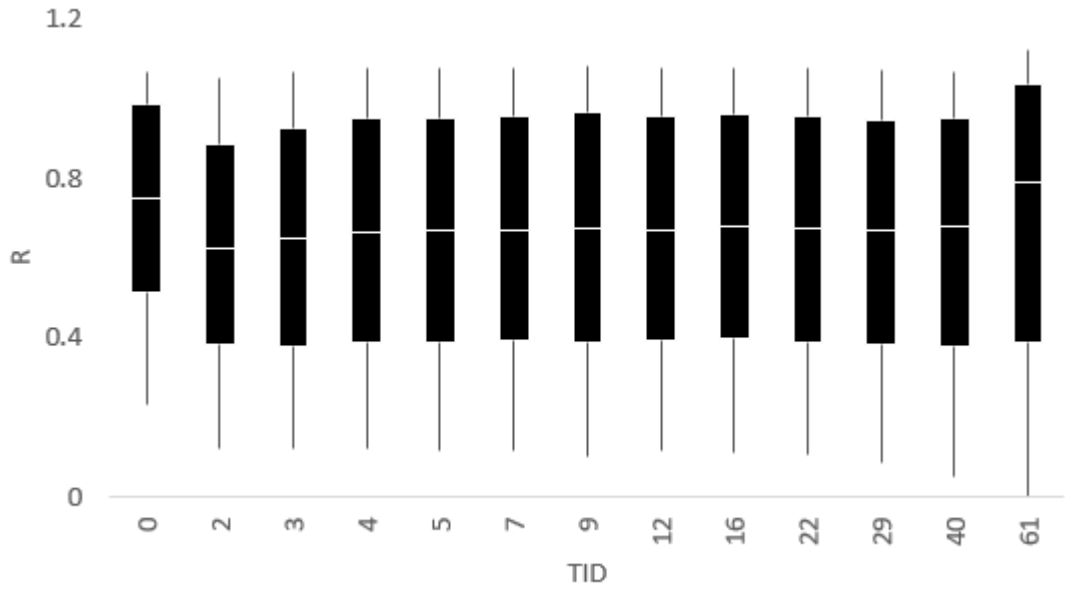

Figure 7. *R distribution through TID domain*

This is an illustration using the Freddie Mac dataset. Vertical axis denotes the domain of the distribution of R, and horizontal axis denotes values of CLTV. Each box and whiskers plot represents a distribution of R where the top and bottom lines are the 5% and 95% percentile, the box contains values within the 20% and 80% percentile, and the middle line (median) is the 50% percentile.

TTR possibly represents the amount of fees and interest charged from default until liquidation of collateral. Analysis shows that both low and high (>4 years) TTR yields high recoveries tend to decrease first when approaching 4 years, demonstrating a nonlinear (parabolic) relationship. A simplistic view ends up with a net negative relationship which may highlight that the amount of expenses increases as the time taken to reach resolution increases.

LP_C and LP_P implies that refinanced loans with cash-out yield a lower recovery after the property has been liquidated compared to a loan issued for purchase of property (no cash out). This makes sense because of the nature of cash-out where EAD increases without necessarily having a higher collateral value.

Having one borrower compared to multiple borrowers ends up with lower recovery amount. This may be attributed to the amount of care allotted by the occupant to the property or the amount of support between co-borrowers, although the estimate is not material. Looking at the data, the average uncured recovery rate drops by a small amount with more than one borrower.

Certain property types yield higher uncured recovery like planned unit developments (more commonly known as subdivisions), followed by manufactured housing, co-op and condominiums, when looking at the first benchmark model. This is also supported under the second benchmark model – single family housing has the least possibility of having non-zero recovery, but manufactured housing, co-op and condominiums yield the least non-zero uncured recovery amount, as these tend to have more buildings and improvements rather than land, which depreciate even during an economic boom.

Owner occupied (OO) properties generally are sold at a higher price given owner occupiers usually take more care of properties; Clauretie & Daneshvary (2011); Do et al. (2020); Qi (2013); Qi & Yang (2009) have also observed this.

Accounts that undergo debt consolidation (modified) have been observed to yield slightly higher recovery, meaning that the strategy of lenders in modifying defaulted loans by debt consolidating works.

For states with non-judgemental foreclosures, there may be savings in fees that drive recovery higher, but properties in states with no statutory right of redemption may have been left in worse condition than those with statutory right of redemption. Lastly, states that prohibit deficiency judgement yield slightly less uncured recovery. Both NJF and PDJ roughly align with Clauretie & Herzog (1990); Do et al. (2020), however NSRR shows a net negative effect, which does not align with Do et al. (2020). Our result does not align with Do et al. (2020) for coefficient of P(A>0), but in that paper, the design was P(loss = 0) which is slightly different.

There is comfort in the fact that NJF and PDJ align roughly with the results of Do et al. (2020), especially for the non-zero recovery, and also in the fact that the coefficient for NSRR is less material than the former two given that NSRR was not aligned with the literature.

Mortgage interest rate at default yielded a negative coefficient which makes sense as higher interest rates hurt customers. From this, it may be hypothesised that customers who struggle to keep repayments may take less care of their properties and make fewer early repayments. This relationship between recovery rate and INT is aligned with the findings of Do et al. (2018) where non-zero LGD has a positive relationship with current interest rate.

Coefficients for FHB is not statistically significant under either benchmark model while SNS highlights yet another non-linear relationship with R: the chance that servicers who are not

originators will get non-zero recoveries is lower, but the actual recovery rate for non-zero cases are higher for these entities.

Aside from those highlighted, there is no material difference between benchmark model M1 and M2.

3.3.2. Alternative models for individual stage of recoveries

3.3.2.1. Modelling for Stage 1

The model for unconditional recovery R1 (#1) is given by:

$$E(R1) = f(X) \quad \text{..... 3.7}$$

where f is a generic linear combination of the matrix of independent variables X containing information defined in **Table 8** and R1 is as defined in **Table 4**.

Table 8. *Table of abbreviations*

Abbreviation	Definition
AUC	Area Under Curve
C&I	Commercial and Industrial
CECL	Current Expected Credit Losses
CLTV	Cross collateralised loan to value ratio at origination
COVID-19	N-Coronavirus 19
CT	Conditional two-step model
DDTI	Default Debt-to-Income Ratio
DICR	Default DTI Change Ratio
DLTV	Default loan to value ratio
DLTVCR	Default LTV Change Ratio
DRD	Portfolio default rate at default
DSCR	Debt Service Coverage Ratio
DTI	Debt-to-Income Ratio
DTI0	Origination Debt-to-Income Ratio
EAD	Exposure at Default
ECL	Expected Credit Losses
FHB	First Home Buyer
FHFA	Federal Housing Finance Agency
FICO	Fair Isaac Corporation Score
GAAP	Generally Accepted Accounting Principles
GFC	Global Financial Crisis
GPI	Global Peace Index
GSE	Government Sponsored Enterprise
HPI	House Price Index
IAS 39	International Accounting Standards, 39th section
IFRS 9	International Financial Reporting Standards, 9th chapter
LC	Liquidity Constraint
LGD	Loss Given Default
LIBOR	London Interbank Offered Rate
LICR	Liquidation DTI Change Ratio

Abbreviation	Definition
LIED	Loss in Event of Default
LLTVCR	Liquidation LTV Change Ratio
LOB	Log of Origination Balance
LP	Loan Purpose
LTV	Loan to Value ratio
MF	Modification Flag
MI	Mortgage Insurance
MIP	MI Percentage
MOB	Months on Book
MSA	Metropolitan Statistical Area
NB	Number of Borrowers
NJF	Non-judicial Foreclosure
NSRR	No Statutory Rights of Redemption
OLS	Ordinary Least Squares
OO	Owner Occupied
P2p	Peer-to-peer
PD	Probability of Default
PDJ	Prohibited Deficiency Judgement
PT	Property Type
RCM	Resolution Correctness Measure
REO	Real Estate Owned
RMSE	Root Mean Squared Error
RWA	Risk Weighted Asset
SICR	Significant Increase in Credit Risk
SNS	Originator (Seller)-not-Servicer
SSE	Sum of Squared Errors
TID	Time in Delinquency
TTR	Time to Resolution
UMP	Unemployment
UPB	Unpaid Principal Balance
US	United States

On the other hand, the model for 2-step unconditional recovery R1 (#1.1) is given by:

$$E(R1) = P(A1 > 0)E(R1|A1 > 0) = P(A1 > 0)E(R1') \quad \text{..... 3.8}$$

where E(R1') is a linear combination of independent variables, P(A1>0) is probability that A1>0. Following Do et al. (2020) in their 2 step model, the structure for modelling R1 in 2 steps allows the correlation matrix of the error terms to be non-identity. This is similar to M2 (benchmark model 2).

Table 9 highlights that R1 is mainly driven by: TTR, MF and LLTVCR.

Table 9. *Standardised regression coefficients for modelled Stage 1 recovery rates*

Parameter	#1: E(R1)	#1.1: P(A1>0)	#1.1: E(R1')
Intercept	0.0043***	-1.1408***	-0.0365***
	(0)	(0.0074)	(0.0003)
CLTV	-0.0005***	-0.0632***	-0.0021***
	(0.0001)	(0.0065)	(0.0002)
DLTVCR	-0.0011***	-0.1637***	-0.0053***
	(0.0001)	(0.0073)	(0.0002)
LLTVCR	-0.0026***	-0.1149***	-0.0037***
	(0.0002)	(0.0049)	(0.0002)
DTIO	-0.0001***	-0.0015	-0.0001
	(0.0001)	(0.0047)	(0.0002)
DICR	0	0.0057	0.0002
	(0.0001)	(0.0053)	(0.0002)
LICR	0.0004***	0.0458***	0.0015***
	(0.0001)	(0.0049)	(0.0002)
FICO	-0.0003***	-0.0891***	-0.0028***
	(0.0001)	(0.005)	(0.0002)
LC	-0.0001	0.023**	0.0009**
	(0.0001)	(0.0111)	(0.0004)
LOB	-0.0005***	-0.049***	-0.0017***
	(0.0001)	(0.0059)	(0.0002)
MOB	-0.001***	-0.2003***	-0.0063***
	(0.0001)	(0.0064)	(0.0002)
MIP	-0.0001**	0.0095	0.0003
	(0.0001)	(0.0062)	(0.0002)
TID	0.0009***	0.1865***	0.0059***
	(0.0001)	(0.006)	(0.0002)
TTR	0.0082***	0.6864***	0.0223***
	(0.0001)	(0.0052)	(0.0002)
LP_C	0	0.0039	0.0001
	(0.0001)	(0.006)	(0.0002)
LP_P	0.0004***	0.0417***	0.0014***
	(0.0001)	(0.0065)	(0.0002)
NB_01	-0.0001***	0.0074	0.0002
	(0.0001)	(0.0048)	(0.0002)
PT_CO	0	-0.0413***	-0.0013***
	(0)	(0.0052)	(0.0002)
PT_PU	0.0002***	-0.001	0
	(0)	(0.0051)	(0.0002)
FHB	0	0	0
	(0.0001)	(0.0054)	(0.0002)
SNS	0.0002***	0.0307***	0.001***
	(0)	(0.0048)	(0.0002)
OO	0.0001***	0.0248***	0.0009***
	(0.0001)	(0.005)	(0.0002)
MF	-0.0042***	-1.1025***	-0.0355***
	(0.0001)	(0.0181)	(0.0006)

Parameter	#1: E(R1)	#1.1: P(A1>0)	#1.1: E(R1')
NJF	0.0016***	0.1549***	0.005***
	(0.0001)	(0.0048)	(0.0002)
NSRR	0.0001**	0.0164***	0.0005***
	(0.0001)	(0.0047)	(0.0001)
PDJ	0	0.002	0.0001
	(0.0001)	(0.0046)	(0.0001)
INT	-0.0007***	-0.0778***	-0.0025***
	(0.0001)	(0.0058)	(0.0002)
UMP	-0.0001**	0.0115**	0.0005**
	(0.0001)	(0.0058)	(0.0002)

*This table reports the average coefficients obtained for models under Stage 1 based on 10 rounds of random sampling without replacement. Level of coefficient significant as follows: *** for p-value < 0.01, ** for p-value < 0.05 and * for p-value < 0.1. Values come in the following order: coefficient, degree of significance/ (standard error). The estimates are based on in-time training sample. 2nd column is for model #1, 3rd and 4th columns contain estimates for each of the two steps under #1.1.*

All LTV variables end up with negative coefficients. This means that the less secured the defaulted loan has become after default, the lower R1 is. This may be attributed to willingness of borrower to pay. Collateral whose value has decreased relative to outstanding loan amount might affect a borrower's willingness to pay, given that the property must have been bought at a price higher than what it is worth at the time of default. LLTVCR is the one leads in terms of LTV variables, because R1 takes place between default date and liquidation date, when applicable. Among all LTV variables, LLTV also carries the most recent information.

LTV holds information on the loan (this amount is relatively stable and rarely experiences shocks in value) and collateral value (where systematic shocks are captured by the FHFA HPI), ensuring LLTV is always relevant. Unfortunately for DTI variables, the DTI index mostly carries information for non-defaulted borrowers, and its use may be ignoring accounts which defaulted because of high actual DTI experiencing unknown income shocks. As such, results obtained for the DTI variables may be less meaningful than they appear.

It may be the same for the FICO score coefficient which has an immaterial negative value, given that FICO score is last updated in origination and so loses its predictive value through a loan's lifecycle.

High origination balances (LOB) yield lower recovery under Stage 1. This may be the other side of the picture for lenders' prioritisation. High origination balances usually mean high default balances which may prove riskier when foreclosure procedures are not immediately exhausted. High origination balances may also contain higher origination LTV which gives

borrowers less hope of being able to pay a mortgage off. Given more drastic measures are about to be at play, this may affect the willingness of borrowers to make early payments.

Accounts that default later in a loan's lifecycle tends to yield lower recovery, possibly due to more options being exhausted before finally defaulting. As a result, fewer older accounts default than younger ones. On the other hand, the longer a defaulted account spent in delinquency, the more the borrower gets used to this status and so more early recoveries are obtained.

A positive TTR is trivial – the longer an account spends between default to resolution, the more chances the lender can recover something prior to property liquidation. This is either due to the effort spent or expectation from the perspective of the lender (i.e., which accounts need to be foreclosed right away because there is very small chance of early recovery).

Accounts that are modified by debt consolidation yields significantly lower R1. The presence of a highly negative coefficient for MF shows that debt consolidation efforts are usually not related to efforts of rehabilitation.

NJF is shown to affect R1 positively, which means that defaulted borrowers with collateral in states with no required court rulings to perform foreclosure may still be making efforts, meeting the lender halfway to pay off some parts of the EAD. This may be because the defaulted borrower intends to keep the collateral.

INT ends up with a negative coefficient which may reflect the willingness of defaulted borrowers to repay given higher interest.[4]

The main differences between 1 and 2-step R1 are that DLTVCR is more material than LLTCCR for P(A1>0) and MF has an even significantly higher estimate for P(A1>0). This implies that the chance at a non-zero R1 is highly dependent on the MSA-level house price index falling at default and the absence of modifications due to debt consolidation.

In general, machine learning models are often challenging to interpret. As such, results of the random forest model are not discussed.

[4] Coefficients for LC, MIP, PT_CO, PT_PU, LP_C, LP_P, NB_01, FHB, OO, NSRR, UMP and PDJ are either statistically insignificant or immaterial.

3.3.2.2. Modelling for Stage 2

To account for the different alternatives of E(R2) presented in M4-M6 under 0, we introduce the following definitions.

Modelling R2 (#2) directly may be the simplest alternative to estimating E(R2).

$$E(R2) = f(X) \quad \text{..... 3.9}$$

where f is a linear combination of independent variables.

Modelling unconditional recovery for Stage 2 under 2- steps (#2.1) means predicting the probability that A2 > 0 and R2'. First, we estimate the probability that A2>0, and then we proceed to estimate R2 in the case that A2 is greater than 0.

$$E(R2) = P(A2 > 0)E(R2|A2 > 0) = P(A2 > 0)E(R2') \quad \text{..... 3.10}$$

As previously mentioned, cases where A2 <= 0 are those that went through charge-off.

where E(R2') is a linear combination of independent variables that explains R2', P(A2>0) is the probability that A2>0, which can be interpreted as the probability of non-charge-off. Following the 2-step model of Do et al. (2020), the structure for modelling R3 in 2 steps also allows for the correlation of the error terms for both terms to be a non-identity matrix.

Modelling the 2-step conditional recovery for Stage 2 (#2.2) means predicting S2, the proportion of recovery under Stage 2 expressed as a proportion of what's left to be recovered after Stage 1, through estimating E(S2).

$$E(S2) = P(A2 > 0)E(S2|A2 > 0) = P(A2 > 0)E(S2') \quad \text{..... 3.11}$$

where E(S2') is a linear combination of independent variables that explains S2'. Following the 2-step model of Do et al. (2020), the structure for modelling S2 in 2 steps also allows for the correlation of the error terms for both terms to not be an identity matrix.

Solving back for E(R2) we get $E(R2) = P(A2 > 0)E(S2')[1 - E(R1)]$, where E(R1) is the estimate obtained from predicting R1 using any model defined under Stage 1.

In **Table 10** we find that R2 coefficients are similar to those of M1, with a few exceptions: MIP, which has turned negative from a high positive in M1, TID, which has become irrelevant for Stage 2, and SNS, which has become positive for Stage 2. Among these, only MIP has a material coefficient.

Table 10. *Standardised regression coefficients for Stage 2 recovery rates*

Parameter	#2: E(R2)	#2.1: P(A2>0)	#2.1: E(R2')	#2.2: P(A2>0)	#2.2: E(S2')
Intercept	0.5415***	2.6282***	0.5674***	2.6281***	0.5702***
	(0.0007)	(0.0193)	(0.0007)	(0.0193)	(0.0007)
CLTV	-0.0718***	-0.1965***	-0.069***	-0.1964***	-0.0698***
	(0.0012)	(0.013)	(0.0009)	(0.013)	(0.0009)
DLTVCR	-0.1222***	-0.2381***	-0.1176***	-0.2388***	-0.1188***
	(0.0009)	(0.0166)	(0.0009)	(0.0166)	(0.0009)
LLTVCR	-0.1016***	-0.2394***	-0.0957***	-0.2401***	-0.098***
	(0.0009)	(0.0101)	(0.0007)	(0.0101)	(0.0007)
DTIO	-0.0033***	0.0103	-0.0042***	0.0103	-0.0043***
	(0.0007)	(0.0089)	(0.0007)	(0.0089)	(0.0007)
DICR	0.0172***	0.1172***	0.0148***	0.1172***	0.0148***
	(0.0008)	(0.0113)	(0.0007)	(0.0113)	(0.0007)
LICR	0.0195***	0.1623***	0.0154***	0.1622***	0.0157***
	(0.0007)	(0.0125)	(0.0007)	(0.0125)	(0.0007)
FICO	0.006***	0.0424***	0.0062***	0.0424***	0.006***
	(0.0008)	(0.0099)	(0.0007)	(0.0099)	(0.0007)
LC	0.0067***	0.0547***	0.0021	0.0548***	0.002
	(0.002)	(0.0174)	(0.0015)	(0.0174)	(0.0016)
LOB	0.1457***	0.9914***	0.1096***	0.9915***	0.1097***
	(0.001)	(0.0142)	(0.0009)	(0.0142)	(0.0009)
MOB	-0.0272***	-0.1647***	-0.0193***	-0.1648***	-0.02***
	(0.0011)	(0.0109)	(0.0009)	(0.0109)	(0.0009)
MIP	-0.0204***	-0.0433***	-0.0208***	-0.0434***	-0.021***
	(0.0009)	(0.0126)	(0.0009)	(0.0126)	(0.0009)
TID	0	-0.0132	0.0007	-0.0131	0.0014
	(0.0011)	(0.0103)	(0.0009)	(0.0103)	(0.0009)
TTR	-0.0632***	-0.3303***	-0.0501***	-0.3299***	-0.0444***
	(0.001)	(0.0094)	(0.0008)	(0.0094)	(0.0008)
LP_C	-0.0154***	-0.0598***	-0.0131***	-0.0597***	-0.0132***
	(0.001)	(0.0114)	(0.0008)	(0.0114)	(0.0009)
LP_P	0.0394***	0.1596***	0.0373***	0.1595***	0.0377***
	(0.001)	(0.0138)	(0.0009)	(0.0138)	(0.0009)
NB_01	-0.0124***	-0.0663***	-0.0117***	-0.0663***	-0.0118***
	(0.0007)	(0.0103)	(0.0007)	(0.0103)	(0.0007)
PT_CO	0.0032***	0.1499***	-0.0048***	0.1498***	-0.0048***
	(0.0007)	(0.0104)	(0.0007)	(0.0104)	(0.0007)
PT_PU	0.0278***	0.2748***	0.0258***	0.2745***	0.026***
	(0.0006)	(0.0241)	(0.0007)	(0.0241)	(0.0007)
FHB	0.0005	-0.0242**	0.0008	-0.0241**	0.0008
	(0.0008)	(0.0112)	(0.0007)	(0.0112)	(0.0007)
SNS	0.0018**	-0.0189*	0.0022***	-0.019*	0.0023***
	(0.0007)	(0.01)	(0.0007)	(0.01)	(0.0007)
OO	0.0348***	0.1523***	0.025***	0.1524***	0.0251***
	(0.0008)	(0.0077)	(0.0007)	(0.0077)	(0.0007)
MF	0.0122***	0.0785***	0.0098***	0.0776***	0.007***
	(0.0008)	(0.0105)	(0.0007)	(0.0104)	(0.0007)

Parameter	#2: E(R2)	#2.1: P(A2>0)	#2.1: E(R2')	#2.2: P(A2>0)	#2.2: E(S2')
NJF	0.0216***	0.114***	0.0185***	0.1142***	0.0198***
	(0.0008)	(0.0091)	(0.0007)	(0.0091)	(0.0007)
NSRR	-0.0107***	0.0223**	-0.01***	0.022**	-0.0099***
	(0.0007)	(0.0097)	(0.0007)	(0.0097)	(0.0007)
PDJ	-0.0069***	-0.0137	-0.0074***	-0.0137	-0.0074***
	(0.0008)	(0.0085)	(0.0007)	(0.0085)	(0.0007)
INT	-0.0133***	-0.0772***	-0.0088***	-0.0772***	-0.0094***
	(0.001)	(0.0098)	(0.0008)	(0.0098)	(0.0008)
UMP	-0.031***	-0.0994***	-0.028***	-0.0994***	-0.028***
	(0.0009)	(0.0111)	(0.0008)	(0.0111)	(0.0008)

*This table reports the average coefficients obtained for models under Stage 2 based on 10 rounds of random sampling without replacement. Level of coefficient significant as follows: *** for p-value < 0.01, ** for p-value < 0.05 and * for p-value < 0.1. Values come in the following order: coefficient, degree of significance/ (standard error). The estimates are based on in-time training sample. 2nd column is for model #2, 3rd and 4th columns contain estimates for each of the two steps under #2.1, then final 5th and 6th columns for #2.2.*

The negative coefficients from MIP mean the existence of insurance and the percentage coverage have a negative effect on property sale price. This may be caused by some biases when selling the collateral since both the homeowner and the lender know that they can claim on shortfalls.

DTIO, on the other hand, turns out to be statistically insignificant for P(A2>0) for both 2-step models and still holds negative coefficients for E(R2), E(R2') and E(S2') which makes more intuitive sense. Immateriality may be attributed to relevance of information given this variable is last updated at origination.

Aside from those highlighted by benchmark models M1 and M2, #2.1: P(A2>0), shows PT_CO and PT_PU to have slightly higher coefficients. This points to the presence of collateral meaning there is less chance for charging off, while E(R2') and E(S2') still end up with less recovery amount for properties like condominiums.

#2.2 shows a similar set of parameter coefficients for S2', and its equivalent P(A2>0) is very similar to that of R2 in #2.1. In aggregate, one difference is with TTR which loses some predictive power when R2 becomes S2 by taking out the effect of R1 when comparing E(S2') vs E(R2').

We see from the previous sections that R2 is driven mainly by the same drivers as that of R in M1, except for MIP which possibly contains biases of not having to sell collateral at a higher price due to the presence of mortgage insurance. While the structure changes with each option, only the level of accuracy changes and not the overall message from the coefficients. TTR has been a weak predictor for R2, and by taking out components unrelated to property and property

disposition, it only became weaker. Note that it is not completely irrelevant because some expenses may be positively correlated with TTR.

Machine learning models are often challenging to interpret. As such, results of the random forest model are not discussed.

From all the models under Stage 2, we end up with the 2-step model for R2' (#2.1) which is the best model so far for each of the 10 independent rounds of sampling performed.

3.3.2.3. Modelling for Stage 3

This section explores options for modelling recovery rate under Stage 3. Modelling R3 directly may be the simplest alternative to estimating E(R3):

$$E(R2) = f(X) \quad \text{..... 3.12}$$

where f is a linear combination of independent variables that explains R3.

For #3.1, modelling R3 in two steps may help improve model accuracy because there is a significant number of defaulted loans where A3 = 0. It comes in two steps: first, we estimate the probability that A3 is larger than 0. Then we model the non-zero A3 in the form of R3'.

$$E(R3) = P(A3 > 0)E(R3|A3 > 0) = P(A3 > 0)E(R3') \quad \text{..... 3.13}$$

where P(A3>0) is the probability that A3>0, E(R3') is a linear combination of independent variables that explain R3'. Following the 2-step model of Do et al. (2020), the structure for modelling R3 in 2 steps allows for the correlation of the error terms for both terms to be not an identity matrix.

Table 11 shows model estimates for R3 using OLS. It can easily be seen that MIP is the strongest driver and affects R3 positively.

***Table 11**. Standardised regression coefficients for recovery rates modelled under Stage 3*

Parameter	#3: E(R3)	#3.1: P(A3>0)	#3.1: E(R3')
Intercept	0.0904***	1.0534***	0.082***
	(0.0002)	(0.0039)	(0.0003)
CLTV	0.0009***	0.0517***	0.004***
	(0.0002)	(0.0047)	(0.0004)
DLTVCR	0.0037***	0.0807***	0.0063***
	(0.0003)	(0.0047)	(0.0004)
LLTVCR	0.0063***	0.1193***	0.0093***
	(0.0003)	(0.0037)	(0.0003)
DTIO	0.0007***	0.0101***	0.0008***
	(0.0002)	(0.0033)	(0.0003)
DICR	-0.0008***	-0.0095***	-0.0007***
	(0.0002)	(0.0036)	(0.0003)
LICR	-0.0023***	-0.0337***	-0.0026***
	(0.0002)	(0.0036)	(0.0003)
FICO	0.0002	-0.0072**	-0.0006**
	(0.0003)	(0.0036)	(0.0003)
LC	0.003***	0.0428***	0.0033***
	(0.0007)	(0.0076)	(0.0006)
LOB	-0.0078***	-0.127***	-0.01***
	(0.0003)	(0.0042)	(0.0003)
MOB	-0.0052***	-0.071***	-0.0056***
	(0.0004)	(0.0044)	(0.0003)
MIP	0.1125***	1.4817***	0.1164***
	(0.0003)	(0.0055)	(0.0003)
TID	0.0014***	0.0227***	0.0018***
	(0.0004)	(0.0044)	(0.0003)
TTR	0.0045***	0.0614***	0.0048***
	(0.0004)	(0.0039)	(0.0003)
LP_C	-0.0007***	0.001	0.0001
	(0.0003)	(0.0042)	(0.0003)
LP_P	-0.0004	-0.0098**	-0.0008**
	(0.0003)	(0.0046)	(0.0004)
NB_01	-0.0002	-0.0022	-0.0002
	(0.0002)	(0.0034)	(0.0003)
PT_CO	-0.0009***	-0.0148***	-0.0012***
	(0.0002)	(0.0034)	(0.0003)
PT_PU	-0.002***	-0.0284***	-0.0022***
	(0.0002)	(0.0034)	(0.0003)
FHB	0.0002	-0.0024	-0.0002
	(0.0003)	(0.0037)	(0.0003)
SNS	-0.0023***	-0.0404***	-0.0032***
	(0.0002)	(0.0033)	(0.0003)
OO	0	-0.0009	-0.0001
	(0.0002)	(0.0034)	(0.0003)
MF	-0.0019***	-0.0294***	-0.0023***
	(0.0003)	(0.0037)	(0.0003)

Parameter	#3: E(R3)	#3.1: P(A3>0)	#3.1: E(R3')
NJF	-0.0047***	-0.0667***	-0.0052***
	(0.0002)	(0.0034)	(0.0003)
NSRR	0.0007***	0.0137***	0.0011***
	(0.0002)	(0.0033)	(0.0003)
PDJ	0.0005**	0.0097***	0.0008***
	(0.0002)	(0.0034)	(0.0003)
INT	0.0034***	0.0482***	0.0038***
	(0.0003)	(0.0038)	(0.0003)
UMP	0.0027***	0.0415***	0.0033***
	(0.0003)	(0.0041)	(0.0003)

*This table reports the average coefficients obtained for models under Stage 3 based on 10 rounds of random sampling without replacement. Level of coefficient significant as follows: *** for p-value < 0.01, ** for p-value < 0.05 and * for p-value < 0.1. Values come in the following order: coefficient, degree of significance/ (standard error). The estimates are based on in-time training sample. 2nd column is for model #3, 3rd and 4th columns contain estimates for each of the two steps under #3.1.*

While CLTV is also a type of LTV which is expected to play a significant role in total recovery, it is shown here to have less influence on recoveries under Stage 3. It is positive but immaterial for both dependent variables under both #3 and #3.1. This view is unique and has only been observed by breaking down recovery rate in three stages. Recoveries in Stage 3 turns out to be positive for DLTVCR and LLTVCR because cases with significant increase in LTV at default and liquidation usually uses more of the mortgage insurance proceeds.

All DTI-related variables have even weaker influence on recovery under stage 3.

FICO score at origination and LC yields extremely weak relationships with recovery under Stage 3 so are ignored. High LC also yields positive result because cases where an account is highly constrained in liquidity also use up more of the mortgage insurance proceeds.

High LOB yields slightly lower R3. This means given shortfall after collateral liquidation in Stage 2, the small A3 is not always able to cover a relatively high EAD. This feature is only seen by splitting uncured recovery rate into three stages.

In the Freddie Mac data set, the mean R3 decreases with higher MOB. Old accounts tend to be well amortised, which means at default, there is seldom a shortfall after collateral disposition. Consequently, less is recovered under R3.

Defaulted loans with high TTR still yields high R3. This may be driven by the need to collect more given expenses have also accumulated beyond Stage 1.

The positive coefficient for PDJ implies that in states where deficiency judgements are allowed, R3 is naturally higher.

High interest rates at default see more recovery under R3, which may be due to higher shortfalls. In these scenarios, lenders tend to maximise the use of insurance proceeds.

High unemployment rate yields high R3 recovery, and this may be due to the negative relationship between UMP and house price movements. As house prices fall, there is a higher chance of shortfall therefore a higher need to collect more under Stage 3.

We see that by decomposing uncured recovery into 3 recovery stages, recoveries from Stage 3 may be explained better by MIP.

3.3.3. Model performance

Table 12 shows results of Stage 1 model performance, in which we evaluated three modelling approaches: simple linear regression (Model #1), a two-step framework (Model #1.1), and a machine learning model using random forests (Model #1.2). Across 10 rounds of random sampling, Model #1.2 consistently outperformed the others, demonstrating superior accuracy in predicting recovery rates. Despite Model #1.1's lower performance, its combination with the optimal Stage 2 model enhanced overall performance, warranting its inclusion in the experimental design.

Table 12. *RMSE for Stage 1 models*

Stage 1 models	Average RMSE		
	#1	**#1.1**	**#1.2**
In-time sampling	0.019	0.021	0.014
Out of sample 2007 in-sample 1999-2006	0.021	0.023	0.017
Out of sample 2008 in-sample 2000-2007	0.021	0.024	0.017
Out of sample 2009 in-sample 2001-2008	0.019	0.022	0.016
Out of sample 2010 in-sample 2002-2009	0.019	0.020	0.014
Out of sample 2011 in-sample 2003-2010	0.018	0.019	0.013
Out of sample 2012 in-sample 2004-2011	0.018	0.019	0.013
Out of sample 2013 in-sample 2005-2012	0.017	0.019	0.013
Out of sample 2014 in-sample 2006-2013	0.017	0.019	0.012
Out of sample 2015 in-sample 2007-2014	0.017	0.019	0.012
Out of sample 2016 in-sample 2008-2015	0.017	0.019	0.012
Out of sample 2017 in-sample 2009-2016	0.016	0.018	0.012
Out of sample after 2017 in-sample 2010-2017	0.016	0.017	0.012

This table reports the average Root Mean Square Errors (RMSE) of the out-of-sample prediction for Stage 1 models across 10 rounds of random sampling with replacement, measured on the validation dataset. Model #1 uses Ordinary Least Squares (OLS), the formulae for model #1.1 is $P(A1 > 0)E(R1|A1 > 0)$ *and #1.2 uses a random forest model. Each row represents a type of out-of-sample prediction using the validation dataset and highlights in bold and red the best performing model. As the choice of best Stage 1 model is performed in this step, RMSE is based on the validation dataset to ensure that there is no bias from the test dataset. While best model by RMSE is #1.2, #1.1 is also chosen for the combination model in consistency with the Stage 1 expected recovery model used to adjust S2 back to R2 form.*

For Stage 2, we explored performance of five models, including linear regression, two-step frameworks, and machine learning models for both R2 and S2 (see **Table 13**). Model #2.3, a random forest model based on S2, emerged as the best performer across all sampling scenarios. Although Model #2.4 showed slightly lower performance, its combination with Model #1.2 was also effective, leading to its inclusion in our design. Given that R is often dominated by R2, the choice of Stage 1 model is contingent on the best Stage 2 model.

***Table 13**. RMSE for Stage 2 models*

Stage 2 models	Average RMSE				
	#2	**#2.1**	**#2.2**	**#2.3**	**#2.4**
In-time sampling	0.236	0.232	0.219	0.219	0.219
Out of sample 2007 in-sample 1999-2006	0.259	0.258	0.252	0.252	0.252
Out of sample 2008 in-sample 2000-2007	0.261	0.259	0.251	0.251	0.251
Out of sample 2009 in-sample 2001-2008	0.253	0.250	0.237	0.237	0.237
Out of sample 2010 in-sample 2002-2009	0.237	0.235	0.221	0.220	0.220
Out of sample 2011 in-sample 2003-2010	0.231	0.228	0.214	0.214	0.214
Out of sample 2012 in-sample 2004-2011	0.227	0.224	0.211	0.211	0.211
Out of sample 2013 in-sample 2005-2012	0.225	0.222	0.210	0.210	0.210
Out of sample 2014 in-sample 2006-2013	0.224	0.221	0.210	0.210	0.210
Out of sample 2015 in-sample 2007-2014	0.224	0.221	0.210	0.209	0.210
Out of sample 2016 in-sample 2008-2015	0.223	0.219	0.209	0.209	0.209
Out of sample 2017 in-sample 2009-2016	0.223	0.219	0.209	0.209	0.209
Out of sample after 2017 in-sample 2010-2017	0.227	0.223	0.214	0.214	0.214

This table reports the average Root Mean Square Errors (RMSE) of the out-of-sample predictions for Stage 2 models measured on validation dataset across 10 independent rounds of random sampling with replacement. As the choice of best Stage 2 model is performed in this step, RMSE is based on the validation dataset to ensure that there is no bias from the test dataset. Model #2 uses Ordinary Least Squares (OLS), the formulae for model #2.1 is $P(A2 > 0)E(R2|A2 > 0)$*, #2.2 is a random forest model* $E(R2_{RF})$*, #2.3 is a random forest model based on S2,* $E(S2_{RF})$ *converted back to R2 using 2-step R1, and #2.4 is also a random forest model based on S2,* $E(S2_{RF})$ *converted back to R2 using random forest R1. Recall that S2 = R2/(1-R1). The best model, #2.4, is highlighted in red and bold. As R is often dominated by R2, the choice of Stage 1 model is to depend on best model for Stage 2.*

We present the performance of Stage 3 models in **Table 14**, in which, two models were assessed: a linear regression model (Model #3) and a random forest model (Model #3.2). Model #3.2 consistently outperformed the others, making it the preferred choice for our experimental setup.

Table 14. *RMSE for Stage 3 models*

Stage 3 models	Average RMSE		
	#3	**#3.1**	**#3.2**
In-time sampling	0.072	0.072	0.067
Out of sample 2007 in-sample 1999-2006	0.096	0.095	0.092
Out of sample 2008 in-sample 2000-2007	0.091	0.091	0.087
Out of sample 2009 in-sample 2001-2008	0.085	0.085	0.081
Out of sample 2010 in-sample 2002-2009	0.076	0.076	0.072
Out of sample 2011 in-sample 2003-2010	0.071	0.071	0.068
Out of sample 2012 in-sample 2004-2011	0.068	0.068	0.065
Out of sample 2013 in-sample 2005-2012	0.066	0.066	0.063
Out of sample 2014 in-sample 2006-2013	0.066	0.066	0.062
Out of sample 2015 in-sample 2007-2014	0.066	0.066	0.062
Out of sample 2016 in-sample 2008-2015	0.066	0.066	0.062
Out of sample 2017 in-sample 2009-2016	0.065	0.065	0.060
Out of sample after 2017 in-sample 2010-2017	0.066	0.065	0.060

This table reports the average Root Mean Square Errors (RMSE) of the out-of-sample prediction for Stage 3 models across 10 independent rounds of random sampling. Model #3 uses Ordinary Least Squares (OLS), the formulae for model #3.1 is $P(A3 > 0)E(R3|A3 > 0)$ *and #3.2 uses a random forest model predicting R3. Each row represents a type of out-of-sample prediction using the validation dataset and highlights in bold and red the best performing model. As the choice of best Stage 3 model is performed in this step, RMSE is based on the validation dataset to ensure that there is no bias from the test dataset. The best model is #3.2. This is used for the combination model.*

We synthesised several model combinations and present their performance results in **Table 15**. The Naïve model (M1) serves as a baseline, while M2 and M3 represent benchmark models using a two-step approach and random forests, respectively. M4 replicates M1 by decomposing the recovery rate into three stages, each modelled with OLS. Models M5 and M6, which incorporate advanced techniques such as two-step and machine learning models, consistently

outperform the benchmarks. M5 demonstrates superior performance across most sampling windows, validating the efficacy of our decomposition approach.

Table 15. *Performance measures for combined models*

Combination models	Average RMSE					
	M1	**M2**	**M3**	**M4**	**M5**	**M6**
In-time sampling	0.232	0.229	0.216	0.232	**0.215**	0.215
Out of sample 2007 in-sample 1999-2006	0.262	0.261	0.263	0.262	**0.261**	0.262
Out of sample 2008 in-sample 2000-2007	0.254	0.247	0.231	0.254	**0.231**	0.232
Out of sample 2009 in-sample 2001-2008	0.23	0.225	0.213	0.23	**0.211**	0.212
Out of sample 2010 in-sample 2002-2009	0.219	0.217	0.206	0.219	0.205	**0.205**
Out of sample 2011 in-sample 2003-2010	0.222	0.22	0.211	0.222	**0.209**	0.210
Out of sample 2012 in-sample 2004-2011	0.228	0.226	0.218	0.228	**0.217**	0.217
Out of sample 2013 in-sample 2005-2012	0.252	0.249	0.24	0.252	**0.239**	0.240
Out of sample 2014 in-sample 2006-2013	0.274	0.272	0.259	0.274	**0.258**	0.259
Out of sample 2015 in-sample 2007-2014	0.281	0.279	0.265	0.281	**0.263**	0.263
Out of sample 2016 in-sample 2008-2015	0.278	0.273	0.262	0.278	0.260	**0.259**
Out of sample 2017 in-sample 2009-2016	0.267	0.26	0.253	0.267	0.250	**0.249**
Out of sample after 2017 in-sample 2010-2017	0.275	0.268	0.257	0.275	0.252	**0.251**
	Average R-square					
	M1	**M2**	**M3**	**M4**	**M5**	**M6**
In-time sampling	0.404	0.420	0.483	0.404	**0.487**	0.487
Out of sample 2007 in-sample 1999-2006	0.323	0.328	0.315	0.323	**0.327**	0.321
Out of sample 2008 in-sample 2000-2007	0.250	0.294	0.380	0.250	**0.383**	0.377
Out of sample 2009 in-sample 2001-2008	0.314	0.347	0.414	0.314	**0.422**	0.420
Out of sample 2010 in-sample 2002-2009	0.375	0.390	0.446	0.375	0.451	**0.451**
Out of sample 2011 in-sample 2003-2010	0.365	0.376	0.425	0.365	**0.434**	0.432
Out of sample 2012 in-sample 2004-2011	0.360	0.371	0.416	0.360	**0.422**	0.422
Out of sample 2013	0.370	0.381	0.425	0.370	**0.430**	0.428

in-sample 2005-2012 Out of sample 2014	0.332	0.338	0.401	0.332	**0.405**	0.402
in-sample 2006-2013 Out of sample 2015	0.340	0.352	0.417	0.340	**0.423**	0.422
in-sample 2007-2014 Out of sample 2016	0.342	0.365	0.416	0.342	0.424	**0.426**
in-sample 2008-2015 Out of sample 2017	0.299	0.335	0.369	0.299	0.384	**0.388**
in-sample 2009-2016 Out of sample after 2017	0.260	0.296	0.350	0.260	0.377	**0.381**
in-sample 2010-2017						

This table reports the average Root Mean Square Errors (RMSE) and r-square of the out-of-sample prediction across 10 independent rounds of random sampling when models from all three stages are combined and compared with benchmarks. Each row represents a type of out-of-sample prediction and highlights in bold and red the best performing model. Model M1 uses Ordinary Least Squares (OLS). It is chosen as a benchmark model due to its popularity in both the literature and industry. M2 is the second benchmark model which is a slightly modified version from Do et al. (2020). M3 is a random forest benchmark based on A. Bellotti et al. (2019). M4 is a full replication of M1 since it decomposes uncured recovery rate R into 3 stages, and each stage is modelled using OLS. M5 uses 2-step for Stage 1, random forests of S2 and R3 for Stages 2 and 3, respectively. The concept of a 2-step model is taken from Do et al. (2020). M6 is similar with M5, except it now uses machine learning R1 instead of 2-steps. Recall that S2 is R2/(1-R1). Best overall model is M5 for most sampling windows as it is consistently better than all 3 benchmark models.

Our findings highlight the value of decomposing recovery rates into distinct stages, allowing for tailored modelling strategies that enhance prediction accuracy. The integration of machine learning techniques, particularly in Stage 2, significantly improves model performance, highlighting the potential of advanced methodologies in LGD modelling.

3.3.4. Appropriate measure for accuracy

RMSE and R-square are used to judge model accuracy. These measures are used in several studies such as Hurlin et al. (2018); Loterman et al. (2012); Qi & Zhao (2011); Thomas et al. (2012) in their reviews of models. There are two key aspects being tested here: the ability of each model to discriminate between good and bad uncured recovery rates (ranking ability), and how accurate each prediction is (error measure). Performance measures like area under curve (AUC) and linear correlation between dependent variable and predictions are most common forms of checks for model discrimination while RMSE and R-square are measures for accuracy. While accurate models have good discrimination, models that discriminate well are not necessarily accurate (Loterman et al., 2012). Given this, performance measures like RMSE and R-square are given priority over others. While RMSE alone is enough to provide performance ranking among models proposed, R-square is also considered for comparison with the literature.

RMSE and R-square are employed to assess the performance of each model by looking at predictions. RMSE focuses on the error term expressed in units of the dependent variable *R*, and R-square looks at the RMSE of a model and measures how much better it is compared to using the mean as predictor. While R-square can theoretically have values less than 0, it can be floored at 0 under the assumption that the mean is used if predictions are less accurate.

3.3.5. Combination of stage models to form overall model for R

Figure 8 and ***Figure* 9** visualize the performance of the combined models using the two accuracy measures discussed in section 4.3, which are RMSE and R-square, respectively. We observe a few observations and generalisations from the two figures as follows:

- M2 and M3 are better than M1
- M4 is indeed a perfect replication of M1, as hypothesised and proven in 3.2.6
- Models M5 and M6, which are the challenger models in this paper, have consistently outperformed M1 and M2 and mostly but marginally outperforms benchmark M3.
- M5 outperforms all models for across all sampling types.

Given that all proposed models have been shown to consistently outperform benchmark models M1, M2 and M3, it can be concluded that this way of splitting recovery rate in stages not only contributes to a better understanding of LGD by revealing hidden relationships and significant drivers but also improves the accuracy of predictions.

Within reason, the combination of "best models" for each stage yields the most accurate model. This is found by performing 2 main types random sampling for 10 independent rounds: in-time and out of time with rolling windows.

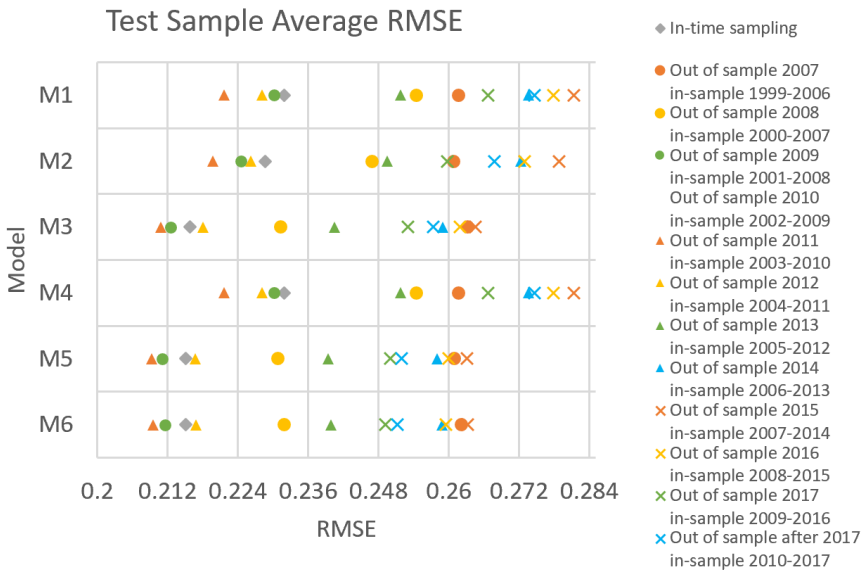

Figure 8. *Average RMSE measured on test sample.*

Y-axis is the model name, X-axis represents average RMSE. M1, M2 and M4 show highest RMSE which means all proposed models outperformed non-machine learning benchmarks. As a machine learning benchmark, M3 is better than models M1, M2 and M4. While M6 is almost like M3 but slightly better, M5 is consistently better than M3 through all types of sampling. This, M5 is the overall best model which outperforms all benchmarks.

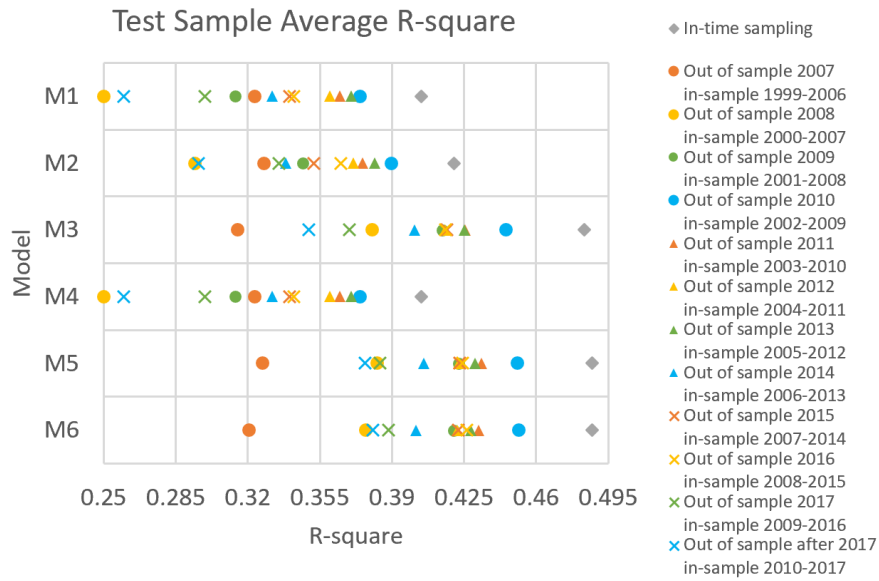

Figure 9*. R-square out-of-time measured on test sample.*

Y-axis is the model name, X-axis represents average R-square. M1, M2 and M4 show lowest r-square which means all proposed models outperformed non-machine learning benchmarks. As a machine learning benchmark, M3 is better than models M1, M2 and M4. While M6 is almost like M3 but slightly better, M5 is consistently better than M3 through all types of sampling. This, M5 is the overall best model which outperforms all benchmarks.

Since R-square is defined as $1 - \frac{[RMSE(combined\ R)]^2}{variance(R)}$, it is not surprising that the ranking of model performance is similar when RMSE is used.

3.4. Conclusion and implications

This study demonstrates the effectiveness of a three-stage decomposition approach in Loss Given Default (LGD) modelling, grounded in both statistical evidence and economic theory. As outlined in gap 1 (Section 1.4), this decomposition aligns with typical bank collection processes.

The empirical results yield several key insights. First, the stage-specific distributions and temporal patterns confirm distinct recovery dynamics across stages, challenging the traditional single-component approach. Second, our correlation analysis reveals unique driver relationships at each stage, with mortgage insurance percentage (MIP) and modification flags (MF) emerging as significant but previously overlooked factors. Third, the superior performance of our combined modelling approach, particularly in out-of-time predictions, validates the decomposition framework's practical utility.

The primary contribution of this three-staged framework lies not in marginal R^2 improvement but in its modularity. Each stage corresponds to a distinct economic mechanism, enabling institutions to adjust or replace individual components independently. This modularity has practical value: when market conditions change, a single stage can be modified without disturbing the remainder of the model. It also enables more precise risk assessment and capital allocation, targeted collection strategies aligned to each recovery stage, and enhanced transparency in loss prediction. The OCC (2011) provides supervisory guidance consistent with this design, noting that model complexity should be proportionate to the risk being managed and the materiality of the exposure. The added complexity of separately modelling pre-disposition recovery, collateral disposition, and post-collateral recovery is justified on the grounds of economic interpretability and modular flexibility rather than marginal R^2 improvement alone. For institutions with simpler mortgage portfolios or limited modelling resources, a reduced-form single-stage approach may remain appropriate, and the modular design permits selective adoption of components where they add the most value.

While superficially attractive, converting R^2 improvements into currency benefits is inadvisable in practice. Such conversions require assumptions about portfolio composition, correlation structure, and capital allocation that are highly institution specific. Presenting

currency-denominated accuracy claims in a general-purpose academic model would risk misleading practitioners and would raise legitimate concerns from model validators, auditors, and regulators, who typically view such precision claims as evidence of conflicts of interest rather than integrity. What the framework does provide is a transparent and modular structure that enables each institution to calibrate the components to their own portfolio and assess the dollar impact within their own capital and provisioning frameworks.

In conclusion, our three-stage decomposition approach represents a significant advancement in LGD modelling, providing the foundation for the resolution-based refinements explored in Essay 2 and the enhanced collateral valuation methods in Essay 3.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Justin Rylie Tang
Name and title of main supervisor:	Dr. David Tripe, Adjunct Professor of Banking

In which chapter is the manuscript/published work?	No

Describe the contribution that the student and members of the supervisory team have made to the manuscript/published work:[1]

This paper proposes a practical LGD modelling approach that first predicts the likelihood of different post-default resolutions, then estimates conditional recovery rates for each resolution type. Applied to Freddie Mac data, it outperforms common benchmarks like OLS and two-step regression, while revealing distinct drivers for resolution propensity and recovery. The contribution is improved accuracy, transparency, and flexibility for banks in LGD forecasting.

Please select one of the following three options:

○ **The manuscript/published work is published or in press**

Please provide the full reference of the research output:

○ **The manuscript is currently under review for publication**

Please provide the name of the journal:

◉ **It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:	Justin Rylie Tang — Digitally signed by Justin Rylie Tang Date: 2026.12.08 06:04:59 +13'00'	Main supervisor's signature:	

This form should be placed at the beginning of each relevant thesis chapter.

[1] Refer to the Massey University Publishing and Authorship guidelines (OneMassey for staff, Stream for students) and/ or Contributor Roles Taxonomy (CRediT) guidelines for guidance.

4. Chapter 4 - Essay 2: Predicting loss severity for residential mortgage loans across distinct default resolution processes

4.1. Introduction

Building on the credit risk framework and LGD decomposition introduced in Section 1.1, Baesens & Smedts (2025) reaffirm this decomposition as central to credit risk modelling across institutions globally.

As discussed in Section 1.3, inaccurately estimated LGDs can lead to inaccurate pricing of retail mortgages, inaccurate provisioning of expected losses and ultimately to systemic vulnerabilities in the housing finance market. From a portfolio manager's perspective, a transparent and resolution-based LGD model also informs strategic decisions around loss mitigation spending, securitisation structuring and hedging tail exposures in mortgage-backed securities. In short, LGD modelling is not merely a statistical exercise - it directly impacts lenders' balance sheets, investors' returns and the stability of the broader financial system.

Yet, LGD modelling remains challenging. Baesens & Smedts (2025) note that LGD models often exhibit low explanatory power, with R^2 values seldom exceeding 20%, despite their linear impact on both expected and unexpected losses. Moreover, the accounting framework used can influence the quality of credit loss estimates. Marton & Runesson (2017) find that loan loss provisions under IFRS (IAS 39) are less predictive of actual losses than those under local GAAP, especially in high-enforcement environments—highlighting the importance of judgment, governance, and regulatory context in credit risk estimation.

Building on gap 2 identified in Section 1.4, as a loan defaults, it goes through processes that eventually achieve resolution.[5] These processes involve rehabilitating loans back to current (i.e., paying), or eventually closing them down and writing them off after recovery efforts by the lender. We identify nine resolutions as the set of possible outcomes for a residential mortgage loan in default, three of which typically result in little or no loss and six where loss is more likely. The resolution for a specific loan depends primarily on the choices made by the

[5] While resolution generally means a method by which a loan is settled or is concluded, resolution in this paper refers in particular to method by which defaulted accounts have eventually been concluded.

lender, especially by their collection operations, and these choices are closely linked to LGD (Bellotti & Crook (2012). The nine possible resolutions are:

- ‘Cure’ – the defaulted loan returns to good standing and continues with the lender;
- ‘Repurchase’ – the defaulted loan is sold back to the originator;
- ‘Attrition/ repaid early’ – the borrower in default repays the loan in full, often by refinancing to another lender;
- ‘Short sale’ – the property is sold by the distressed homeowner with permission of lender;
- ‘Third party sale’ – the property is sold via a third party in a foreclosure auction;
- ‘REO sale’ – the lender takes possession of the property used as collateral (“real estate owned") who sells it on its own terms;
- ‘Note sale’ – the defaulted loan is sold, often at a discount, to another willing lender;
- ‘Reperforming sale’ – the loan is sold to another lender after it has started performing again, possibly with modification to the loan terms; and
- ‘Charge-off’ – the lender closes their position by writing off the entire outstanding amount without taking possession of the property used as collateral or selling it.

The main motivation of this paper is that different resolution outcomes exhibit distinct loss characteristics. Extending the stage-based decomposition from Essay 1, by leveraging resolution-related information, we can enhance the accuracy of LGD prediction. Using available loan data, we can estimate the probability of each resolution outcome for a given loan. This enables a modular approach to LGD estimation, where different modelling techniques can be applied to predict LGD under each resolution scenario. The rationale behind this motivation is discussed in more detail below.

Defaulted loans with unsuccessful short-term recovery attempts usually indicate risks to full recovery. What typically follows are more severe measures, such as realisation of collateral. Several approaches are available for collateral/contract realisation. As expense from marketing often grows with time (Clauretie & Daneshvary, 2011), lenders often prefer foreclosure through (i) third party sales (i.e., foreclosure auctions). With lender approval, there are cases where the distressed owner sells (ii. short sale) the collateral themselves. While this process is often more expensive due to the longer time it takes for the seller to be satisfied with an offer, the cost can at times be shouldered by the seller rather than the lender. A third option is for the lender to take ownership of the collateral to dispose of on their own terms (iii) real-estate owned

(a.k.a. REO). As an alternative to selling collateral, loan contracts may also be sold to another lender as once a non-performing loan starts performing again with (iv. reperforming sale) or without modifications. As a more generic term, a reperforming sale falls under note sale, which simply refers to a loan contract being sold by one lender to another, regardless of the status of a loan.

Because each resolution may have intricacies within each process, it follows that these differences may also lead to cash flows with different sources, intensities and patterns. In their paper, Clauretie & Daneshvary (2011) compared different resolutions: short sale, REO sale and foreclosure sale (i.e., foreclosure auction). They found that during their data period (Dec 2007 – Dec 2008), collateral was often sold at a higher price at the expense of incurring higher selling cost from longer time in the market. It was the other way around for the other two resolutions. We intend to build on the principle of having multiple possible resolutions and use this as a feature in building a transparent and more accurate LGD model.[6]

Our investigations have identified that (i) recovery rate distributions, and (ii) mean recovery rates through time vary across different resolutions, (iii) the proportion of defaulted loans going to different resolutions vary over time, and (iv) recovery rates correlate strongly to different sets of drivers under each resolution.

As noted in Section 1.4, earlier residential mortgage LGD studies focussed on the use of Ordinary Least Square (OLS) (Calem & LaCour-Little, 2004; Clauretie & Herzog, 1990; Crawford & Rosenblatt, 1995; Goodman & Zhu, 2015; Lekkas et al., 1993; Pennington-Cross, 2003; Qi & Yang, 2009; Zhang et al., 2010). In the past decade, the literature has expanded to include more sophisticated regression-based and machine learning approaches, including An & Cordell (2021); Higgins et al. (2022); Miu & Ozdemir (2017); Ozdemir & Huang (2021). These methods, ranging from quantile regression to gradient boosting and neural networks, have demonstrated incremental improvements in LGD prediction accuracy.

A recent contribution by Fabozzi et al. (2025) offers a comprehensive review of the intersection between financial modelling and operations research, highlighting the growing role of machine

[6] It should be noted that selecting a resolution only on the basis of its historical recovery rate can introduce selection bias: servicers naturally devote more resources to the strategies they expect to perform best, thereby inflating observed recoveries. This concern may be addressed by validating the resolution methods through independent, randomised trials and rigorous performance testing.

learning and high-dimensional statistics in credit risk modelling. Their work emphasizes the potential of integrating behavioural factors, macroeconomic variables, and advanced simulation techniques to enhance LGD estimation, particularly in the context of mortgage portfolios. This aligns with emerging trends in the literature that advocate for hybrid models combining traditional econometrics with AI-driven methods.

We contribute to the literature by providing a commercially viable modelling framework that decomposes expected recovery into two components: (a) the probability of occurrence of each resolution and (b) the conditional recovery rate under each resolution. This two-component structure complements the stage-based decomposition in Essay 1 by offering an alternative that is both more transparent and more practical than single-equation alternatives. When the mix of resolutions shifts, as it did post-GFC when note sales became more prevalent, only the resolution probability module requires recalibration. The framework further enables transparency by highlighting varying dominant drivers across resolutions. The primary audience for these models are lending institutions.

The remainder of the paper is organised as follows. Section 2 reviews LGD literature and mortgage-specific models. Section 3 motivates our proposed LGD modelling approach. Section 4 describes the Freddie Mac dataset, its quirks and derived variables. Section 5 outlines the methodology, sampling strategy and accuracy metrics. Section 6 presents the empirical results and key findings.

4.2. Literature review

4.2.1. 4.1 Building on Multi-Step LGD Models

As discussed in the main literature review, multi-step LGD models have evolved from simple binary splits like Bijak & Thomas (2015); Do et al. (2020); Leow & Mues (2012) to more sophisticated conditional approaches. This essay extends this framework by incorporating resolution type as identified in gap 2 (Section 1.4).

4.2.2. LGD and different resolution types

Calem & LaCour-Little (2004) built an expected loss model which considered transition probabilities to both default and prepay. Building on the workout LGD models in the main literature review, this was carried out while adjusting for the probability that short sale might happen in lieu of foreclosure. They claimed that this leads to better risk-sensitive capital estimates. More recently, Gabriel et al. (2020) provided empirical evidence that foreclosure

imposes significantly higher credit losses compared to alternative resolution methods such as short sales and modifications. Their analysis based on regulatory data and policy shifts in California supports the notion that incorporating resolution type probabilities, especially short sales, can materially improve loss forecasting and capital adequacy modelling. We build on this principle by using multiple possible resolutions as the basis for a more transparent LGD model.

Clauretie & Daneshvary (2011) compared different resolutions: short sale, REO sale and foreclosure sale. They found that two main factors affect net losses: property sale price and time to sell. They also found that different resolutions tend to have different values for both factors: for example, short sales resulted in the highest selling prices due to longer time of listing in the market. However, longer time in the market also means higher expense. They built a model to predict each of the two factors. In this framework, all available information predicts sale price and predicted sale price predicts time to sale. While there is no way to know how long a property has been in the market using Freddie Mac data, we will attempt to address biases from this by using time to resolution as a control variable.

A related stream of literature examines the role of mortgage servicers as agents whose incentives may diverge from those of investors. Adelino, Gerardi, and Willen (2013) show that servicer renegotiation decisions are shaped by the structure of securitisation agreements rather than borrower characteristics alone. Agarwal et al. (2011) provide evidence that bank-affiliated servicers modify loans at different rates compared to independent servicers, reflecting institutional constraints on loss mitigation. Piskorski, Seru, and Vig (2010) demonstrate that securitised loans experience different foreclosure rates than portfolio-held loans, suggesting servicer behaviour is endogenous to contractual arrangements. These findings are relevant because resolution type, the key dependent variable in this essay, is partly determined by servicer decisions that reflect unobservable factors such as borrower cooperativeness and property condition. This motivates the careful interpretation of resolution-type coefficients as associations rather than causal effects, as discussed in the limitations section below.

4.2.3. Models built considering foreclosure probability

Extending the multi-step approaches outlined in the main literature review, Jiang & Zhang (2025) propose a covariate-driven, two-stage finite-mixture framework that first models "recoverability" via a logistic or multinomial regression and then captures "severity" with a flexible parametric mixture - demonstrating notably improved fit in the upper LGD quantiles

and 10–15% better capital estimates under Basel III stress tests. This intention was carried out by considering the propensity that loss rates may be <=0 or >=100% in the absence of an actual resolution. While it makes sense to consider possibility of different outcomes, purely looking at the range may falsely cluster different resolutions with different loss distributions together just because of coincidental similarity in some resulting LGDs.

Leow & Mues (2012) introduced a 2-step LGD model: probability of repossession and a subsequent haircut model for cases which had undergone repossession. Application on two UK banks have shown that this model performs better than directly modelling LGD in a single step. Using the same model, Leow et al. (2014) found that adding interest rate at default improves LGD prediction significantly. From an economic perspective, this makes sense as interest rates tend to have distinct historical correlations with house prices for any suburb. While the relationship is negative, they have used default interest rate as a proxy to house prices.

We intend to build on this concept and decompose the recovery process into the "choice" of resolution and recovery under each resolution. Note that "choice" here refers to actualisation of a resolution from a list of possible resolutions.

4.2.3.1. Loss models considering outcome probabilities

In the previous sections, it has been established that the duration of foreclosure greatly affects LGD. Using a multinomial logit model, Pennington-Cross (2010) modelled duration of foreclosure by considering different propensities for each outcome (i.e., resolution) to occur. They found that duration of foreclosure is affected by factors like current housing market conditions, previous delinquency states of a loan and state laws.

Thomas et al. (2012) built LGD models that decompose LGD by value and method of collection. In-house collection would be separated into zero LGD vs LGD > 0 and those that went to 3rd party collections would be decomposed by LGD = 1 and LGD < 1. However, while the outcomes with LGD = 0 or LGD = 1 may have distinct interpretations, they may still contain combinations of different resolution types.

We intend to contribute to the literature by incorporating concepts discussed above to build a more transparent and accurate LGD model. This is done in two steps; a multinomial logit model is used to estimate propensity of each type of resolution, and a regression model is used to predict the conditional recovery rates for each resolution.

4.3. Motivation for modelling recovery by resolution

Given that recovery rates from different resolutions may be driven by different intentions and scenarios, their time series trends, distribution and relationship with available drivers are also expected to be different between different resolutions. There may be intellectual inconsistencies where models based on LGD understood as a single resolution are not going to be accurate in the long run when the proportion going to specific resolutions and the level of recovery rate per resolution move in different directions through time. As such, structuring them in separate resolutions is a logical approach.

4.3.1. Varying conditional recovery rates distribution across resolutions

This approach addresses the multimodality issues identified in the main literature review by decomposing recovery into resolution-specific components.

Figure 10 shows that the distribution of recovery rate R varies dramatically among different resolutions. It shows a bimodal distribution for charge-off composed of a massive concentration around 0% recovery and a few recovering 100% which are mostly from mortgage insurance and early repayments. Note or reperforming sale shows a unimodal distribution concentrated around 100% recovery, with heavy left tail insinuating that the outcome varies greatly between each case. REO disposition, short sale and third-party sale appear to be bell-shaped curves with means around 55%, 65% and 50%, respectively, with extra spikes at 100% recovery, making them bimodal in nature. As defaulted repurchased happens when bad accounts are bought back by the originator, most of them have 100% recovery, except for a few small loan amounts where costs presumably dominate total amount owed. Two additional resolutions are excluded from these charts: defaulted attrition and cure which have 100% recovery for all cases.

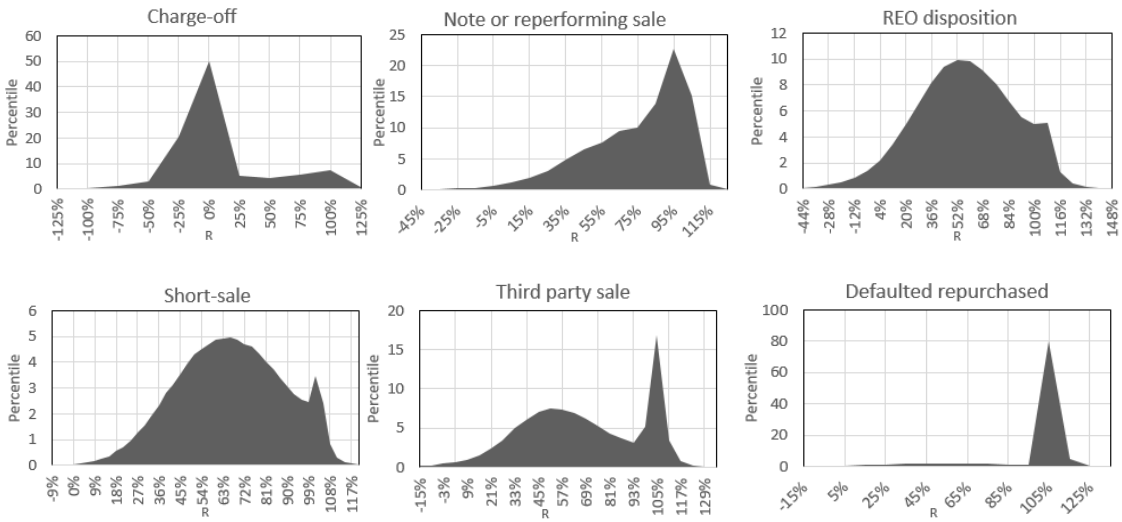

Figure 10. *R distribution under various resolutions.*

This is a histogram of non-discounted R using the Freddie Mac dataset. Vertical axis denotes frequency of observation and horizontal axis denotes values of R. High frequency around 0% are cases mostly written-off, net of costs. These are usually loans with smaller exposures at default which makes foreclosure expenses look relatively more material. Those with 100% R have recovered everything owed. A few exceeded 100% covering additional foreclosure expenses outside of property or contract disposition. For some resolutions like REO disposition, short sale and third-party sale have partial recoveries around means of 55%, 65% and 50%, respectively. Defaulted repurchased has massive concentration around 100%, with a few exceptions of lower recovery net of other fees and relevant costs dominating amount owed due to low remaining balances. Others excluded from the charts are defaulted attrition and cure as they both have 100% R for all cases.

4.3.2. Varying recovery rate trends per resolution

Typically, LGD increases (recovery rate decreases) during economic downturns. **Figure 11** exhibits this through the 2008 GFC and a slow recovery up to 2017, and each resolution were affected by the GFC in varying degrees. Here, REO disposition seems to have been affected the most and its GFC effect started earlier (2004-2005). As this point may be controversial, it will be investigated and discussed further for this research. It can also be seen that both short sale and third-party sale were severely and abruptly affected by the GFC - noting that third party sale has recovered better than the others. Due to the nature of note sale/reperforming sale, it does not seem to show signs of the GFC. This may be due to an incomplete picture given limited access to information post sale of debt.

These observations support the idea that modelling LGD separately for each resolution may contribute to increasing the level of overall accuracy.

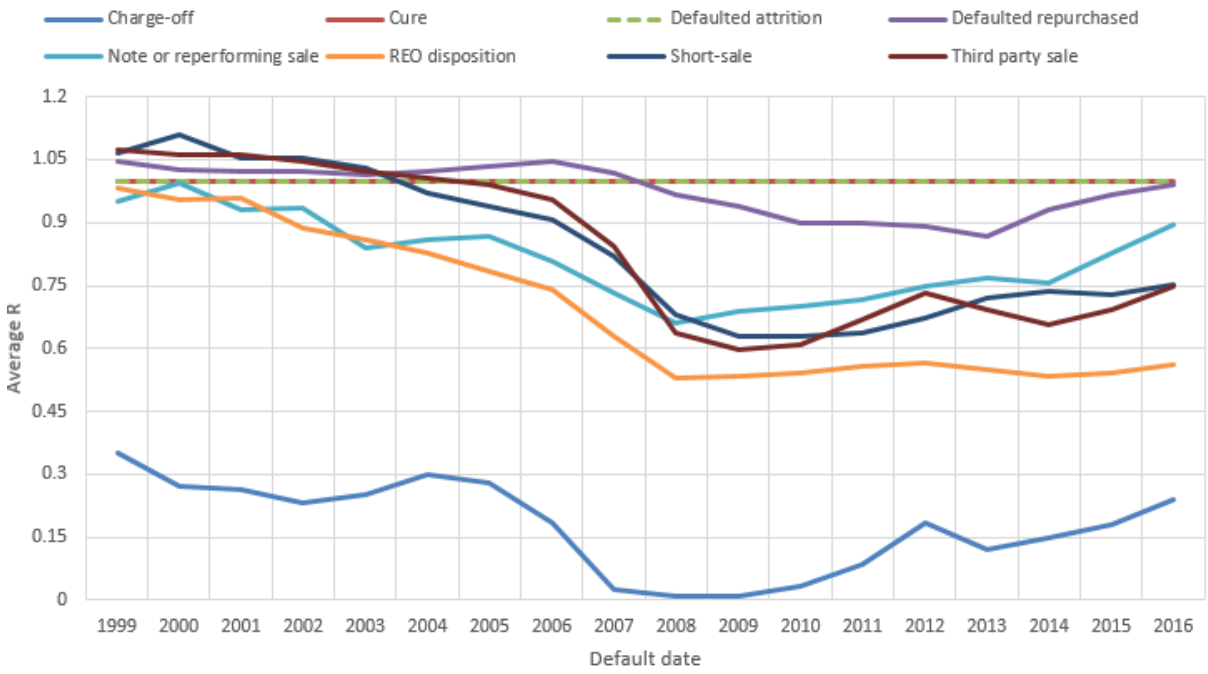

Figure 11. *Account weighted average uncured recovery rates time series.*

Vertical axis denotes recovery rate (for average R line charts) and horizontal axis denotes date of default. In the Freddie Mac dataset, each line shows different levels of recovery for each resolution.

4.3.3. Proportion of each resolution changes through time

Figure 12 shows that the proportion of defaulted loans going to each resolution varies through time. This shows the types of choices performed by lenders as a reaction to changing situations. Resolutions that result to higher recovery rates like defaulted attrition reduced during the GFC while other resolutions such as defaulted repurchased started to become more dominant. In the recovery period shortly following GFC, resolutions like short sale gained more share. This may be due to capacity issues for the industry to process a suddenly inflated number of foreclosures auctions. Other resolutions such as cure, third party sale, and note/reperforming sale started becoming more important following this recovery period.

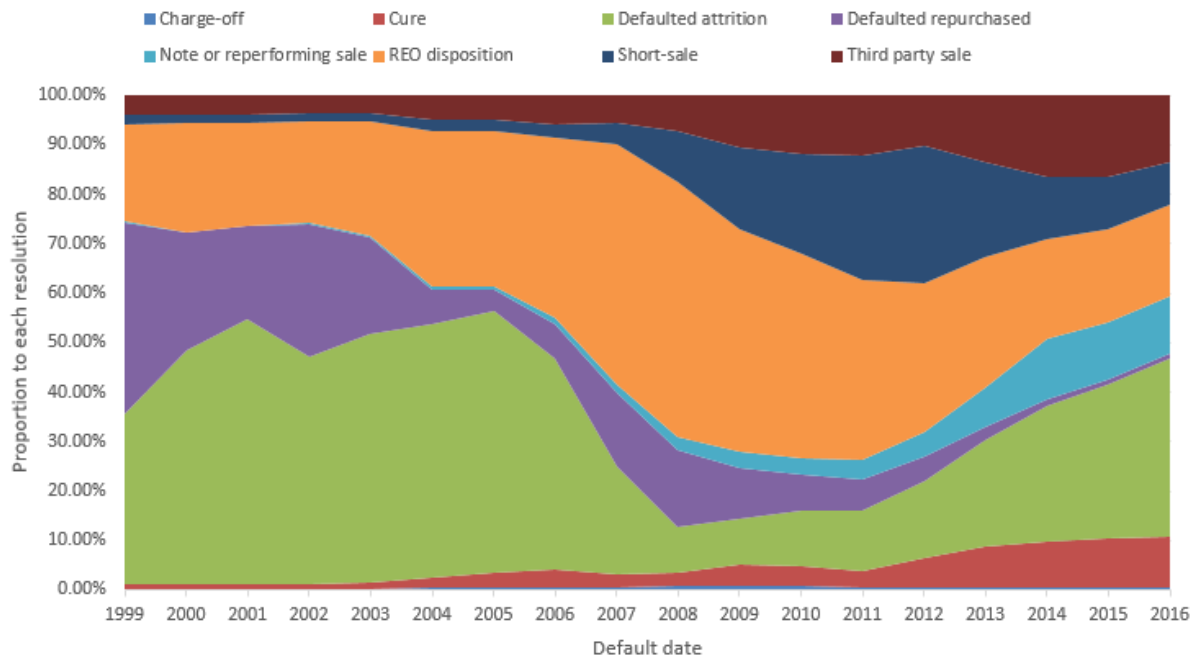

Figure 12. *Time series of proportions going to each resolution.*

Vertical axis denotes proportion going to resolutions and horizontal axis denotes date of default. Each line shows different proportions allocated to each resolution in the Freddie Mac dataset. Proportions sum up to 1 for each date.

These observations support the idea that modelling LGD separately for each resolution may contribute to increasing the level of risk-sensitivity, thereby contributing to betterment of overall accuracy.

4.3.4. Recovery rates correlated with different drivers across resolutions

Because recovery rates have different historical trends for each resolution, it follows that they may have different primary drivers under each resolution. **Table 16** illustrates that R is only strongly dependent on DLTV under note/reperforming sale, and third-party sale. Time in delinquency (TID) is highly correlated with R under note/reperforming sale and time to resolution (TTR) – the latter implying that accounts that have been in delinquency for extended (shorter) periods of time also take a long (short) period to achieving resolution. Charged-off loans have high correlation between LOB and R, supporting the view that decisions made during charge-off may be dependent on loan amount. Percentage coverage in MI (MIP) is also shown to have strong relationship with R under the REO disposition and short sale. Portfolio default rate is negatively correlated with R under all resolutions, but it is especially strong under third party sale. In terms of external indicators, unemployment rate in the previous

quarter of default month is strongly negatively correlated with R under third party sale and short sale, while GDP change from origination to a quarter leading to default month is only highly correlated with R under third party sale. Without decomposing to different resolutions, the overall picture would only have been similar to Third Party Sale. Even then, there are still minor differences, like MOB which has opposite signs for overall if compared between the two.

Table 16. *Correlation table between available drivers in Freddie Mac dataset and recovery rates*

R correlation	Overall	Charge-off	Defaulted repurchased	Note or reperforming sale	REO disposition	Short-sale	Third party sale
DLTV	-25%	2%	-2%	-21%	-3%	-10%	-36%
DDTI	4%	4%	-2%	-4%	8%	6%	0%
FICO	-11%	4%	2%	-11%	-2%	-5%	-4%
TID	12%	-1%	-10%	28%	4%	3%	11%
TTR	-21%	-6%	-14%	27%	-16%	-17%	-12%
LC	-7%	2%	4%	-2%	-5%	-13%	-13%
LOB	-1%	31%	-1%	-5%	18%	7%	-10%
MOB	-2%	-7%	-19%	19%	-2%	-1%	15%
MIP	14%	-4%	3%	9%	30%	35%	13%
DT	-7%	1%	14%	-10%	0%	-12%	-14%
FHB	0%	2%	1%	-1%	6%	3%	-2%
INT	-3%	-11%	16%	-7%	-7%	-3%	-5%
OO	11%	13%	-5%	12%	14%	14%	11%
LP_C	-7%	-1%	-8%	-5%	-17%	-12%	-8%
LP_P	3%	0%	7%	0%	15%	6%	0%
MF	-1%	2%	-7%	5%	-1%	-3%	-4%
NB_01	-6%	-3%	7%	-8%	-9%	-10%	-5%
NJF	5%	13%	8%	19%	16%	8%	5%
NSRR	5%	5%	-2%	0%	2%	6%	-2%
PDJ	4%	5%	-1%	6%	3%	10%	5%
SNS	0%	0%	-7%	13%	-5%	-7%	0%
PT_CO	-7%	8%	2%	-7%	-6%	-10%	-5%
R_SS	-16%	-4%	-17%	-6%	-9%	-4%	-10%
R_TP	9%	4%	1%	12%	6%	2%	8%
DRD	-28%	-10%	-14%	-19%	-17%	-18%	-26%
UMP_L3	-25%	-5%	-17%	-12%	-16%	-26%	-23%
GDPCY_L3	21%	6%	3%	14%	17%	16%	23%

4.4. Freddie Mac dataset: description and transformations

4.4.1. Freddie Mac dataset

This research utilises U.S. single-family fixed rate prime loan-level dataset from Freddie Mac.[7] The data tables contain accounts approved from 1999 to 2019 with observations ranging from 1999 to 2020. Performance and origination information are matched for analyses carried out in this paper. Performance information comes in the form of quarterly updates of monthly snapshots for each outstanding loan. Data captured for these updates include interest rates, payment status, maturity date, and outstanding debt amount for each performing loan, and information for loans that default such as net collateral sale proceeds, associated foreclosure costs, reason for early performance ending marks an account's last transaction status. Mortgage insurance (MI) recovery and non-MI recovery may also be available when a bad account has been resolved and/ or is purged from the dataset. The origination portion contains all information taken at time of application such as, the Fair Isaac score (FICO) score, LTV for the specific loan, combined LTV (CLTV) of the customer, application balance, application date, location where property is located (state or MSA), number of borrowers, property type, occupancy status, mortgage insurance coverage, debt-to-income (DTI) ratio, name of lending entity that sold the loan to Freddie Mac, and name of loan servicer.

4.4.2. Variables derived from additional datasets

In addition to the main data table, four more additional and external data tables are used: house price index (HPI) from Federal Housing Finance Agency (FHFA), foreclosure-related state laws, DTI index, and state-level monthly seasonally adjusted unemployment rate throughout the United States.

Origination LTV and combined origination LTV are updated to reflect new information held at month of default. To update the denominator of LTV (i.e., collateral value), MSA-level FHFA HPI was used to estimate change in collateral values from origination to default month. On the other hand, the numerator (i.e., loan amount) was updated with principal balance (UPB) at default month. To illustrate, DLTV = (CLTV x origination HPI/default HPI x default UPB). This is considered common practice and was used by Do et al. (2020); Leow & Mues (2012); Qi & Yang (2009); Tong et al. (2013). To account for MSA updates and to ensure the most

[7] Data tables available at: http://www.freddiemac.com/research/datasets/sf_loanlevel_dataset.page.

accurate matching HPI, several MSAs were updated to account for old and inactive MSA locations that were already replaced with new ones while ensuring continuity in the dataset through time.

The DTI (debt-to-income) ratio was adjusted at MSA level at the time of default, to ensure that the original DTI ratio is still relevant at default. A DTI index provided by the Federal Reserve was utilized to carry out this process. A new variable, DDTI (default-adjusted DTI), was defined as DTI multiplied by the index at default and then divided by the index at origination. Although this approach is not as widespread, it is based on the same reasoning as the LTV (loan-to-value) ratio above. In most states[8], if a borrower defaults on their loan and there is a statutory right of redemption in place, they are allowed to regain their collateral by paying the full amount owed. If all other attempts at negotiating a solution have failed, this may be the borrower's last chance to get their property back. If this fails, it becomes an additional cost that is subtracted from the overall amount recovered. One might expect that states where statutory right of redemption is not allowed would result in fewer pre-collateral recoveries, but Do et al. (2020) found the opposite to be true. They pointed out that a statutory right of redemption can lead to more processing, which takes time and money. This decreases the chance of a zero loss and increases the non-zero LGD.

In some US states[9], lenders must obtain a court ruling to foreclose on a defaulted borrower, which adds to the cost of recovery. Recovery on shortfalls may only be possible if it is allowed by law, as deficiency judgments are prohibited in a few states[10] in the US. This is consistent with the findings of Clauretie & Herzog (1990); Do et al. (2020), where LGD is higher when lenders are unable to recover the total outstanding amount after a foreclosure sale.

Recovery rates are expected to decrease as the unemployment rate increases, while both GDP variables are expected to increase recovery rates. While the unemployment rate is used as is,

[8] Relevant states include Arizona, Connecticut, Delaware, Hawaii, Illinois, Iowa, Louisiana, Maryland, Massachusetts, Mississippi, Montana, New Hampshire, New York, Oklahoma, Pennsylvania, South Carolina, Texas, Washington D.C., and West Virginia *United States Foreclosure Laws* (2020)

[9] Relevant states include Connecticut, Delaware, Florida, Illinois, Indiana, Iowa, Kansas, Kentucky, Louisiana, Maine, Nebraska, New Jersey, New Mexico, North Dakota, Ohio, Pennsylvania, and South Carolina *United States Foreclosure Laws* (2020)

[10] Relevant states include Delaware, Iowa, Massachusetts, Mississippi, Missouri, Nebraska, West Virginia *United States Foreclosure Laws* (2020)

the year-on-year change in GDP a quarter before the default date provides the most accurate improvement.

These values are detailed in **Table 18** for different variations of LTV and DTI, while the definitions of state laws, GDP, and unemployment rate are provided in **Table 17**.

Table 17. *Table definition for Freddie Mac*

Variable	Definition
CLTV	Cross collateralised loan to value ratio at origination
DTI_0	Origination Debt-to-Income ratio. Valid values range from 0 to 65%. Calculated using sum of borrower's monthly debt payment & housing expense at time of delivery to Freddie Mac / total monthly income at origination
FHB	First home buyer flag. 1 if loan is a mortgage for borrower's first home. 0 otherwise.
FICO	FICO score at origination
LC	Liquidity Constraint. Balance at default/scheduled balance – 1; capped at 5 to control for outliers
LOB	log of origination UPB
LP	Purpose for the mortgage loan: · Purchase – used to purchase a property · Refinance with cash out – no specific purpose to the loaned amount. Was not used to secure property. · Refinance with no cash out - limited to paying off first mortgage, or loans for other properties used to secure current mortgage, and cash out of min (2% of refinance amount, $2000) · Refinance but not specified · Unknown
MF	Flag for modifications that resulted in unpaid principal balance increase. If total modification is positive, 1, else 0
MIP	MI percentage. If MI = 1, MI is the percentage of the shortfall covered by insurance provider on the event of default after collateral disposition
MOB	Months on book. The age of the loan when it defaulted

Variable	Definition
NB	Number of borrowers: 1 2+ blank
NJF	States with non-judicial foreclosure: Connecticut, Delaware, Florida, Illinois, Indiana, Iowa, Kansas, Kentucky, Louisiana, Maine, Nebraska, New Jersey, New Mexico, North Dakota, Ohio, Pennsylvania, South Carolina (*United States Foreclosure Laws* (2020)
NSRR	States with Statutory right of redemption: Arizona, Connecticut, Delaware, Hawaii, Illinois, Iowa, Louisiana, Maryland, Massachusetts, Mississippi, Montana, New Hampshire, New York, Oklahoma, Pennsylvania, South Carolina, Texas, Washington D.C., and West Virginia (*United States Foreclosure Laws* (2020)
OO	1 if property under mortgage is owner occupied, 0 if not.
PDJ	Prohibited deficiency judgement. The following states prohibits deficiency judgement: Delaware, Iowa, Massachusetts, Mississippi, Missouri, Nebraska, West Virginia (*United States Foreclosure Laws* (2020)
PT	Property type · CO = Condo, Co-op or manufactured housing · PU = PUD · SF = Single Family homes
SNS	Seller (originator) name is not the same as servicer name (1) else 0
TID	Months between first 30 DPD and default date
TTR	Time to resolution. Months between default date to property disposition, which is also when A2 and A3 are realised
INT	Mortgage rate at the time of default
UMP_L3	State level unemployment rate one quarter prior to time of default
GDPCY_L3	State level GDP year-on-year change one quarter prior to default

Variable	Definition
R_SS	Proportion allocated to short sale resolution out of 5 resolutions excluding cure, attrition and repurchase, taken at month of default
R_TP	Proportion allocated to third party sale resolution out of 5 resolutions excluding cure, attrition and repurchase, taken at month of default
DRD	Default rate of the entire Freddie Mac portfolio at the time of default
DT	Indicator for non-standard dataset. Loans where this has a value of 1 are from an additional dataset also provided by Freddie Mac.

4.4.3. Variable transformations

Several transformations were made to create key variables in modelling the recovery rate R. The first transformation involved defining default, which was considered to occur when an account becomes at least 90 days past due or goes into REO sale, whichever comes first. REO sale is a type of resolution where the lender gains ownership of the collateral through court proceedings and sells it to cover the amount owed. This definition of default was chosen to align with regulatory standards such as those mentioned in Bank for International Settlements (2006) and common banking practice. If an account defaults, cures, and then re-defaults, only the first default is used in this study.

The exposure at default (EAD) is defined as the total outstanding amount at default after future modifications from debt consolidation have been added. Modifications are defined as any instance when the outstanding principal balance increases over time, and to control for possible bias from modifications, a dummy independent variable is included in the model to serve as a flag for cases when modifications occurred.

Monthly records are kept for account performance, but data quality issues may result in missing records. Therefore, the unpaid principal balance right before collateral disposition must also include the modification amount. In the case of the Freddie Mac dataset, EAD is defined as the higher value of default or pre-collateral liquidation balance. Accrued interest during delinquency is not considered, similar to Goodman & Zhu (2015). To simplify the models, an account is defined as cured if there is no record of 30+ delinquency for 12 consecutive months.

The definition of these quantities is laid out in **Table 18**.

Table 18. *Freddie Mac concepts, given and derived quantities*

Concept	definition
Default	90 days past due or any loss experienced, in alignment with both IFRS9 and Basel II regulations
Date of default (t)	The date when an account first hits 90 days past due, if sudden loss without delinquency, then the loss date will be recognised as date of default
Cure	A once defaulted account which has never been 30+ days past due for 12 consecutive months, for an account whose performing instance has concluded.
Scheduled balance	Scheduled balance if no modification or advance payment has been made since origination
Balance at default	Total unpaid balance as at date of default (90 days past due or earlier if went to REO earlier)
Total Modification	Pre-collateral balance – balance at default
Pre-collateral balance	Unpaid principal balance before zero balance code is assigned. Collateral recovery and Stage 3 recovery happen in zero balance code assignment. This is zero if an account cures.
EAD	Exposure at default; max (balance at default, Pre-collateral balance)
Accrued interest	(First delinquency unpaid balance – non-interest-bearing unpaid balance) * (Current Interest rate – 0.35) * (Months between Last Principal & Interest paid to date and zero balance date) * 30/360/100. This formula is defined in the Freddie Mac User Guide in Freddie Mac (2019). This quantity will not be used as prescribed by Goodman & Zhu (2015)
Origination UPB	Origination unpaid principal balance
MI recovery	Dollar amount recovered from MI
Non-MI recovery	Dollar amount recovered from non-collateral and non-MI source post collateral disposition
MV	Net proceed from property sale. The amount credited to the lender when collateral has been sold; net of allowable expenses related to selling.

Concept	definition
C_0	Origination collateral value extracted from origination LTV
DDTI	Dynamic debt-to-income ratio. DTI at origination x DTI_t/DTI_0. DTI_t is DTI index at time of default, DTI_0 is DTI index at time of origination. DTI index matched by MSA.
DLTV	Default UPB / (C_0 / HPI_0 * HPI_t), HPI_t is from FHFA at default and matched by MSA
MI	Indicator whether an account has mortgage insurance or not. 1 if MI present, 0 otherwise.

This table defines core modelling quantities, concepts and their formula, if applicable, as they are used in Freddie Mac dataset. Some quantities like DDTI and DLTV are not solely obtained from the Freddie Mac dataset. DDTI is DTI at origination scaled by a DTI index obtained from the Federal Reserve. DLTV on the other hand is CLTV where the collateral value is scaled using FHFA HPI and loan amount is the updated unpaid principal balance at default.

4.4.4. Definition: dependent variables

The study includes outliers, which are boundary values containing plausible cases of recovery rate (R) for training models. These outliers are shown in **Table 19** and are excluded from the training dataset, as they mostly involve loans with small outstanding amounts where expenses dominate the net recovery. However, they are included in assessing model performance as they form part of real-world observations.

***Table 19**. Derived dependent variables definition for Freddie Mac*

Variable	definition
A	EAD - balance prior to collateral liquidation + asset net sale proceeds + MI recoveries + non-MI recoveries - expenses
EAD	the higher of balance at default and balance prior to collateral liquidation
R	A/EAD. R is filtered to [-5,5] for charged off and REO, at (-inf,2] for short sale, and [-2.5,2.5] for third party sale to control for extreme outliers while training the models. No R level filters used for the test datasets
R’	R if A > 0 and 0 otherwise

This table defines dependent variables for modelling and their formula as they are used in Freddie Mac dataset.

4.4.5. Data filters

Certain observations such as accounts that are still performing after a default event are treated as inconclusive and are excluded to ensure that the model is well-informed. **Table 20** presents the reasons for exclusions and the number of observations excluded for each category. After

applying the necessary filters, the study was left with 1,544,336[11] defaulted loans out of a starting number of 3,026,608.

Table 20. *Table of exclusions*

Exclusion Description	Accounts
Excluded due to inconclusive resolutions	1,161,684
Property net sale amount marked as unknown	914
Mortgage insurance (MI) percentage coverage unknown	5,113
Components of LC missing so can't be calculated	5,550
Recovery rate R or EAD changed by more than 1% from the last Freddie Mac version so can't be trusted	45,698
GDP at state level missing for quarter prior to default date	196,965
Recovery percentage missing for some of the 5 main resolutions at default date	324
Property type unknown/ unavailable. Only done for training dataset. Quantity only an illustration using the first in-time sample with replacement.	13
One of the components of DLTV is missing so can't be calculated. Only done for training dataset. Quantity only an illustration using the first in-time sample with replacement.	18,529
One of the components of DDTI missing so can't be calculated. Only done for training dataset. Quantity only an illustration using the first in-time sample with replacement.	10,930
Interest rate above 50% - data quality error. Only done for training dataset. Quantity only an illustration using the first in-time sample with replacement.	2
Redefault may cause confusion while in training. Only done for training dataset. Quantity only an illustration using the first in-time sample with replacement.	8,647
Observation post November 2016 are likely to change significantly. Only done for training dataset. Quantity only an illustration using the first in-time sample with replacement.	27,583
Outliers identified for recovery rate R. Values differ by resolution. Absolute value bounds for charge-off/ REO = 5, Short sale = 2, and third-party sale = 2.5. Anything beyond are excluded. Only done for training dataset. Quantity only an illustration using the first in-time sample with replacement.	3
Few cases where LC appears to be too high. This is set at > 5. Only done for training dataset. Quantity only an illustration using the first in-time sample with replacement.	317

[11] Exact number changes depending on sampling window and round. This is for 1st sampling round for in-time.

This table exhaustively reports the data clean-up process taken prior to modelling. Exact number changes depending on sampling window and round. This is for 1st sampling round for in-time.

It should be noted that even though 1.5 million defaulted loans are left for modelling after exclusions, the actual data used for modelling may still vary depending on the type of test conducted. To ensure the reliability of the model, we perform random sample selections with replacement and out-of-time predictions, which are repeated 10 times. This process is explained further in section 4.5.9.

4.4.6. Discussion of variable correlations

We now refer to **Table 16** which) between dependent and independent variables. Because we start with simple linear regression models, this matrix may serve as a point of validation for coefficient values. R here is shown to be strongly negatively correlated with DRD, UMP_L3, DLTV and TTR, and strongly positively correlated with GDPCY_L3 and MIP.

FICO score, on the other hand, has a relatively flat relationship against R if it were not for note or reperforming sale. Nevertheless, the mostly slight negative correlation is confirmed by Do et al. (2020) through their probability of zero loss model, which played a major part in their LGD estimation. According to them, when high FICO customers face negative life events like loss of job or divorce, they first try and fix their situations and only give up when all hope is lost. One motivation for this is that damage to FICO scores from defaulting and foreclosing affects a customer for at least the next 10 years. While this is the case, they mentioned that the relationship is positive once non-zero LGD is involved instead of total LGD, aligning with the intuition of positive correlation between LGD and PD.

While it is widely known that the distribution of LGD is bimodal (Hlawatsch & Ostrowski, 2016; Hurlin et al., 2018; Krüger & Rösch, 2017; Witzany et al., 2012), transformations on LGD prior to modelling often lead to poorer accuracy (Loterman et al., 2012). On the other hand, if done properly, results are still at par with raw counterparts (Qi & Zhao, 2011). As such, no transformations are performed to address the non-normality of uncured recovery rates for this exercise.

4.5. Methodology

Figure 13 illustrates a sequence of events post a default event.

There are three resolutions that yield to almost, if not exactly, 100% recovery of outstanding balance: these are:

- Cure[12] - a defaulted account may return to, and stay performing
- Repurchased – occur when the loan originator or servicer is contractually obliged to buy the mortgage back—typically because of a breach of Freddie Mac's representations and warranties (for example, incomplete documentation, underwriting defects or misstatements about borrower creditworthiness). These repurchases are an administrative remedy, not a sign that the borrower's credit has been restored, which is why we treat them separately from true cures.
- Repaid early/ refinanced – if available, the borrower may utilise funds and other resources to prepay the defaulted loan

If uncured, not repurchased and not refinanced/undergoing attrition, the defaulted loan may go through any of the following processes to liquidate the contract/collateral:

- Charge-off – after some assessment, the lender decides to close of its position by writing off the entire outstanding amount
- Short sale – property is sold by the distressed homeowner
- Third party sale – property may be sold via third party in a foreclosure auction
- Note sale – the loan is sold by the lender, often at a discount, to another willing lender
- Reperforming sale – a type of note sale where the loan being sold has started performing again – with or without modification
- REO sale – the lender may choose to take ownership of the property and sell it right away

While the expected recovery rate for the first three resolutions is mostly 1, recovery rate can be modelled for each of the remaining resolutions. We define these conditional recovery rates

[12] Note that a full (100 %) recovery is not the same as a cure. We define a cure as a defaulted loan that ultimately returns to performing status—either by being repurchased or by prepayment—so that no loss is recorded. By contrast, a 100 % recovery simply means all outstanding cash flows were collected, even if the loan never formally "cures." Due to the dataset's limited observation window, very few defaults have had time to resume performing, making genuine cures hard to identify. Furthermore, certain repurchases or prepayments may reflect administrative processes rather than an obligor's restored creditworthiness; these should therefore be modelled separately, not conflated with true cures. Finally, we exclude two unresolved categories from our analysis:

- Defaults that have begun to recover funds but whose resolution remains incomplete.
- Loans still classified as performing, whose ultimate status is undetermined.

as R_{res_i} where *res_i* may be one of the 8 resolutions where some of them may simply take values of 1.

As there is no way to determine what resolution a defaulted account goes through other than in hindsight, predicting LGD would mean simply multiplying the expected conditional recovery rate R_{res_i}, $E(R_{res}|res)$, by the proportion of accounts expected to go through specific resolution *res*. These proportions, $P(res)$, may be modelled independently of $E(R_{res}|res)$, and together they sum up to 100%.

In short, expected recovery rate, E(R), is the weighted average of different recovery rates conditional on each type of resolution, which may follow different distributions.

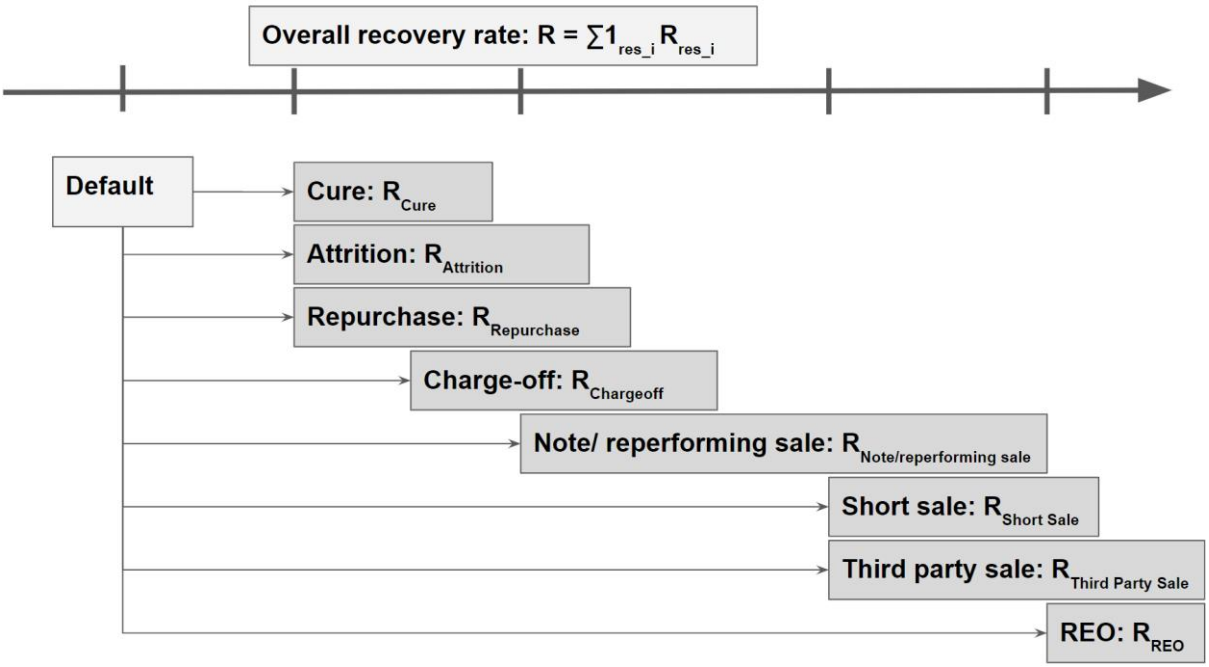

Figure 13: *Resolution diagram for recovery rate.*

A defaulted account goes through a list of decision points acted on by lenders. While the resolution is always among the 8 listed above, the usual time it takes to reach each resolution varies. In particular, it is only when the first three resolutions are identified to be unachievable when lenders resort to more drastic measures such as foreclosure alternatives (third party sale, note/ reperforming sale and short sale)– that is, if it is expected to be worth the effort. Otherwise, an account is simply charged off. If foreclosure alternatives are not achievable but collaterals are present, REO is called upon where the lender takes ownership of the property and disposes immediately. In aggregate, overall recovery rate R is simply $\sum_{i=1}^{N} 1_{res_i} R_{res_i}$*, where* res_i *exhausts each of the 8 resolutions,* 1_{res_i} *is an indicator function taking on a value of 1 if the resolution is called upon, and 0 otherwise. Note this is only possible in hindsight, thus predicting recovery rate* $E(R) = \sum_{i=1}^{all\ resolutions} P(res_i)E(R|res_i)$.

4.5.1. Definitions

Three key concepts are defined to model recovery.

Default happens when either the bank deems the obligor unable to pay back a loan or the obligor goes more than 90 days past due for a material amount owed (Bank for International Settlements (2006). When a defaulted account is non-delinquent for 12 consecutive months, it is considered cured. This study ignores subsequent defaults. This definition is aligned with Dermine & de Carvalho (2006) as they caution against biases when considering each default case for multiple default customers. The definitions are summarised in **Table 21**.

Most mortgage LGD studies like Leow et al. (2014); Leow & Mues (2012); Tong et al. (2013) assume foreclosure or repossession as the point of default. While this may seem to result to more accurate predictions, it misses out on accounts which have defaulted but have not been repossessed or cured/repurchased/repaid early. This study explores a more comprehensive sample and by doing so predictive power may not be directly comparable with the literature. The models proposed predict LGD of all impaired accounts, regardless of what resolution they may eventually end up with.

Table 21. *Glossary of key concepts*

Variable	Definition
Default	A retail mortgage account defaults when it reaches 90 days past due for the first time or when it goes through foreclosure proceedings, whichever comes first.
Default date	The date in which default happened
Cure	When a retail mortgage account defaults but eventually starts to perform again. An account is considered to have recovered if it has been performing without at least 30 days past due for 12 consecutive months. If default happens again, this instance is a new default event performed by a different account. Because the previous instance has been considered resolved, it is treated as cured.

Table 22 enumerates all core quantities important for modelling in this research. EAD is an important core quantity given that LGD is a proportion of this. When a defaulted account eventually prepays or matures, and the lending relationship ends without any write-offs, it is considered to have been cured, repaid early or repurchased. By definition, LGD is 1 – R, where R is the total amount recovered as a proportion of EAD.

R is the sum of R_{res_i} through all possible resolutions. While there may be 8 possible resolutions, each account can only go through one. As such, there is only one specific value of res where R_{res} may have a non-zero value and the rest are 0.

Table 22. *Definitions of core quantities and ratios*

Variable	Definition	Formulae
EAD	Exposure at default. The amount owed at the default date.	
LGD	Loss given default. The proportion of EAD that is considered a loss.	$LGD = \begin{cases} 0 & \text{if cured} \\ 1 - R & \text{if non-cured} \end{cases}$
R	The proportion of EAD recovered, for non-cured accounts.	$R = \sum_{i=1}^{all\ res} R_i$
R_{res_i}	The total dollar amount collected, net of costs.	$R_{res\_i} = \begin{cases} R & \text{if resolution = res} \\ 0 & \text{otherwise} \end{cases}$
Res_i	The specific resolution a defaulted account eventually falls under	

This table defines core modelling quantities and their formula.

4.5.2. Overview of modelling narrative

The hypothesis for this paper is that modelling recovery rate in separate resolutions, with distinct models for each resolution, and aggregated together through weighting by modelling the propensity of each resolution, can result in a better prediction for the overall recovery rate R than directly modelling R. However, there are some subtleties in exactly what quantities are modelled. With this, steps taken to ensure best modelling practice must be carefully considered. This section provides an overview of the model structure. Then succeeding sections discuss each component.

- As the first benchmark for the purposes of explanation, we define our **naïve** model as an OLS model of R directly.

- We add a 2nd benchmark model, a **2-step model**, which claims to have outperformed the OLS model.

- We define a **perfect resolution model**, $E(R) = E(R|resolution)$ assuming resolution is known. The accuracy measure for this model is used as the best performance if resolutions are predicted with 100% accuracy. This is used in later steps when choosing which recovery rate model should be used for specific resolutions. For simplicity, this is to be performed using simple average, OLS and the 2-step model mentioned above. Due to its definition, these are also called the **conditional recovery models**.

- We propose the first **fixed proportion model** where the probability of resolution is a fixed proportion. As an example, if 10% of all defaulted loans went to short sale, then the probability that the resolution is short sale is 10%. The sum of all probabilities for each resolution is 100%.

- Next, we propose the use of more **advanced resolution models** to estimate probability of resolution such that available drivers up until the time of default are used to estimate probability of each resolution. As with the previous model, the sum of all probabilities for each resolution is 100%.

- Finally, we choose the advanced resolution model and **combine** it with the best conditional recovery models for each resolution.

4.5.3. Benchmark: Naïve OLS model

The naïve model is used as the first benchmark for the purposes of explanation and performance. It is simply a direct model of the recovery rate R as a single quantity, for example a simple ordinary least squares (OLS) model of R against several independent variables.

4.5.4. Benchmark: 2-step benchmark model

The second benchmark is a 2-step model taken from Do et al. (2020). It is chosen as a benchmark for its simplicity, recency, and its claim of outperformance against OLS. The underlying concept of the model is as follows:

$$R = \begin{cases} 0 \; when \; A = 0 \\ R \; when \; A > 0 \end{cases} \quad \text{..... 4.1}$$

Such that

$$E(R) = P(A > 0)E(R|A > 0) \quad \text{..... 4.2}$$

Where $A = R \times EAD$.

4.5.5. Conditional resolution models P1-P3

The conditional resolution model predicts recovery rate for each perfectly predicted resolution. Among these predicted recovery rates, only one is chosen given an assumption that resolution is already known. This obviously is not realistic and can't be used for real world prediction. However, this:

- sets the highest level of performance if the probability of resolution is to be modelled, assuming the same stand-alone linear regression models predicting recovery rates under each resolution, and;
- may be combined with a model that predicts the probability of resolution separately at a later stage.

Three models are introduced:

- P1 *simple EAD-weighted average*:
- P2 *ordinary least squares*:
- P3 *2-step benchmark*:

For each resolution, a best model among P1-P3 is chosen for the validation dataset and used in section 4.6.2.3 in an attempt at a more accurate model.

4.5.6. Fixed proportion to each resolution

Given the highest level of accuracy for a probability of resolution model, a simple model is introduced to serve as a worst-case model. That is, if any proposed model had an accuracy worse than this, then this model would be best used. Defining both best and worst possible levels of accuracy helps with finding the best proposed model.

4.5.7. Advanced resolution model

Given the best- and worst-case measures for model accuracy, this section proposes the use of multinomial logit model to estimate the probability for each resolution to happen. This involves estimation of 1 set of parameter coefficients per resolution. The structure for the probability terms takes the following form:

$P(res_i) = e^{F_i(x)}P(res_1)$ for i = 2 to N and N is the index for the last resolution. For simplicity, i=1 is chosen to be the "reference" for this such that $P(res_1) = \frac{1}{1+\sum_{j=2}^{all\ resolutions} e^{F_j(x)}}$ $\frac{1}{1+\sum_{j=2}^{N} e^{F_j(x)}}$. This way, the sum of all probabilities is always 1. x represents a set of parameters. While random forest models are likely to achieve more accurate results, these models are least preferred by lending institutions due to their lack of transparency.

4.5.8. Combination: p(resolution) and E(R|resolution)

Once combined, the ultimate model takes the following form:

$E(R) = \sum_{i=1}^{all\ resolutions} P(res_i)E(R|res_i)$ 4.3

Where $P(res_i)$ is the probability that resolution i happens as estimated above, $E(R|res_i)$ is the most accurate conditional recovery model for resolution res_i and res_i is the ith resolution.

4.5.9. Robustness: sample selection methodology

The Freddie Mac dataset is sampled in two ways under this study:

In-time sampling. A random 30% is set aside as test dataset, and another independent random 35% of the dataset is used as training sample. The rest is used as validation sample if needed for modelling. The need for a validation dataset is specific to each model.

Out-of-time sampling. Test dataset is chosen by calendar year from 2007 to 2015. Due to low volume, all defaults post 2016 will be merged with 2016 test dataset. A random 50% of the remaining data is used for training and the rest is for validation, if needed.

Each method of sample selection is performed 10 times on a 1/10 sub-sample of the dataset and accuracy is measured using the 10 test datasets. An illustration of the two types of sampling is presented in Figure 14.

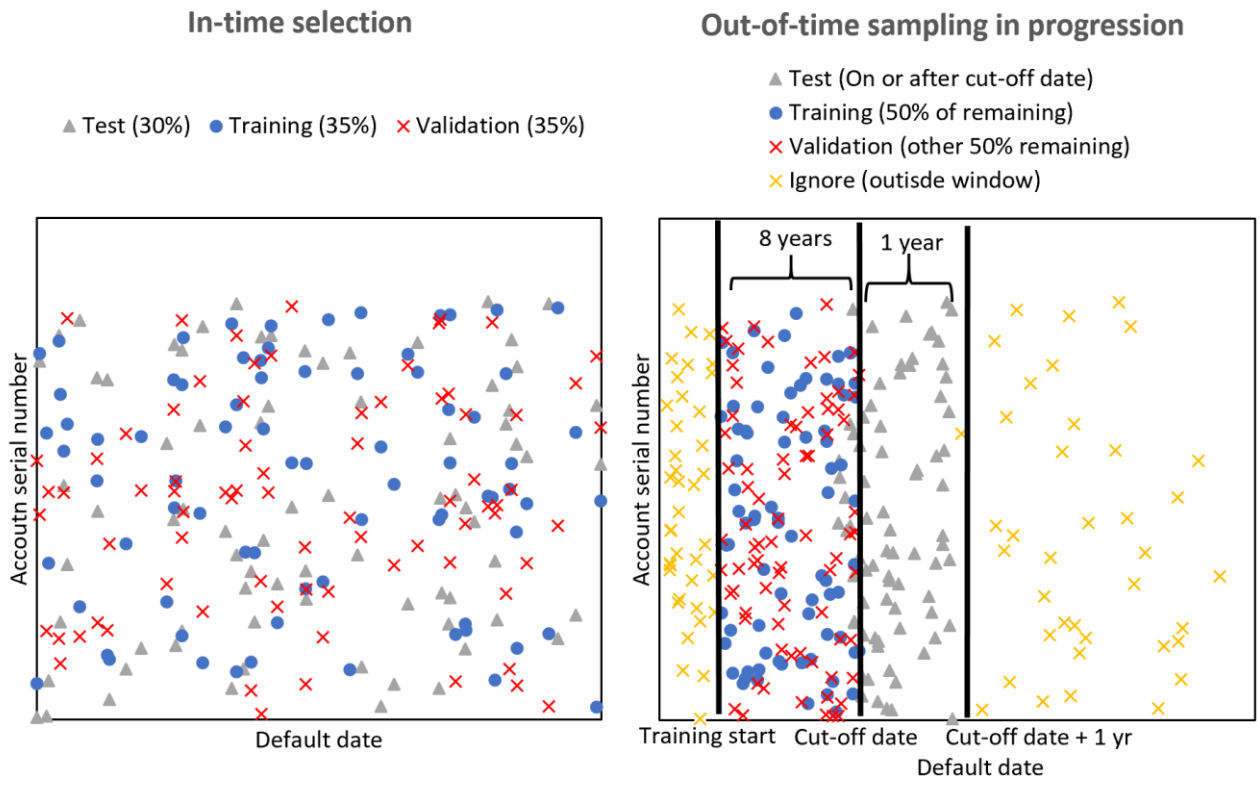

Figure 14. *Visual representation of two kinds of sampling employed in this paper.*

Vertical axis in both charts are illustrations of a unique account identifier, artificially created for easier visualisation, while horizontal axis denotes default date. For each chart, 1 model is built using data composed

of the round (blue) points and RMSE/r-square are measured by applying each model on dataset composed of triangles (grey). Left chart shows in-time sample selection and right chart shows out of time sample selection. The rest (red) are ignored. Note that the source of randomness for out-of-time sampling only applies to the selection for model build sample. The test dataset is exactly the same in the 10 independent rounds of random sampling with replacement performed.

4.5.10. Overall accuracy measure

RMSE and r-square are used to judge model accuracy. These measures are used in a few studies such as Hurlin et al. (2018); Loterman et al. (2012); Qi & Zhao (2011); Thomas et al. (2012) in the review of models they covered. There are two key aspects being tested here: the ability of each model to discriminate between good and bad uncured recovery rates (ranking ability), and how accurate each prediction is (error measure). Performance measures like area under curve (AUC) and linear correlation between dependent variable and predictions are most common forms of checks for model discrimination while RMSE and r-square are measures for accuracy. While accurate models have good discrimination, models that discriminate well are not necessarily accurate (Loterman et al., 2012). Given this, performance measures like RMSE and r-square are given priority over others. While RMSE alone is enough to provide performance ranking among models proposed, r-square is also considered for comparison with the literature.

RMSE and r-square are employed to assess the performance of each model by looking at predictions. RMSE focuses on the error term expressed in units of the dependent variable R, and r-square looks at the RMSE of a model and measures how much better it is compared to using the mean as predictor. While r-square can theoretically have values less than 0, it can be floored at 0 under the assumption that the mean is used if predictions are less accurate.

4.5.10.1. Performance measure for p(resolution) model

In addition to RMSE and r-square, accuracy of predicting resolutions can also be measured by getting the percentage predicted correctly. This is given by:

$$RCM = \frac{\sum_{obs=1}^{ALL}\sum_{i=1}^{N} P(res_{i,obs})\mathbb{I}_{res_{i,obs}}}{ALL} \quad \text{..... 4.4}$$

Where RCM stands for Resolution Correctness Measure, N is the total number of resolutions, $\mathbb{I}_{res_{i,obs}}$ has a value of 0 unless the realised resolution for observation *obs* is res_i, and ALL is the total number of observations. Given the definition of $P(res_{i,obs})$, this measurement ranges from 0 to 1, where 1 means perfect resolution prediction and 0 means nothing is predicted

correctly. This definition may be performed on subsets of the data (i.e. resolutions) to determine how the model has performed in specific cohorts.

Now that performance measures are defined, we move to discussion of model coefficients.

4.6. Empirical results

Continuous independent variables are first standardised prior to regression for easier comparison. Each independent variable is subtracted by its own mean, and the result is divided by the respective variance. This way, the magnitude of coefficients may be used to compare the influence of each independent variable on recovery rate. Categorical variables are expanded into X-1 dummies, where X is the finite number of categories. This transforms categorical variables into continuous form. Because the purpose of presenting coefficients is to point out key relationships not highlighted by benchmark models, this exercise will not cover any non-linear relationships.

4.6.1. Benchmark models 1 and 2: naïve and 2 step model

In predicting uncured recovery rate, one of the simplest model structures is a linear regression model or ordinary least squares (OLS). Here, the expected value of uncured recovery rate R is modelled as:

$$E(R) = f(X) \quad \text{..... 4.5}$$

where f is a generic linear combination of X, a matrix of independent drivers defined in **Table 17** and R is as defined in **Table 21**. Theoretically, the quality of prediction from a linear regression model increases with more independent nominated drivers, but too many drivers come at a risk of overfitting/over specification. On the other hand, a new generation of models with mathematical structures like multi-step conditional approaches came about after exploration of new variables in predicting LGD have slowed down. Several iterations were explored to accommodate the multimodality of LGD. Leow et al. (2014) built separate models for probability of repossession and a 'haircut model' for repossessed accounts and this multi-step model has become standard.

Using subprime US lending data provided by a private institution, Do et al. (2020) built a two-step conditional probability model but with a dependent structure and found that the chances of zero losses are driven by borrower liquidity constraints and positive equity while non-zero loss values are driven by negative equity. Given its recency and the soundness of this concept, the two-step conditional framework of Do et al. (2020) is used as the second benchmark model.

We apply this framework to build a two-step conditional framework consisting of probability of non-positive (i.e., less than or equal to 0) recovery and expected value of positive recovery (i.e., greater than 0) using the Freddie Mac dataset:

$$E(R) = P(R > 0)E(R|R > 0) \quad \text{4.6}$$

Here, $P(A > 0)$ is equivalent to $P(R > 0)$ because EAD is always positive. This quantity is a probit function, while $E(R|R > 0)$ will simply be a linear function without censoring. They are allowed to have dependency on each other.

Table 23 shows the average standardised in-time coefficient estimates for R and $P(A > 0)$ using both benchmark models. We find five dominant factors for benchmark model 1: DLTV, TTR, LOB, MIP and DRD. The five continuous variables align with **Table 16** except LOB. Among those highlighted with high correlation, unemployment and GDP year on year change interestingly did not turn out to have material coefficients, which may indicate that other independent variables are far better predictors of R and that serving as control variables have removed weights materially for both variables.

Table 23. *Standardised regression coefficients for R using Benchmark Models 1 and 2.*

Parameter	B1: E(R)	B2: P(A>0)	B2: E(R')
Intercept	0.7379***	2.7068***	0.7544***
	(0.0004)	(0.009)	(0.0004)
DLTV	-0.0976***	-0.1995***	-0.0952***
	(0.0005)	(0.0066)	(0.0005)
DDTI	0.0093***	0.0338***	0.008***
	(0.0004)	(0.0047)	(0.0004)
FICO	-0.0185***	-0.0066	-0.0178***
	(0.0004)	(0.0052)	(0.0004)
TID	0.0121***	0.0298***	0.0111***
	(0.0005)	(0.0048)	(0.0004)
TTR	-0.0381***	-0.2357***	-0.0287***
	(0.0004)	(0.0044)	(0.0004)
LC	0.0022***	0.0455***	0.0003
	(0.0005)	(0.0079)	(0.0005)
LOB	0.0722***	0.6991***	0.0467***
	(0.0005)	(0.0063)	(0.0004)
MOB	-0.0262***	-0.1539***	-0.0201***
	(0.0005)	(0.0063)	(0.0005)
MIP	0.0568***	0.145***	0.0527***
	(0.0004)	(0.0064)	(0.0004)
DT	-0.0221***	-0.0004	-0.0216***
	(0.0004)	(0.0056)	(0.0004)
FHB	-0.0059***	-0.0184**	-0.0055***

Parameter	B1: E(R)	B2: P(A>0)	B2: E(R')
	(0.0005)	(0.0067)	(0.0004)
INT	-0.0264***	-0.0876***	-0.0232***
	(0.0004)	(0.006)	(0.0004)
OO	0.0264***	0.1504***	0.0189***
	(0.0004)	(0.0041)	(0.0004)
LP_C	-0.0143***	-0.0458***	-0.0122***
	(0.0005)	(0.0057)	(0.0004)
LP_P	0.0284***	0.1283***	0.0261***
	(0.0005)	(0.0076)	(0.0005)
MF	-0.002***	0.0134**	-0.0026***
	(0.0004)	(0.0054)	(0.0004)
NB_01	-0.0145***	-0.0993***	-0.0123***
	(0.0004)	(0.0056)	(0.0004)
NJF	0.0149***	0.1004***	0.0126***
	(0.0004)	(0.0049)	(0.0004)
NSRR	0.0036***	-0.0011	0.004***
	(0.0004)	(0.0051)	(0.0004)
PDJ	0.0045***	0.0362***	0.0029***
	(0.0004)	(0.0051)	(0.0004)
SNS	0.0017***	-0.0148**	0.0015***
	(0.0004)	(0.0055)	(0.0004)
PT_CO	-0.0056***	0.1164***	-0.0099***
	(0.0004)	(0.0062)	(0.0004)
R_SS	0.0009*	0.0002	0.0013***
	(0.0005)	(0.006)	(0.0004)
R_TP	0.0005	-0.0048	0.0007
	(0.0004)	(0.0055)	(0.0004)
DRD	-0.049***	-0.1066***	-0.0466***
	(0.0005)	(0.0067)	(0.0005)
UMP_L3	-0.0188***	-0.0908***	-0.016***
	(0.0005)	(0.0073)	(0.0005)
GDPCY_L3	0.0041***	0.0398***	0.0033***
	(0.0004)	(0.0078)	(0.0004)

*This table reports the coefficients obtained for models under the 2 benchmark models. Level of coefficient significant as follows: *** for p-value < 0.01, ** for p-value < 0.05* and * for p-value < 0.1. Values come in the following order: coefficient, degree of significance/ standard error. The estimates are based on in-time training sample. *B1 is chosen as a benchmark model due to its popularity in both the literature and industry. B2 is the second benchmark model which is a slightly modified version from Do et al. (2020). It has been chosen for being relatively recent, simple and for claiming outperformance against OLS.*

4.6.2. Proposed model framework

Two benchmark models are built and compared with results obtained from combining and permuting direct application of OLS and alternative models introduced in 4.4.2 to predict R. The models and their corresponding names are defined for convenience in the succeeding sections.

The models and their corresponding names are defined for convenience in the succeeding sections.

There are 3 ways to model each conditional recovery rate per resolution: (1) simple average (CA), linear regression (CL) and a similar two-step model as that of B2 (CT). Their resolution-level accuracies are illustrated in **Table 24**. This also serves as the best performance per resolution had the probability of resolution model been 100% accurate. We define this as step 1. With this, there is flexibility to have a different best model per sampling type and resolution. For simplicity, only one best model is determined for all 10 sampling rounds.

Table 24. *RMSE for Step 1 models, by resolution*

In-time	**CA**	**CL**	**CT**
Charge-off	2.9603	**2.9578**	2.9745
Short sale	0.2244	**0.1839**	0.1840
Third party sale	0.3107	0.2711	**0.2678**
REO	0.3300	0.2820	**0.2792**
Note/ reperforming sale	**0.2587**	0.3225	0.3202
Defaulted repurchased	0.2193	**0.2070**	0.2070
Training 1999-2006 Test 2007	**CA**	**CL**	**CT**
Charge-off	0.5776	**0.5062**	0.5159
Short sale	0.1959	**0.1556**	0.1608
Third party sale	0.1865	**0.1680**	0.1745
REO	0.3279	0.2549	**0.2532**
Note/ reperforming sale	0.2553	**0.2002**	0.2513
Defaulted repurchased	0.1338	0.1321	**0.1321**
Training 2000-2007 Test 2008	**CA**	**CL**	**CT**
Charge-off	0.5273	**0.4749**	0.5128
Short sale	0.2385	**0.1803**	0.1853
Third party sale	0.2482	**0.2088**	0.2092
REO	0.3347	0.2705	**0.2689**
Note/ reperforming sale	0.2997	**0.2250**	0.2303
Defaulted repurchased	0.1563	**0.1526**	0.1527
Training 2001-2008 Test 2009	**CA**	**CL**	**CT**
Charge-off	0.4621	**0.4386**	0.4541
Short sale	0.2656	**0.1981**	0.1982
Third party sale	0.3268	**0.2403**	0.2404
REO	0.3259	0.2666	**0.2648**
Note/ reperforming sale	0.3233	0.2412	**0.2409**
Defaulted repurchased	0.1948	**0.1879**	0.1882
Training 2002-2009 Test 2010	**CA**	**CL**	**CT**
Charge-off	0.4349	**0.4198**	0.4388
Short sale	0.2482	**0.1962**	0.1962
Third party sale	0.3155	0.2439	**0.2432**
REO	0.3121	0.2594	**0.2578**
Note/ reperforming sale	0.3257	0.2490	**0.2478**
Defaulted repurchased	0.2173	**0.2070**	0.2072
Training 2003-2010 Test 2011	**CA**	**CL**	**CT**
Charge-off	0.4219	**0.4077**	0.4350
Short sale	0.2387	**0.1922**	0.1923
Third party sale	0.3066	0.2448	**0.2441**
REO	0.3081	0.2575	**0.2557**
Note/ reperforming sale	0.3263	0.2560	**0.2550**
Defaulted repurchased	0.2369	**0.2241**	0.2242
Training 2004-2011 Test 2012	**CA**	**CL**	**CT**
Charge-off	0.4216	**0.4050**	0.4358
Short sale	0.2315	**0.1886**	0.1887
Third party sale	0.3025	0.2496	**0.2487**
REO	0.3075	0.2580	**0.2561**

Note/ reperforming sale	0.3271	0.2649	**0.2640**
Defaulted repurchased	0.2495	**0.2352**	0.2353
Training 2005-2012 Test 2013	**CA**	**CL**	**CT**
Charge-off	0.4338	**0.4135**	0.4487
Short sale	0.2265	**0.1858**	0.1858
Third party sale	0.3032	0.2562	**0.2553**
REO	0.3079	0.2594	**0.2573**
Note/ reperforming sale	0.3260	0.2699	**0.2692**
Defaulted repurchased	0.2564	**0.2413**	0.2413
Training 2006-2013 Test 2014	**CA**	**CL**	**CT**
Charge-off	0.4313	**0.4151**	0.4448
Short sale	0.2254	0.1857	**0.1857**
Third party sale	0.3057	0.2653	**0.2638**
REO	0.3085	0.2607	**0.2585**
Note/ reperforming sale	0.3220	0.2702	**0.2692**
Defaulted repurchased	0.2622	**0.2466**	0.2466
Training 2007-2014 Test 2015	**CA**	**CL**	**CT**
Charge-off	0.4461	**0.4333**	0.4557
Short sale	0.2255	0.1854	**0.1854**
Third party sale	0.3080	0.2725	**0.2703**
REO	0.3088	0.2603	**0.2580**
Note/ reperforming sale	0.3212	0.2730	**0.2721**
Defaulted repurchased	0.2657	**0.2503**	0.2503
Training 2008-2015 Test 2016 onwards	**CA**	**CL**	**CT**
Charge-off	0.4550	**0.4482**	0.4672
Short sale	0.2249	0.1855	**0.1854**
Third party sale	0.3101	0.2770	**0.2745**
REO	0.3105	0.2617	**0.2593**
Note/ reperforming sale	0.3208	0.2707	**0.2692**
Defaulted repurchased	0.2707	**0.2550**	0.2550

This table reports the average Root Mean Square Errors (RMSE) of the out-of-sample prediction for Step 1 conditional resolution models across 10 independent rounds of random sampling, segmented by resolution. Model CA uses a fixed EAD weighted average recovery rate, CL uses ordinary Least Squares (OLS), CT uses the same 2-step model as the 2nd benchmark (B2). Each row represents a type of out-of-sample prediction and highlights in bold and red the best performing model. These models are used in the combined model.

As proof of concept, we introduce two models to estimate probability of default: (1) fixed proportion (RF), and (2) multinomial logit model (RM). Their accuracy measures are illustrated in **Table 25**. We define this as step 2. RM has clearly outperformed RF given the nature of the model.

Table 25. *RCM for Step 2 models*

Step 2 models	Average RCM	
	Fixed resolution (RF)	**Multinomial Logit resolution (RM)**
In-time	0.23%	**35%**
Training 1999-2006 Test 2007	0.04%	**34%**
Training 2000-2007 Test 2008	0.05%	**41%**
Training 2001-2008 Test 2009	0.11%	**36%**
Training 2002-2009 Test 2010	0.19%	**35%**
Training 2003-2010 Test 2011	0.23%	**36%**
Training 2004-2011 Test 2012	0.25%	**36%**
Training 2005-2012 Test 2013	0.24%	**34%**
Training 2006-2013 Test 2014	0.21%	**31%**
Training 2007-2014 Test 2015	0.20%	**33%**
Training 2008-2015 Test 2016 onwards	0.18%	**32%**

This table reports the average Resolution Correctness Measure (RCM) of the out-of-sample prediction for Step 2 models across 10 independent rounds of random sampling. Model RF is a set of constant proportions of exposures going to each resolution, and RM uses multinomial logit regression predicting the probability of each resolution. Each row represents a type of out-of-sample prediction and highlights in bold and red the best performing model. RM model clearly outperforms RF for obvious reasons.

Permuted, overall accuracies are shown in **Table 26** where model RM x CT and RM x best step 1 has outperformed the rest for in-time and up to 9 out of 10 types of sampling across 10 independent rounds. This is an outperformance[13] against benchmarks B1 and B2.

Table 26. *RMSE for combined models*

Sampling type	B1	B2	RM x CT	RM x best step 1
In-time	0.277	0.276	**0.267**	0.268
Training 1999-2006 Test 2007	0.346	0.345	0.339	**0.339**
Training 2000-2007 Test 2008	**0.341**	0.344	0.359	0.355
Training 2001-2008 Test 2009	0.316	0.314	**0.308**	0.315
Training 2002-2009 Test 2010	0.306	0.305	**0.296**	0.296
Training 2003-2010 Test 2011	0.278	0.277	**0.270**	0.270
Training 2004-2011 Test 2012	0.340	0.338	0.332	**0.332**
Training 2005-2012 Test 2013	0.363	0.362	**0.358**	0.358
Training 2006-2013 Test 2014	0.296	0.295	**0.292**	0.292
Training 2007-2014 Test 2015	0.287	0.285	**0.282**	0.282
Training 2008-2015 Test 2016 onwards	0.415	0.413	**0.411**	0.411

This table reports the average Root Mean Square Errors (RMSE) of the out-of-sample prediction for combined step 1 and 2 models compared against benchmark performance across 10 independent rounds of random sampling. Model B1 uses OLS, B2 uses a 2-step model, RM x CT combines multinomial logit regression in predicting probability of resolution with a linear conditional recovery model build by resolution. Finally, we combine RM with the best average conditional recovery model as the final model. To recall, choices for best conditional recovery model are CA: fixed EAD weighted average recovery rate, CL: ordinary Least Squares (OLS), CT: the same 2-step model as the 2nd benchmark (B2). Each row represents a type of out-of-sample prediction and highlights in bold and red the best performing model. For most sampling windows except the one where test window is 2008, combined steps 1 and 2 models outperformed both benchmark models B1 and B2, which supports proof of concept.

Figure 15 and **Figure 16** illustrate RMSE and r-square results for each round of random sampling. These were measured for out of sample datasets.

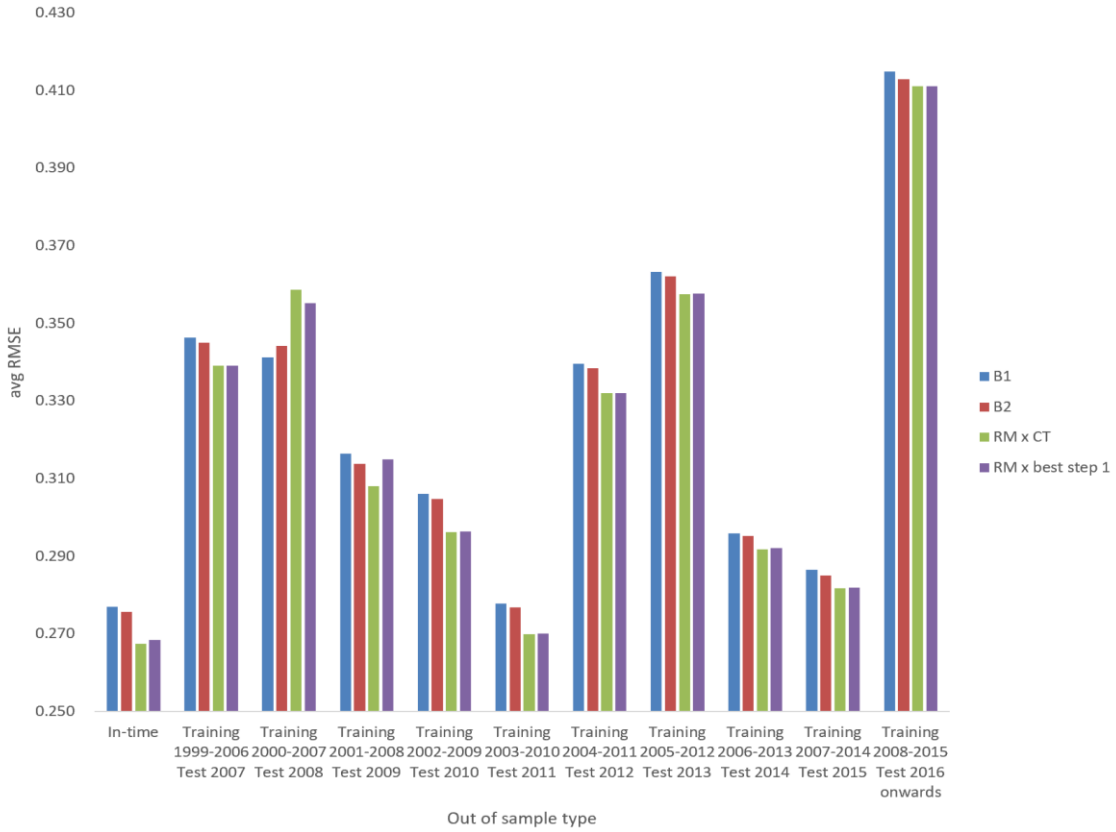

Figure 15. *RMSE out-of-time measured on test sample.*

Y-axis is RMSE, X-axis represents the type of sampling window. Test performed 10 times using random sampling with replacement. The two combined models exhibit the lowest RMSE which means both proposed models outperformed both benchmarks up to 9 out of 10 sampling windows.

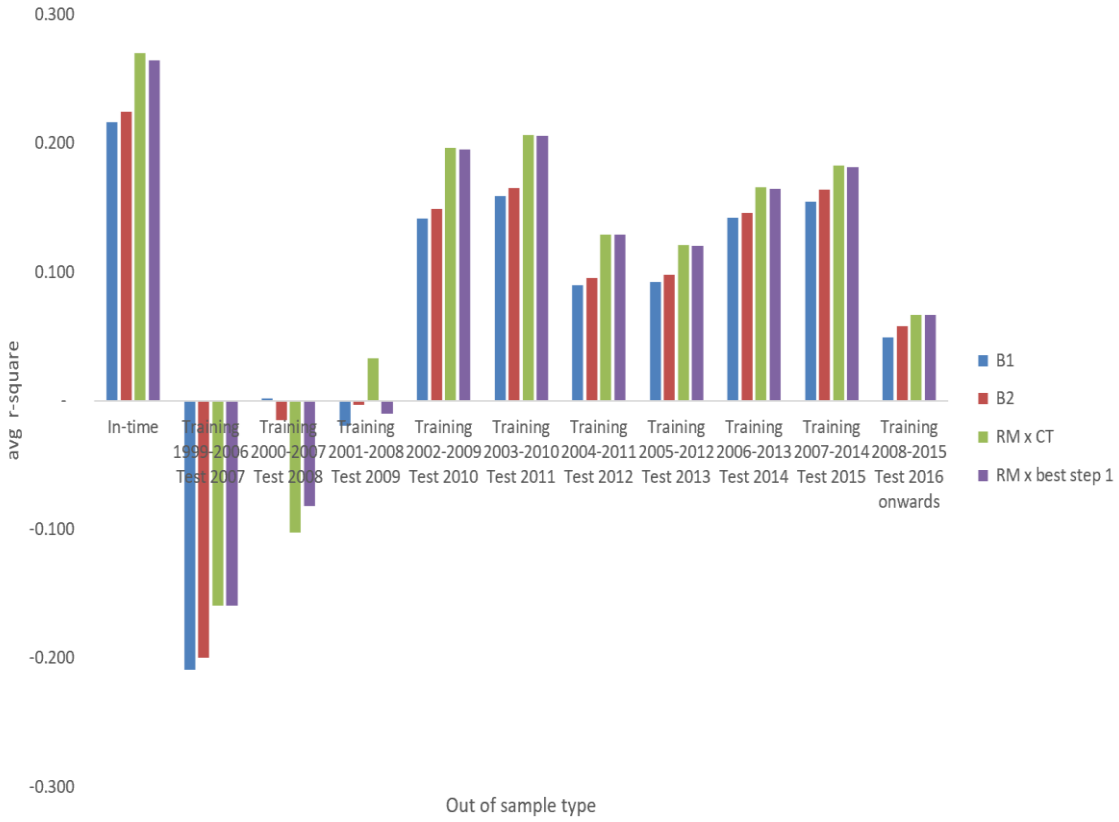

Figure 16. *R-square out-of-time measured on test sample.*

Y-axis is r-square, X-axis represents the type of sampling window. Test performed 10 times using random sampling with replacement. The two combined models exhibit the highest r-square which means both proposed models outperformed both benchmarks up to 9 out of 10 sampling windows. Where combined models didn't work, no model has worked. This is indicated by a negative r-square. A negative r-square simply means that an average could have done a better job. However, since average can't be measured until after the recovery event has transpired, it still can't be used. Given this, the combined models work in 100% of all valid tests.

Methods for sampling are explained in detail under section 4.5.9.

4.6.2.1. Step 1: Conditional resolution models

As the focus of the paper is delivering a proof of concept that the combined models provide a more accurate recovery rate prediction than benchmarks, discussion about parameters is limited to relative materiality of driver coefficients to other independent variables within the same model, provided they are statistically significant.

Table 27. *Standardised regression coefficients for modelled step 1: CL recovery rates.*

Parameter	Charge off	short sale	3rd party sale REO		Note sale	Defaulted repurchased
Intercept	0.295***	0.6834***	0.7049***	0.5963***	0.4941***	0.9016***
	(0.0142)	(0.001)	(0.0013)	(0.0007)	(0.0042)	(0.0016)
DLTV	-0.043***	-0.0625***	-0.1203***	-0.0862***	-0.035***	-0.0134***
	(0.0096)	(0.0008)	(0.0014)	(0.001)	(0.0022)	(0.0017)
DDTI	0.0058	0.006***	0.0064***	0.0084***	-0.0053**	0.0027**
	(0.0066)	(0.0007)	(0.0012)	(0.0006)	(0.0018)	(0.001)
FICO	0.0021	0.002*	-0.0053***	0.0036***	-0.0048*	-0.0064***
	(0.0081)	(0.0008)	(0.0013)	(0.0007)	(0.0019)	(0.0011)
TID	0.0015	0.0065***	0.0026*	0.0035***	0.0133***	0.016***
	(0.0064)	(0.0009)	(0.0012)	(0.0008)	(0.0015)	(0.002)
TTR	-0.0037	-0.054***	-0.019***	-0.0444***	0.0943***	-0.0828***
	(0.0081)	(0.0011)	(0.0014)	(0.0008)	(0.0015)	(0.0017)
LC	0.0319***	-0.0004	-0.0003	-0.0007	-0.0014	-0.0023
	(0.0055)	(0.0012)	(0.002)	(0.0009)	(0.0028)	(0.0022)
LOB	0.1314***	0.0585***	0.059***	0.1028***	0.0508***	0.0105***
	(0.0082)	(0.001)	(0.0015)	(0.0008)	(0.0021)	(0.0011)
MOB	-0.0145	-0.0101***	-0.0186***	-0.0248***	0.0685***	-0.0603***
	(0.0097)	(0.0012)	(0.0016)	(0.001)	(0.0025)	(0.0022)
MIP	-0.0088	0.0869***	0.063***	0.0945***	-0.0038	-0.0051***
	(0.0093)	(0.0008)	(0.0015)	(0.0007)	(0.0024)	(0.001)
DT	-0.0118	-0.0366***	-0.0443***	-0.0161***	-0.0276***	0.0009
	(0.0081)	(0.0007)	(0.0013)	(0.0007)	(0.0019)	(0.001)
FHB	-0.0026	-0.0051***	-0.006***	-0.0056***	-0.0034	-0.0049***
	(0.0102)	(0.0007)	(0.0014)	(0.0007)	(0.0022)	(0.001)
INT	-0.0139	-0.0006	-0.0075***	-0.0174***	-0.0211***	-0.004***
	(0.0093)	(0.0008)	(0.0013)	(0.0008)	(0.0019)	(0.0012)
OO	0.0202**	0.0243***	0.0291***	0.0332***	0.0276***	0.0026**
	(0.0059)	(0.0007)	(0.0012)	(0.0006)	(0.0022)	(0.0009)
LP_C	0.0005	-0.0119***	-0.0227***	-0.0217***	-0.0224***	-0.0012
	(0.0085)	(0.0009)	(0.0014)	(0.0008)	(0.002)	(0.0013)
LP_P	0.0089	0.0178***	0.0252***	0.0341***	-0.0089***	0.0081***
	(0.011)	(0.0009)	(0.0016)	(0.0009)	(0.0026)	(0.0014)
MF	-0.0035	0.0013	-0.0038**	-0.0048***	-0.0018	0.006***
	(0.009)	(0.0009)	(0.0013)	(0.0007)	(0.0012)	(0.0009)
NB_01	-0.0077	-0.0148***	-0.0127***	-0.0208***	-0.0023	-0.0019
	(0.0086)	(0.0007)	(0.0012)	(0.0006)	(0.0018)	(0.0011)
NJF	0.0188**	0.0167***	0.0296***	0.0353***	0.0396***	0.0002
	(0.0079)	(0.0007)	(0.0013)	(0.0007)	(0.0017)	(0.001)
NSRR	0.0122	-0.0027***	-0.0132***	0.0009	-0.015***	-0.0005
	(0.0089)	(0.0007)	(0.0012)	(0.0006)	(0.0017)	(0.001)
PDJ	0.0173	0.0108***	0.0022	0.0068***	0.0112***	-0.0026*
	(0.0077)	(0.0009)	(0.0013)	(0.0006)	(0.0016)	(0.0011)
SNS	-0.0165*	-0.0079***	-0.0072***	-0.0063***	-0.0159***	-0.001
	(0.0082)	(0.0007)	(0.0012)	(0.0006)	(0.0022)	(0.001)
PT_CO	0.0338**	-0.0111***	-0.0019	-0.0071***	-0.002	-0.0002

Parameter	Charge off	short sale	3rd party sale	REO	Note sale	Defaulted repurchased
	(0.0108)	(0.0006)	(0.0012)	(0.0006)	(0.0021)	(0.0009)
R_SS	0.0012	0.003***	0.002	0.0008	0.0216***	-0.0112***
	(0.0096)	(0.0008)	(0.0014)	(0.0007)	(0.0021)	(0.0011)
R_TP	-0.0024	0.0012	-0.0034**	0.0016**	0.0222***	-0.0041***
	(0.0087)	(0.0007)	(0.0012)	(0.0007)	(0.0018)	(0.0011)
DRD	-0.0503***	-0.0211***	-0.05***	-0.0415***	-0.0815***	-0.0062***
	(0.011)	(0.001)	(0.0017)	(0.0008)	(0.0027)	(0.0013)
UMP_L3	0.0014	-0.0345***	-0.0103***	-0.0085***	0.0266***	-0.0108***
	(0.0107)	(0.0009)	(0.0016)	(0.0009)	(0.0027)	(0.0014)
GDPCY_L3	0.0123	0.0152***	0.014***	0.0108***	-0.0217***	0.0047***
	(0.0091)	(0.0021)	(0.0016)	(0.0007)	(0.0039)	(0.0014)

*This table reports the coefficients obtained for model CL under Stage 1. Level of coefficient significant as follows: *** for p-value < 0.01, ** for p-value < 0.05 and * for p-value < 0.1. Values come in the following order: coefficient, degree of significance/ (standard error). The estimates are based on in-time training sample. Each column is for one resolution.*

Table 28. *Standardised regression coefficients for modelled step 1: CT: P(A>0).*

Parameter	Charge off	short sale	3rd party sale REO		Note sale	Defaulted repurchased
Intercept	0.0559	3.5239	3.1604***	2.6286***	1.8134***	2.8483***
	(0.0452)	(2.2579)	(0.0436)	(0.0133)	(0.0737)	(0.0518)
DLTV	-0.1749***	-0.0202	-0.3227***	-0.1422***	0.0139	-0.0212
	(0.0313)	(0.0554)	(0.0225)	(0.0103)	(0.0366)	(0.0436)
DDTI	0.0115	0.0481	0.0303	0.0269***	-0.0137	0.0282
	(0.0215)	(0.0397)	(0.0176)	(0.0065)	(0.0253)	(0.0254)
FICO	0.0151	-0.0109	-0.0003	0.0339***	-0.0401	-0.0151
	(0.0262)	(0.0465)	(0.0199)	(0.0075)	(0.0271)	(0.0286)
TID	0.0129	-0.0104	0.0221	0.0135*	0.0402	0.0951**
	(0.0208)	(0.0475)	(0.0145)	(0.0068)	(0.0239)	(0.04)
TTR	0.0775**	-0.5916***	-0.3456***	-0.3425***	0.2274***	-0.3903***
	(0.0256)	(0.041)	(0.0178)	(0.0074)	(0.0273)	(0.0287)
LC	0.0292	0.0571	0.0978*	0.017	0.1505	0.2075*
	(0.0174)	(0.1318)	(0.0407)	(0.0116)	(0.0652)	(0.1077)
LOB	0.3063***	0.5764***	0.8087***	0.7583***	0.5174***	0.4269***
	(0.0268)	(0.0526)	(0.0259)	(0.0092)	(0.0298)	(0.0304)
MOB	-0.0348	-0.0357	-0.2721***	-0.2111***	0.2415***	-0.1515***
	(0.0314)	(0.0648)	(0.0225)	(0.0093)	(0.0388)	(0.0446)
MIP	-0.0196	0.2389***	0.1095***	0.3026***	-0.1014**	0.001
	(0.03)	(0.0629)	(0.0257)	(0.0092)	(0.0353)	(0.0301)
DT	-0.0341	-0.0102	0.0257	0.0003	-0.0634*	-0.0028
	(0.0267)	(0.045)	(0.0222)	(0.0077)	(0.0284)	(0.0278)
FHB	-0.0043	-0.0452	-0.0138	-0.0021	-0.0438	-0.045
	(0.0329)	(0.0512)	(0.0261)	(0.0093)	(0.0313)	(0.0296)
INT	-0.0386	-0.0164	0.0163	-0.0542***	-0.1779***	-0.0529
	(0.0301)	(0.043)	(0.0221)	(0.0086)	(0.0291)	(0.0312)
OO	0.0527	0.1461***	0.1498***	0.1591***	0.1494***	0.0807***
	(0.0193)	(0.0322)	(0.0156)	(0.0057)	(0.0221)	(0.0216)
LP_C	0.0273	-0.0262	-0.0759***	-0.0487***	-0.0703	-0.0203
	(0.0277)	(0.0506)	(0.0212)	(0.0077)	(0.0319)	(0.0339)
LP_P	0.079	-0.0027	0.1472***	0.1255***	-0.0652	0.0631
	(0.0359)	(0.0613)	(0.0292)	(0.0105)	(0.0413)	(0.0389)
MF	-0.0718*	0.1711***	0.0767***	0.0308***	0.0721**	0.0065
	(0.0296)	(0.0466)	(0.0205)	(0.0076)	(0.0228)	(0.0246)
NB_01	-0.0243	-0.1701***	-0.0537**	-0.0915***	-0.0457	-0.1064***
	(0.0279)	(0.0522)	(0.0209)	(0.0077)	(0.0286)	(0.0329)
NJF	0.0528	0.0416	0.0931***	0.139***	0.1187***	0.0665**
	(0.0253)	(0.0416)	(0.0183)	(0.0068)	(0.0263)	(0.0263)
NSRR	0.0641**	-0.015	-0.1154***	-0.0448***	-0.0094	-0.0123
	(0.0282)	(0.0441)	(0.018)	(0.0071)	(0.0259)	(0.0273)
PDJ	0.0483	0.4421	0.0505**	0.0421***	0.1177***	0.0433
	(0.0241)	(9.2148)	(0.0198)	(0.0068)	(0.036)	(0.0311)
SNS	-0.0398	-0.0767	-0.0223	-0.0116	-0.1133***	-0.0222
	(0.0269)	(0.0497)	(0.0218)	(0.0074)	(0.0331)	(0.0273)
PT_CO	0.0772*	0.2145***	0.1442***	0.1143***	0.0642*	0.0676*

Parameter	Charge off	short sale	3rd party sale	REO	Note sale	Defaulted repurchased
	(0.0337)	(0.0608)	(0.023)	(0.0082)	(0.0304)	(0.0278)
R_SS	-0.0057	-0.0699	-0.0363	-0.02*	0.0868*	-0.0383
	(0.0309)	(0.0458)	(0.0225)	(0.0081)	(0.0357)	(0.029)
R_TP	-0.0082	-0.0118	-0.0096	-0.0144*	0.0188	0.0017
	(0.0287)	(0.0464)	(0.0206)	(0.0074)	(0.026)	(0.0303)
DRD	-0.1733***	-0.0285	0.0009	-0.0311***	-0.3477***	-0.0332
	(0.0357)	(0.0573)	(0.0298)	(0.0008)	(0.0416)	(0.0351)
UMP_L3	-0.0179	0.0074	0.0406	-0.0085***	0.1801***	-0.1537***
	(0.0348)	(0.0592)	(0.0302)	(0.0009)	(0.0433)	(0.0392)
GDPCY_L3	0.0105	0.139	0.1042	0.0108***	0.3959	0.2485
	(0.0282)	(0.1906)	(0.0643)	(0.0007)	(0.4911)	(0.1978)

*This table reports the coefficients obtained for model CA: P(A>0) under Stage 1. Level of coefficient significant as follows: *** for p-value < 0.01, ** for p-value < 0.05 and * for p-value < 0.1. Values come in the following order: coefficient, degree of significance/ (standard error). The estimates are based on in-time training sample. Each column is for one resolution.*

Table 29. *Standardised regression coefficients for modelled step 1: CT: E(R|A>0).*

Parameter	Charge off	short sale	3rd party sale	REO	Note sale	Defaulted repurchased
Intercept	0.6211***	0.6861***	0.7156***	0.6161***	0.5358***	0.9121***
	(0.1041)	(0.0006)	(0.0012)	(0.0006)	(0.0039)	(0.0015)
DLTV	-0.0636***	-0.0627***	-0.1165***	-0.0891***	-0.0369***	-0.0148***
	(0.0193)	(0.0005)	(0.0013)	(0.0009)	(0.002)	(0.0016)
DDTI	0.0082	0.0059***	0.005***	0.0071***	-0.0049**	0.0021*
	(0.0101)	(0.0004)	(0.0011)	(0.0006)	(0.0016)	(0.0009)
FICO	-0.0037	0.0021***	-0.004***	0.0031***	-0.0037*	-0.0057***
	(0.012)	(0.0005)	(0.0012)	(0.0007)	(0.0017)	(0.001)
TID	-0.007	0.0066***	0.0017	0.003***	0.0114***	0.0143***
	(0.0097)	(0.0006)	(0.0012)	(0.0008)	(0.0014)	(0.0018)
TTR	0.0381**	-0.0511***	-0.0074***	-0.0269***	0.0882***	-0.0727***
	(0.0122)	(0.0007)	(0.0014)	(0.0008)	(0.0014)	(0.0016)
LC	-0.004	-0.0005	-0.002	-0.0023**	-0.0046	-0.0039
	(0.0085)	(0.0007)	(0.0018)	(0.0008)	(0.0025)	(0.002)
LOB	0.0789**	0.0571***	0.0371***	0.0718***	0.0284***	0.0038***
	(0.0282)	(0.0006)	(0.0014)	(0.0007)	(0.0019)	(0.001)
MOB	0.0234	-0.0099***	-0.0112***	-0.0116***	0.0624***	-0.0566***
	(0.0139)	(0.0007)	(0.0015)	(0.0009)	(0.0022)	(0.002)
MIP	0.0049	0.0866***	0.0596***	0.0854***	-0.0025	-0.0053***
	(0.0145)	(0.0004)	(0.0014)	(0.0006)	(0.0021)	(0.001)
DT	-0.0102	-0.0366***	-0.0422***	-0.0151***	-0.0255***	0.0009
	(0.013)	(0.0004)	(0.0012)	(0.0006)	(0.0017)	(0.001)
FHB	-0.0023	-0.005***	-0.0058***	-0.0056***	-0.0021	-0.0044***
	(0.0152)	(0.0004)	(0.0013)	(0.0006)	(0.002)	(0.0009)
INT	-0.0059	-0.0004	-0.0048***	-0.0128***	-0.0159***	-0.0026**
	(0.014)	(0.0005)	(0.0012)	(0.0007)	(0.0017)	(0.0011)
OO	0.0248	0.0239***	0.0242***	0.0254***	0.0163***	0.0011
	(0.0122)	(0.0004)	(0.0011)	(0.0006)	(0.002)	(0.0009)
LP_C	-0.005	-0.0119***	-0.0213***	-0.0193***	-0.0203***	-0.001
	(0.013)	(0.0005)	(0.0013)	(0.0007)	(0.0017)	(0.0012)
LP_P	-0.0037	0.0178***	0.0234***	0.0324***	-0.0059**	0.0076***
	(0.0186)	(0.0005)	(0.0015)	(0.0008)	(0.0023)	(0.0013)
MF	0.0071	0.0007	-0.0053***	-0.0063***	-0.0033**	0.0061***
	(0.0152)	(0.0005)	(0.0012)	(0.0006)	(0.0011)	(0.0008)
NB_01	0.0041	-0.0146***	-0.0121***	-0.0184***	-0.0019	-0.0005
	(0.0128)	(0.0004)	(0.0011)	(0.0006)	(0.0016)	(0.001)
NJF	-0.0151	0.0169***	0.0306***	0.0327***	0.0381***	-0.0003
	(0.0125)	(0.0004)	(0.0012)	(0.0006)	(0.0015)	(0.0009)
NSRR	0.0261	-0.0027**	-0.0109***	0.0032***	-0.0126***	-0.0003
	(0.013)	(0.0004)	(0.0011)	(0.0006)	(0.0015)	(0.0009)
PDJ	0.0029	0.0106***	0.0002	0.004***	0.0084***	-0.0035**
	(0.0116)	(0.0005)	(0.0013)	(0.0006)	(0.0014)	(0.001)
SNS	-0.0339**	-0.0079***	-0.0067***	-0.0072***	-0.0118***	-0.001
	(0.0128)	(0.0004)	(0.0011)	(0.0006)	(0.0019)	(0.0009)
PT_CO	0.0135	-0.0114***	-0.0056***	-0.0122***	-0.0055**	-0.0015

Parameter	Charge off	short sale	3rd party sale	REO	Note sale	Defaulted repurchased
	(0.0159)	(0.0003)	(0.0011)	(0.0006)	(0.0018)	(0.0008)
R_SS	0.0223	0.0032***	0.002	0.0015**	0.0197***	-0.0109***
	(0.0137)	(0.0005)	(0.0013)	(0.0007)	(0.0019)	(0.001)
R_TP	-0.0059	0.0012	-0.003**	0.0022***	0.022***	-0.0042***
	(0.0138)	(0.0004)	(0.0011)	(0.0006)	(0.0016)	(0.001)
DRD	-0.0337	-0.0214***	-0.0495***	-0.042***	-0.0714***	-0.0058***
	(0.0233)	(0.0006)	(0.0016)	(0.0008)	(0.0024)	(0.0012)
UMP_L3	-0.0244	-0.0346***	-0.0099***	-0.0073***	0.0233***	-0.0088***
	(0.0155)	(0.0005)	(0.0015)	(0.0008)	(0.0024)	(0.0013)
GDPCY_L3	0.0178	0.015***	0.0129***	0.0086***	-0.0221***	0.004***
	(0.0115)	(0.0014)	(0.0015)	(0.0007)	(0.0034)	(0.0013)

*This table reports the coefficients obtained for model CA: E(R|A>0) under Stage 1. Level of coefficient significant as follows: *** for p-value < 0.01, ** for p-value < 0.05 and * for p-value < 0.1. Values come in the following order: coefficient, degree of significance/ (standard error). The estimates are based on in-time training sample. Each column is for one resolution.*

Under CL, DLTV is material only for 4 resolutions: charge-off, short sale, third party sale and REO. This is trivial as these 4 resolutions are the only ones that involve attempts to sell securities. Coefficient materiality is similar for CT: $E(R|A > 0)$. For CT: $P(A > 0)$, DLTV is further used for charge-off and third-party sale, indicating that the probability that something is recovered is lower for high DLTV. This observation would not have been made without a decomposition model.

Though not known at time of default, TTR expectations may be used to determine the magnitude of losses in each resolution. Experienced debt collectors are likely to have a more accurate view of this. TTR is material for short sale, defaulted repurchased and note sale, although it had a positive coefficient for note sale. Due to the nature of the resolution, short sales may lead to lower asking prices for longer times to sale. TTR affects recovery rate negatively for default-repurchased loans and this may depend entirely on the party buying the loan back due to adverse selection. For note/ reperforming sale, a longer time horizon typically gives the internal collection team to perform more early recovery amounts. Coefficient materiality is similar for CT: E(R| A>0). However, coefficients of TTR are material for all resolutions except for charge-off, default-attrition and default-cured for CT: P(A>0), indicating that the probability of having something recovered is materially reduced for these 5 resolutions. As these resolutions are quite typical, most of them may be confirmed by the findings of Do et al. (2020) who indicated that the probability of cure (0 loss in their case) is negatively affected by time to resolution (DefaultToEOO, in their case).

Aside from default-attrition, default-repurchased and default-cured loans, LOB coefficients are material for CL. As the 5 resolutions are mostly within the control of lenders, this reflects the way that they prioritise collecting on loans that started with higher lending amount. Looking into the two-step model CT, it turns out that LOB affects both the possibility of recovering something and the amount of recovery for cases with a positive amount recovered. This once again affirms how lenders prioritise high outstanding loans. High outstanding balances are also likely loans that started out with a high principal balance.

The intention in including DRD as a driver is to quantify lender efforts in collecting. The effort may be influenced by internal policies and state of collection queues. DRD is material for charge-off, third party sale and note sale, affecting recovery rate negatively for all 3 resolutions under CL. Cases where DRD are relatively high may indicate that collections teams are under stress and/ or facing challenges in staying on top of queues. While delays typically do not directly affect recovery amount, delays in contacting customers in distress during a systemic crisis may mean further deterioration of their finances – which includes less amount allocated for repayment, if available. Only charge-off and note sale were material for CT: P(A>0) which may indicate trends inherent in the way collections procedures are carried out within these two resolutions. For CT: E(R|A>0), third party sale and note sale had material coefficients pointing to the same direction.

MIP results to higher recovery, especially for eligible cases. It is not difficult to imagine the typical resolutions (short sale, third party sale and REO disposition) would usually fall within eligible cases for mortgage insurances. This is seen both in CL, and both parts of CT models, which indicates higher insurance coverage doesn't only lead to more chances of having something recovered but also leads to a higher recovery amount.

MOB is negative for typical resolutions except for note sale. As intrinsically bad accounts are typically flushed out within 12-18 months in a typical mortgage lifecycle, loans that default later are usually those that have no other choice. As such, it is easy to see why they may result in poorer quality recoveries than those that default earlier in the lifecycle. For note sale, however, contracts may still be able to hold some value provided that the borrower is willing to negotiate terms. Borrowers that default after 2 years are likely to be more willing to negotiate repayment terms given that they have invested more in the security through past repayments. For CL, MOB is material for note sale and default-repurchased. For CT: P(A>0), third party sale, REO and note sale had material coefficients for MOB while short sale, note sale and

default-repurchased resolutions had material coefficients under $E(R|A > 0)$. Due to the nature of the resolution, default-repurchased loans may have significantly more information with the original lenders which may shed light to this trend. Unfortunately, this is not accessible to users of the Freddie Mac dataset.

There are specific cases when an additional driver turns out to have material and statistically significant coefficient. This is: GDPCY_L3 for note sale, which is an indicator that a high year-on-year state level GDP change in the quarter leading to default is a good indicator of whether the default event is in an economic crisis or not.

4.6.2.2. Step 2: probability of resolution models

Fixed proportion models are simply taken using the training dataset. Due to the materially poor level of RCM accuracy for this model on the validation dataset, this model did not proceed further for the combined model. With this, there is only 1 model used in calculating probability of resolution.

Material and statistically significant coefficients for the multinomial logit resolution models are discussed in this section. As charge-off is indexed as the first resolution, there are no coefficients estimated for this resolution.

Table 30. *Standardised regression coefficients for modelled step 2: probability of resolution.*

Parameter	short sale	3rd party sale	REO	Note sale	Defaulted attrition	Defaulted repurchase	Cure
Intercept	3.7954***	3.9972***	5.3301***	1.6397***	4.079***	3.2625***	1.9903***
	(0.0349)	(0.0347)	(0.0344)	(0.0379)	(0.0347)	(0.0353)	(0.0376)
DLTV	0.4685***	0.0773**	0.0814**	0.272***	-1.3209***	-0.0398	0.3321***
	(0.028)	(0.0281)	(0.0276)	(0.0295)	(0.0282)	(0.0288)	(0.0291)
DDTI	0.049**	0.0229	0.0651***	0.0366*	0.0922***	0.0752***	0.1859***
	(0.0171)	(0.0171)	(0.0167)	(0.0186)	(0.0169)	(0.0174)	(0.0182)
FICO	0.1775***	0.0939***	0.0683***	-0.0091	-0.1125***	-0.1497***	-0.4275***
	(0.0199)	(0.0199)	(0.0194)	(0.0214)	(0.0196)	(0.0201)	(0.021)
TID	-0.111***	-0.0182	-0.0431**	0.0589***	0.1526***	0.1568***	-0.1067***
	(0.0183)	(0.0179)	(0.0174)	(0.0187)	(0.0175)	(0.0199)	(0.019)
TTR	-1.391***	-0.1973***	-0.0802***	0.9722***	-0.783***	-0.8284***	-0.0748***
	(0.021)	(0.0198)	(0.0191)	(0.0201)	(0.0198)	(0.0208)	(0.0204)
LC	0.5298***	0.467***	0.5478***	0.5152***	0.514***	0.5509***	0.573***
	(0.0301)	(0.0306)	(0.0295)	(0.0319)	(0.0291)	(0.0311)	(0.0301)
LOB	1.8429***	1.3102***	1.0771***	1.6098***	1.4231***	1.4385***	1.3392***
	(0.0221)	(0.022)	(0.0214)	(0.0238)	(0.0217)	(0.0222)	(0.0234)
MOB	0.159***	0.2309***	-0.1004***	0.7736***	-0.3773***	-1.2494***	0.4076***
	(0.0253)	(0.0251)	(0.0245)	(0.0268)	(0.0248)	(0.0271)	(0.0263)
MIP	-0.1235***	-0.4082***	-0.0039	-0.5795***	-0.1061***	0.0868***	-0.0745**

Parameter	short sale	3rd party sale	REO	Note sale	Defaulted attrition	Defaulted repurchase	Cure
	(0.0225)	(0.0227)	(0.0221)	(0.0248)	(0.0224)	(0.0226)	(0.0235)
DT	0.1283***	0.0393	0.0594**	-0.1395***	-0.053	0.136***	-0.2145***
	(0.0208)	(0.0208)	(0.0203)	(0.0223)	(0.0206)	(0.0209)	(0.0219)
FHB	0.0401	0.0268	0.013	0.1239***	-0.0174	0.0323	0.0707**
	(0.0255)	(0.0256)	(0.0251)	(0.0272)	(0.0255)	(0.0256)	(0.0266)
INT	-0.2554***	-0.2387***	-0.0372	-0.4771***	-0.1648***	-0.0449	-1.1791***
	(0.0246)	(0.0246)	(0.0242)	(0.0256)	(0.0245)	(0.0249)	(0.0252)
OO	0.2319***	0.2705***	0.2482***	0.2007***	0.3691***	0.1238***	0.4856***
	(0.0147)	(0.0148)	(0.0142)	(0.0173)	(0.0146)	(0.0148)	(0.0186)
LP_C	-0.0268	-0.0325	-0.034	-0.0071	-0.087***	-0.0149	0.0883***
	(0.0216)	(0.0216)	(0.0211)	(0.023)	(0.0214)	(0.022)	(0.0227)
LP_P	0.1066***	0.1771***	0.1565***	0.0256	0.4275***	0.2676***	0.1347***
	(0.0278)	(0.0279)	(0.0273)	(0.0299)	(0.0277)	(0.0281)	(0.0294)
MF	0.1212***	0.0356	-0.0198	-0.0037	0.0894***	0.2363***	-1.0709***
	(0.0229)	(0.0228)	(0.0224)	(0.0232)	(0.0227)	(0.0228)	(0.0296)
NB_01	-0.2256***	-0.1547***	-0.1753***	-0.0904***	-0.3086***	-0.0804***	-0.1695***
	(0.0215)	(0.0215)	(0.0211)	(0.0227)	(0.0213)	(0.0218)	(0.0224)
NJF	-0.3393***	0.0468**	0.0455**	-0.0803***	-0.0711***	-0.2355***	-0.0921***
	(0.019)	(0.0191)	(0.0186)	(0.0202)	(0.0188)	(0.0192)	(0.02)
NSRR	0.1411***	0.2331***	0.1157***	0.3263***	0.2726***	0.1534***	0.2812***
	(0.0216)	(0.0216)	(0.0212)	(0.0226)	(0.0214)	(0.0217)	(0.0224)
PDJ	0.0438*	-0.0122	0.0328	0.0244	0.0535**	0.0268	0.0254
	(0.0192)	(0.0192)	(0.0186)	(0.0202)	(0.0188)	(0.0193)	(0.02)
SNS	-0.0098	-0.055**	-0.0981***	-0.1587***	-0.0337	-0.1021***	0.9541***
	(0.0207)	(0.0207)	(0.0203)	(0.0225)	(0.0206)	(0.0208)	(0.0253)
PT_CO	0.365***	0.2538***	0.2577***	0.2317***	0.2114***	0.3385***	0.1136***
	(0.0262)	(0.0263)	(0.026)	(0.0276)	(0.0262)	(0.0263)	(0.0275)
R_SS	0.3425***	0.1225***	0.1455***	0.4052***	0.2512***	0.2664***	-0.007
	(0.0253)	(0.0255)	(0.025)	(0.0266)	(0.0252)	(0.0255)	(0.0267)
R_TP	0.0575**	0.0815***	0.0252	0.2369***	0.0544**	0.0618**	-0.025
	(0.0221)	(0.022)	(0.0216)	(0.0229)	(0.0219)	(0.0221)	(0.023)
DRD	-0.0781***	-0.2472***	-0.0169	-0.5131***	-0.5733***	-0.1481***	-0.1606***
	(0.0251)	(0.0252)	(0.0245)	(0.0274)	(0.025)	(0.0253)	(0.0266)
UMP_L3	-0.296***	-0.2746***	-0.5659***	-0.5806***	-0.5237***	-0.4229***	-0.6105***
	(0.0267)	(0.0267)	(0.0262)	(0.0288)	(0.0266)	(0.0271)	(0.0282)
GDPCY_L3	-0.0358	0.0204	0.0192	-0.2038***	0.0522*	-0.0516	0.0835***
	(0.0267)	(0.025)	(0.0242)	(0.0315)	(0.0242)	(0.0251)	(0.0254)

*This table reports the coefficients obtained for model CA: P(A>0) under Stage 1. Level of coefficient significant as follows: *** for p-value < 0.01, ** for p-value < 0.05 and * for p-value < 0.1. Values come in the following order: coefficient, degree of significance/ (standard error). The estimates are based on in-time training sample. Each column is for one resolution.*

DLTV is materially significant for predicting probability of attrition following default. High LTV loans at default date leads to less chances of being refinanced elsewhere. This makes intuitive sense given the existence of credit bureaus.

High TTR tends to lead to more chances of going to note sale. On the other hand, high TTR reduces the chances of leading to other resolutions. Note, however, that it was not determined whether this is pre-determined (i.e. lenders expect a note sale which just to happens to have a long time to resolution) or because the lender expects a long time to resolution hence loans are eventually put under note sale.

Because LC affects the chances of each of the 7 resolutions almost similarly, it reduces the chances of being charged off by deduction. Liquidity constraints, by definition, build up through time. By nature, charge-off cases are typically those that result from sudden defaults.

Similarly, LOB also reduces the chances of being charged off by deduction. This makes sense given that lenders wouldn't want to charge off materially larger loans.

A higher MOB leads to more chances of being resolved under short sale and note sale, while reducing the chances of ending up being repurchased following default. Accounts that default at a later stage to the usual "flushing" of a loan lifecycle usually indicates higher equity, and so contracts become easier to sell if unsold via short sale. Naturally, accounts that stick around for longer periods are also harder to justify as falling under adverse selection, which is why it reduces the chances of going through default-repurchased.

Both existence of insurance and insurance coverage (MIP) decreases the chance of being under note sale, which makes sense post fact as note sales typically have higher proportion of early recoveries compared to other resolutions.

A low DRD significantly increases the chances of a defaulted account to be resolved under note sale and default-attrition. Note sales significantly increased post GFC, which indicates that this has started to become a preferred resolution after crisis. Similarly, more lenders started refinancing, even for defaulted accounts.

UMP_L3 indicates more chances of being charged-off. This may be true especially for low balance loans. Assuming that the distribution of LVR is similar through different starting loan balances, this makes sense given that the cost of liquidating a property may outweigh the expected recovery amount and that the defaulted borrower may no longer have income to support negotiated repayments.

High interest rate at origination indicates the least chance of being cured following a default. High interest rate environments often occur simultaneously as slowing down of house prices. This may have influences of strategic defaults, although it lacks further evidence.

Owner occupied loans have the most to lose given the borrowers not only gain a record in the credit bureaus but also lose the roof that they are living under. As such, it is not difficult to associate owner occupied borrowers to higher chances of a defaulted loan to go back to non-arrears status for a significant period (i.e. cure).

4.6.2.3. Combination of both components for overall R model

Both combined models outperform both benchmarks for in-time sampling and in 8 out of 10 sampling windows performed for out-of-time tests. This alone is an indicator of the success of the proof of concept.

While the way that a best model has been defined in step 1 may have undermined the performance of the 2nd combined model (RM x best step 1), it still outperforms the both benchmarks for in-time and 8 out of 10 out-of-time tests performed on rolling windows.

It may be understandable that sophisticatedly structured models may not work as well for test windows happening when GFC is starting (i.e. 2008). In which case, the best model is the simplest one – B1. Even then, the overall negative r-square indicates that no model is good enough to predict recoveries within this period. Once the GFC indicators were picked up, r-square once again started having positive values. This started for testing window 2010 onwards.

4.7. Research findings and conclusion

Our proposal to structure LGD modelling around the unique characteristics of each resolution builds on the stage-based decomposition in Essay 1 and addresses gap 2 (Section 1.4). The results strongly support resolution-specific modelling, revealing distinct drivers behind each resolution path.

Once adopted, this framework also delivers clear operational benefits. First, its modular design greatly simplifies model maintenance. Rather than recalibrating an entire LGD model when conditions change, a bank need only monitor module-level performance, then update or replace the underperforming component—yielding substantial cost savings over time.

Second, it offers powerful value-add for teams managing early- and late-stage collections (or for investors in distressed debt). This complements the stage-based insights from Essay 1 by providing resolution-specific recovery expectations. By segmenting accounts according to their most promising recovery streams, institutions can decide, for example, to in-house only

those loans with the highest projected recoveries and sell off the rest, or to time bad-debt sales more precisely using forward-looking recovery curves.

However, it may be tempting simply to prioritise the resolution with the highest observed recoveries. We caution that this can lead to a circular logic: servicers tend to apply the strategies they already believe will work best, so elevated recoveries may reflect operator bias as much as genuine recoverability. To avoid this, resolutions should be selected and validated through independent and randomised trials and performance tests, not solely by their historical payoffs.

An important limitation of this framework is that resolution type is treated as exogenous conditional on the included covariates. In practice, servicers exercise discretion in steering borrowers towards particular resolution paths based on factors not fully captured in the data, such as borrower willingness or degree of cooperation, property condition, and internal cost-recovery assessments. For example, a servicer may prefer to negotiate a short sale with a cooperative borrower because it reduces carrying costs and avoids the uncertainty of foreclosure auction, while pursuing REO disposition for an uncooperative borrower or a property in poor condition. Because these servicer assessments are based partly on information not available in the Freddie Mac dataset, the resolution type coefficients in our multinomial logit model capture not only the true effect of resolution on recovery but also the selection effect from servicer steering. The likely direction of this bias is upward for short-sale coefficients: loans that enter short sale are mostly those where servicers expected a relatively favourable outcome.

Introduction of a variable as an indicator may in address this theoretically, but credible ones that affect resolution assignment without directly affecting recovery are difficult to identify in this setting within the Freddie Mac dataset. Servicer identity, for instance, correlates with both resolution choice and recovery outcomes. We therefore interpret the resolution type coefficients as associations that are useful for prediction rather than as causal effects. For the purpose of this study, this is sufficient: the framework is designed to improve predictive accuracy and operational transparency, not to estimate the causal effect of choosing one resolution over another. Practitioners applying this framework should be aware that the resolution-specific recovery estimates reflect historical patterns of servicer behaviour, and that changing servicer practices would alter the estimated relationships.

A related limitation is that the Freddie Mac dataset does not capture borrower behavioural factors such as strategic default motives, financial stress from sources outside the mortgage, or

borrower expectations of house price recovery. These factors may influence both the likelihood of different resolutions and the level of recovery achieved. While variables such as delinquency history and debt-to-income ratio serve as partial proxies for borrower financial stress, they do not capture the full range of behavioural heterogeneity that may affect outcomes. Incorporating credit bureau data at different points of the default process or borrower survey information could address this gap in future work, though such data is not available within the Freddie Mac public dataset used here.

Lastly, the transparency and granularity of our approach extend its relevance beyond collections into origination decisions, risk appetite frameworks, credit policies and pricing. From a provisioning standpoint, resolution-level models provide the granularity needed for IFRS 9 and CECL compliance, where forward-looking estimates benefit from understanding how the mix of resolutions is expected to shift under different macroeconomic scenarios. From a portfolio management standpoint, the framework supports decisions around securitisation structuring and loss mitigation spending by quantifying the expected recoveries across resolutions. This resolution-based framework, when combined with the enhanced valuation approaches in Essay 3, provides a comprehensive toolkit for mortgage LGD modelling.

MASSEY UNIVERSITY
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Justin Rylie Tang
Name and title of main supervisor:	Dr. David Tripe, Adjunct Professor of Banking
In which chapter is the manuscript/published work?	No

Describe the contribution that the student and members of the supervisory team have made to the manuscript/published work:[1]

Accurate Loss Given Default (LGD) models are crucial for effective risk management in residential mortgage lending. This study introduces an innovative approach to predicting Loss Given Default (LGD) by incorporating a Distressed House Price Discount Index (DHPDI) as a function of a reconstructed House Price Index (HPI) by using Freddie Mac data. By using DHPDI, we aim to capture market conditions more precisely, improving LGD predictions. Our analysis demonstrates that the inclusion of DHPDI significantly enhances model performance, offering better insights into market-driven loss factors and supporting more robust risk management strategies.

Please select one of the following three options:

○ **The manuscript/published work is published or in press**

Please provide the full reference of the research output:

○ **The manuscript is currently under review for publication**

Please provide the name of the journal:

◉ **It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:	Justin Rylie Tang — Digitally signed by Justin Rylie Tang Date: 2025.12.08 05:07:21 +13'00'	Main supervisor's signature:	

This form should be placed at the beginning of each relevant thesis chapter.

[1] Refer to the Massey University Publishing and Authorship guidelines (OneMassey for staff, Stream for students) and/or Contributor Roles Taxonomy (CRediT) guidelines for guidance.

5. Chapter 5 - Essay 3: Enhancing Loss Given Default Models for Residential Mortgage Loans: Integrating a Distressed House Price Discount Index

5.1. Introduction

Effective credit risk management in mortgage lending relies heavily on accurately predicting Loss Given Default (LGD). Building on gap 3 identified in Section 1.4, for most times, this quantity is highly dependent on the value realised from collateral disposition. Traditional LGD models often fail to account for complexities from the real-world: property sales under distressed conditions such as foreclosure, short sale, or REO transaction. These resolution types, detailed in Essay 2, commonly transact at prices below fair market value due to factors including forced sale dynamics, market stigma, inferior property conditions, and potentially negative spillover effects on surrounding properties.

A growing body of research highlights how the presence of distressed transactions can bias house price indices (HPIs) and subsequently distort loss estimates and market assessments - effects that become especially pronounced during periods of economic stress or in localised market downturns. The magnitude and drivers of distressed sale discounts are multifaceted, involving both property-specific factors such as maintenance or vandalism, neighbourhood-level effects, and broader liquidity and/ or crisis dynamics. Recent debates in the literature have also questioned the extent to which traditional estimation methods may overstate the size of these discounts, emphasizing the importance of robust empirical methodologies and appropriate controls.

This essay focuses specifically on Stage 2 recovery (collateral disposal) from Essay 1's framework, developing enhanced valuation approaches for distressed sales. The material impact of using different methodologies to value collateral is illustrated in **Figure 17** below. It compares three approaches for measuring median property prices using Freddie Mac loan data from 1999–2020. The blue line shows the actual Freddie Mac median sale price each quarter for all property sales, which includes both foreclosed and regular sales. The orange line uses the first (June 1999) point from the blue line, and projects it into the future using the national seasonally adjusted FHFA index (i.e., orange line at September 2020 is blue line at June 1999 divided by FHFA index at June 1999 multiplied by FHFA index at September 2020). The grey line modifies the blue line by applying a distressed house price discount index (DHPDI), reflecting realised prices under distressed sale conditions. This reveals two critical insights:

- Standard HPIs (orange line) can dramatically overstate realised prices at disposition date, especially following sharp recoveries, while a regular median measurement (blue line) gives a more moderate trend.
- The distressed-adjusted values (grey line), is consistently and materially lower than both, quantifying the discount experienced on actual distressed transactions. This aligns with the literature showing substantial distress discounts (Aroul & Hansz, 2014; Doerner & Leventis, 2015) and directly motivates the need for a DHPDI for LGD prediction

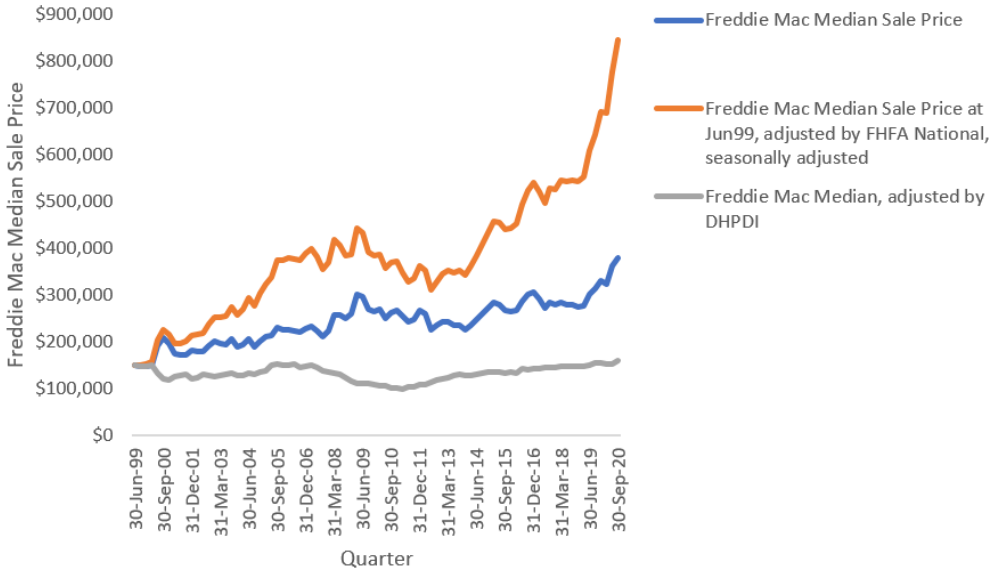

Figure 17. *Property price estimates using three methodologies.*

The horizontal axis is the date of sale of homes. The blue line shows actual median of all single-family homes originated within Freddie Mac. The orange line takes the first point from the under the blue line, adjusted by growth ratios from the first period to each quarter into the future. The grey line adjusts the blue line down using the DHPDI. The first two lines measure the aggregate house prices, while the third line focuses on those under distress. The difference between the first two lines is attributed to difference in methodology, which may include difference in population between the entirety of USA and what Freddie Mac handles. It is important that the third line be measured on a like-for-like basis and applied as close to the truth as possible, thus used as an adjustment to the blue line and not the orange one.

The DHPDI developed here addresses the crisis-sensitive modelling challenges identified in gap 3 of Section 1.4, providing a tool to capture how external drivers of LGD behave differently under stress. A detailed literature review follows, examining empirical findings on distressed sale discounts and the methodological implications for LGD model development.

This essay contributes a Distressed House Price Discount Index that isolates the discount experienced on distressed property transactions from generic house price movements. The practical contribution is that DHPDI provides a crisis-sensitive adjustment that traditional HPIs miss, a distinction that matters most precisely when accurate LGD estimation matters most, during market downturns when distressed sales become common.

5.2. Literature Review

Building on the LGD modelling framework established in the main literature review and addressing gap 3 (Section 1.4), this essay focuses on incorporating market dynamics during economic stress periods. As discussed in the main literature review, while OLS approaches have been foundational, these models often fall short in accounting for the dynamic nature of housing markets, particularly during periods of economic stress.

While the main literature review noted advances in machine learning approaches, the challenge remains to incorporate real-time market conditions effectively into LGD models.

House Price Indices (HPIs) play a crucial role in understanding market trends and their impact on property values. This extends the property-value based models discussed in the main literature review (Andersson & Mayock, 2014; Biswas et al., 2020; Pelizza & Schenk-Hoppé, 2019). While indices from sources like Zillow, FHFA, or even modelled ones such as Doumpos et al., 2021; Rodriguez-Serrano, 2024) provide a broad overview, they often lack the granularity needed to reflect local market fluctuations, especially in distressed sales scenarios (Goodman & Zhu, 2015). Recent works have shown that the inclusion of distressed sales such as foreclosures and short sales in HPIs can significantly bias these indices, especially during housing downturns. Doerner & Leventis (2015) found that standard HPIs which include distressed sales systematically understate price recovery and overstate declines in the aftermath of a crisis, with important implications for risk estimation and capital management. Their work demonstrates that constructing HPIs with and without distressed transactions can yield materially different trends and insights, underlining the importance of adjusting or segmenting indices when used for LGD modelling.

A continuing debate in the literature concerns the true magnitude and drivers of the price discounts observed on distressed transactions. Methodological choices can have a substantial impact on the estimates produced. For example, Conklin et al. (2023) introduced an innovative approach that uses appraisal fixed effects to control for both observable and unobservable property characteristics. Using this more rigorous method, they find foreclosure and short sale

discounts average around 5%, much lower than the double-digit figures cited in earlier studies. Critically, their work attributes most of the remaining discount to the stigma attached to distressed sales, rather than to differences in property condition or location. This cautions against overestimation of distress discounts in risk modelling and highlights the need for robust empirical controls.

At the same time, contextual factors must be considered. Aroul & Hansz (2014), studying the Fresno, California market during the US housing bubble and subsequent crash following 2008, find much higher average discounts (around 20% for foreclosures and 13% for short sales) even after accounting for possible biases such as time on market and self-selection. Importantly, these discounts vary over time, peaking during the height of the liquidity crisis and then falling as markets recover. These findings illustrate that distress discounts are not only present and economically meaningful, but also highly dynamic, reinforcing the need for LGD models and house price indices to incorporate a time-varying and context-sensitive approach to distressed sales.

While there has been acknowledgement of the important role property values play by incorporation of house price indices in modelling LGD in the form of origination price adjustments to foreclosure date (Do et al., 2018; Qi & Yang, 2009), recent studies made clear that the use of aggregate HPIs without adjustment for distress can lead to significant model error. There is an evident gap in the literature for a dedicated discount index - one that systematically adjusts property price indices for the empirically validated impact of distress and is flexible to both time and market conditions as indicated by recent findings.

Our study builds on above literature by introducing a Distressed House Price Discount Index (DHPDI) alongside a reconstructed HPI. This approach directly integrates insights from Aroul & Hansz (2014); Conklin et al. (2023); Doerner & Leventis (2015), offering an empirically robust and context-aware adjustment for distressed sales, and providing a more detailed and accurate perspective on property value dynamics relevant for LGD prediction. This approach aligns with the findings of Do et al. (2020), who emphasised the importance of considering market-driven variables in LGD modelling.

This aligns with the stage-based decomposition in Essay 1, where Stage 2 represents collateral disposition. Our research extends this concept by assuming the non-property sale-related quantity is known, allowing us to concentrate on the more variable property sale component, which corresponds to Stage 2 recovery in Essay 1's framework.

In summary, the construction of a standalone DHPDI intended to be used in conjunction with an aggregate median-based HPI in predicting LGD represents a significant advancement in the field. This addresses the crisis-sensitive modelling challenges identified in gap 3 (Section 1.4) and provides enhanced drivers for the collateral disposition stage explored in Essays 1 and 2. By doing this, our study offers a practical tool for improving risk management strategies in the mortgage lending industry.

While the dataset used in this study extends into the early COVID-19 period, it is important to note that mortgage foreclosure outcomes require considerable time to materialise. Loans that entered delinquency during 2020 would typically take 12 to 36 months or longer to reach resolution, depending on the state and resolution type. As a result, the COVID-19 period in our data is effectively right-censored: we observe the onset of economic stress but not the eventual foreclosure and recovery outcomes that would follow. This means the models estimated here do not capture the full loss experience of the pandemic period. Once sufficient post-pandemic disposition data becomes available, a natural extension of this work would be to examine how COVID regime policy interventions affected the composition of resolutions and the resulting loss severities. Such analysis could also explore whether the policy response created regime shifts in the relationship between macroeconomic stress and LGD that differ materially from patterns around the GFC period.

5.3. Methodology

5.3.1. Data Sources

Freddie Mac, alongside Fannie Mae, was created in response to the 1930s banking crisis, when banks were hesitant to offer long-term loans without matching funding. This led to a system that supported long-term mortgage lending. The introduction of securitisation by Ginnie Mae in 1968 further enabled long-term loans, making the 30-year fixed-rate mortgage a standard in the U.S. As explained by Green (2013), securitisation allowed for the pooling and selling of mortgage loans as securities, providing liquidity and stability to the housing finance system. Adjustable-rate mortgages, which began in the 1980s, offer a series of fixed rates over shorter periods, providing an alternative to the traditional fixed-rate model.

Our research uses the Freddie Mac dataset, focusing on U.S. single-family fixed-rate prime loans. This dataset spans loans from 1999 to 2019, with performance data extending to 2020.

The Freddie Mac dataset contains two interconnected components with distinct data structures and update frequencies: the origination and performance data.

5.3.1.1. Freddie Mac Performance Data

The performance data delivers quarterly snapshots of loan status, capturing critical information including current delinquency status, outstanding unpaid principal balance, modification activities, and remaining maturity periods. Most crucially for LGD modelling, the performance data records detailed loss and resolution information when applicable. When loans proceed to foreclosure or other credit default events, the dataset captures net property sale proceeds, collateral-related costs and expenses, recovery sources and amounts, and account closure reasons, referred to as "Zero Balance Code" within Freddie Mac (2019).

The dataset distinguishes between mortgage insurance (MI) recoveries and non-MI recoveries, providing transparency into loss allocation across different types of outcomes.

5.3.1.2. Freddie Mac Origination Data

The origination dataset captures comprehensive loan, borrower, and property characteristics at the time of loan origination. Credit risk indicators include FICO credit scores, original loan-to-value (LTV) ratios, combined loan-to-value (CLTV) ratios for cases where borrowers have other mortgages with another lender, and debt-to-income (DTI) ratios. Loan characteristics encompass original unpaid principal balance, origination rate, original loan term, origination date, number of borrowers, loan purpose classification and property information.

Property-level information includes Metropolitan Statistical Area (MSA) identification, state and ZIP code location, number of units, occupancy status (owner-occupied or investor), and property type classification. The dataset also captures mortgage insurance coverage percentages, servicer and seller identification codes, and various loan feature flags that may impact performance or recovery outcomes.

5.3.1.3. Transformed Freddie Mac Data

We use the Freddie Mac data in two parts:

- first, we construct a regular house price index using all origination data, employing origination security values.
- then, we conduct analyses on the 944,827 eligible defaulted observations that had the necessary data fields to predict Loss Given Default (LGD).

We implement several data quality filters to ensure reliable LGD calculations. First, we require complete loss and recovery information, including net sale proceeds, foreclosure costs, and any insurance recoveries. Second, we exclude loans with data inconsistencies, such as negative loss amounts or recovery amounts exceeding original principal balances. Third, we remove loans with implausible timing patterns, such as resolution dates preceding default dates.

Out of 3,026,608 default observations, 944,827 qualified for eligible resolutions[14] and had the necessary data fields for use of the HPI and DHPDI in predicting LGD.

5.3.1.4. External Data

5.3.1.4.1 Federal Housing Finance Agency (FHFA) Data

We supplement the Freddie Mac dataset with quarterly House Price Index data from the Federal Housing Finance Agency, providing market-level price appreciation benchmarks for validation purposes. The FHFA HPI covers all nine U.S. census divisions and individual states, enabling geographic price trend analysis and serving as a comparison metric for our reconstructed indices.

5.3.1.4.2 Federal Reserve Economic Data (FRED)

State-level unemployment rates are sourced from the Federal Reserve Bank of St. Louis Economic Database. This variable is matched to loan records based on property location and default timing.

5.3.1.4.3 Legal and Regulatory Data

We incorporate state-level foreclosure law classifications, distinguishing between judicial and non-judicial foreclosure states based on legal research and industry sources. These classifications remain relatively stable over our analysis period, though we account for any legislative changes that occurred within our timeframe.

The table below presents our list of final variables, selected through preliminary statistical analysis and economic theory. The Distressed Loan-to-Value (DLTV) ratio represents our primary measure of collateral coverage, calculated by adjusting the original CLTV ratio using house price appreciation from origination to default measured using the HPI from FHFA. LOB

[14] For this paper, eligible resolutions refer to "zero balance reasons" in Freddie Mac that mostly resulted to property disposition, namely: REO disposition, charge-off, third party sale, short sale.

is simply the log of origination unpaid principal balance. Variable transformations using Log is standard practice in credit risk modelling, especially for large values.

The Distressed House Price Discount Index (DHPDI) represents our key innovation, capturing the systematic discount applied to distressed property sales relative to regular property sale transactions.

***Table 31**. Definition of independent variables*

Variable	Definition
DLTV	Cross collateralised loan to value ratio at origination, adjusted using HPI to default date
LOB	log of balance at origination
LP	Purpose for the mortgage loan: · Purchase – used to purchase a property · Refinance with cash out – no specific purpose to the loaned amount. Was not used to purchase property. · Refinance with no cash out - limited to paying off first mortgage, or loans for other properties used to secure current mortgage, and cash out of min (2% of refinance amount, $2000) · Refinance but not specified · Unknown
NB	Number of borrowers: 1 2+ blank
NJF	States with non-judicial foreclosure: Connecticut, Delaware, Florida, Illinois, Indiana, Iowa, Kansas, Kentucky, Louisiana, Maine, Nebraska, New Jersey, New Mexico, North Dakota, Ohio, Pennsylvania, South Carolina (United States Foreclosure Laws, 2020)
OO	1 if property under mortgage is owner occupied, 0 if not.
UMP	State level unemployment rate at time of default
DHPDI	Distressed House Price Discount Index

8 independent variables were shortlisted based on a preliminary analysis showing that these are the top 8 strongest predictors of LGD/ recovery rates.

5.3.2. Theoretical Framework and Motivation

The construction of accurate house price indices presents fundamental challenges in real estate economics, particularly when attempting to capture the heterogeneous nature of property transactions. Traditional approaches often fail to distinguish between regular transactions and distressed sales, potentially biasing price level estimates and obscuring important market dynamics. Our methodology addresses this limitation by constructing two complementary

indices that separately capture general market conditions (which includes distressed sale) and distressed sale discounts.

The theoretical foundation for our approach rests on the recognition that distressed property sales, including foreclosures, short sales, and REO dispositions, systematically differ from voluntary market transactions. These differences manifest in several ways: reduced marketing time and effort, motivated sellers with limited negotiating power, potential property deterioration during the distress period, and buyer perceptions of elevated transaction risk. Consequently, distressed sales typically transact at discounts to comparable regular property sale transactions, creating a systematic bias in price indices that fail to account for transaction type.

Our dual-index approach enables the separate identification of general market price movements and distressed sale discounts. This separation is crucial for accurate LGD modelling, as it allows us to distinguish between losses attributable to overall market decline versus those resulting from the distressed nature of the sale process itself.

5.3.3. Index Construction

5.3.3.1. Justification for Median Use

As a measure of central tendency, the median (50th percentile) offers robust advantages when dealing with mortgage origination data that can be heavily influenced by outliers and skewed distributions. Real estate transactions often exhibit wide dispersion of property values and occasional spikes in high-value properties, all of which can shift or distort an average more than a median.

First, the median is less sensitive to outliers. In the Freddie Mac data, a few exceptionally large loan balances or property prices can skew the mean significantly, potentially misrepresenting typical market conditions. By relying on the median, we ensure that extremely high or low property values do not disproportionately distort the index trend and level.

Second, using the median quantifies the "typical" property price more intuitively, especially in contexts where data are not symmetrically distributed, which applies to the case illustrated in **Figure 18** below. This characteristic makes the median well suited for analysing property prices, where skewness may be inherent from regional variations and cyclical factors.

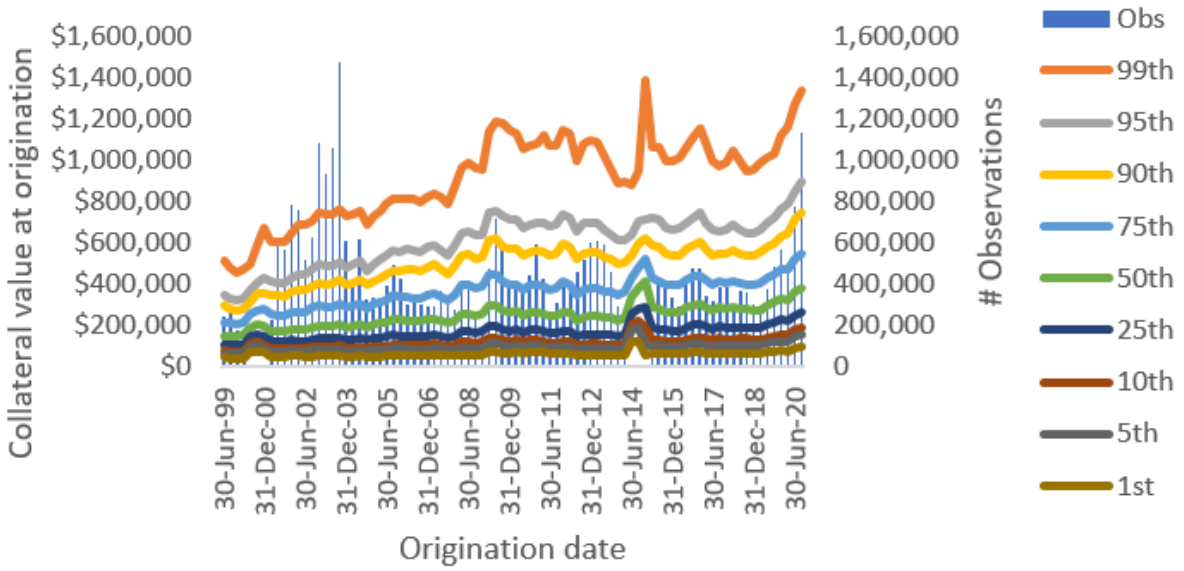

Figure 18. *Freddie Mac single-family origination property prices.*

Quarter level median of implied property prices from loan-to-value (LTV) ratios at origination. The blue bars represent the number of observations. Around mid-2014, fewer mortgages were originated (lower blue bars), and those that were originated tended to be for higher-value properties (increased values at most percentiles). This shift is attributed to two key Freddie Mac policies: 1) the new MOU for vacant properties per Bulletin 2014-7 @ Guide.Freddiemac.Com (n.d.), which likely reduced the number of lower-value distressed sales entering the market, and 2) the unchanged conforming loan limit of $417,000 per Fhfa-Announces-Fannie-Mae-and-Freddie-Mac-Conforming-Loan-Limits-for-2014 @ Www.Fhfa.Gov (n.d.), which may have influenced the distribution of loan amounts and potentially limited transactions for higher-priced homes.

Finally, median-based calculations are particularly appropriate for the construction of subnational indices at the MSA or state level, where fewer data points might exist in certain quarters. In these cases, the median often provides greater stability than the mean, ensuring that random spikes in either direction do not obscure genuine trends.

Two indices are constructed in this study:

5.3.3.2. Base HPI

First, a house price index (HPI) is created by taking the median (50th percentile) of all security values at origination using Freddie Mac data, calculated at the Metropolitan Statistical Area (MSA) level on a quarterly basis. This is achieved by dividing the loan origination value by the loan-to-value ratio (LVR).

The first index, defined at MSA and quarter level, is as follows:

$$HPI_{i,t} = M\left(P_{i,t,j}\right) \quad \text{..... 5.1}$$

Where M() is the median function across all observed Freddie Mac originated loans indexed j for the ith MSA observed on quarter t. In addition:

$$P_{i,t,j} = \frac{B_{i,t,j}}{LTV_{i,t,j}} \quad \text{..... 5.2}$$

Where $B_{i,t,j}$ is the balance at origination for the jth loan which was originated at quarter t and listed under the ith MSA and $LTV_{i,t,j}$ is the loan-to-value ratio of the same loan.

A national level index is presented in **Figure 18**.

5.3.3.3. DHPDI

Constructing a distressed property price index has proven to be a challenge given the change in distribution of distressed sales at each point in time and sparsity of data for some periods even at a national level. Since the purpose was to acquire a better understanding of being in distress has in relation to regular house price expectations, it only made sense to focus on the relative discount in the data.

The second index is constructed by calculating the centralised discount between two prices: the actual distressed sale price and the HPI-adjusted value from origination collateral values using the previously constructed HPI above. Those marked as "C" (stands for "Covered") in the Freddie Mac dataset are assigned a discount of 0. This new index is also constructed using medians at the quarter sale date and MSA level.

The second index, defined as $DHPDI_{i,t}$, is constructed as follows:

$$DHPDI_{i,t} = \begin{cases} M\left(\frac{S_{i,t,j}}{P_{i,t,j}\frac{HPI_{i,t}}{HPI_{i,0}}} - 1\right) \; if \; S_{i,t,j} \; is \; available \\ 0 \; if \; S_{i,t,j} \; is \; provided \; as \; \text{"C" in Freddie Mac} \end{cases} \quad \text{..... 5.3}$$

Where M() is the median function all observed Freddie Mac originated loans indexed j for the ith MSA observed on quarter t, $HPI_{i,t}$ and $P_{i,t,j}$ are as above, and $HPI_{i,0}$ is the HPI value at the quarter right before origination date of loan j, and $S_{i,t,j}$ is the actual price of the property mortgaged under loan j located in the ith MSA and sold at quarter t.

The DHPDI represents our key methodological innovation, designed to capture the systematic discount applied to properties sold through distressed channels relative to their expected market value. This index addresses a critical gap in existing LGD modelling approaches, which typically rely on generic house price indices that may not adequately reflect the unique characteristics of distressed sales.

To illustrate, this is what the national level DHPDI looks like:

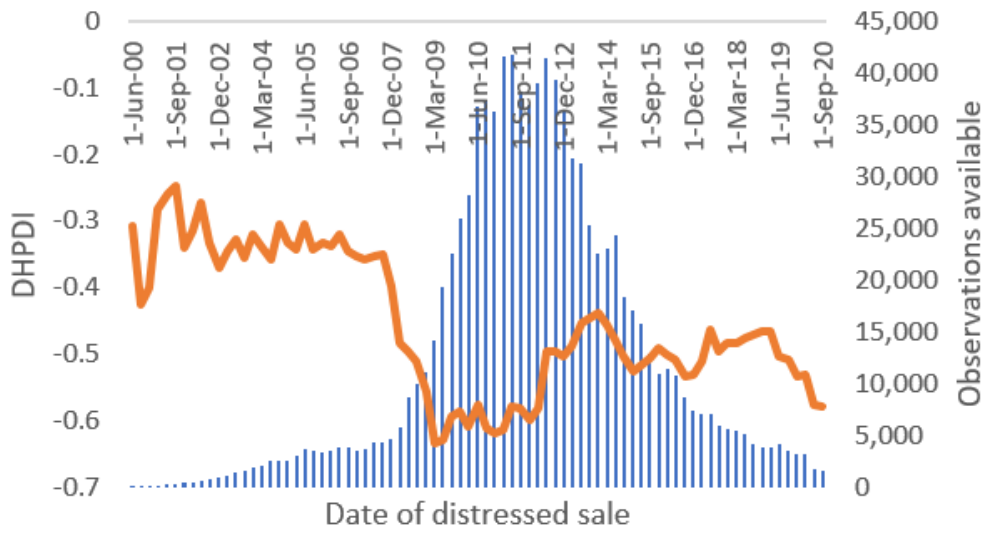

Figure 19. DHPDI index vs volume of distressed sale.

Blue bars represent volume of distressed sale, and the orange line represent the median of the centralised discount index from the blue line in Figure 17. A negative 40% value, as an example, means a 40% discount to the median property price while 0 means no adjustment. Given the nature of distressed sale, it is trivial to get negative values for this index.

5.3.3.4. Use Case

Note that **Figure 18** and Figure 19 were only presented for illustration purposes. Actual indices used in modelling LGD are evaluated at MSA level on a similar quarterly pattern as both figures illustrated.

5.3.3.5. General Variations across indices

Naturally, due to data concentration, some MSAs would have more robust measurements that others. While it may be a weakness for this methodology, we have proceeded with the use of it with an assumption that this study is about establishment of a proof of concept that the methodology works with sufficient data. To further enhance its commercial viability, practices around cluster analysis may be required for MSAs with low volumes, especially for benign periods.

5.3.4. Model Development

5.3.4.1. Definitions of Key Acronyms

Table 32. *Dependent variable definition*

Variable	Definition	Formulae
EAD	Exposure at default. The amount owed at the default date.	
LGD	Loss given default. The proportion of EAD that is considered a loss.	$LGD = \begin{cases} 0 & \text{if no loss} \\ 1 - R & \text{if loss present} \end{cases}$
R	The proportion of EAD recovered, for non-cured accounts.	$R = RNP + RP$
ANP	The total dollar amount collected from non-property sale proceeds/ expenses. Each cashflow discounted to the default date if there is a discount rate.	
AP	The total dollar amount collected from property sale proceeds. Each cashflow discounted to the default date if there is a discount rate.	
RNP	The recovery rate for non-property sale proceeds/ expenses expressed as a proportion of EAD.	$RNP = \frac{ANP}{EAD}$
RP	The recovery rate for property sale proceeds/ expenses expressed as a proportion of EAD.	$RP = \frac{RP}{EAD}$

This table defines core modelling quantities and their formula. In general, R is used for recovery rate (e.g., RNP, RP), and A is used to define dollar amounts (e.g., ANP, AP).

5.3.4.2. Framework and Approach

In our approach, we break down Loss Given Default (LGD), or recovery rates (R), into two sections: property sale related, and non-property sale related. We assume that the non-property sale related quantity is known. Since the most material part of the recovery rate R is found in collateral value disposition, it makes sense to make this assumption to enable the right focus.

We incorporate the Distressed House Price Discount Index (DHPDI) as an independent variable in a multivariate regression framework, assessing their impact on LGD alongside traditional borrower and loan characteristics. Two model methodologies are used and tested: one with a standard set of characteristics and another with the constructed DHPDI used as an additional predictor.

Table 33. *Model definitions:*

Model name	Description
B1	OLS model with a standard set of characteristics
B2	2-step model with a standard set of characteristics
B3	OLS model with a standard set of characteristics + DHPDI
B4	2-step model with a standard set of characteristics + DHPDI

Models B1-B4 have dependent variable RP. B1 and B2 are benchmark models, while B3 and B4 are the models utilising DHPDI as an additional independent variable. B1 and B3 use a naïve OLS model, while B2 and B4 use the same 2-step model introduced by Do et al. (2020).

The idea behind this structure that regardless of the sophistication of a model, the biggest uplift always come from the additional information incorporated if used right. This is evidenced in Bellotti & Crook (2012). Layering this logic on top of the structure defined above, if models B3 and B4 have produced more accurate results than models B1 and B2, then the use of DHPDI as an additional predictor for LGD/ recovery rate is justified.

5.3.4.3. OLS models B1 and B3

The use of Ordinary Least Squares (OLS) model in predicting LGD has always been standard practice. If not to illustrate superiority of incorporating additional information (Bellotti & Crook, 2012; Jankowitsch et al., 2014), it is often used as a benchmark (Bellotti et al., 2019; Tobback et al., 2014). Due to its simplicity, it is also considered a standard methodology to explain relationship between multiple variables (Ben-David, 2019). As such, the use of OLS in this study ensures alignment and preservation of relevance with the industry.

5.3.4.4. 2-Step models B2 and B4

The 2-step model from Do et al. (2020) is one among multiple models in the literature explored to better predict LGD. With the key variables they used, they claimed significant outperformance against the OLS. The use of this model is an attempt at illustrating that the idea of incorporating useful information appropriately also works for models more sophisticated (and more accurate in predicting LGD) than OLS.

5.3.5. Sampling method

Our study employs two sampling methods:

In-Time Sampling: We randomly allocate 30% of observations as test dataset, while an independent 35% is used for training. The rest are used for validation, if any, otherwise discarded.

Out-of-Time Sampling: We treat half of the defaulted observations in an 8-year window as training dataset, the other half as validation, if needed, and use observations in the following

year as test dataset. The rest are discarded. This process is repeated annually, advancing the start and end dates for both training/validation and testing, a technique known as rolling sampling window.

Figure below illustrates these sampling methods. The vertical axis represents account IDs, while the horizontal axis shows default dates. In the left chart, blue points indicate in-time sampling, and in the right chart, grey triangles represent out-of-time sampling with rolling windows. Training begins in January 1999 and continues until January 2010, with cut-off dates from December 2006 to December 2017, and corresponding test datasets dated 2007 up to 2016. Due to limited data after 2019, all observations from January 2018 to December 2018 are included in the final test sample, trained on data from January 2010 to December 2017. Remaining data (yellow) is ignored for irrelevance.

5.3.6. Accuracy assessment

Relative Mean Square Error (RMSE) and R-square are widely used in studies such as Hurlin et al. (2018); Loterman et al. (2012); Qi & Zhao (2011); Thomas et al. (2012) to assess model performance.

Performance measures like the area under the curve (AUC) and linear correlation between predictions and actual values are common for assessing model discrimination. However, RMSE and R-square are prioritised for accuracy. While models with high accuracy often have good discrimination, the reverse is not always true (Loterman et al., 2012). Therefore, RMSE and R-square are emphasised in our analysis, with R-square also serving as a benchmark for comparison with existing literature.

RMSE evaluates the error term in the units of the dependent variable, R, while R-square measures how much better the model performs compared to using the mean as a predictor.

To provide a counterbalance between RMSE's accuracy and over-specification from having too many predictors, adjusted R-square is used in place of R-square. Adjusted R-square provides a more accurate reflection of model performance by penalising the addition of unnecessary variables, thus maintaining model integrity and preventing overfitting.

By measuring RMSE and adjusted R-square, we ensure a comprehensive evaluation of model accuracy, balancing precision with robustness.

RMSE and adjusted R-square are measured on R, as a function of modelled quantity RP and assumes RNP is known.

5.4. Empirical Results

5.4.1. Performance

Our findings indicate that the inclusion of DHPDI significantly improves LGD model accuracy for 10 out of 11 sampling windows. The uplift is illustrated using performance measure discussed: adjusted R-square. Given that the sampling window (test dataset at year 2014) where models B3 and B4 failed to outperform their equivalent benchmarks due to explainable reasons mentioned in **Figure 18**, it is considered justified. The index captures market stress more effectively when used alongside traditional metrics, highlighting its value in predicting losses at the start of economic downturns.

Table 34. *Model performance*

	Adj r-square on test dataset				**RMSE on test dataset**				**Relative improvement vs best benchmark**		
Test type	B1	B2	B3	B4	B1	B2	B3	B4	adj r-square	RMSE	Best
In time	0.34	0.35	0.43	0.43	0.25	0.25	0.24	0.24	23%	6%	B4
In time = 1999-2006/ Out of time = 2007	0.15	0.16	0.14	0.18	0.28	0.28	0.28	0.28	13%	1%	B4
In time = 2000-2007/ Out of time = 2008	-0.29	-0.28	0.16	0.15	0.34	0.34	0.27	0.27	-156%	19%	B3
In time = 2001-2008/ Out of time = 2009	0.20	-0.03	0.34	0.31	0.25	0.29	0.23	0.24	68%	9%	B3
In time = 2002-2009/ Out of time = 2010	0.44	0.43	0.51	0.51	0.23	0.23	0.22	0.21	15%	6%	B4
In time = 2003-2010/ Out of time = 2011	0.45	0.45	0.50	0.50	0.22	0.22	0.21	0.21	11%	5%	B3
In time = 2004-2011/ Out of time = 2012	0.19	0.20	0.22	0.22	0.29	0.29	0.28	0.28	9%	1%	B3
In time = 2005-2012/ Out of time = 2013	0.19	0.22	0.32	0.33	0.27	0.27	0.25	0.25	53%	8%	B4
In time = 2006-2013/ Out of time = 2014	0.23	0.25	0.17	0.19	0.29	0.28	0.30	0.29	N/A	N/A	B2
In time = 2007-2014/ Out of time = 2015	0.33	0.35	0.38	0.38	0.29	0.29	0.28	0.28	11%	3%	B4
In time = 2008-2015/ Out of time = 2016	0.41	0.42	0.44	0.45	0.30	0.30	0.29	0.29	8%	3%	B4
Overall	0.24	0.23	0.33	0.33	0.27	0.28	0.26	0.26	38%	6%	B4

Measured in RMSE and adjusted R-square. As covered in Figure 18, 2014 had a policy change hence should be disqualified from being a test dataset. For the rest of the sampling windows, candidate models using DHPDI at MSA level outperform their equivalent benchmarks for both performance measures.

5.4.2. A note for negative R-square

There are times when R-squares yield negative values. When this happens, the practical implication (given the formula) is that the model is better off predicting using the mean of the dependent variable (recovery rate, in this case). This may typically occur when sample sizes are small and distributions are more random than explainable, or when information in the training dataset is almost irrelevant to the test dataset. Knowing this before carrying out modelling exercises is helpful in identifying whether there is a need (and benefit) from incorporating more sophisticated models. In this case, testing windows in 2008 and 2009 yielded negative R-squared values for some models particularly because of both reasons:

distressed property sale volumes were lower in 2000-2007, and 2008-2009 is the period where the GFC outcomes started materialising.

5.4.3. Inference

The standardised parameter estimates across models B1 to B4 provide valuable insights into the relative importance and direction of influence of various predictors on the outcome variable, with a focus on the Distressed House Price Discount Index (DHPDI). The positive coefficients for DHPDI, observed in models B3 and B4 with values of 0.07, suggest that higher distressed house price discounts are associated with improved recovery outcomes. This relationship indicates that when properties are acquired at significant discounts, there is potential for enhanced financial recovery, likely due to the ability to sell these properties at a profit once market conditions improve. The role of DHPDI as a key variable in the analysis highlights its importance in understanding the dynamics of loan recovery. By capitalising on market conditions that allow for advantageous property transactions, lenders can potentially mitigate losses and improve overall recovery rates. This finding aligns with the broader literature on distressed asset management, where strategic purchasing and selling decisions are crucial for financial optimisation.

In addition to DHPDI, other variables such as DLTV and LOB also play significant roles. The negative standardised estimates for DLTV, ranging from -0.08 to -0.10, underscore the risk associated with higher loan-to-value ratios. This finding is consistent with the established understanding that higher LTV ratios are linked to increased default risk, as highlighted by Greve & Hahnenstein (2016); Somers & Whittaker (2007). The consistent negative effect of DLTV underscores the importance of managing LTV ratios to mitigate risk, as higher ratios often indicate less borrower equity and greater vulnerability to market fluctuations.

Conversely, the positive estimates for LOB, ranging from 0.06 to 0.08, indicate a positive relationship with the outcome variable. This suggests that loans with higher origination balances may be prioritised for recovery, consistent with findings by Calem & LaCour-Little (2004); Pennington-Cross (2003), who noted that larger loans often receive more attention in recovery efforts due to their potential impact on overall financial outcomes. The prioritisation of larger loans for recovery efforts may reflect lenders' strategies to maximise returns on significant investments, thereby reducing overall portfolio risk.

The positive estimates for OO, ranging from 0.03 to 0.04, indicate that owner-occupied properties are associated with better outcomes. This is likely due to better maintenance and

higher resale values, as owner-occupiers tend to take more care of their properties, a finding supported by Clauretie & Daneshvary (2011); Qi & Yang (2009). Owner-occupied homes are often better maintained and more desirable in the market, leading to higher recovery values and reduced loss severities.

The slight negative impact of UMP_L3, with estimates from -0.02 to -0.03, highlights the influence of broader economic conditions on loan performance. Higher unemployment rates are typically associated with increased default risk, reflecting economic stress and reduced borrower capacity to meet financial obligations. This finding aligns with the broader economic literature that links macroeconomic indicators to credit risk.

The consistent positive estimates of 0.02 for NJF suggest that properties in non-judicial foreclosure states may experience quicker foreclosure processes, potentially leading to more efficient recoveries. The expedited process in these states can reduce holding costs and expedite the return of capital, enhancing recovery outcomes.

The negative estimates of -0.02 for NB_01 indicate that loans with a single borrower may be more vulnerable to financial distress compared to those with multiple borrowers, who can provide additional financial stability. This finding suggests that the presence of multiple borrowers can mitigate risk by diversifying income sources and financial responsibility, reducing the likelihood of default.

The estimates for LP_C and LP_P are small, indicating a minimal impact on the outcome variable. However, the positive estimate for LP_P suggests that purchase loans might perform slightly better than cash-out refinance loans, aligning with the notion that purchase loans are often more conservatively underwritten. This conservative underwriting can lead to better loan performance and reduced default risk.

An important caveat when interpreting results from the later years of the sample is that the dataset's performance window extends only to 2020. While this captures the onset of the COVID-19 economic shock, mortgage foreclosure outcomes typically require 12 to 36 months or longer to materialise from the point of serious delinquency, depending on the state jurisdiction and resolution pathway. Loans that entered default during the pandemic period have therefore not yet reached resolution within the observation window, meaning the COVID-era loss experience is effectively right-censored in our data.

This has two implications. First, the models estimated here are not materially influenced by pandemic-era loss outcomes because those outcomes had not yet crystallised. Second, the COVID-19 period introduced a policy environment, including mandatory forbearance and foreclosure moratoriums (Cherry et al., 2021), that directly interrupted the disposition pipeline. These interventions would have altered both the timing and composition of resolutions in ways that cannot be separated from the underlying economic shock using the current data.

Once sufficient post-pandemic resolution data becomes available through future Freddie Mac data releases, it would be valuable to examine whether the relationship between macroeconomic stress and LGD during the COVID-19 period differs structurally from GFC patterns, and whether the staged framework developed here can accommodate such regime shifts through recalibration of the resolution probability and conditional recovery modules.

Table 35. *Parameter estimates*

Standardised estimates	**Model**			
Parameter	**B1**	**B2**	**B3**	**B4**
DLTV	-0.0844***	-0.0886***	-0.0925***	-0.0967***
	(0.0003)	(0.0003)	(0.0005)	(0.0005)
LOB	0.0573***	0.0652***	0.0689***	0.0759***
	(0.0004)	(0.0004)	(0.0005)	(0.0005)
DHPDI			0.0721***	0.074***
			(0.0003)	(0.0005)
OO	0.0263***	0.0286***	0.0336***	0.0367***
	(0.0003)	(0.0003)	(0.0004)	(0.0005)
UMP_L3	-0.0198***	-0.0211***	-0.0298***	-0.0321***
	(0.0003)	(0.0003)	(0.0005)	(0.0005)
NJF	0.0151***	0.0154***	0.0166***	0.0169***
	(0.0003)	(0.0003)	(0.0004)	(0.0004)
NB_01	-0.0158***	-0.0163***	-0.0211***	-0.0216***
	(0.0003)	(0.0003)	(0.0004)	(0.0005)
LP_C	-0.0213***	-0.0218***	-0.0231***	-0.0238***
	(0.0004)	(0.0004)	(0.0005)	(0.0005)
LP_P	0.0063***	0.0066***	0.0136***	0.0141***
	(0.0004)	(0.0004)	(0.0006)	(0.0006)

Standardised scores are presented in descending order of influence on RP.

5.5. Conclusion

Our research demonstrates the effectiveness of incorporating a dual-index approach to enhance LGD predictions for residential mortgage loans. Building on gap 3 (Section 1.4), this approach specifically addresses the challenges of modelling LGD during economic downturns when

distressed sales become prevalent. By integrating a house price index (HPI) with a Distressed House Price Discount Index (DHPDI), we capture the dynamic nature of housing markets more accurately.

Our proposed indices significantly outperform traditional models. This enhanced accuracy is particularly valuable for Stage 2 recoveries (collateral disposition) identified in Essay 1 and for resolution-specific valuations in Essay 2.

The dual-index approach offers distinct advantages across different applications:

For credit underwriting, financial institutions can leverage the DHPDI to better estimate potential losses in default scenarios. By incorporating the distressed-specific discount into underwriting models, banks obtain a more accurate picture of potential recovery values, especially in riskier lending environments. This supports more granular pricing of mortgage products through improved exposure-at-default estimates.

For regulatory compliance, regulators and internal risk managers often require conservative yet realistic LGD estimates to determine capital adequacy under stress-test scenarios. DHPDI-enhanced models offer a fit-for-purpose metric that isolates distress-driven discounts. This isolation proves to be particularly actionable when setting loss reserves and credit appetite, originating loans, or adjusting capital buffers.

For macroeconomic analysis, the comprehensive HPI remains more suitable for assessing long-term housing market trajectories, evaluating property portfolios, and informing central bank policies.

The strong performance of the DHPDI observed in this study relies on both indices being derived from similar underlying datasets. Additionally, the median-based approach employed for the DHPDI may be particularly effective because our HPI is also computed using median values. Users should ensure comparable data sources and aligned statistical measures when implementing this approach.

In conclusion, the integration of HPI and DHPDI into LGD models represents a significant advancement in understanding and predicting mortgage losses. Combined with the decomposition approaches presented in Essays 1 and 2, this provides a comprehensive framework for managing mortgage credit risk in both stable and stressed market conditions.

6. Chapter 6 - Conclusion

This thesis set out to develop and empirically test new methods for modelling Loss Given Default (LGD) in residential mortgage portfolios. Each of the three essays contributed a unique perspective on deconstructing and refining LGD estimation. Collectively, these essays highlight the advantages of a more granular and flexible modelling framework—one that accommodates both the complex pathways to recovery and the nuances of housing market dynamics.

6.1. Decomposing LGD into Three Cash Flow Components (Essay 1)

Essay 1 demonstrated the benefits of breaking down LGD into three stages of cash flows, each modelled separately. This decomposition yielded improved accuracy compared to traditional single-stage or purely OLS-based approaches. By isolating factors such as mortgage insurance coverage, modification flags, and time to resolution, the models more effectively captured the distinct drivers of recovery at each phase of default resolution. Decomposition also brought transparency: banks could better identify where and how interventions—such as foreclosure strategies, loan modifications, or short sales—might reduce overall losses.

6.2. Integrating Resolution Pathways into LGD Modelling (Essay 2)

Essay 2 further refined the focus on resolution pathways by modelling propensity of each default resolution and equivalent corresponding expected recovery rates separately. This modular design simplifies ongoing model maintenance, allowing banks to recalibrate specific modules when performance diverges, rather than overhauling the entire LGD framework. Moreover, operational teams gain actionable insights into early- and late-stage collections strategies. From selectively in-housing loans expected to yield higher recoveries, to timing distressed debt sales, this method bridges the gap between risk modelling and practical collections management. One caution, however, is to avoid the "winner's curse" of self-fulfilling forecasts, whereby the resolution with the highest observed recovery might be favoured merely because it is already heavily resourced. Randomised trials or independent performance validations are essential to keep such biases in check when in full implementation.

6.3. Incorporating a Dual-Index Approach for Housing Market Dynamics (Essay 3)

In essay 3, the thesis introduced a Distressed House Price Discount Index (DHPDI) to capture the price discounts inherent in distressed sales. When used appropriately in conjunction with a median-based House Price Index (HPI), DHPDI significantly enriched LGD models by

isolating distressed-sale discount effects from broader market movements. Empirical tests showed material improvements in prediction accuracy compared to benchmarks. While the HPI on its own remains valuable for macro-level forecasting and policy analysis, the DHPDI offers more precise insights for loan originations, provisioning, capital optimisation, and stress testing, particularly during systemic downturns. This essay also highlighted certain limitations, notably that both the DHPDI and HPI should come from comparable underlying data sources for maximum reliability, and that aligning median-based approaches across both indices preserves consistency.

6.4. Contribution to Literature and Industry

Taken together, these three essays offer a cohesive strategy for modelling LGD in residential mortgages:

- Granular Decomposition: essays 1 and 2 illustrate how splitting LGD into logical stages or resolution pathways illuminates hidden drivers and yields more accurate, actionable results.
- Market Sensitivity: essay 3 introduces a dual-index framework that not only refines LGD predictions further but also addresses how distressed sales behave under different market conditions, a key concern during systemic crises. While the dataset captures the onset of the COVID-19 period, foreclosure outcomes from this period had not yet materialised within the observation window, and this remains an important area for future study.
- Practical Relevance/ Commercial Viability: From heightened transparency and better model recalibration to improved stress-test reliability, the proposed methods are designed to help banks manage defaulted mortgages more effectively, ultimately minimizing credit losses and volatility.

More broadly, the thesis proposes a modelling philosophy in which LGD estimation is decomposed into economically and practically meaningful stages. This piece-wise approach reflects how mortgage losses actually materialise: a loan must first reach resolution, the resolution pathway determines the recovery mechanism, and recovery rate is then conditional on the specific disposition methodology. In addition, the source of cash flow differs for each stage of foreclosure and type of resolution. The framework is designed so that individual modules can be adopted selectively. An institution with limited modelling resources might implement only the Stage level decomposition from Essay 1, while a more

sophisticated institution could layer on the 2-step resolution models from Essay 2 and the dual-index approach from Essay 3.

6.5. Limitations and Future Research

These findings, while promising, do come with constraints. First, the data used (Freddie Mac prime single-family loans) limits generalisability to other jurisdictions or subprime segments. Second, the success of both the decomposition approach (essays 1 and 2) and the DHPDI (essay 3) relies on suitable data granularity and accurate capturing of distressed sales. As agencies, regulators, and private institutions gather richer datasets, subsequent research can explore whether additional segmentation by property type, borrower characteristics, or even macroeconomic indicators, offers further gains. Moreover, cross-validation of DHPDI techniques across geographies or with different measures of central tendency (beyond the median) would help confirm the robustness of the index.

For practitioners looking to implement this framework, several components require institution and industry-specific customisation. The resolution categories themselves may differ: the resolutions identified in Freddie Mac data reflect US-specific processes, and institutions in other markets would need to define resolution categories consistent with their own legal, regulatory and operational environment. The stage boundaries (pre-disposition, disposition, post-disposition) are likely to transfer more directly, as they reflect the general structure of secured lending recovery. The coefficients, however, would need to be re-estimated on local data. For institutions in full-recourse jurisdictions such as New Zealand and Australia, the Stage 3 (post-disposition) component becomes particularly important, as lenders can pursue borrowers for shortfalls, a recovery channel that is limited in many US states by non-recourse provisions or restrictions on deficiency judgments.

A natural extension of this framework is the joint modelling of PD, LGD, and EAD within a unified system. Regulatory and industry interest in integrated credit risk modelling has grown, particularly in the context of IFRS 9 and CECL, where the interaction between default probability and loss severity under different macroeconomic scenarios is central to provisioning calculations. The staged framework developed here provides a foundation for such integration: the resolution propensity model, for instance, could be linked to PD models through shared macroeconomic drivers, while the conditional recovery models could be combined with EAD models to produce joint loss distributions. This extension is left for future research.

6.6. Closing Remarks

Across these three studies, this thesis demonstrates the advantages of designing LGD models around the real complexities of mortgage default and recovery. By dissecting LGD into subcomponents and capturing distressed-sale discounts explicitly, financial institutions gain sharper, more commercially viable models that adapt more readily to evolving market landscapes. In turn, regulators benefit from more transparent and accurate risk measures. Ultimately, the lessons learned from leveraging cash flow sources, resolution pathways, and a dual-index approach serves as a platform for future research, inviting continued refinements that strengthen both scholarly understanding and industry practice around mortgage credit risk.

Appendix

Further details about the datasets used

Statistical Evidence and Economic Rationale

The decomposition of recovery rates into distinct stages requires both statistical justification and economic reasoning. This approach extends the multi-step conditional approaches (Do et al., 2018, 2020; Leow et al., 2014) discussed in Section 2 by aligning decomposition with the actual recovery process. This section presents empirical evidence supporting our three-stage framework (essay 1) using the Freddie Mac Single Family Loan-Level dataset, demonstrating how different recovery channels exhibit unique characteristics and respond to different economic drivers.

We begin by examining the distributional properties of recovery rates across stages. For defaulted accounts that do not cure through prepayment or repurchase, Stage 1 represents internal collection efforts prior to collateral disposition.

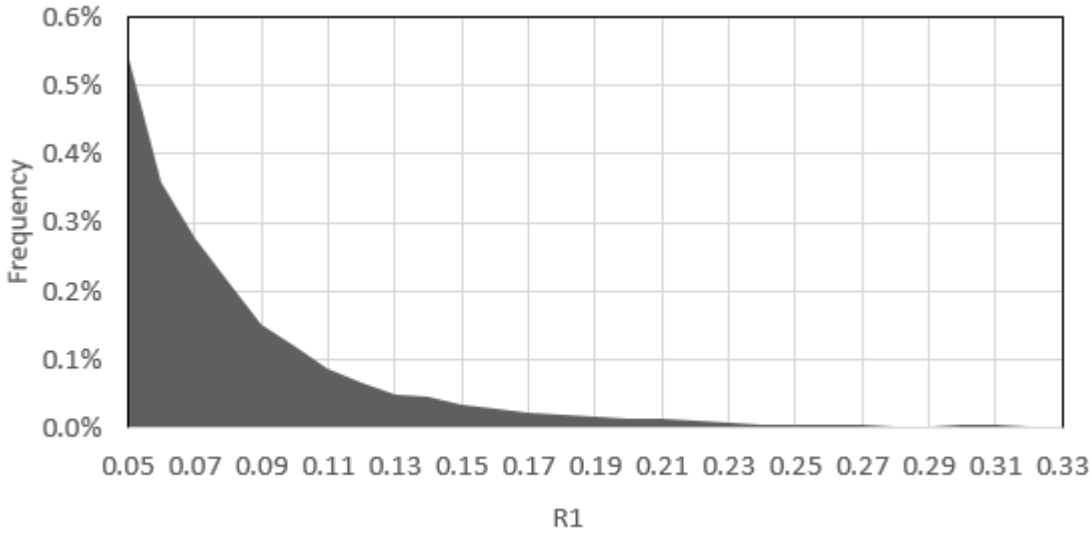

Figure 20. *Non-zero R1 distribution.*

For a defaulted account that does not cure (prepay or repurchase), it goes through internal in-house collection (Stage1) where A1 is recovered, expressed as % of total amount owed EAD, R1. Cases when R1 = 0 are removed for illustration purposes. This is a histogram of non-discounted R1 using the Freddie Mac dataset. Vertical axis denotes frequency of observation and horizontal axis denotes values of R1.

Figure 20 reveals the right-skewed distribution of non-zero Stage 1 recovery rates (R1), reflecting the varying success of early collection efforts. The distribution suggests that while most early recoveries are modest, some accounts achieve substantial recovery through borrower payments before proceeding to collateral disposition.

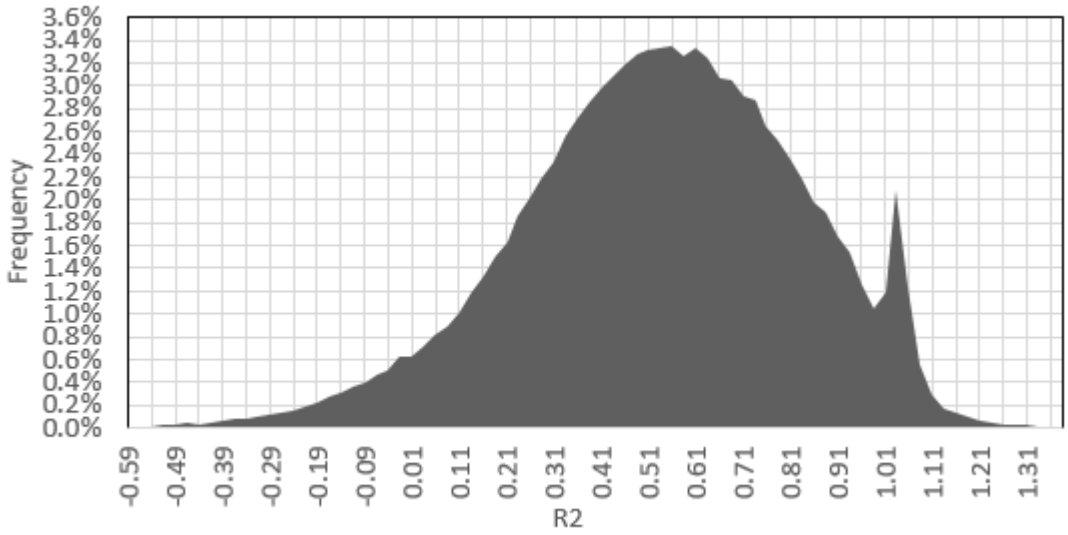

Figure 21. *R2 distribution.*

In the presence of shortfall, collection process proceeds to Stage 2 where collateral is liquidated, or contract is transferred (case of note/ reperforming sale) after deducting for allowable property/contract disposition expenses, A2, expressed once again as a proportion of EAD, R2. This is a histogram of non-discounted R2 using the Freddie Mac dataset. Vertical axis denotes frequency of observation and horizontal axis denotes values of R2. High frequency around 100% are those that have everything owed recovered with relatively lower additional expenses. While lenders are not allowed to recognise profit, total outflow of cash from lenders may exceed that of EAD if expenses appear to be material, like cases when EAD is small.

Stage 2 recovery rates (R2), shown in **Figure 21**, demonstrate noticeably different characteristics. The distribution ranges from approximately -50% to 130%, with notable concentrations around 60% and 100%. The negative values typically represent cases where small exposure amounts are overwhelmed by disposition costs, while values exceeding 100% reflect scenarios where sale proceeds cover both the exposure and associated expenses. This pattern aligns with the economic intuition that collateral-based recovery is primarily driven by property market conditions and disposition efficiency.

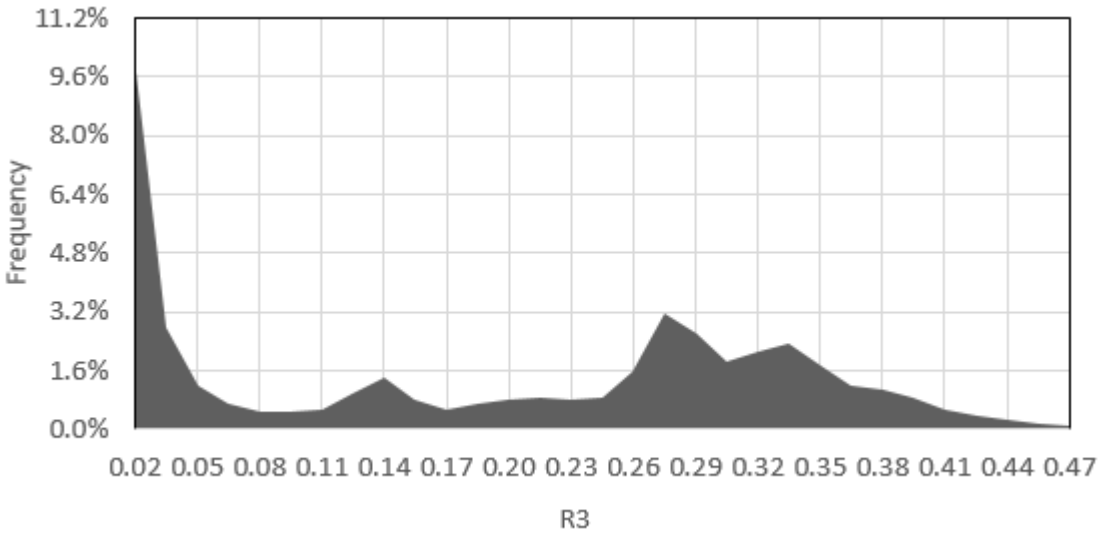

Figure 22. *Non-zero R3 distribution.*
If shortfall is still present, process goes to Stage 3 where insurance proceeds and other miscellaneous cash flows are involved. These quantities add up to A3. Expressed as a proportion of EAD, we have R3. This is a histogram of non-discounted non-zero R3 using the Freddie Mac dataset. Vertical axis denotes frequency of observation and horizontal axis denotes values of R3.

The distribution of non-zero Stage 3 recovery rates (R3), illustrated in **Figure 22**, exhibits a distinctive multimodal pattern. The four observed peaks correspond to typical mortgage insurance coverage levels (0%, 13-14%, 27-28%, and 34%), with slight variations due to additional non-insurance recoveries. This structure reflects the institutional nature of Stage 3 recoveries, particularly the standardized insurance coverage levels in the mortgage market.

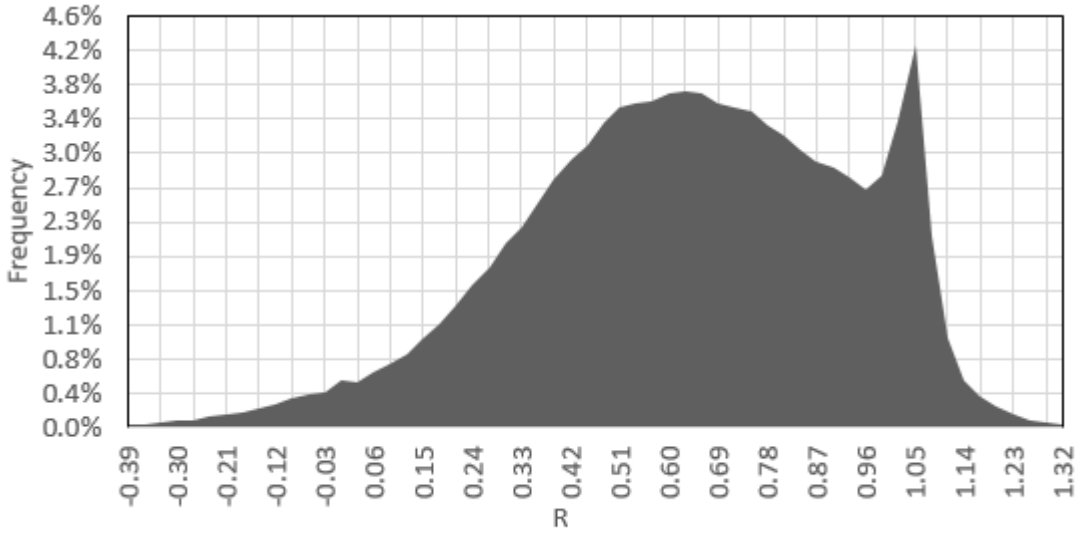

Figure 23. R distribution.

This is a histogram of non-discounted R using the Freddie Mac dataset. Vertical axis denotes frequency of observation and horizontal axis denotes values of R. High frequency around 100% are those that have recovered everything owed. While lenders are not allowed to recognise profit, total outflow of cash from lenders may exceed that of EAD if expenses appear to be material, like cases when EAD is small.

When examining the aggregate recovery rate distribution in Figure 23, we observe characteristics that blend elements from all three stages. While the overall shape resembles the Stage 2 distribution, reflecting the dominance of collateral-based recovery, the more pronounced peak at 100% demonstrates the complementary effects of Stage 1 and Stage 3 recoveries. This aggregation masks important stage-specific patterns and potentially obscures the underlying drivers of recovery performance.

The distinct distributional characteristics across stages are complemented by varying temporal patterns and economic relationships. Stage 1 recoveries show sensitivity to servicing practices and borrower circumstances, while Stage 2 recoveries closely track real estate market conditions. Stage 3 recoveries, predominantly driven by insurance proceeds, exhibit more stable patterns but vary systematically with initial loan-to-value ratios and insurance coverage levels.

These empirical patterns support three key conclusions. First, the recovery process comprises distinct channels with unique statistical properties. Second, these channels respond to different economic drivers, suggesting the need for stage-specific modelling approaches. Third, the aggregation of recovery rates masks important patterns that could inform both prediction and policy.

The evidence presented here provides a foundation for our modelling approach, which explicitly accounts for these stage-specific characteristics. This framework not only improves predictive accuracy but also offers practical insights for loan servicing and risk management strategies.

Temporal Patterns in Recovery Rates

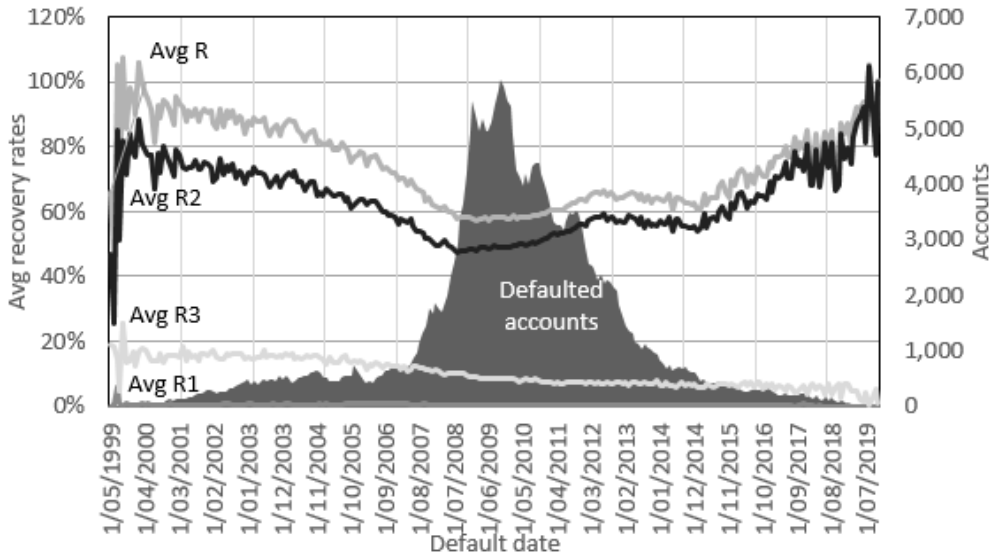

Figure 24. *EAD weighted average recovery rates time series.*

Left vertical axis denotes recovery rate (for avg R, avg R1, avg R2 and avg R3 line charts), right vertical axis denotes default volume (for defaulted accounts area chart), and horizontal axis denotes date of default.

The temporal patterns of recovery rates across stages reveal distinct behaviours, particularly during economic stress periods. **Figure 24** illustrates the EAD-weighted average recovery rates and default volume from 1999 to 2019. During the 2008 Global Financial Crisis (GFC), we observe a sharp increase in default volume accompanied by a decline in overall recovery rates (Avg R). While default volumes eventually returned to pre-GFC levels, recovery patterns show interesting stage-specific variations. Stage 2 recoveries (Avg R2) have demonstrated remarkable resilience, returning to and even exceeding pre-GFC levels (~80%). In contrast, Stage 1 recoveries (Avg R1) have remained consistently low throughout the period, while Stage 3 recoveries (Avg R3) show a persistent decrease post-GFC, primarily reflecting reduced mortgage insurance recoveries. The divergent patterns across stages, particularly evident during periods of economic stress, underscore the importance of stage-specific modelling approaches rather than treating recovery as a single aggregate measure.

Driver Analysis and Cross-Stage Relationships

Table 36. *Correlation table between available drivers in Freddie Mac dataset and recovery rates*

Variable Definition	Abbreviation	R	R1	R2	R3
Combined loan to value (LTV) ratio at origination	CLTV	0.22	-0.03	-0.04	0.48
LTV change from origination to default	DLTVCR	-0.3	-0.12	-0.21	-0.18
LTV change from default to liquidation of	LLTVCR	-0.26	-0.06	-0.26	0.03
Debt-to-income (DTI) ratio at origination	DTIO	0.04	-0.01	0.02	0.03
DTI change from origination to default	DICR	0.09	0.04	0.05	0.08
DTI change from default to liquidation of collateral	LICR	0.06	-0.06	0.06	-0.04
FICO score	FICO	-0.01	-0.15	0.05	-0.12
Liquidity constraints	LC	0.01	0.02	0.02	-0.04
Log of original loan balance	LOB	0.19	-0.04	0.27	-0.17
Months on book	MOB	-0.01	-0.09	0.04	-0.07
Percentage of mortgage insurance (MI)	MIP	0.3	0.01	-0.09	0.69
Time in delinquency	TID	0.05	0.24	0.02	0.07
Time to resolution	TTR	-0.14	0.36	-0.19	0.12
Mortgage interest rate	INT	-0.09	0.03	-0.17	0.16
Unemployment rate	UMP	-0.24	-0.05	-0.18	-0.11

This table reports the Pearson correlation between R, R1, R2, R3 and nominal drivers constructed from Freddie Mac, the Federal Housing Finance Agency (FHFA), and the Federal Reserve state-level DTI data. A cell is highlighted in green when the correlation is positive and red when negative. The colour is darker when the absolute value of correlation is closer to 1.

Table 36 presents the Pearson correlations between recovery rates and key drivers, revealing distinct patterns across recovery stages. The aggregate recovery rate (R) shows strongest correlations with loan-to-value metrics (CLTV, DLTVCR, LLTVCR), mortgage insurance coverage (MIP), and unemployment rate (UMP). However, decomposing into stages unveils more nuanced relationships. Stage 1 recoveries (R1) are primarily driven by timing factors, showing strong positive correlations with time in delinquency (TID, 0.24) and time to resolution (TTR, 0.36). Stage 2 recoveries (R2) demonstrate stronger relationships with collateral-related variables, particularly LTV changes (DLTVCR, -0.21; LLTVCR, -0.26) and original loan balance (LOB, 0.27). Stage 3 recoveries (R3) show the strongest correlation with mortgage insurance coverage (MIP, 0.69) and original LTV (CLTV, 0.48), reflecting the nature of mortgage insurance requirements for high-LTV loans. These distinct correlation patterns provide statistical support for our stage-specific modelling approach, suggesting that different economic and operational factors drive recoveries at each stage of the collection process.

Relationships between the covariates and recovery components

In this supplementary section, we discuss the relationship between selected explanatory variables and the components of recoveries. We base on the estimation output obtained from benchmark model 1 (i.e., the one-step OLS regression) and the benchmark model 2 (i.e., the two-step selection model of Do et al., 2020). Other competing models show consistent economic relationship.

Continuous independent variables are standardised for easier comparison. Each non-categorical/binary independent variable is subtracted from its own mean and the result is divided by its variance. This way, the magnitude of coefficients may be used to compare the influence of each independent variable on recovery rate.

References

Adelino, M., Gerardi, K., & Willen, P. S. (2013). Why don't lenders renegotiate more home mortgages? Redefaults, self-cures and securitization. *Journal of Monetary Economics*, 60(7), 835–853. https://doi.org/10.1016/j.jmoneco.2013.08.002

Agarwal, S., Amromin, G., Ben-David, I., Chomsisengphet, S., & Evanoff, D. D. (2011). The role of securitization in mortgage renegotiation. *Journal of Financial Economics*, 102(3), 559–578. https://doi.org/10.1016/j.jfineco.2011.07.005

Altman, E. I. (2003). Default recovery rates and LGD in credit risk modelling and practice: An updated review of the literature and empirical evidence. In *Advances in Credit Risk Modelling and Corporate Bankruptcy Prediction* (pp. 175–206). Cambridge University Press. https://doi.org/10.1017/CBO9780511754197.008

Altman, E. I., Brady, B., Resti, A., & Sironi, A. (2005). The Link between Default and Recovery Rates: Theory, Empirical Evidence, and Implications. *CFA Digest*, *36*(3), 19–21. https://doi.org/10.2469/dig.v36.n3.4214

Ambrose, B. W., Buttimer, R. J. JR., & Capone, C. A. (1997). Pricing Mortgage Default and Foreclosure Delay. *Journal of Money, Credit and Banking*, *29*(3), 314–325.

An, X., & Cordell, L. (2021). Mortgage loss severities: What keeps them so high? *Real Estate Economics*, *49*(3), 809–842. https://doi.org/10.1111/1540-6229.12334

Andersson, F., & Mayock, T. (2014). Loss severities on residential real estate debt during the Great Recession. *Journal of Banking and Finance*, *46*(1), 266–284. https://doi.org/10.1016/j.jbankfin.2014.05.010

Aroul, R. R., & Hansz, J. A. (2014). The Valuation Impact on Distressed Residential Transactions: Anatomy of a Housing Price Bubble. *Journal of Real Estate Finance and Economics*, *49*(2), 277–302. https://doi.org/10.1007/s11146-013-9425-0

Asarnow, E., & Edwards, D. (1995). Measuring Loss Defaulted Bank Loans: A 24-Yr Study. *The Journal of Commercial Lending*.

Bade, B., Rösch, D., & Scheule, H. (2011). Default and recovery risk dependencies in a simple credit risk model. *European Financial Management*, *17*(1), 120–144. https://doi.org/10.1111/j.1468-036X.2010.00582.x

Baesens, B., & Smedts, K. (2025). Boosting credit risk models. *British Accounting Review*, *57*(4). https://doi.org/10.1016/j.bar.2023.101241

Bajari, P., Chu, C. S., & Park, M. (2008). AN EMPIRICAL MODEL OF SUBPRIME MORTGAGE DEFAULT FROM 2000 TO 2007. *NATIONAL BUREAU OF ECONOMIC RESEARCH*.

Bank for International Settlements. (2006). *International convergence of capital measurement and capital standards : a revised framework, comprehensive version*. Bank for International Settlements.

Basel Committee on Banking Supervision. (2001). Potential Modifications to the Committee's Proposals. *Bank for International Settlements. BIS*, *November*, 1–4.

Basel Committee on Banking Supervision. (2021). *The Basel Framework*. https://www.bis.org/basel_framework/index.htm?export=pdf&pdfid=15549727552898657

Bastos, J. A. (2010). Forecasting bank loans loss-given-default. *Journal of Banking and Finance*, *34*(10), 2510–2517. https://doi.org/10.1016/j.jbankfin.2010.04.011

Bellotti, A., Brigo, D., Gambetti, P., & Vrins, F. D. (2019). Forecasting Recovery Rates on Non-Performing Loans with Machine Learning. *SSRN Electronic Journal*, *xxxx*. https://doi.org/10.2139/ssrn.3434412

Bellotti, T., & Crook, J. (2007). Modelling and predicting Loss Given Default for credit cards. *Credit Scoring and Credit Control XI Conference*, 18. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Modelling+and+predicting+Loss+Given+Default+for+credit+cards#0

Bellotti, T., & Crook, J. (2009). Loss Given Default models for UK retail credit cards. *Credit Research Centre University Of*, *January 2009*, 1–28. http://www.crc.man.ed.ac.uk/publications/papers/workingpapers/workingpaper09-1.pdf

Bellotti, T., & Crook, J. (2012). Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting*, *28*(1), 171–182. https://doi.org/10.1016/j.ijforecast.2010.08.005

Ben-David, I. (2019). High Leverage and Willingness to Pay: Evidence from the Residential Housing Market. *Real Estate Economics*, *47*(3), 643–684. https://doi.org/10.1111/1540-6229.12234

Bijak, K., & Thomas, L. C. (2015). Modelling LGD for unsecured retail loans using Bayesian methods. *Journal of the Operational Research Society*, *66*(2), 342–352. https://doi.org/10.1057/jors.2014.9

Biswas, A., Fout, H., & Pennington-Cross, A. (2020). Mortgage Losses under Alternative Property Disposition Approaches: Deed-in-Lieu, Short Sales, and Foreclosure Sales. *Journal of Real Estate Finance and Economics*. https://doi.org/10.1007/s11146-020-09785-2

Black, F., & Cox, J. C. (1976). Valuing Corporate Securities. *Journal of Finance*, *XXXI*(2), 351–367.

Bluhm, C., Overbeck, L., & Wagner, C. (2016). Introduction to Credit Risk Modeling. *Introduction to Credit Risk Modeling*, 1–23. https://doi.org/10.1201/9781584889939

Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency. (2011). Supervisory guidance on model risk management (SR Letter 11-7/OCC Bulletin 2011-12). https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm

Bonini, S., & Caivano, G. (2016). Estimating loss-given default through advanced credibility theory. *European Journal of Finance*, *22*(13). https://doi.org/10.1080/1351847X.2013.870918

Brady, B., Chang, P., Miu, P., Ozdemir, B., & Schwartz, D. C. (2011). Discount Rate for Workout Recovery: An Empirical Study. *SSRN Electronic Journal*, 1–45. https://doi.org/10.2139/ssrn.907073

Breeden, J. L. (2016). Incorporating lifecycle and environment in loan-level forecasts and stress tests. *European Journal of Operational Research*, *255*(2), 649–658. https://doi.org/10.1016/j.ejor.2016.06.008

Bulletin 2014-7 @ guide.freddiemac.com. (n.d.).

Calabrese, R. (2014a). Predicting bank loan recovery rates with a mixed continuous-discrete model. *Applied Stochastic Models in Business and Industry*, *30*(2), 99–114. https://doi.org/10.1002/asmb.1932

Calabrese, R. (2014b). Downturn Loss Given Default: Mixture distribution estimation. *European Journal of Operational Research*, *237*(1), 271–277. https://doi.org/10.1016/j.ejor.2014.01.043

Calabrese, R., & Zenga, M. (2010). Bank loan recovery rates: Measuring and nonparametric density estimation. *Journal of Banking and Finance*, *34*(5), 903–911. https://doi.org/10.1016/j.jbankfin.2009.10.001

Calem, P. S., & LaCour-Little, M. (2004). Risk-based capital requirements for mortgage loans. *Journal of Banking and Finance*, *28*(3), 647–672. https://doi.org/10.1016/S0378-4266(03)00039-6

Carty, L. V, & Lieberman, D. (1996). Defaulted Bank Loan and Bond Prices. *Moody's Investor Service*, *November*.

Chalupka, R., & Kopecsni, J. (2009). Modeling bank loan LGD of corporate and SME segments: A case study. *Finance a Uver - Czech Journal of Economics and Finance*, *59*(4), 360–382.

Chava, S., Stefanescu, C., & Turnbull, S. (2011). Modeling the Loss Distribution. *Management Science*, *57*(7), 1267–1287. https://doi.org/10.1287/mnsc.1110.1345

Chen, H., Ma, Y., Chen, M., Tang, Y., Wang, B., Chen, M., & Yang, X. (2009). Recovery Discrimination based on Optimized-Variables Support Vector Machine for Nonperforming Loan. *Systems Engineering - Theory & Practice*, *29*(12), 23–30. https://doi.org/10.1016/s1874-8651(10)60088-9

Chen, H. Z. (2018). A new model for bank loan loss given default by leveraging time to recovery. *Journal of Credit Risk*, *14*(3), 1–29. https://doi.org/10.21314/JCR.2017.237

Chen, J., & Zhang, J. (2011). Modeling Commercial Real Estate Loan Credit Risk: An Overview. *Moody's Analyitcs*, *May*.

Chen, R., & Wang, Z. (2013). Curve Fitting of the Corporate Recovery Rates: The Comparison of Beta Distribution Estimation and Kernel Density Estimation. *PLoS ONE*, *8*(7), 1–9. https://doi.org/10.1371/journal.pone.0068238

Chen, R., Zhou, H., Jin, C., & Zheng, W. (2019). Modeling of recovery rate for a given default by non-parametric method. *Pacific Basin Finance Journal*, *57*(June 2018), 101085. https://doi.org/10.1016/j.pacfin.2018.10.014

Chen, T. H., & Chen, C. W. (2010). Application of data mining to the spatial heterogeneity of foreclosed mortgages. *Expert Systems with Applications*, *37*(2), 993–997. https://doi.org/10.1016/j.eswa.2009.05.076

Cheng, D., & Cirillo, P. (2018). A reinforced urn process modeling of recovery rates and recovery times. *Journal of Banking and Finance*, *96*, 1–17. https://doi.org/10.1016/j.jbankfin.2018.08.014

Cheng, D., & Cirillo, P. (2019). An Urn-based nonparametric modeling of the dependence between PD and LGD with an application to mortgages. *Risks*, *7*(3). https://doi.org/10.3390/risks7030076

Cherry, S. F., Jiang, E. X., Matvos, G., Piskorski, T., & Seru, A. (2021). Government and private household debt relief during COVID-19 (NBER Working Paper No. 28357). National Bureau of Economic Research. https://doi.org/10.3386/w28357

Clauretie, T. M., & Daneshvary, N. (2011). The Optimal Choice for Lenders Facing Defaults: Short Sale, Foreclose, or REO. *Journal of Real Estate Finance and Economics*, *42*(4), 504–521. https://doi.org/10.1007/s11146-009-9201-3

Clauretie, T. M., & Herzog, T. (1990). The Effect of State Foreclosure Laws on Loan Losses: Evidence from the Mortgage. In *Journal of Money, Credit and Banking* (Vol. 22, Issue 2).

Conklin, J. N., Edward Coulson, N., Diop, M., & Mota, N. (2023). An Alternative Approach to Estimating Foreclosure and Short Sale Discounts. *Journal of Urban Economics*, *134*(February). https://doi.org/10.1016/j.jue.2023.103546

Crawford, G., & Rosenblatt E. (1995). Efficient Mortgage Default Option Exercise: Evidence from Loss Severity. *Journal of Real Estate Research*, *10*.

Crook, J., & Bellotti, T. (2010). Time varying and dynamic models for default risk in consumer loans. In *J. R. Statist. Soc. A* (Issue 2).

Cui, H., Tan, K. S., & Yang, F. (2024). Portfolio credit risk with Archimedean copulas: asymptotic analysis and efficient simulation. *Annals of Operations Research*, *332*(1–3), 55–84. https://doi.org/10.1007/s10479-022-04717-0

DeFranco, R. (2001). *Unifying models of severity on defaulted mortgage*.

Deloitte. (2016). *IFRS 9: Financial Instruments – high level summary*.

Demyanyk, Y., & Hasan, I. (2010). Financial crises and bank failures: A review of prediction methods. *Omega*, *38*(5), 315–324. https://doi.org/10.1016/j.omega.2009.09.007

Demyanyk, Y. S., Koijen, R. S. J., & Van Hemert, O. (2012). Determinants and Consequences of Mortgage Default. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1699783

Dermine, J., & de Carvalho, C. N. (2006). Bank loan losses-given-default: A case study. *Journal of Banking and Finance*, *30*(4), 1219–1243. https://doi.org/10.1016/j.jbankfin.2005.05.005

Djeundje, V. B., & Crook, J. (2019). Dynamic survival models with varying coefficients for credit risks. *European Journal of Operational Research*, *275*(1), 319–333. https://doi.org/10.1016/j.ejor.2018.11.029

Do, H. X., Rösch, D., & Scheule, H. (2018). Predicting loss severities for residential mortgage loans: A three-step selection approach. *European Journal of Operational Research*, *270*(1), 246–259. https://doi.org/10.1016/j.ejor.2018.02.057

Do, H. X., Rösch, D., & Scheule, H. (Harry). (2020). Liquidity Constraints, Home Equity and Residential Mortgage Losses. *Journal of Real Estate Finance and Economics*. https://doi.org/10.2139/ssrn.2833145

Doerner, W. M., & Leventis, A. V. (2015). Distressed Sales and the FHFA House Price Index. *Journal of Housing Research*, *24*(2), 127–146. https://doi.org/10.1080/10835547.2015.12092100

Doumpos, M., Papastamos, D., Andritsos, D., & Zopounidis, C. (2021). Developing automated valuation models for estimating property values: a comparison of global and locally weighted approaches. *Annals of Operations Research*, *306*(1–2), 415–433. https://doi.org/10.1007/s10479-020-03556-1

Duffie, D., & Singleton, K. J. (1999). Modeling Term Structures of Defaultable Bonds. In *The Review of Financial Studies* (Vol. 12, Issue 4). Ramaswamy and Sundaresan.

Düllmann, Klaus., & Trapp, Monika. (2004). *Systematic risk in recovery rates an empirical analysis of US corporate credit exposures*. Dt. Bundesbank.

Economic Research, F. R. B. of S. L. (2021). *Real Estate Loans: Residential Real Estate Loans, All Commercial Banks*.

Elul, R., Souleles, N. S., Chomsisengphet, S., & Glennon, D., Hunt, R., 2010. (2010). What "Triggers" Mortgage Default? *American Economic Review*, *10*. www.philadelphiafed.org/research-and-data/publications/working-papers/.

Fabozzi, F. J., Recchioni, M. C., & Renò, R. (2025). Fifty years at the interface between financial modeling and operations research. In *European Journal of Operational Research*. Elsevier B.V. https://doi.org/10.1016/j.ejor.2025.01.001

Feldman, D., & Gross, S. (2005). Mortgage default: Classification trees analysis. *Journal of Real Estate Finance and Economics*, *30*(4), 369–396. https://doi.org/10.1007/s11146-005-7013-7

fhfa-announces-fannie-mae-and-freddie-mac-conforming-loan-limits-for-2014 @ www.fhfa.gov. (n.d.).

Foote, C. L., Gerardi, K., & Willen, P. S. (2008). Negative equity and foreclosure: Theory and evidence. *Journal of Urban Economics*, *64*(2), 234–245. https://doi.org/10.1016/j.jue.2008.07.006

Francesca, G. (2012). A Discrete-Time Hazard Model for Loans : Some Evidence from Italian Banking System. *American Journal of Applied Sciences*, *9*(9), 1337–1346.

Freddie Mac. (2019). *Single Family Loan-Level Dataset General User Guide December*. *December*.

Frontczak, R., & Rostek, S. (2015). Modeling loss given default with stochastic collateral. *Economic Modelling*, *44*, 162–170. https://doi.org/10.1016/j.econmod.2014.10.006

FSI Connect. (2018). *IFRS 9 and expected loss provisioning - Executive Summary* (Issue July 2014).

Gabriel, S., Iacoviello, M., & Lutz, C. (2020). A Crisis of Missed Opportunities? Foreclosure Costs and Mortgage Modification During the Great Recession. *Finance and Economics Discussion Series*, *2020.0*(53). https://doi.org/10.17016/feds.2020.053

Geske, R. (1977). The Valuation of Corporate Liabilities as Compound Options. *The Journal of Financial and Quantitative Analysis*, *12*(4), 541–552.

Ghent, A. C., & Kudlyak, M. (2011). Recourse and residential mortgage default: Evidence from US states. *The Review of Financial Studies*, 24(9), 3139–3186. https://doi.org/10.1093/rfs/hhr055

Goodman, L., & Zhu, J. (2015). Loss Severity on Residential Mortgages. *The Journal of Fixed Income*, *February*. https://doi.org/10.3905/jfi.2015.25.2.048

Green, R. K. (2013). Introduction to Mortgages & Mortgage Backed Securities. In *Introduction to Mortgages & Mortgage Backed Securities*. https://doi.org/10.1016/C2012-0-00208-6

Greve, C., & Hahnenstein, L. (2016). Benchmarking the LGD Parameter for Mortgage Loan Portfolios Under Stress. *Journal of Credit Risk*, *12*(4), 79–107. https://doi.org/10.2139/ssrn.2578002

Guégan, L., Rebreanu, M., Clifford, A., & Kengla, T. (2018). *IFRS 9 Expected Credit Loss*.

Gupton, G. M., Gates, D., & Carty, L. V. (2000). Bank-Loan Loss Given Default. *Moody's*.

Gupton, G. M., & Stein, R. M. (2005). LossCalc V2: Dynamic Prediction of LGD. *Moodys Investors Service*, *January*, 1–44.

Gupton, G. M., Stein, R. M., & Bren, D. (2002). Losscalc Tm : Model for Predicting Loss Given Default (Lgd) Modeling Methodology. *Moody's*, *February*, 1–32. https://doi.org/10.1111/j.0391-5026.2005.00149.x

Hamilton, D. T., Carty, L. V, & Heckman, S. (1999). Debt recoveries for corporate bankruptcies. *Moody's Investor Service*, *June*.

Hao, C., & Ala, M. M. (2010). Review of the literature on credit risk modeling: Development of the past 10 years. *Banks and Bank Systems*, *5*(3), 43–60.

Harrison, I., & Mathew, C. (2008). A structural approach to the understanding and measurement of residential mortgage lending risk. *Reserve Bank of New Zealand*.

Higgins, E., Yavas, A., & Zhu, S. (2022). Private mortgage securitization and loss given default. *Real Estate Economics*, *50*(5), 1334–1359. https://doi.org/10.1111/1540-6229.12376

Hlawatsch, S., & Ostrowski, S. (2016). Simulation and estimation of loss given default. *Credit Risk*, *7*(3), 39–73.

Höcht, S., & Zagst, R. (2014). Loan Recovery Determinants -- A Pan-European Study. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2443724

Hoskin, K., & Irvine, S. (2009). Quality of bank capital in New Zealand. *Reserve Bank of New Zealand: Bulletin*, *72*(3).

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

Hurt, L. (1998). *Measuring Loss on Latin American Defaulted Bank Loans: A 27-Year Study of 27 Countries Acknowledgment: This study would not have been possible without the cooperation, assistance, and contribution of many people throughout Latin America and Citibank's Eme*.

Hwang, R. C., Chung, H., & Chu, C. K. (2016). A Two-Stage Probit Model for Predicting Recovery Rates. *Journal of Financial Services Research*, *50*(3). https://doi.org/10.1007/s10693-015-0231-0

Jankowitsch, R., Nagler, F., & Subrahmanyam, M. G. (2014). The determinants of recovery rates in the US corporate bond market. *Journal of Financial Economics*, *114*(1), 155–177. https://doi.org/10.1016/j.jfineco.2014.06.001

Jarrow, A., Turnbull, R. A. ., & Stuart, M. . (1995). American Finance Association Pricing Derivatives on Financial Securities Subject to Credit Risk. *The Journal of Finance*, *50*(1), 53–85.

Jiang, E. X., & Zhang, A. L. (2025). Collateral value uncertainty and mortgage credit provision $. *Journal of Financial Economics*, *169*(March), 104054. https://doi.org/10.1016/j.jfineco.2025.104054

Karwański, M., Gostkowski, M., & Jałowiecki, P. (2015). Loss given default modeling: An application to data from a Polish bank. *Journal of Risk Model Validation*, *9*(3), 23–40. https://doi.org/10.21314/JRMV.2015.139

Khieu, H. D., Mullineaux, D. J., & Yi, H. C. (2012). The determinants of bank loan recovery rates. *Journal of Banking and Finance*, *36*(4), 923–933. https://doi.org/10.1016/j.jbankfin.2011.10.005

KPMG. (2014). *First Impressions: IFRS 9 Financial Instruments*.

Krüger, S., & Rösch, D. (2017). Downturn LGD modeling using quantile regression. *Journal of Banking and Finance*, *79*, 42–56. https://doi.org/10.1016/j.jbankfin.2017.03.001

Lekkas, V., Quigley, J. M., & Van Order, R. (1993). Loan Loss Severity and Optimal Mortgage Default. *Journal of the American Real Estate and Urban Economics Association*, *21*(4), 353–371.

Leow, M., & Mues, C. (2012). Predicting loss given default (LGD) for residential mortgage loans: A two-stage model and empirical evidence for UK bank data. *International Journal of Forecasting*, *28*(1), 183–195. https://doi.org/10.1016/j.ijforecast.2011.01.010

Leow, M., Mues, C., & Thomas, L. (2014). The economy and loss given default: Evidence from two UK retail lending data sets. *Journal of the Operational Research Society*, *65*(3), 363–375. https://doi.org/10.1057/jors.2013.120

Piskorski, T., Seru, A., & Vig, V. (2010). Securitization and distressed loan renegotiation: Evidence from the subprime mortgage crisis. *Journal of Financial Economics*, 97(3), 369–397. https://doi.org/10.1016/j.jfineco.2010.04.003

Li, X., Qi, M., & Zhao, X. (2012). The Extent of Strategic Defaults Induced by Mortgage Modification Programs. *Office of the Comptroller of the Currency*.

Lin, Z., Rosenblatt, E., & Yao, V. W. (2009). Spillover effects of foreclosures on neighborhood property values. *Journal of Real Estate Finance and Economics*, *38*(4), 387–407. https://doi.org/10.1007/s11146-007-9093-z

Longstaff, F. A. ., & Schwartz, E. S. . (1995). A Simple Approach to Valuing Risky Fixed and Floating Rate Debt. *The Journal of Finance*, *50*(3).

Loterman, G., Brown, I., Martens, D., Mues, C., & Baesens, B. (2012). Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting*, *28*(1), 161–170. https://doi.org/10.1016/j.ijforecast.2011.01.006

Marton, J., & Runesson, E. (2017). The predictive ability of loan loss provisions in banks – Effects of accounting standards, enforcement and incentives. *British Accounting Review*, *49*(2), 162–180. https://doi.org/10.1016/j.bar.2016.09.003

Mei, Y., Boyle, P., & Li, J. S. H. (2019). Improving Risk Sharing and Borrower Incentives in Mortgage Design. *North American Actuarial Journal*, *23*(4), 485–511. https://doi.org/10.1080/10920277.2019.1634594

Merton, R. C. (1974). On the Pricing of Corporate Debt: the Risk Structure of Interest Rates. *The Journal of Finance*, *29*(2), 449–470. https://doi.org/10.1111/j.1540-6261.1974.tb03058.x

Miu, P., & Ozdemir, B. (2017). Adapting the Basel II advanced internalratings- based models for international financial reporting standard 9. *Journal of Credit Risk*, *13*(2), 53–83. https://doi.org/10.21314/JCR.2017.224

Nazemi, A., & Fabozzi, F. J. (2018). Macroeconomic variable selection for creditor recovery rates. *Journal of Banking and Finance*, *89*. https://doi.org/10.1016/j.jbankfin.2018.01.006

Nazemi, A., Fatemi Pour, F., Heidenreich, K., & Fabozzi, F. J. (2017). Fuzzy decision fusion approach for loss-given-default modeling. *European Journal of Operational Research*, *262*(2). https://doi.org/10.1016/j.ejor.2017.04.008

Nazemi, A., Heidenreich, K., & Fabozzi, F. J. (2018). Improving corporate bond recovery rate prediction using multi-factor support vector regressions. *European Journal of Operational Research*, *271*(2). https://doi.org/10.1016/j.ejor.2018.05.024

Ozdemir, B., & Huang, E. (2021). A prudent loss given default estimation for mortgages. II. *Journal of Risk Model Validation*, *15*(4), 1–27.

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

Park, Y. W., & Bang, D. W. (2014). Loss given default of residential mortgages in a low LTV regime: Role of foreclosure auction process and housing market cycles. *Journal of Banking and Finance*, *39*(1), 192–210. https://doi.org/10.1016/j.jbankfin.2013.11.015

Pelizza, M., & Schenk-Hoppé, K. R. (2019). Pricing defaulted Italian mortgages. *Journal of Risk and Financial Management*. https://ssrn.com/abstract=3491178

Pennington-Cross, A. (2003). Subprime and Prime Mortgages: Loss Distributions. In *Office of Federal Housing Enterprise Oversight*.

Pennington-Cross, A. (2010). The duration of foreclosures in the subprime mortgage market: A competing risks model with mixing. *Journal of Real Estate Finance and Economics*, *40*(2), 109–129. https://doi.org/10.1007/s11146-008-9124-4

PWC. (2016). *IFRS 9, Financial Instruments Understanding the basics*.

Qi, M. (2013). Credit Portfolio Risk Measurement. In D. Rosch & H. Scheule (Eds.), *Credit Securitizations and Derivatives: Challenges for the Global markets* (pp. 35–52). John Wiley & Sons Ltd.

Qi, M., & Yang, X. (2009). Loss given default of high loan-to-value residential mortgages. *Journal of Banking and Finance*, *33*(5), 788–799. https://doi.org/10.1016/j.jbankfin.2008.09.010

Qi, M., & Zhao, X. (2011). Comparison of modeling methods for Loss Given Default. *Journal of Banking and Finance*, *35*(11), 2842–2855. https://doi.org/10.1016/j.jbankfin.2011.03.011

Qian, X., Cai, H. H., Innab, N., Wang, D., Ciano, T., & Ahmadian, A. (2024). A novel deep learning approach to enhance creditworthiness evaluation and ethical lending practices in the economy. *Annals of Operations Research*, *346*(2), 1597–1619. https://doi.org/10.1007/s10479-024-05849-1

Rapisarda, G., & Echeverry, D. (2016). A nonparametric approach to incorporating incomplete workouts into loss given default estimates. *The Journal of Credit Risk*, *9*(2). https://doi.org/10.21314/jcr.2013.159

Renault, O., & Scaillet, O. (2004). On the way to recovery: A nonparametric bias free estimation of recovery rate densities. *Journal of Banking and Finance*, *28*(12), 2915–2931. https://doi.org/10.1016/j.jbankfin.2003.10.018

Rodriguez-Serrano, J. A. (2024). Prototype-based learning for real estate valuation: a machine learning model that explains prices. *Annals of Operations Research*, *344*(1), 287–311. https://doi.org/10.1007/s10479-024-06273-1

Roldán, J. M., & Saurina, J. (2012). Old and New Lessons of the Financial Crisis for Risk Management. In *Banks at Risk: Global Best Practices in an Age of Turbulence* (pp. 84–101).

Rösch, D., & Scheule, H. (2014). Forecasting probabilities of default and loss rates given default in the presence of selection. *Journal of the Operational Research Society*, *65*(3), 393–407. https://doi.org/10.1057/jors.2012.82

Saurina, J. (2009). Loan loss provision in Spain. a working macroprudential tool. *BANCO DE ESPAÑA*.

Scheule, H., & Jortzik, S. (2020). Benchmarking LGD Discount Rates. *Journal of Risk Model Validation*, *August*.

Schuermann, T. (2004). What do We Know about Loss Given Default? In *Wharton Financial Institutions Center Working Paper*. https://doi.org/10.2139/ssrn.525702

Seidler, J., & Jakubík, P. (2009). The Merton Approach to Estimating Loss Given Default: Application to the Czech Republic. *Czech National Bank*.

Siao, J. S., Hwang, R. C., & Chu, C. K. (2016). Predicting recovery rates using logistic quantile regression with bounded outcomes. *Quantitative Finance*, *16*(5), 777–792. https://doi.org/10.1080/14697688.2015.1059952

Siddiqi, N. A., & Zhang, M. (2004). *A General Methodology for Modeling Loss Given Default*.

Sigrist, F., & Stahel, W. A. (2010). Using The Censored Gamma Distribution for Modeling Fractional Response Variables with an Application to Loss Given Default. *ASTIN Bulletin*. https://doi.org/10.2143/AST.41.2.2136992

Somers, M., & Whittaker, J. (2007). Quantile regression for modelling distributions of profit and loss. *European Journal of Operational Research*, *183*(3), 1477–1487. https://doi.org/10.1016/j.ejor.2006.08.063

Sopitpongstorn, N., Silvapulle, P., & Gao, J. (2017). Local Logit Regression for Recovery Rate. *SSRN Electronic Journal*, *November*. https://doi.org/10.2139/ssrn.3053774

S&P Global. (2017). *IFRS 9 Implementation Top Five Concerns*. Credit Analysis. https://www.spglobal.com/marketintelligence/en/news-insights/blog/ifrs-9-implementation-top-five-concerns

Starosta, W. (2021). Loss given default decomposition using mixture distributions of in-default events. *European Journal of Operational Research*, *292*(3), 1187–1199. https://doi.org/10.1016/j.ejor.2020.11.034

Stokes, J., & Gloy, B. (2007). Mortgage delinquency migration: An application ofmaximum entropy econometrics. *Journal of Real Estate Portfolio Management*, *13*(2), 153–160.

Tanoue, Y., Kawada, A., & Yamashita, S. (2017). Forecasting loss given default of bank loans with multi-stage model. *International Journal of Forecasting*, *33*(2), 513–522. https://doi.org/10.1016/j.ijforecast.2016.11.005

Thomas, L. C., Matuszyk, A., & Moore, A. (2012). Comparing debt characteristics and LGD models for different collections policies. *International Journal of Forecasting*, *28*(1), 196–203. https://doi.org/10.1016/j.ijforecast.2010.11.004

Thomas, Lyn C., Matuszyk, A., So, M. C., Mues, C., & Moore, A. (2016). Modelling repayment patterns in the collections process for unsecured consumer debt: A case study. *European Journal of Operational Research*, *249*(2). https://doi.org/10.1016/j.ejor.2015.09.013

Thomas, Lyn C, Mues, C., & Matuszyk, A. (2010). Modelling LGD for unsecured personal loans: Decision tree approach. *Journal of the Operational Research Society*.

Tobback, E., Martens, D., Van Gestel, T., & Baesens, B. (2014). Forecasting Loss Given Default models: Impact of account characteristics and the macroeconomic state. *Journal of the Operational Research Society*, *65*(3), 376–392. https://doi.org/10.1057/jors.2013.158

Tomarchio, S. D., & Punzo, A. (2019). Modelling the loss given default distribution via a family of zero-and-one inflated mixture models. *Journal of the Royal Statistical Society, Series A*, *182*(4), 1247–1266. https://doi.org/10.1111/rssa.12466

Tong, E. N. C. (2015). *Mixture Models for Consumer Credit Risk*.

Tong, E. N. C., Mues, C., & Thomas, L. (2013). A zero-adjusted gamma model for mortgage loan loss given default. *International Journal of Forecasting*, *29*(4), 548–562. https://doi.org/10.1016/j.ijforecast.2013.03.003

United States Foreclosure Laws. (2020). http://www.foreclosurelaw.org/

Witzany, J., Rychnovský, M., & Charamza, P. (2012). Survival Analysis in LGD Modeling. *European Financial and Accounting Journal*, *7*(1), 6–27.

Wood, R. M., & Powell, D. (2017). Addressing probationary period within a competing risks survival model for retail mortgage loss given default. *Journal of Credit Risk*, *13*(3), 47–66. https://doi.org/10.21314/JCR.2017.228

XRB. (2014). *New Zealand Equivalent to International Financial Reporting Standard 9 Financial Instruments (NZ IFRS 9)*.

Yao, X., Crook, J., & Andreeva, G. (2015). Support vector regression for loss given default modelling. *European Journal of Operational Research*, *240*(2). https://doi.org/10.1016/j.ejor.2014.06.043

Yao, X., Crook, J., & Andreeva, G. (2017). Enhancing two-stage modelling methodology for loss given default with support vector machines. *European Journal of Operational Research*, *263*(2), 679–689. https://doi.org/10.1016/j.ejor.2017.05.017

Yashkir, O., & Yashkir, Y. (2013). Loss Given Default Modelling: Comparative Analysis. *Journal of Risk Model Validation*, *7*(1). http://mpra.ub.uni-muenchen.de/46147/

Zhang, Y., Ji, L., & Liu, F. (2010). Local Housing Market Cycle and Loss Given Default: Evidence from Sub-Prime Residential Mortgages. *IMF Working Paper*, *10*(167).

Zhang, Z., Gao, G., & Shi, Y. (2014). Credit risk evaluation using multi-criteria optimization classifier with kernel, fuzzification and penalty factors. *European Journal of Operational Research*, *237*(1), 335–348. https://doi.org/10.1016/j.ejor.2014.01.044