Erasmus University Rotterdam
Erasmus School of Economics
Master Thesis Quantitative Finance

Predicting Future Recoveries of Defaulted Loans: a Random Survival Forest Approach

Claire van der Wal (497103)

Supervisor:	Marina Khismatullina	Supervisors Deloitte:	Arwyn Goos,
Second assessor:	Chen Zhou		Benjamin Chroneer
Date final version:	4th April 2024		

Abstract

Lenders are required to estimate the Loss Given Default (LGD) accurately to adhere to the Basel Accord. An accurate LGD estimation can help in determining a more precise capital buffer size, which in turn can absorb losses and free up capital for investments. Traditional LGD forecasting techniques, however, can only use resolved cases in the modelling process, while resolved and unresolved cases may exhibit different recovery behaviour. Therefore, these techniques may result in biased estimates. We propose a machine learning-based Survival Analysis model; the Random Survival Forest, and compare this model to the traditional Regression-Based model and the semi-parametric Survival Analysis model; the Cox Proportional Hazards. The main advantage of these Survival Analysis models is that they can handle unresolved cases. We predict the final and twelve-monthly LGDs on a set of American mortgages from Freddie Mac. The results show that before calibration, the Random Survival Forest was the only model that could capture the bimodal distribution of the LGD. Furthermore, it had high discriminatory power, but low calibration power. After calibrating the model via a binning method, we found that the calibration power improved, while the discriminatory power remained the same. The Calibrated Random Survival Forest model outperformed the (Calibrated) Regression-Based model and (Calibrated) Cox Proportional Hazards model based on the Loss Capture Ratio and Mean Squared Error and also outperformed all models based on the Mean Absolute Error and Loss Shortfall for the twelve-monthly predictions. Specifically, when one prioritizes high discriminatory power and the reduction of large errors in LGD prediction, or when one wants to perform short-term LGD predictions, the Calibrated RSF model is an appropriate model.

Contents

1 Introduction

Following the Basel Accord, lenders are required to hold a minimum amount of capital based on their estimated exposure to credit, market and operational risk (European Banking Authority, 2018). Credit risk is based on the Probability of Default (PD), Exposure at Default (EAD) and Loss Given Default (LGD). The latter is the loss for the lender, for example, a bank, corresponding to a default, that is, a non-performing loan. Defaulted loans can either be resolved or unresolved. A resolved default is a defaulted loan for which the workout period, i.e. the duration during which the lender attempts to recover as much value of the defaulted loan as possible, is over and no more recoveries by the client are being made. In other words, the recovery cash flow is determined. These resolved defaults are also known as closed cases. Unresolved defaults, however, are defaulted loans for which the client is still in the process of repaying his or her loan, i.e. further recoveries are still expected. In this case, it is unclear how large the final LGD will be.

It is important to be able to estimate the LGD accurately to adhere to the Basel Accord. An accurate LGD estimation can help determine a more precise capital buffer size. A capital buffer is required such that lenders are able to absorb losses while still being able to operate. Setting the minimum amount of capital too low can result in bankruptcy for the lender. On the other hand, setting the minimum amount too high, implies that the lenders cannot use the surplus to, for example, invest. In this case, it is an opportunity cost to hold more capital than needed. Another reason for accurate LGD estimation is that the lender can roughly estimate its total expected loss and avoid bankruptcy by setting the terms and interest rates for future credit transactions.

This research will try to correctly predict the expected future recovery cash flows of defaulted loans. We will focus on finding a new method to predict the LGD more accurately. The main problem with current techniques is that only resolved cases can be used in the modelling process (Zhang & Thomas, 2012). However, resolved and unresolved cases may exhibit different recovery behaviour, and thus, these techniques may result in biased estimates. Furthermore, current techniques use linear regression, which means that only a linear relationship between the dependent and independent variables can be modelled. Current LGD forecasting techniques include the Direct Approach and the Component-Based Approach (Hurlin et al., 2018; Miller & Töws, 2018; Zhang & Thomas, 2012). The Direct Approach is modelled with a linear regression, in an LGD context this model is also known as the Regression-Based model. The Component-Based Approach splits the LGD estimation into two parts. The most common method splits it based on the default outcome; cure or non-cure, i.e. loans that returned to the performing state or loans that were not able to return to the performing state, respectively. Loss given cure and loss given non-cure are modelled via linear regression, and then, a probability of cure, modelled via logistic regression, is used to combine the two models into a final LGD prediction. This paper will try to apply new methods that can use both resolved and unresolved cases, and, that do not assume a specific relationship between the dependent and independent variables to obtain more accurate LGD predictions.

We will use Survival Analysis (SA) to predict the expected future recovery cash flows of defaulted loans. SA is a common technique to forecast the time until a certain event happens, in our case, a recovery cash flow, and has the major advantage that it can handle unresolved cases. Particularly, the Cox Proportional Hazards (Cox PH) model and the Random Survival Forest (RSF) model will be used. The first is a semi-parametric SA model and the latter is a technique that combines machine learning, namely, Random Forest, with SA. These two models will be compared to the traditional Regression-Based model. We will investigate which method performs best in terms of LGD prediction.

The Freddie Mac Loan Level data set will be used for this research (Freddie Mac, 2023a). It contains loan characteristics and performances of American mortgages that were sold to Freddie Mac, an American company that operates in the US secondary mortgage market. They buy loans from lenders and pool them into securities which they again sell to investors around the world.

Past research papers that address SA for time-to-event modelling find that SA models combined with machine learning can outperform traditional statistic methods such as the Regression-Based model and the Cox PH model. However, these machine learning-based SA models have only been applied to making predictions in medical and PD cases, i.e. binary events. This paper contributes to the current literature as we investigate whether these enhanced SA models, specifically the RSF, also outperform the traditional Regression-Based and Cox PH method in predicting expected future recovery cash flows, i.e. non-binary events, of defaulted loans.

The results following this research indicate that before calibrating the RSF model, the RSF was the only model that could capture the bimodality of the LGD. Furthermore, it had the best Loss Capture Ratio (LCR) and Loss Shortfall (LS) implying that it was able to differentiate between the severity of losses better and could capture the total loss better. The latter is important when lenders need to set their capital buffer. However, the Cox PH model slightly outperformed the Regression-Based model and RSF model based on the Mean Absolute Error (MAE) and Mean Squared Error (MSE), implying that the Cox PH model was able to make more accurate LGD predictions. We also found that all three models were able to differentiate well between resolved and unresolved cases. Furthermore, the Regression-Based model obtained better values for the MAE, MSE and LS when we predicted the LGD at twelve-month intervals. However, the twelve-monthly predictions were more cumbersome to calculate for the Regression-Based model. Lastly, we calibrated the three models via a binning method. We found that the Calibrated RSF model outperformed the (Calibrated) Regression-Based model and (Calibrated) Cox PH model based on the LCR and MSE. It also outperformed all models based on all four measures for the twelve-monthly predictions. Specifically, when one prioritizes high discriminatory power and the reduction of large errors in LGD prediction, or when one wants to perform short-term LGD predictions, the Calibrated RSF model is an appropriate model.

The remainder of this paper is structured as follows. In Section 2, we discuss past research

papers regarding LGD models and (machine learning-based) SA models. Section 3 contains a description of the data set used in this research. We discuss the methodology in Section 4, with the corresponding results presented in Section 5. Lastly, a conclusion and future research ideas are given in Section 6.

2 Literature Review

The following section gives a brief overview of several relevant papers on the prediction of future recovery cash flows. First, current LGD modelling techniques are discussed. Subsequently, some alternative methods are examined. Finally, the contribution of this research to the literature is explained.

Currently, there are two general LGD forecasting techniques, the Direct Approach and the Component-Based Approach (Hurlin et al., 2018; Miller & Töws, 2018; Zhang & Thomas, 2012). The Direct Approach models the LGD with a linear regression, in an LGD context this model is also known as the Regression-Based model. The main advantage of this parametric model is that it is very interpretable. The Component-Based Approach splits the LGD estimation into two parts. The most common method splits it based on the default outcome; cure or non-cure, i.e. loans that returned to the performing state or loans that were not able to return to the performing state, respectively. Loss given cure and loss given non-cure are modelled via linear regression, and then, a probability of cure, modelled via logistic regression, is used to combine the two models into a final LGD prediction. These techniques, however, have a few disadvantages. First, linear regression has several assumptions, including that the error term follows a normal distribution with mean zero. This assumption is, however, often violated as the distribution of the LGD tends to be bimodal. Furthermore, only resolved cases can be used in the modelling process of these two approaches. Resolved and unresolved cases may, however, follow different recovery patterns, and thus, these techniques may result in biased estimates. Nevertheless, due to their simplicity, both approaches are still commonly used in the prediction of future recovery cash flows. The Regression-Based model is easier to interpret, and is, therefore, appropriate to use as a benchmark model in this paper.

A good alternative to the Regression-Based model is time-to-event modelling which is commonly used in medical research (George et al., 2014). These models capture more information than whether an event has occurred or not. An example of time-to-event modelling is Survival Analysis (SA). In medical research, SA is used to, for example, compare the risk of death or recovery from a disease between or among population groups receiving different medications or treatments (Liu, 2012). The results provide information on which medication or treatment performs better. An advantage of these SA models is that they can handle censored observations, that is, observations that have not experienced the event before the end of the study yet (right-censored) and observations that had already experienced the event at the start of the study (left-censored) (George et al., 2014). Methods such as linear regression ignore these censored observations. A common SA technique is the Cox Proportional Hazards (Cox PH) model which is expressed by the hazard function, which estimates the rate at which events occur at a certain

time. It is a semi-parametric model and can examine the effect of several predictor variables on the time-to-event. In this research, we have that the event is a recovery cash flow, thus, unlike in the medical field, we have a non-binary event. This requires an adjustment in the modelling procedure, which we will need to apply.

Recent literature has shown that SA models also have become more popular in credit risk modelling. Witzany et al. (2012), Zhang and Thomas (2012) and Prívara et al. (2013) investigated how SA models, such as the Cox PH model, performed in modelling the recovery process of defaulted loans. They regarded the target variable as how much has been recovered before the end of the collection period. Witzany et al. (2012) and Prívara et al. (2013) found that the Cox PH model outperformed the traditional regression method. An advantage of SA models is that they can handle unresolved cases, that is, cases for which clients are still in the process of repaying their loans. Furthermore, no distributional assumptions are required for the Cox PH model, which is beneficial as the size of the recovery rate does not follow a normal distribution (Miller & Töws, 2018). For these reasons, the Cox PH model will be used as our second model to predict future recovery cash flows. The main difference between this research and the ones mentioned is that Witzany et al. (2012), Zhang and Thomas (2012) and Prívara et al. (2013) look at unsecured loans, i.e. loans with no collateral, whereas this paper investigates loans with collateral, this could influence the results. Loans that are not protected by collateral generally carry a higher risk for the lender (Brock, 2023). This paper will contribute to the literature by investigating whether the Cox PH model can also outperform the traditional regression method in terms of LGD prediction for secured loans.

Doan et al. (2022) enhanced the traditional SA model by applying three machine learning techniques to SA for both clinical and transcriptomic data. These three techniques included Random Forest, Gradient Boosting and Support Vector Machine. Random Forest is a machine learning method that uses bagging and feature randomness to create uncorrelated decision trees (Yingchun, 2014). Predictions are then based on the average of the individual trees' predictions. The uncorrelated decision trees result in a model with low variance, meaning the model is a good method against overfitting. Furthermore, the Random Forest has the additional advantage of being able to handle both continuous and categorical variables, however, it is memory and time-intensive (Yingchun, 2014). Gradient Boosting is another machine learning technique that uses a series of weak learners, such as decision trees, to improve the model sequentially (Friedman, 2001). Each tree minimizes a loss function and learns from the residual of its successor's prediction. The final prediction is based on the weighted average of the prediction of all the individual trees. As predictions are based on various estimations of weak learners, Gradient Boosting provides robust estimates, moreover, it can also handle both continuous and categorical variables, however, it is memory and time-intensive (Friedman, 2001). Another common machine learning model is the Support Vector Machine. The objective of this model is to classify data points by finding a separating hyperplane that maximizes the margin between groups and minimizes misclassification (Cervantes et al., 2020). This model only uses a subset of the training data making it less prone to overfitting. Furthermore, it is robust to noise in data. However,

it does not perform well in large data sets (Cervantes et al., 2020). Doan et al. (2022) show that these machine learning-based SA models, the Random Survival Forest (RSF), Gradient Boosted Survival (GBS) and Survival Support Vector Machine (SSVM), can outperform the traditional statistic methods, such as the Cox PH model. These models combine the advantages of machine learning techniques, which are well known for their ability to handle high-dimensional data, non-linear relationships and interaction effects, and the advantages of SA models, which can handle censored data. RSF computes a cumulative hazard function for each tree after which all functions are averaged to obtain the final cumulative hazard function. GBS uses the partial likelihood function of the Cox PH model as the loss function and SSVM uses an asymmetric penalty function to handle survival data. Based on the results found in Doan et al. (2022), these machine learning-based SA models seem to perform better than the traditional statistic methods in the medical field. However, these techniques have not been investigated in the prediction of future recovery cash flows yet. The main difference between Doan et al. (2022)'s research and this research, is that we have a non-binary event in the credit risk field, whereas Doan et al. (2022) investigated the prediction of a binary event in the medical field. Again, we will need to adjust the modelling procedure to handle this difference which could affect the results.

The RSF and GBS models are also investigated by Xia et al. (2021). They found that a variation of GBS outperformed other models in terms of the dynamic predictions on probability of default (PD) under out-of-sample validation. However, they also found that this variation did not significantly outperform the other methods under out-of-time validation. This variation of GBS is called the Survival Extreme Gradient Boosting (SurvXGBoost) model, which combines Extreme Gradient Boosting, an advanced version of Gradient Boosting that uses a more regularized model formalization to control overfitting, with survival models. Extreme Gradient Boosting is known to outperform Gradient Boosting in terms of accuracy and speed (Wade & Glynn, 2020). Nevertheless, Bhakta et al. (2021) mention that it does not perform as well as the Random Forest in large data sets, furthermore it is more prone to overfitting. SurvXGBoost was compared to the traditional Cox PH model, RSF, GBS and a time-varying Cox PH model. Again, this comparison has not been performed on the prediction of future recovery cash flows yet. Xia et al. (2021) investigated the prediction of a binary event, whereas we have a non-binary event. An adjustment in the modelling procedure will need to be made to manage this difference which could affect the performance results. As RSF performs better than SSVM and SurvXGBoost, which itself performs better than GBS, in large data sets, we will investigate RSF in the prediction of future recovery cash flows. This research will contribute to the literature by applying a machine learning-based SA model in the prediction of recovery cash flows.

To summarize, based on the discussed literature, this paper contributes to the current literature. To the best of our knowledge, a comparative study of different prediction techniques for future recovery cash flows has not been performed yet. The main contribution of this paper is to propose a machine learning-based SA model in the prediction of recovery cash flows. Traditional SA models have already been used in LGD prediction, and machine learning-based SA models have already been used in binary-event predictions, such as in the medical field and for PD pre-

diction. However, machine learning-based SA models have not been investigated in non-binary events, specifically LGD, prediction yet.

3 Data

In this section, the data set is analysed. First, a general description of the data and its source is given. Then, the procedure of how the data is cleaned is explained, after which the risk drivers are discussed. Next, the summary statistics are given. Lastly, the division of the data set into a training and test set is explained.

3.1 Mortgage Data

The Freddie Mac Loan Level data set will be used for this research. It contains loan characteristics and performances of American mortgages that were sold to Freddie Mac, an American company that operates in the US secondary mortgage market (Freddie Mac, 2023a). They buy loans from lenders and pool them into securities which they again sell to investors around the world. Data is available with mortgage start dates in 1999-2022. The original data set contains approximately 52.4 million mortgages. As this is computationally challenging, a subsample is also provided; 50,000 mortgages are randomly selected per year from the original set. For this research, the subsample is used with start dates in 2005-2022, this way, both the financial crisis in 2007-2008 and the Covid-19 pandemic in 2020-2023 are included. The data set includes 899,968 unique mortgages.

The data set consists of two parts; a yearly origination file and a monthly loan performance set. The first consists of 32 fields regarding loan, borrower and property characteristics at the start date. Examples include loan-to-value, credit score, debt-to-income ratio, property type and postal code. The monthly loan performance set consists of 32 fields regarding monthly performance metrics for each loan. These fields include the loan delinquency status, interest rate and monthly loan balance. A unique identification number for each loan links the two parts. A description of all the variables can be found in Freddie Mac (2023b) and a description of the important variables for this research can be found in Table A1.

3.2 Data Cleaning

As this research focuses on predicting future recoveries of defaulted loans, we only include mortgages that have defaulted (approximately 4.5% of the 899,968 unique mortgages). A mortgage is considered to be in default if it has been delinquent for 90 days or more (European Banking Athority, 2018). In the data set, a delinquency status of 0 represents a loan that has been delinquent for 0-29 days. A value of 1 corresponds to 30-59 days delinquent, etc. Therefore, all loans included in the data set have a delinquency status of at least 3 at least once. Moreover, loans already in default at the start of the observation window are removed as well as information at the time of default is unknown. The workout period then starts on the date the mortgage goes into default and ends on the date a zero balance event occurs or on the performance cutoff date. The latter is the last date for which performance data is available for any loan in the data set.

The zero balance event date is the loan termination date, i.e. the date the loan is considered cured or non-cured, which will be explained more in Subsection 3.3.

Furthermore, we exclude observations that were repurchased, modified and deferred. We provide a more in-depth elaboration on the decisions made in the process of removing these observations in Section A.2. Moreover, there are approximately 450 loans that have been terminated but without a reason given. We will assume these loans have not been cured. Lastly, loans that are terminated in the same month as they went into default, are deleted as well. After deleting these observations, we remain with 17,439 unique mortgages. An overview of the data removal is also provided in Table A2.

Table 1 summarizes the procedure for removing certain variables. We remove variables based on their correlation with other variables, missing values, whether they only have one unique value, and their definition. The full list of variables with their reason for removal, if removed, can be found in Table A1. We also provide a more in-depth explanation of this procedure in Section A.2. We end with 32 variables, of which 29 are risk drivers.

Table 1: Variable requirements

Requirement	Reason
Correlation <70%	Strong correlation affects the predictor standard error negatively
Missing Values <70%	Variables with too many missing values provide no useful information
More than One Value	Variables with only one unique value don't provide additional relevant information
Loan Characteristic	Only loan properties are of interest
Normal Loans Characteristic	Variables related to the deferred payment plan and modified loans are irrelevant

Moreover, only the first digit of the three-digit *Postal Code* is used, this digit corresponds to one of ten regions in the US. Figure 1 shows the postal code division. The ten regions will be named Northeasteast, Northeast, East, Southeast, Mideast, North, Mid, South, Midwest, and West, for first digit 0-9 respectively. Lastly, before the second quarter of 2018, the variable *Number of Borrowers* could only take two values; 1, if there was only one borrower, and 2, if there was more than one borrower. From the second quarter of 2018 onwards, the variable takes the value equivalent to the number of borrowers, that is, it can also take a value larger than 2. To handle this difference, we use the first definition and set all values larger than 2, equal to 2.

3.3 Risk Drivers

Additional variables are created for the calculation of the LGD. The LGD is given by

$$LGD_{i,T} = 1 - RR_{i,T} = 1 - \frac{1}{EAD_i} \sum_{t=1}^{T} \frac{CF_{i,t}}{(1+r_i)^{t-t_0}}, \tag{1}$$

where $RR_{i,T}$ is the recovery rate at time T and EAD_i is the Exposure at Default, equal to the outstanding amount of the mortgage at the time of default, t_0, for observation i. The cash flows of observation i at time $t > t_0$, $CF_{i,t}$, are discounted by discount rate r_i, for $t - t_0$ months after the default date. A cash flow is defined as the difference between the *Current Actual Unpaid*

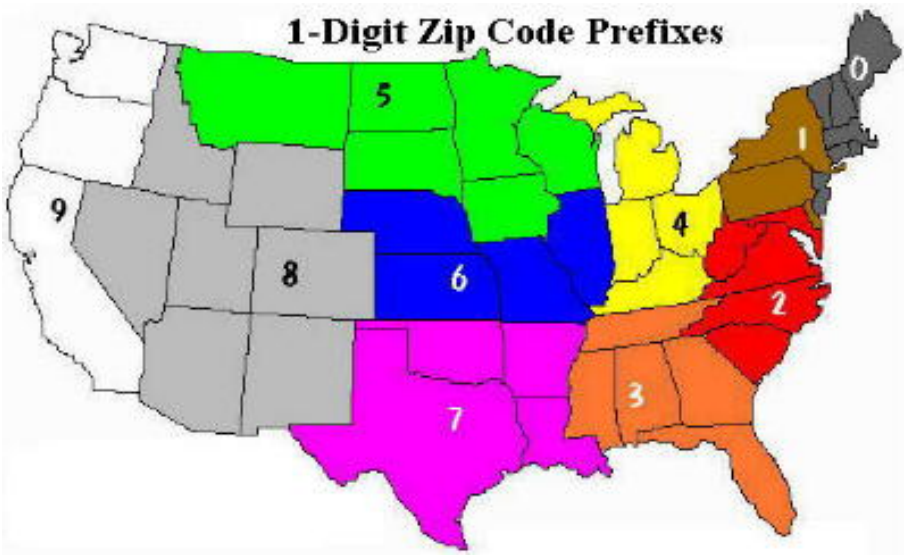


Figure 1: Postal code division

Principal Balance (UPB) of a certain month with the previous month. We have T cash flows, where T is the number of months after the default date at which the LGD is calculated, and we use the *Current Interest Rate* for the discount rate. Note that in practice the 500bps plus three-month Euro Interbank Offered Rate (3M Euribor) is used for discounting (European Banking Authority, 2017), however, for simplicity, the *Current Interest Rate* will be used. This may affect the size of the estimated losses, as higher rates result in lower present values of future cash flows, and therefore, higher loss estimates, and vice versa. Nevertheless, as the same discount rate is used consistently over the whole data set, using the *Current Interest Rate* should not impact the internal model performances, however, one may need to adjust the rates when the models are performed on other data sets. Furthermore, the name *Current Interest Rate* suggests the variable is time-varying. However, we find that the variable is loan specific, and does not vary with time.

Moreover, the LGD must be adjusted for loans that have been cured (European Banking Authority, 2017). For these loans, an artificial cash flow is added to the calculation at the time of termination. This artificial cash flow is discounted in the same way as a normal cash flow, however, this results in non-zero LGDs for cured losses. Nevertheless, this way reflects the economic loss, which we are interested in, rather than the accounting loss, obtained by setting the LGD to zero for cured cases. The *Zero Balance Code* indicates whether the loan is cured or not, Table 2 gives an overview of which Zero Balance Event belongs to a cured loan. When a defaulted loan is cured, it means that the client was able to recover the loan, non-cured otherwise.

There are several categorical variables in the data set. If a categorical variable has M different values, $M - 1$ dummy variables are created to avoid the "dummy variable trap" (Hirschberg & Lye, 2001). If all M values were constructed as dummy variables, perfect multicollinearity, and thus, inaccurate results, could occur. The following values of the categorical variables are excluded: *First Time Homebuyer Flag - No*, *Occupancy Status - Primary Residence*, *Property*

Table 2: Overview of zero balance events

	Zero Balance Event	Definition
Cured	Prepaid or Matured (Voluntary Payoff)	The outstanding amount is fully paid
	Reperforming Loan Securitizations	The loan is sold once it has been cured
Non-Cured	Third Party Sale	The loan is sold to a third party
	Short Sale or Charge Off	The loan is sold at a lower price
	Real Estate Owned Disposition	The loan is sold to the lender
	Whole Loan Sale	The loan is sold in the secondary market

Type - Single-Family, *Loan Purpose - Purchase*, *Prepayment Penalty Mortgage (PPM) Flag - No* and *Postal Code - Northeasteast*. Note that the choice of which value is to be removed is arbitrary.

Lastly, all numeric variables are normalized via the min-max scaling technique (de Amorim et al., 2023). Let X_{ij} be the jth variable of loan i, the normalized variable then becomes

$$X_{ij,scaled} = \frac{X_{ij} - X_{j,min}}{X_{j,max} - X_{j,min}}, \tag{2}$$

with $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, P\}$, where N and P are the number of loans and explanatory variables, respectively. The minimum and maximum value over all loans i of variable j is denoted by $X_{j,min}$ and $X_{j,max}$, respectively. Normalizing variables ensures each feature is of equal importance. By the min-max scaling, each variable is scaled to a value between 0 and 1, this allows for a better comparison of the coefficients. When features have different scales, the feature with a larger range of values can influence the results more, although it is not necessarily more important as a predictor (Vidiyala, 2020).

3.4 Summary Statistics

Table 3 contains the summary statistics of the remaining non-categorical risk drivers before normalizing the variables. The credit score is an indication of the borrower's likelihood to repay loans in a timely manner. Scores generally range from 300 to 850, a higher score means the likelihood is higher. In the data set, borrowers score quite high on their credit score, the average credit score is 707.43. The average mortgage insurance percentage is 5.89%, however, at least 75% of the mortgages have no mortgage insurance. Mortgage insurances are intended to protect the lender, higher mortgage insurance percentages lead to lower LGD since the recoveries are higher. Thus, it is expected that a higher mortgage insurance percentage is negatively related to LGD. Similarly, at least 50% of the mortgage notes are to be repaid by one borrower, and nearly all mortgages are for one-unit properties. It is expected that the number of borrowers is also negatively related to the LGD as the mortgage can be repaid by more people, this hypothesis is also supported in Section 5. The number of units, on the other hand, is likely to be positively related to the LGD. Multi-unit properties are more complex to handle and come with higher operational costs (Trion Properties, 2023).

In Table 4, the summary statistics of categorical risk drivers are given. Most borrowers (86.41%)

Table 3: Summary statistics[1] of non-categorical risk drivers

Risk Driver	Mean	Std	Min	p25	p50	p75	Max
Exposure at Default	183,067.61	109,744.92	35.26	99,990.16	159,635.53	245,671.94	916,535.07
Credit Score	707.43	58.00	300	667	708	752	839
MI[2] Percentage (%)	5.89	11.11	0.00	0.00	0.00	0.00	42.00
Number of Units	1.04	0.25	1	1	1	1	4
Original CLTV[3]	80.12	23.14	8	70	80	90	529
Original DTI[4] Ratio (%)	38.50	10.70	2	32	39	45	65
Current Interest Rate (%)	5.59	1.06	1.88	4.75	5.88	6.38	8.75
Original Loan Term	338.36	56.74	96	360	360	360	480
Months in Default[5]	26.22	30.73	0	6	14	33	205
Number of Borrowers	1.37	0.48	1	1	1	2	2

[1] Statistics are of variables before normalizing the data
[2] MI = Mortgage Insurance
[3] CLTV = Combined Loan-to-Value
[4] DTI = Debt-to-Income
[5] Statistics of Months in Default includes unresolved cases

are not a first-time homebuyer. It is expected that first-time homebuyers have more difficulty repaying their mortgage, and therefore, have a higher LGD. As the majority of borrowers are not a first-time homebuyer, this could mean the average LGD is slightly lower than if there were more first-time homebuyers. However, in Section 5, we find that there is a negative relation between the two. An explanation of this negative relation will be provided in Section 5. Furthermore, nearly all properties (88.04%) are occupied by the owner of the mortgage, and most properties (71.70%) are single-family homes. The loan purpose types are fairly evenly distributed, and very few loans (0.10%) are prepayment penalty mortgages. Lastly, the least number of properties (4.29%) are located in the North of America, and the most (17.73%) are located in the Southeast. Highly imbalanced categorical variables may affect the estimation performance (Moerbeek & van Schie, 2016). We have some risk drivers with very few observations belonging to one outcome, therefore, the outcome may contribute minimal information for prediction.

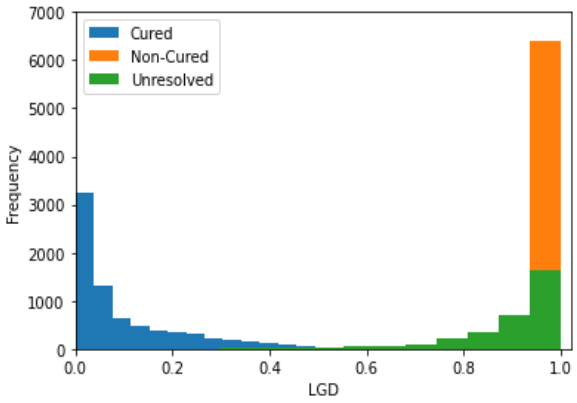


Figure 2: Histogram of the loss given default of cured, non-cured and unresolved loans

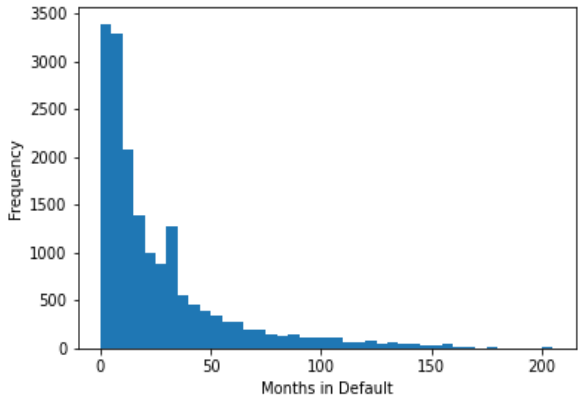


Figure 3: Histogram of the number of months in default

As this paper focuses on LGD, we also look at the LGD distribution of the data set and its development over time. In Figure 2, the histogram of the LGD of cured, non-cured and unresolved loans is given. Here, it is seen that the LGD is bimodally distributed. Most loans experience

Table 4: Summary statistics of categorical risk drivers

Risk Driver	Categories	N	%
First Time Homebuyer Flag	No	15,069	86.41
	Yes	2,370	13.59
Occupancy Status	Primary Residence (P)	15,354	88.04
	Investment Property (I)	1,444	8.28
	Second Home (S)	641	3.68
Property Type	Single-Family (SF)	12,504	71.70
	PUD (PU)	3,084	17.69
	Condo (CO)	1,647	9.44
	Manufactured Housing (MH)	167	0.96
	Co-op (CP)	37	0.21
Loan Purpose	Purchase (P)	6,358	36.46
	Refinance - No Cash Out (N)	5,592	32.07
	Refinance - Cash Out (C)	5,489	31.47
PPM[1] Flag	No	17,421	99.90
	Yes	18	0.10
Postal Code	Southeast (3)	3,092	17.73
	West (9)	2,879	16.51
	Midwest (8)	1,757	10.07
	Mideast (4)	1,754	10.06
	East (2)	1,646	9.44
	Northeasteast (0)	1,452	8.33
	Mid (6)	1,403	8.05
	South (7)	1,359	7.79
	Northeast (1)	1,348	7.73
	North (5)	749	4.29

[1] PPM = Prepayment Penalty Mortgage

no loss or a complete loss, this may influence the performance of some models when predicting the LGD. Figure 3 shows the histogram of the number of months in default of all loans. The majority of the loans are less than 3 years in default. Interesting to see is that there is a small peak at approximately 30 months in default. This means fewer loans are slightly shorter and slightly longer than approximately 30 months in default. From Table 5, we see that the average number of months in default is slightly more than 26 months, with an average LGD of 59%. However, if we only look at cases that are closed, the average is slightly more than 23 months, with an average LGD of 51%. Approximately 54% of the resolved loans are cured. We also see that the average LGD is approximately 1 for non-cured cases, that is, if a loan is considered non-cured the owner was hardly able to repay the loan. From Figure 2 and Table 5, we also see that the highest LGD for cured cases is 57%, though only very few loans experience an LGD of more than 40% if the loan was considered cured. Similarly, the smallest LGD for non-cured cases is 5%, though only very few non-cured loans experienced an LGD of less than approximately 100%. The smallest LGD for unresolved cases is 4%, though only very few unresolved cases have an LGD of less than 60%.

Lastly, in Figure 4, the cumulative LGD is plotted against the number of months in default. We

see that the LGD goes towards 59% when we include both resolved and unresolved cases, this is the average LGD of all mortgages, as was also seen in Table 5. Note, however, that the exact number (59.63%) is slightly higher than the exact average LGD (58.60%) observed in Table 5. This is because the calculation in the cumulative LGD is based on the sum of cash flows per month in default divided by the total exposure at default. It is, therefore, a weighted LGD, whereas the average LGD in Table 5 is unweighted. From Figure 4, we can also clearly see that resolved and unresolved loans follow a different LGD pattern. However, this is mainly caused by the last cash flow of a loan. Often the largest cash flow is observed at the end of the workout period. Excluding these cash flows results in a similar LGD pattern as the unresolved cases. Nevertheless, to avoid biased predictions this means we must use both resolved and unresolved loans in the modelling process.

Table 5: Summary statistics[1] of resolved and unresolved cases

			Months in Default			Loss Given Default		
		N	Mean	Min	Max	Mean	Min	Max
Resolved	Cured + Non-Cured	14,111	23.46	1	203	0.51	0.00	1.00
	Cured	7,597	27.89	1	203	0.10	0.00	0.57
	Non-Cured	6,514	18.29	1	145	1.00	0.05	1.00
Unresolved		3,328	37.95	0	205	0.90	0.04	1.00
All		17,439	26.22	0	205	0.59	0.00	1.00

[1] Statistics are of variables before normalizing the data

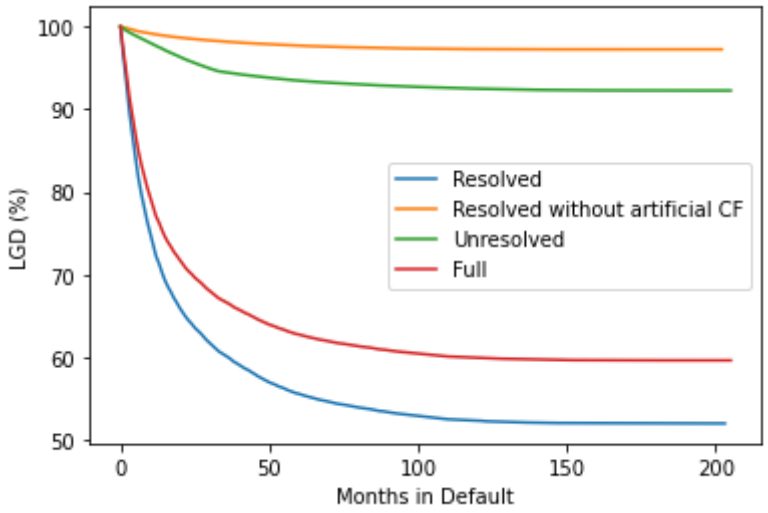


Figure 4: Cumulative loss given default against the number of months in default

3.5 Training and Test Set

To test the performance of the models, the data set is split into a training and test set. For this research, 75% of the loans with their monthly performances are put into the training set, and the other 25% is put into the test set for performance validation. Loans are divided randomly. The models are estimated on the in-sample set, i.e. the training set. Validation is then performed

on the out-of-sample set, in this case, the test set, to measure the consistency of performance. For the Random Survival Forest, we split the data another time. Trees are created by randomly selecting features from the training set (the in-bag data), and then, with the remaining data (approximately 37% (Qu et al., 2020)) of the training set (the out-of-bag data), each tree is validated. The test data set is again used as an overall performance measure of the model.

4 Methodology

This research will evaluate the semi-parametric Cox PH model and the machine learning-based RSF model against the traditional Regression-Based model. The latter will be considered the benchmark model. In the next section, first, the three different models are explained. Then, the variable selection method will be discussed. Lastly, an overview of the different performance measures is given.

4.1 Regression-Based Model

The Regression-Based method models the LGD with a linear regression (Hurlin et al., 2018; Miller & Töws, 2018; Zhang & Thomas, 2012). It assumes that the error term follows a normal distribution with mean zero. This assumption is often, however, violated as the distribution of the LGD tends to be bimodal (Figure 2). This model, therefore, could result in weak predictive performances. Furthermore, only resolved cases can be used in the modelling process. However, resolved and unresolved cases may exhibit different recovery behaviour (Figure 4), and thus, this technique may result in biased estimates. Nevertheless, this approach is still commonly used in the prediction of future recovery cash flows, and is, thus, appropriate to use as a benchmark model. It uses Ordinary Least Squares (OLS) to model the LGD of loan $i \in \{1, \ldots, N\}$, where N is the number of loans, on the explanatory variables in a linear combination,

$$LGD_i = \beta_0^{RB} + \sum_{p=1}^{P} \beta_p^{RB} X_{ip} + \epsilon_i, \tag{3}$$

where β_0^{RB} is the regression's constant, β_p^{RB} the slope coefficient of explanatory variable X_{ip}, P the number of explanatory variables and ϵ_i the residual. Let $\beta^{RB} = (\beta_0^{RB}, \beta_1^{RB}, \ldots, \beta_P^{RB})$, the parameters are estimated by minimizing the residual sum of squares (RSS):

$$\hat{\beta}^{RB} = \underset{\beta^{RB}}{\operatorname{argmin}}\, RSS = \underset{\beta^{RB}}{\operatorname{argmin}} \sum_{i=1}^{N} \left(LGD_i - \beta_0^{RB} - \sum_{p=1}^{P} \beta_p^{RB} X_{ip} \right)^2. \tag{4}$$

4.2 Survival Analysis

Survival Analysis is a popular technique for time-to-event modelling. In this research, we have that the event, also known as a failure, is a recovery cash flow. SA models are able to handle right- and left-censored observations, for this research, however, only right-censored observations in the data set will be considered. Right-censored observations, in the context of this research, are clients who are in default but whose workout period has not ended yet, that is, there might

still be recovery cash flows in the future. This section provides the basic principle of SA and also introduces two types of modified SA models, namely the Cox Proportional Hazards and Random Survival Forest model.

Let T be the non-negative random variable that represents the time the event happens. It has probability density function $f(t)$ and cumulative distribution function $F(t)$. The survival function, $S(t) = 1 - F(t)$, is the cumulative probability of the subject not having encountered the event until time $t \geq 0$, with $t = 0$ the start time. In our case, we have that the survival function represents the cumulative probability of not receiving a cash flow by the end of the study period, and t is the number of months the loan has been in default, with $t = 0$ the time the loan defaulted. Then, the hazard function, which is defined as

$$h(t) = \lim_{\tau \to 0} \frac{\mathbb{P}(t \leq T \leq t + \tau | T > t)}{\tau} = \frac{f(t)}{S(t)}, \tag{5}$$

is the rate at which the events happen exactly at time t given survival, that is, the event has not happened yet, until t. The second equality holds by definition of the conditional density function and the derivative of a function. The cumulative hazard rate is derived as follows

$$H(t) = \int_0^t h(s)\, ds. \tag{6}$$

Lastly, by the chain rule,

$$-\frac{\partial}{\partial t} \log S(t) = -\frac{\frac{\partial}{\partial t} S(t)}{S(t)} = -\frac{\frac{\partial}{\partial t}[1 - F(t)]}{S(t)} = -\frac{-f(t)}{S(t)} = h(t), \tag{7}$$

therefore, the survival function takes the following form

$$S(t) = \exp\left[-\int_0^t h(s)\, ds\right] = \exp[-H(t)] = \mathbb{P}(T > t). \tag{8}$$

The above is the general concept of SA. The following two subsections will explain two methods for the implementation.

4.2.1 Cox Proportional Hazards Model

The Cox Proportional Hazards model can examine how specific factors influence the rate of a particular event happening at a certain point in time. The hazard function takes the form

$$h(t|X_i) = h_0(t) \exp(X_i' \beta^{CPH}), \tag{9}$$

where $h_0(t)$ is a non-parametric baseline hazard, and $\exp(X_i' \beta^{CPH})$ denotes a parametric relative risk function, with $X_i = (X_{i1}, \ldots, X_{iP})'$ a vector of P explanatory variables of loan $i \in \{1, \ldots, N\}$, with corresponding coefficients $\beta^{CPH} = (\beta_1^{CPH}, \ldots, \beta_P^{CPH})'$. The Cox PH model is, thus, a semi-parametric model. The cumulative baseline hazard and baseline survival

function are, respectively, defined as

$$H_0(t) = \int_0^t h_0(s)\, ds; \tag{10}$$

$$S_0(t) = \exp[-H_0(t)]. \tag{11}$$

Consider two individuals i and j, $i \neq j$, the ratio between their hazard rates is

$$\frac{h(t|X_i)}{h(t|X_j)} = \frac{h_0(t)\exp(X_i'\beta^{CPH})}{h_0(t)\exp(X_j'\beta^{CPH})} = \frac{\exp(X_i'\beta^{CPH})}{\exp(X_j'\beta^{CPH})} = \kappa, \tag{12}$$

where $\kappa > 0$ is a constant independent of time. In other words, we assume that an individual is always a certain factor riskier to experience the event compared to another individual at every point in time. This assumption is tested with the Schoenfeld Residuals statistic (Winnett & Sasieni, 2001). Violation of this assumption could lead to weak predictive performances.

Cox PH models are estimated via the partial likelihood function (Hosmer et al., 2008). This method depends only on the parameters of interest and the estimation of the baseline function is disregarded. Including the baseline function will result in difficulties when using the log-likelihood function. The partial likelihood, nevertheless, obtains the same distributional properties of the parameter estimates as the full likelihood. The proof can be found in Slud (1982).

Let $R(t_i)$ denote the set of defaulted loans at time t_i, where t_i is the time that observation $i \in \{1, \ldots, N\}$ experiences the event. Furthermore, let δ_i indicate whether observation i exited from default by curing or writing off. If $\delta_i = 1$, observation i is uncensored at time t_i, and $\delta_i = 0$ otherwise. In the latter case, the survival time is longer than the observed event time of the loan; the loan is right-censored. The partial likelihood, is, therefore, denoted by

$$\begin{aligned}
\mathcal{L}(\beta^{CPH}) &= \prod_{i=1}^{N} \left(\frac{h(t_i|X_i)}{\sum_{l \in R(t_i)} h(t_i|X_l)} \right)^{\delta_i} \\
&= \prod_{i=1}^{N} \left(\frac{h_0(t_i)\exp(X_i'\beta^{CPH})}{\sum_{l \in R(t_i)} h_0(t_i)\exp(X_l'\beta^{CPH})} \right)^{\delta_i} \\
&= \prod_{i=1}^{N} \left(\frac{\exp(X_i'\beta^{CPH})}{\sum_{l \in R(t_i)} \exp(X_l'\beta^{CPH})} \right)^{\delta_i}.
\end{aligned} \tag{13}$$

Here, we assume that all observations are independent. Although loans are generally independent of each other, the loss rates might be correlated due to, for example, bad market conditions to sell collateral. This could result in weak predictive performances. The log-likelihood then becomes

$$\log \mathcal{L}(\beta^{CPH}) = \sum_{i=1}^{N} \delta_i \left[X_i'\beta^{CPH} - \log \left(\sum_{l \in R(t_i)} \exp(X_l'\beta^{CPH}) \right) \right], \tag{14}$$

which is maximized with respect to β^{CPH}.

The above formulation is based on the assumption that events do not happen at the exact same time. Survival times are continuous random variables, so to a certain extent this assumption is valid, however, due to discrete observation moments, cash flows can be received at the same time. Breslow (1974) and Efron (1977) derived an approximation of the above model to solve this problem.

When we obtain the estimates of β^{CPH}, $\hat{\beta}^{CPH}$, we can estimate the cumulative baseline hazard rate. Peng and Dear (2000) proposed the following estimate

$$\hat{H}_0(t) = \sum_{t_i \leq t} \frac{d_i}{\sum_{l \in R(t_i)} \exp(X_l' \hat{\beta}^{CPH})}, \tag{15}$$

where d_i is the number of events at time t_i. From equations (8), (9), (10) and (15), the estimated survival function can be computed by

$$\begin{aligned}
\hat{S}(t) &\overset{(8),(9)}{=} \exp\left[- \int_0^t \hat{h}_0(s) \exp(X_i' \hat{\beta}^{CPH}) \, ds \right] \\
&= \exp\left[- \exp(X_i' \hat{\beta}^{CPH}) \int_0^t \hat{h}_0(s) \, ds \right] \\
&\overset{(10)}{=} \exp\left[- \exp(X_i' \hat{\beta}^{CPH}) \hat{H}_0(t) \right] \\
&\overset{(15)}{=} \exp\left[- \sum_{t_i \leq t} \frac{d_i}{\sum_{l \in R(t_i)} \exp(X_l' \hat{\beta}^{CPH})} \exp\left(X_i' \hat{\beta}^{CPH} \right) \right].
\end{aligned} \tag{16}$$

Furthermore, with the relation $S(t) = 1 - F(t)$, the estimated probability of receiving a cash flow, $\hat{\mathbb{P}}(T \leq t) = \hat{F}(t)$, is denoted by

$$\hat{\mathbb{P}}(T \leq t) = \hat{F}(t) = 1 - \hat{S}(t) = 1 - \exp\left[-\hat{H}_0(t) \exp\left(X_i' \hat{\beta}^{CPH} \right) \right]. \tag{17}$$

4.2.2 Random Survival Forest Model

Random Survival Forest is an extension of the Random Forest method that includes SA. It uses bagging and feature randomness to create uncorrelated decision trees (Yingchun, 2014). Bagging, or bootstrap aggregation, is the random selection of data from the training set with replacement, that is, data points can be selected more than once. Feature randomness, or feature bagging, is the random selection of features. Predictions are then based on the average of the individual trees' predictions. RSF extends this by creating survival trees by splitting nodes based on the log-rank splitting rule that maximizes the survival difference between children nodes (Doan et al., 2022). A cumulative hazard function (CHF) is then computed for each tree, after which all functions are averaged to obtain the final CHF. The main advantage of RSF over Cox PH is that RSF does not assume proportional hazard rates (Nasejje & Mwambi, 2017). Furthermore, it is non-parametric, that is, it does not assume a specific form of the hazard function. However, this does imply that the coefficients of the covariate effects are less

interpretable.

Survival trees are binary trees grown by recursive splitting of tree nodes (Ishwaran et al., 2008). A node is split based on the survival difference between its two daughter nodes. The survival difference is the difference in the survival probability curves of observations. The best split is the split for which the largest survival difference is obtained, found by searching over all $k \leq P$ randomly selected predictor variables of loan $i \in \{1, \ldots, N\}$ and split values c. This way, the differences between trees are maximized, while differences within trees are minimized. A survival tree reaches its saturation point when no new daughters can be formed because of the criterion that each node must contain a minimum number of events. Let $\mathcal{P}$ denote the set of the P predictor variables, and $\mathcal{K}$ the set of the k randomly selected predictor variables, $\mathcal{K} \subseteq \mathcal{P}$. In the remainder of this subsection, let X_i denote the k predictor variables in subset $\mathcal{K}$ of loan i and X_{ip} the pth covariate of X_i. The log-rank statistic of X_{ip}, with splitting value c is given by

$$L(X_{ip}, c) = \left| \frac{\sum_{i=1}^{N} (d_{i,1} - \frac{d_i}{Y_i} Y_{i,1})}{\sqrt{\sum_{i=1}^{N} \frac{Y_{i,1}}{Y_i}(1 - \frac{Y_{i,1}}{Y_i})(\frac{Y_i - d_i}{Y_i - 1}) d_i}} \right|, \tag{18}$$

where $d_{i,j}$ is the number of events in daughter node $j = 1, 2$ at time t_i, $d_i = d_{i,1} + d_{i,2}$. Similarly, $Y_{i,j}$ is the number of individuals who are alive in daughter node $j = 1, 2$ at time t_i, $Y_i = Y_{i,1} + Y_{i,2}$. The split value c determines whether an observation X_{ip} goes to the left ($X_{ip} \leq c$) or the right ($X_{ip} > c$) daughter node.

Let h denote a terminal node of a saturated tree, and $t_{1,h} < t_{2,h} < \cdots < t_{m(h),h}$ the distinct event times within node h. Furthermore, let $d^*_{l,h}$ and $Y^*_{l,h}$ denote the number of cash flows and individuals at risk at time $t_{l,h}$, respectively (Xia et al., 2021). The CHF and survival function for terminal node h are estimated using the Nelson-Aalen and Kaplan-Meier estimators, respectively,

$$H^*_h(t) = \sum_{t_{l,h} \leq t} \frac{d^*_{l,h}}{Y^*_{l,h}}; \tag{19}$$

$$S^*_h(t) = \prod_{t_{l,h} \leq t} \left(1 - \frac{d^*_{l,h}}{Y^*_{l,h}}\right). \tag{20}$$

Equation (19) implies that for a given tree, the hazard estimate for node h is the ratio of events to individuals at risk, summed across all unique event times. Given a new sample i with features X_i, X_i will be assigned to a unique terminal node h due to the binary nature of a tree. The CHF and survival function for X_i are given by the Nelson-Aalen and Kaplan-Meier estimator for X_i's terminal node:

$$H^*(t|X_i) = H^*_h(t),\ i \in \{1, \ldots, N\}; \tag{21}$$

$$S^*(t|X_i) = S^*_h(t),\ i \in \{1, \ldots, N\}. \tag{22}$$

The in-bag ensemble CHF and survival function are derived by dropping the sample through all B trees and averaging the hazard and survival function from each tree. Let $H^*_b(t|X_i)$ and $S^*_b(t|X_i)$ denote the CHF and survival function of the bth survival tree, $b \in \{1, \ldots, B\}$ of loan

i. The ensemble estimators are computed as

$$\bar{H}_e^*(t|X_i) = \frac{1}{B}\sum_{b=1}^{B} H_b^*(t|X_i),\ i \in \{1,\ldots,N\}; \tag{23}$$

$$\bar{S}_e^*(t|X_i) = \frac{1}{B}\sum_{b=1}^{B} S_b^*(t|X_i),\ i \in \{1,\ldots,N\}. \tag{24}$$

The out-of-bag (OOB) estimates for case i are given by

$$H^{OOB}(t|X_i) = H_h^*(t),\ \text{if } X_i \in h \text{ and case } i \in \{1,\ldots,N\} \text{ is OOB}; \tag{25}$$

$$S^{OOB}(t|X_i) = S_h^*(t),\ \text{if } X_i \in h \text{ and case } i \in \{1,\ldots,N\} \text{ is OOB}. \tag{26}$$

Let O_i denote the trees where case i is OOB. The ensemble CHF and survival function for individual i are then given by

$$\bar{H}_e^{OOB}(t|X_i) = \frac{1}{|O_i|}\sum_{b\in O_i} H_b^*(t|X_i),\ \text{if } i \in \{1,\ldots,N\} \text{ is OOB}; \tag{27}$$

$$\bar{S}_e^{OOB}(t|X_i) = \frac{1}{|O_i|}\sum_{b\in O_i} S_b^*(t|X_i),\ \text{if } i \in \{1,\ldots,N\} \text{ is OOB}. \tag{28}$$

Note that the Nelson-Aalen and Kaplan-Meier estimators do not use risk drivers to fit the model. However, the risk drivers are used in growing the trees. Furthermore, as RSF makes use of Random Forest, interpreting the results is not very straightforward. Generally, the Variable Importance (VIMP) measure is used to make the model more explainable. The VIMP measures how much the predictive accuracy score decreases when a variable is removed. For RSF the predictive accuracy score is given by the Concordance index, which measures the rank correlation between correctly ordered (concordant) pairs to comparable pairs (Pölsterl, 2023b). Samples i and j are comparable and concordant if the sample with lower observed time (let's assume sample i) experienced an event and the corresponding survival risk of i is higher than j. The RSF algorithm can be found in Algorithm 1.

4.2.3 Censoring

SA works slightly differently in the LGD context compared to the context of, for example, the medical field. In the latter, we have that the event, for example, is that the client died, i.e. we have a binary event, and no more data is observed. As you can only die once, each client encounters the event only once. However, in LGD modelling, the event is that a cash flow is received, i.e. we have a non-binary event, and clients can experience the event multiple times. To handle this, weights are assigned to the observations such that the sum of weights for each client equals one (Witzany et al., 2012). The weight at time t is defined as $d_{i,t} = \frac{CF_{i,t}}{EAD_i}$, with $CF_{i,t}$ and EAD_i the cash flow and EAD, respectively, corresponding to observation i. If the sum of weights for a loan equals one it is comparable to the event that a client dies in medical research.

Algorithm 1 Random Survival Forest Algorithm

Input: Let B be the number of trees in the forest and k the number of predictors used for splitting each node.

Output: A Random Survival Forest

1: Draw B independent bootstrap samples from the training data. Use the in-bag data (on average 63% per bootstrap sample, a formal proof can be found in Qu et al. (2020)) to grow the tree and for prediction, and use the out-of-bag data for cross-validation.

2: Grow a survival tree for each bootstrap sample $b \in \{1, \ldots, B\}$:

- At each tree node, randomly select k candidate predictor variables.
- Compute the log-rank statistic, $L(X_{ip}, c)$ (equation (18)), for each candidate variable and splitting value c.
- Split the node based on the log-rank splitting rule that maximizes the survival difference between children nodes.
- Grow the tree to full size under the constraint that the number of event observations (cash flows) in each node is greater than a predefined minimum terminal node size.

3: Compute the CHF for each tree. Then, average all functions to estimate the ensemble CHF (equations (19), (21), (23)).

4: Use out-of-bag data to calculate the prediction error for the ensemble CHF (equation (27)).

However, if a case is censored in a credit modelling context, it can still have an outstanding amount. For SA this means that the weight is smaller than one. To make it equal to one, we set the remaining weight to an artificial observation at a maximum time (Witzany et al., 2012). If the sum of weights is less than one, it practically means the client never "dies", which is not possible. When we add an artificial observation with the remaining weight this means that the client is still alive, and only "dies" at the maximum time. A payment can be censored due to two reasons (Prívara et al., 2013):

1 Loss Event: The sum of payments of a resolved case is less than the EAD. In this case, an artificial observation with weight $d_{i,t_{max}} = \frac{EAD_i - \sum_{t=1}^{t_{max}} CF_{i,t}}{EAD_i}$, which is censored at t_{max}, the maximum length of the recovery process, is included. In this research, we will set t_{max} to the maximum number of months a cured loan was in default, from Table 5, we see this is 203 months;

2 Unresolved Event: The workout period lasts longer than the study period and the sum of payments is less than the EAD. An artificial observation censored at time t_{end} with weight $d_{i,t_{end}} = \frac{EAD_i - \sum_{t=1}^{t_{end}} CF_{i,t}}{EAD_i}$ is then included, where t_{end} is the last known date of the concerning loan.

The sample weights indicate the importance of an observation, an observation with a larger weight means the observation is more important in the prediction. Furthermore, for each observation, we need to specify whether an event has occurred or not. There are three scenarios (Witzany et al., 2012):

1 Repayment Event: A cash flow at time t is received, $event_{i,t} = 1$;

2 Loss Event: The sum of payments of a resolved case is less than the EAD, if a cash flow at

t_{end} is received then $event_{i,t_{end}} = 1$ and $event_{i,t_{end}} = 0$ otherwise. Its artificial observation, as previously explained, censored at t_{max}, has $event_{i,t_{max}} = 0$;

3 Unresolved Event: The workout period lasts longer than the study period and the sum of payments is less than the EAD, if a cash flow at t_{end} is received then $event_{i,t_{end}} = 1$ and $event_{i,t_{end}} = 0$ otherwise. Its artificial observation, as previously explained, censored at time t_{end}, has $event_{i,t_{end}} = 0$,

where $event_{i,t} = 1$ means the event has occurred at time t, i.e. a cash flow is received, and $event_{i,t} = 0$ otherwise.

4.3 Variable Selection

Not all features are necessary in the estimation and prediction procedure. Some features may be insignificant, and, therefore, not provide sufficient explanatory power. Moreover, if the number of risk drivers used is large, there is a risk of overfitting the data, and the computation time can be long. To select relevant variables, the stepwise selection method will be used. This is a common method in LGD modelling (Zhang & Thomas, 2012) and is based on a selection criterion, such as the Akaike Information Criterion (AIC),

$$\text{AIC} = 2k - 2\log \mathcal{L}(\hat{\beta}), \tag{29}$$

where k is the number of estimated parameters and $\mathcal{L}(\hat{\beta})$ is the maximized value of the likelihood function for the model with parameters $\hat{\beta}$. The AIC makes a trade-off between the number of variables in the model and the log-likelihood achieved with the corresponding set of features. There are many other information criteria, however, for this research, only the AIC is considered as it is known to select the model with the most predictive power (Song, 2020).

The stepwise selection method goes as follows. Start with an empty set of risk drivers. For all variables, check if including it in the set improves the AIC, i.e. it decreases, if so, the variable that improves the AIC the most is added to the set of risk drivers. Then, for each variable in the set, test if the AIC improves if the variable is removed from the set. Remove the variable for which the AIC improves the most. Repeat these steps for all features. Lastly, for the remaining variables in the set, remove the insignificant variables based on a significance level α.

4.4 Performance Measures

There are two types of performance measures. The first is discrimination, which refers to how well the model can correctly assign rank orders. A commonly used measure of discrimination in LGD modelling is the Loss Capture Ratio (LCR) (Li et al., 2009). It explains how well the model captures the final observed loss amount. The loss capture curve (LCC) is defined as the cumulative observed loss amount captured from the highest to the lowest expected LGD. The LCR is then the ratio of the area between the model LCC and the random LCC (Area B in

Figure 5) to the area between the ideal LCC and the random LCC (Area A + B in Figure 5),

$$\text{LCR} = \frac{B}{A+B}. \tag{30}$$

The ideal model is the model that can perfectly predict LGDs, and the random model represents a hypothetical model that randomly assigns LGDs. The LCR takes values between 0 and 1, where a value close to 0 means the model can poorly estimate LGDs, and a value close to 1 means the model's LGD estimation is a good representation of the observed LGDs.

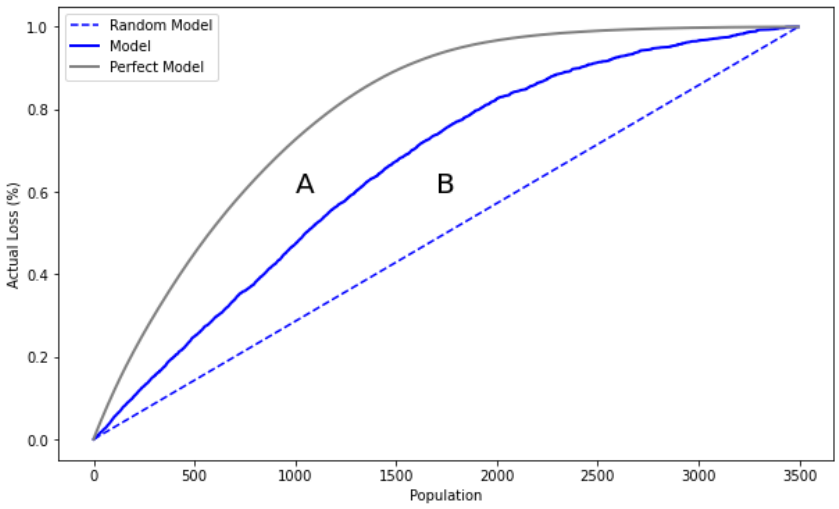


Figure 5: The loss capture curves

Performance can also be measured by calibration metrics, that is, metrics that reflect the accuracy of the LGD estimations. The Mean Absolute Error (MAE) and the Mean Squared Error (MSE) are two calibration measures commonly used in LGD modelling (Witzany et al., 2012; Zhang & Thomas, 2012; Prívara et al., 2013; Miller & Töws, 2018). The first measures the average absolute deviation between the actual and predicted LGD. The latter measures the average squares of the errors, which means that larger deviations are punished harder. For both, a smaller value corresponds to a better model. The MAE and MSE are defined, respectively, as

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |LGD_i - \hat{LGD}_i|; \tag{31}$$

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N} (LGD_i - \hat{LGD}_i)^2, \tag{32}$$

where $\hat{LGD}_i$ is the predicted final LGD of loan $i \in \{1, \ldots, N\}$.

The t-test is another common calibration method to assess the predictive ability of LGD models (European Central Bank, 2019). The t-test compares predicted LGD with actual LGD under the null hypothesis that the predicted LGD is equal to the actual LGD. The test statistic is asymptotically Student-t distributed with $N - 1$ degrees of freedom under the null hypothesis,

with N the number of observations. The t-test statistic is defined as

$$\mathrm{T} = \sqrt{N}\frac{\frac{1}{N}\sum_{i=1}^{N}(LGD_i - L\hat{G}D_i)}{\sqrt{s^2_{L\hat{G}D}}}, \tag{33}$$

where

$$s^2_{L\hat{G}D} = \frac{\sum_{i=1}^{N}\left((LGD_i - L\hat{G}D_i) - \frac{1}{N}\sum_{j=1}^{N}(LGD_j - L\hat{G}D_j)\right)^2}{N-1}, \tag{34}$$

with $L\hat{G}D_i$ again defined as before, similarly $L\hat{G}D_j$ of loan $j \in \{1, \ldots, N\}$. If the corresponding p-value is lower than the significance level, we reject the null hypothesis which states that the true LGD is equal to the predicted LGD.

Lastly, the Loss Shortfall (LS) indicates how well the model can capture the total loss (Li et al., 2009). It measures the relative error in predicting losses and is defined as

$$\mathrm{LS} = 1 - \frac{\sum_{i=1}^{N}(L\hat{G}D_i \times EAD_i)}{\sum_{i=1}^{N}(LGD_i \times EAD_i)}, \tag{35}$$

where, $L\hat{G}D_i$ again defined as before and EAD_i the EAD corresponding to loan i. A value close to zero is an indication that the model can capture the total loss of clients well.

5 Results

In the following section, first, the selected risk drivers are discussed. Then, the estimates of the Regression-Based, Cox PH and RSF model are given. Next, the models are compared using different performance measures. Lastly, we discuss an interesting finding discovered during the process of predicting LGD with SA. All coding was done in Python.

5.1 Variable Selection

The stepwise selection procedure is applied to the training set and the AIC is based on the model fitted by OLS. We use a significance level of 5%. After going through all regressors, we end with 19 risk drivers. These can be found in the tables in the following subsections. We use these variables for all three different models for best comparison. These regressors result in the best performance for the benchmark model, the Regression-Based model. Therefore, if the Cox PH and RSF models perform better than the benchmark with the same variables, we can conclude that the optimal performance of these models is at least the performance found with the benchmark optimal risk drivers.

Note that when testing the significance of the coefficients of the remaining risk drivers, multiple hypotheses are tested simultaneously (Herzog et al., 2019). This leads to an increased risk of making at least one Type I error, i.e. rejecting a true null hypothesis of the risk driver's coefficient being equal to zero. To address this problem, we use the Bonferroni Correction. We adjust the significance level of 5% by dividing it by the total number of tests conducted during

the stepwise procedure, that is, the number of remaining risk drivers after the stepwise selection procedure. Therefore, the adjusted significance level, in our case, is equal to 0.26%.

Furthermore, we find that not all dummy variables for the categorical variables are included in the optimal set of risk drivers. This means that these risk drivers do not significantly improve the model's predictive power. They may not provide any additional information. Consider the following scenarios:

1 The base category is significant:

   - Another category is significant: this category has a significantly different effect on the dependent variable compared to the base category;
   - Another category is insignificant: the effect of this category on the dependent variable does not significantly differ from the base category.

2 The base category is insignificant:

   - Another category is significant: this category has a significantly different effect on the dependent variable compared to the base category;
   - Another category is insignificant: the effect of this category on the dependent variable is not significant.

In all scenarios, changing the base category does not change the relationship between variables or the statistical significance of the model's coefficients (Gujarati & Porter, 2009). Only the interpretation of the coefficients changes as the coefficients depend on the base category.

5.2 Regression-Based Model

The Regression-Based model consists of four steps:

1 Only resolved loans are included in the data set. Each monthly performance of a loan is treated as a different loan. However, as this implies that observations are very similar and the data set will be highly correlated, only loans at month in default $t = 0, 12, 24, \ldots$ are included in the training set. The dependent variable can then be defined as the expected future recovery rate. We define the expected future recovery rate at time t' as the sum of cash flows till the loan is closed (time T), divided by the current outstanding loan amount,

$$\tilde{RR}_{i,t'} = \frac{\sum_{t=t'+1}^{T} CF_{i,t}}{COA_{i,t'}}, \tag{36}$$

   where $COA_{i,t'}$ and $CF_{i,t}$ are the outstanding loan amount at current time t' and cash flow at time t of observation i, respectively. OLS is performed and the model is fitted.

2 Only unresolved loans are included in the data set. To estimate the final LGD of the unresolved loans, for each last observation of the unresolved loans, we predict the future recoveries by multiplying the current outstanding loan amount with the predicted future recovery rate obtained from the fitted model in the previous step. The total sum of cash

flows is then equal to the sum of the predicted future recoveries and the current sum of cash flows. The final predicted LGD can be calculated as the complement of the total sum of cash flows divided by the exposure at default,

$$\tilde{LGD}_i = 1 - \frac{\hat{RR}_{i,t_{end}} \times COA_{i,t_{end}} + \sum_{t=0}^{t_{end}} CF_{i,t}}{EAD_i} = 1 - \frac{\sum_{t=0}^{T} CF_{i,t}}{EAD_i}, \tag{37}$$

where $\hat{RR}_{i,t_{end}}$ is the fitted recovery rate from Step 1 for observation i at time t_{end}, the last observation date.

3 Only the first observation for all resolved and unresolved observations is included in the data set. The data set is split into a training and test set, and the variable selection procedure is performed. Note that we exclude the risk driver *Months in Default*, as we only look at the first monthly performance of each defaulted loan. With the obtained best features, we fit the model on the training set for the final predicted LGD,

$$\tilde{LGD}_i = \begin{cases} 1 - \tilde{RR}_{i,0}, & \text{if case } i \text{ is resolved.} \\ 1 - \frac{\hat{RR}_{i,0} \times COA_{i,0}}{EAD_i}, & \text{otherwise.} \end{cases} \tag{38}$$

We exclude $CF_{i,0}$ in the above equation, as we do not observe any cash flow at time of default, $CF_{i,0} = 0$.

4 We predict the final LGD for resolved cases of the test set using the fitted model obtained in the previous step. Note that we predict based on LGD estimations of unresolved loans. In other words, we are using processed unresolved cases based on resolved cases, thus, uncertainty may be introduced. These unresolved loans are, however, necessary in the modelling process to include more recent data.

The result of the fitted model in the third step can be found in Table 6. Here, we see that at a Bonferroni-adjusted significance level of 0.26%, all except four risk drivers are significant. At a Bonferroni-adjusted significance level of 0.53% (original level of 10%), *Number of Borrowers*, *Postal Code - East (2)* and *Property Type - Co-op (CP)* are insignificant, we, therefore, cannot say anything about these risk drivers. Most significant risk drivers have an effect in the direction as expected. Notable is the direction of the *First Time Homebuyer Flag - Yes* coefficient (significant at a level of 0.53%). A first-time homebuyer is expected to have fewer resources to cover the mortgage payments and thus, has a higher LGD. However, in the USA it is common to provide aid for first-time homebuyers (Araj, 2023). Grants and no-interest loans for the down payment, tax reductions and several federal, state or local programs are available for first-time homebuyers to assist them in buying a property. This could be an explanation of the negative relation between the *First Time Homebuyer Flag - Yes* risk driver and the LGD. In Table 6, we also see that the *Original CLTV* has the largest positive effect on the expected LGD. For a one-unit increase in the *Original CLTV*, the expected LGD increases on average by approximately 1.533 percentage points, ceteris paribus. This positive effect is as expected, the *Original CLTV* is the ratio between all outstanding loans and the property's value. A higher value indicates the

borrower has a higher total loan value, or the property value is lower. This means it is harder for the borrower to repay his or her loan, and the recovery rate is lower, i.e. the LGD is higher. The largest negative effect is caused by *Postal Code - South (7)*. The expected LGD decreases on average by approximately 0.162 percentage points if the property lies in the Postal Code Area South, compared to if the property lies in Postal Code Area Northeasteast (the base category), ceteris paribus.

Table 6: Estimates of the Regression-Based model

	Coefficient	s.e.	t-value	p-value
Constant	-0.428	0.020	-21.801	0.000**
Current Interest Rate	1.101	0.021	52.948	0.000**
First Time Homebuyer Flag - Yes	-0.032	0.011	-2.989	0.003*
Loan Purpose - Refinance - Cash Out (C)	0.035	0.009	3.871	0.000**
Loan Purpose - Refinance - No Cash Out (N)	0.057	0.010	5.726	0.000**
Number of Borrowers	-0.014	0.007	-2.163	0.031
Occupancy Status - Investment Property (I)	-0.056	0.011	-4.851	0.000**
Original Combined Loan-to-Value (CLTV)	1.533	0.054	28.327	0.000**
Original Debt-to-Income (DTI) Ratio	0.113	0.011	9.876	0.000**
Original Loan Term	0.172	0.023	7.442	0.000**
Postal Code - Northeast (1)	-0.045	0.012	-3.701	0.000**
Postal Code - East (2)	-0.029	0.011	-2.638	0.008
Postal Code - Southeast (3)	0.055	0.009	6.261	0.000**
Postal Code - North (5)	-0.079	0.016	-5.017	0.000**
Postal Code - South (7)	-0.162	0.012	-13.124	0.000**
Postal Code - Midwest (8)	0.078	0.011	7.061	0.000**
Property Type - Condo (CO)	0.099	0.011	9.130	0.000**
Property Type - Co-op (CP)	-0.133	0.064	-2.074	0.038
Property Type - Manufactured Housing (MH)	-0.100	0.032	-3.130	0.002**
Exposure at Default	0.138	0.025	5.584	0.000**

** $p < 0.0026$ Bonferroni-adjusted significance level with original $\alpha = 0.05$
* $p < 0.0053$ Bonferroni-adjusted significance level with original $\alpha = 0.10$

Lastly, we plot the realized LGDs against the predicted LGDs (Figure 6a). We find that the Regression-Based model's distribution is slightly right-skewed, and has a very small peak at an LGD of approximately zero. However, overall, it is not able to capture the bimodality of the LGD.

5.3 Cox Proportional Hazards Model

The Cox PH model can estimate the effect of an individual risk driver on the hazard rate, or, the risk of a cash flow occurring. It uses all observations in the training set. First, we check the proportionality requirement with the Schoenfeld Residuals test. We find that the Cox PH model violates the proportionality assumption, i.e. the hazard ratios are not constant. This could imply that the estimated coefficients are biased. The model may not accurately capture the true relationship between covariates and the hazard ratio, therefore, the estimates need to be interpreted with caution. The effect of a risk driver on the hazard rate changes over time, making it difficult to provide a single magnitude for the entire study period, nevertheless, the

direction of the effect can still be discussed.

The estimates of the Cox PH model can be found in Table 7, here, we see that only ten risk drivers are significant at a Bonferroni-adjusted significance level of 0.26% and only eleven at a Bonferroni-adjusted significance level of 0.53%. For all other risk drivers, we cannot conclude anything about their estimates. Although we cannot say anything about the magnitude of the effects of the (significant) risk drivers, the next part will interpret the estimates as if the proportionality assumption were not violated. This way, we can indicate how the risk drivers affect the hazard rate. The largest positive effect is caused by *Postal Code - South (7)*. The risk of a cash flow is $\exp(0.433) = 1.542$ times for clients who have their property in Postal Code Area South, compared to if the property were in the Northeasteast Area (the base category), ceteris paribus. The largest negative effect is caused by *Original Combined Loan-to-Value (CLTV)*. With a one-unit increase of this variable, the expected hazard decreases by approximately 100%, ceteris paribus.

For all significant variables, we see similar effects to the LGD as compared to the Regression-Based model. When we have a positive effect in the benchmark model, we see a negative effect on the expected hazard in the Cox PH model, and vice-versa. A negative effect on the expected hazard implies the client has a smaller probability of receiving a cash flow, i.e. its LGD is higher. Accordingly, the two risk drivers with the largest absolute effect on the LGD are the same for both models and with an effect in the same direction.

Lastly, we plot the realized LGDs against the predicted LGDs (Figure 6b). We find that the Cox PH model's distribution is slightly right-skewed. Compared to the Regression-Based model, it is skewed slightly more to the right but does not have a peak at an LGD of approximately zero. Overall, it is not able to capture the bimodality of the LGD.

5.4 Random Survival Forest Model

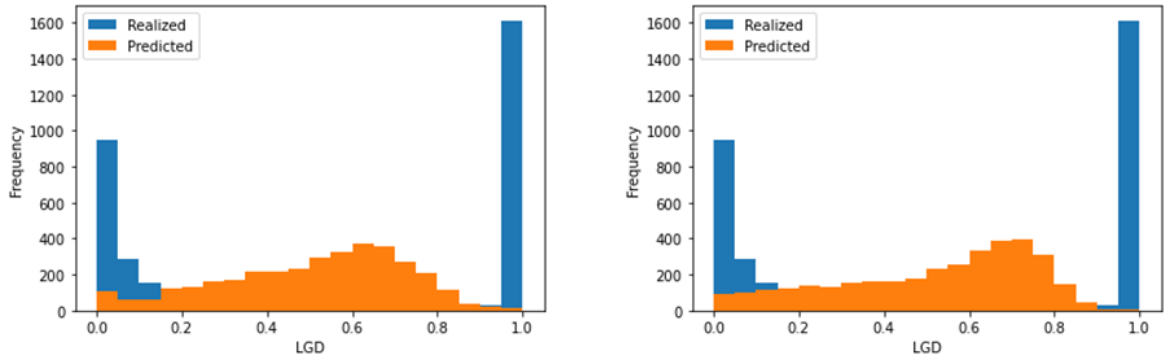
As explained in Subsection 4.2.3, SA requires the usage of sample weights. The RSF package in Python, however, is not able to incorporate these sample weights yet (Pölsterl, 2023a). To still be able to implement the RSF as accurately as possible, we propose the following method. The sample weights indicate the importance of an observation, an observation with a larger weight means the observation is more important in the prediction. We duplicate each observation in the training set *sample weight* number of times, where we convert the sample weights into numbers between 1 and 1,000, and take its integer value. We also tried using sample weights with values in the range of 1 to 100, although the computation time was shorter, we found that the model performed poorly. Here, a trade-off between accuracy and speed had to be made. A larger range for the sample weights would result in better accuracy, however, training the model would take significantly longer. With sample weights with values in the range of 1 to 1,000, the following results presented are adequate, but not optimal.

Furthermore, RSF is a machine learning-based model that requires hyperparameter tuning for

Table 7: Estimates of the Cox Proportional Hazards model

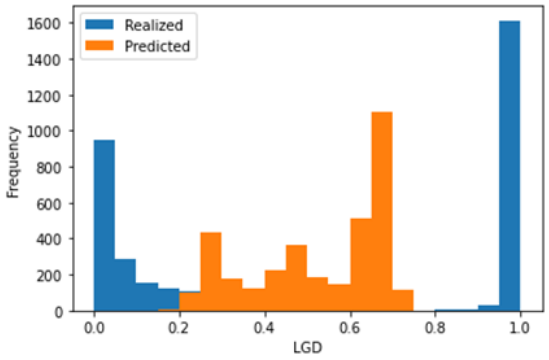
	Coefficient	exp(Coeff)	s.e.	z-score	p-value
Current Interest Rate	-3.789	0.023	0.093	-40.827	0.000**
First Time Homebuyer Flag - Yes	0.100	1.105	0.047	2.112	0.035
Loan Purpose - Refinance - Cash Out (C)	-0.250	0.779	0.039	-6.332	0.000**
Loan Purpose - Refinance - No Cash Out (N)	-0.279	0.757	0.043	-6.423	0.000**
Number of Borrowers	0.065	1.068	0.029	2.281	0.023
Occupancy Status - Investment Property (I)	0.235	1.265	0.051	4.637	0.000**
Original Combined Loan-to-Value (CLTV)	-9.893	0.000	0.316	-31.278	0.000**
Original Debt-to-Income (DTI) Ratio	-0.201	0.818	0.050	-3.983	0.000**
Original Loan Term	0.071	1.074	0.090	0.790	0.429
Postal Code - Northeast (1)	0.093	1.098	0.052	1.790	0.073
Postal Code - East (2)	0.078	1.081	0.047	1.654	0.098
Postal Code - Southeast (3)	-0.105	0.901	0.041	-2.551	0.011
Postal Code - North (5)	0.267	1.305	0.064	4.154	0.000**
Postal Code - South (7)	0.433	1.542	0.048	9.045	0.000**
Postal Code - Midwest (8)	-0.179	0.836	0.053	-3.358	0.001**
Property Type - Condo (CO)	-0.219	0.804	0.054	-4.075	0.000**
Property Type - Co-op (CP)	0.219	1.245	0.266	0.824	0.410
Property Type - Manufactured Housing (MH)	0.389	1.475	0.139	2.801	0.005*
Exposure at Default	0.246	1.279	0.106	2.326	0.020

** $p < 0.0026$ Bonferroni-adjusted significance level with original $\alpha = 0.05$
* $p < 0.0053$ Bonferroni-adjusted significance level with original $\alpha = 0.10$



(a) The Regression-Based model

(b) The Cox Proportional Hazards model



(c) The Random Survival Forest model

Figure 6: Histogram of the realized vs predicted LGDs for resolved cases

better performance. However, due to the large set of data, hyperparameter tuning takes a substantial amount of time, therefore, we chose to perform the grid search on only a few hyper-

parameter values. The hyperparameter tuning was done on the original set (pre-duplication) and we used the grid search shown in Table B1. After the grid search, we trained the model. In total, this took approximately four days.

In Table 8, the importance values of the features are given. As RSF grows trees using different bootstrap samples of the data and considers a random subset of variables at each split, variable important scores can vary between different runs of the algorithm. For a more robust result, we computed the feature importance three times and took the average. Again, this was a trade-off between computation time and robustness, as computing the importance measures three times took eight hours in Python. We find that the *Current Interest Rate* is the most important feature. If its relation to survival time is removed, the Concordance index decreases on average by approximately 0.043561 points. This is similar to the benchmark and Cox PH model, where the *Current Interest Rate* also had a large effect on the dependent variable, measured by the regression coefficient and hazard ratio, respectively. A negative or zero-importance value indicates that the risk driver has no predictive power. Interesting to see is that eight out of 19 variables do not contribute to the predictive power in the RSF model. This could, however, be caused by the fact that the model's risk drivers are based on the optimal features of the Regression-Based model. This may imply that the set of optimal features of the RSF model is very different compared to the optimal set of features of the benchmark model. RSF models can handle complex relationships, while the Regression-Based model is simpler and based on linear relationships between the dependent variable and the risk drivers. The data may have complex patterns, which the RSF can capture better, leading to differences in the selected features that contribute to their predictive performance.

Table 8: Variable importance Random Survival Forest model

	Importance mean	Importance s.d.
Current Interest Rate	0.043561	3.9463 e-04
Postal Code - Southeast (3)	0.000321	4.6536 e-05
Postal Code - North (5)	0.000277	1.5635 e-05
Original Debt-to-Income (DTI) Ratio	0.000166	1.2152 e-04
Exposure at Default	0.000143	1.4095 e-04
First Time Homebuyer Flag - Yes	0.000127	9.3205 e-06
Property Type - Condo (CO)	0.000101	2.2881 e-05
Postal Code - East (2)	0.000099	1.6760 e-06
Postal Code - Midwest (8)	0.000064	2.2162 e-05
Occupancy Status - Investment Property (I)	0.000011	2.0536 e-06
Number of Borrowers	0.000003	2.6708 e-07
Property Type - Co-op (CP)	0.000000	0.0000 e+00
Property Type - Manufactured Housing (MH)	0.000000	0.0000 e+00
Postal Code - Northeast (1)	-0.000004	2.2930 e-06
Postal Code - South (7)	-0.000054	2.6371 e-04
Loan Purpose - Refinance - Cash Out (C)	-0.000130	2.9042 e-06
Loan Purpose - Refinance - No Cash Out (N)	-0.000151	1.4978 e-05
Original Loan Term	-0.004680	2.7054 e-04
Original Combined Loan-to-Value (CLTV)	-0.009805	2.5451 e-04

Lastly, we plot the realized LGDs against the predicted LGDs (Figure 6c). We find that the RSF model's distribution has three peaks. The two largest are at an LGD of approximately 0.30 and 0.70. This shows that the RSF is somewhat able to capture the bimodality of the LGD, however, the peaks are slightly closer to each other. Furthermore, it also has a slightly smaller peak at an LGD of approximately 0.5. The fact that the RSF model is somewhat able to capture the bimodality is very interesting. As seen in the previous sections, common modelling techniques are not able to capture this distribution.

5.5 Model Comparison

Table 9 presents the performance results for the three different models. First, we compare the backtesting results of the test set with resolved cases only. The Regression-Based model has a good LCR of 0.6743, which is higher than the Cox PH model (0.5744), however, the RSF has the highest LCR (0.7442), indicating that the RSF is able to differentiate between the severity of losses better. The three curves are also shown in Figure 7. Here we see that all curves are very similar, but the RSF model curve is the closest to the perfect model curve. Furthermore, the MAE and MSE of the Regression-Based model and RSF model are worse compared to the Cox PH. This indicates that the Cox PH has the most accurate LGD estimations. From Figure 8, we see that the RSF has fewer almost perfect predictions as compared to the Regression-Based and Cox PH model, however, the RSF has more observations at the peaks ($\pm$ 0.3). We also find that the RSF has the best LS (0.0132). This implies that the RSF model can capture the total loss better than the Regression-Based and Cox PH model. Lastly, although the t-test is a commonly used calibration metric, we find that it is not relevant to our data set. T-tests are based on the assumption that the distribution of the data is normally distributed (Kim & Park, 2019), while the LGD of this data set is bimodally distributed (Figure 2). Furthermore, when we plot the distribution of the differences between actual and predicted LGD we see that for resolved cases this is bimodal, with the two peaks at a difference of $\pm$ 0.3 (Figure 8). This implies that on average, these two cancel each other out, resulting in a very good average predicted LGD. We, therefore, cannot conclude anything from the t-test and exclude this measure when comparing the models. Overall, we find that the RSF model has high discriminatory power but low calibration power. RSF is based on an ensemble of individual trees. While it is often able to capture complex relationships in the data, leading to high discriminatory power, the predictions are based on averaging the individual trees, making it more difficult to obtain LGD values close to zero and one (Oleszak, 2023), as was also seen in Figure 6c. This could be a reason for the low calibration power.

Next, we compare the results of when we train and test on only resolved cases with when we train on resolved and unresolved cases but test on only resolved cases (Table 9). If the results for both methods are similar, that is, both methods exhibit comparable levels of predictive power, we can conclude that the model can handle unresolved cases well. We find that the results for most performance measures are very similar for both methods for all models. Interesting to see is that the LCR for the Regression-Based model is slightly better when we train on both resolved and unresolved cases, similarly for the LS for all the models. A possible explanation could be

Table 9: Model performances of final LGD predictions

Test Set	Model	LCR[1]	MAE[2]	MSE[3]	LS[4]
Resolved	Regression-Based	0.6743	0.3486	0.1527	0.0341
	Cox Proportional Hazards	0.5744	0.3426	0.1514	0.0377
	Random Survival Forest	**0.7442**	0.3713	0.1551	0.0132
	Calibrated Regression-Based	0.6700	**0.3302**	0.1492	**0.0027**
	Calibrated Cox Proportional Hazards	0.5737	0.3346	0.1499	0.0376
	Calibrated Random Survival Forest	**0.7442**	0.3328	**0.1448**	0.0039
Resolved - Resolved[5]	Regression-Based	0.6444	0.3411	0.1515	0.0351
	Cox Proportional Hazards	0.5789	0.3409	0.1517	0.0564
	Random Survival Forest	**0.7597**	0.3604	0.1508	0.0380
	Calibrated Regression-Based	0.6399	0.3313	0.1495	0.0149
	Calibrated Cox Proportional Hazards	0.5784	0.3356	0.1504	0.0366
	Calibrated Random Survival Forest	**0.7597**	**0.3220**	**0.1440**	**0.0009**

[1] LCR = Loss Capture Ratio
[2] MAE = Mean Absolute Error
[3] MSE = Mean Squared Error
[4] LS = Loss Shortfall
[5] Both training and testing on resolved cases only
Note: Bold values indicate the corresponding model performs the best

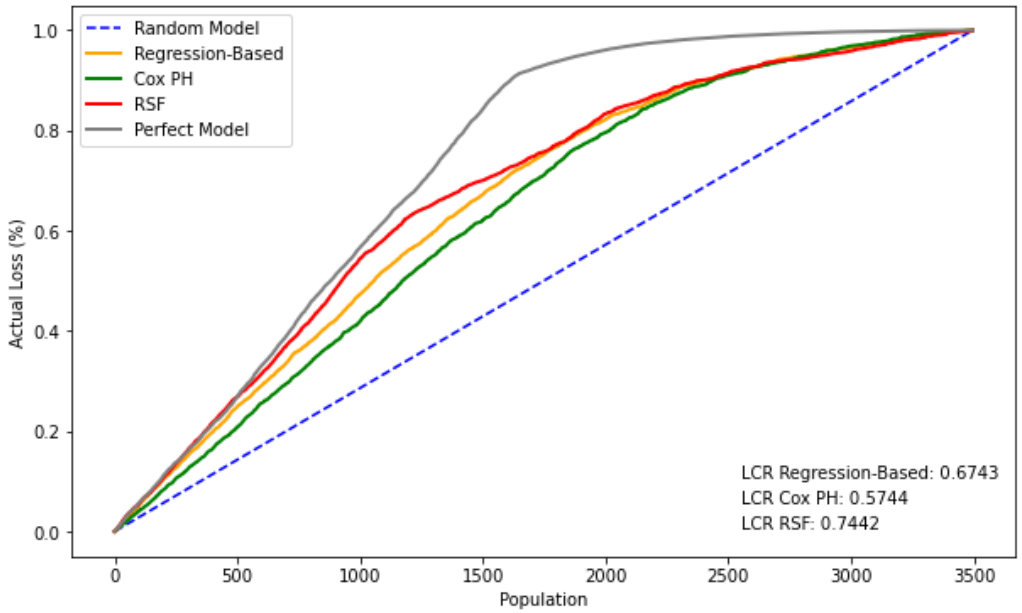
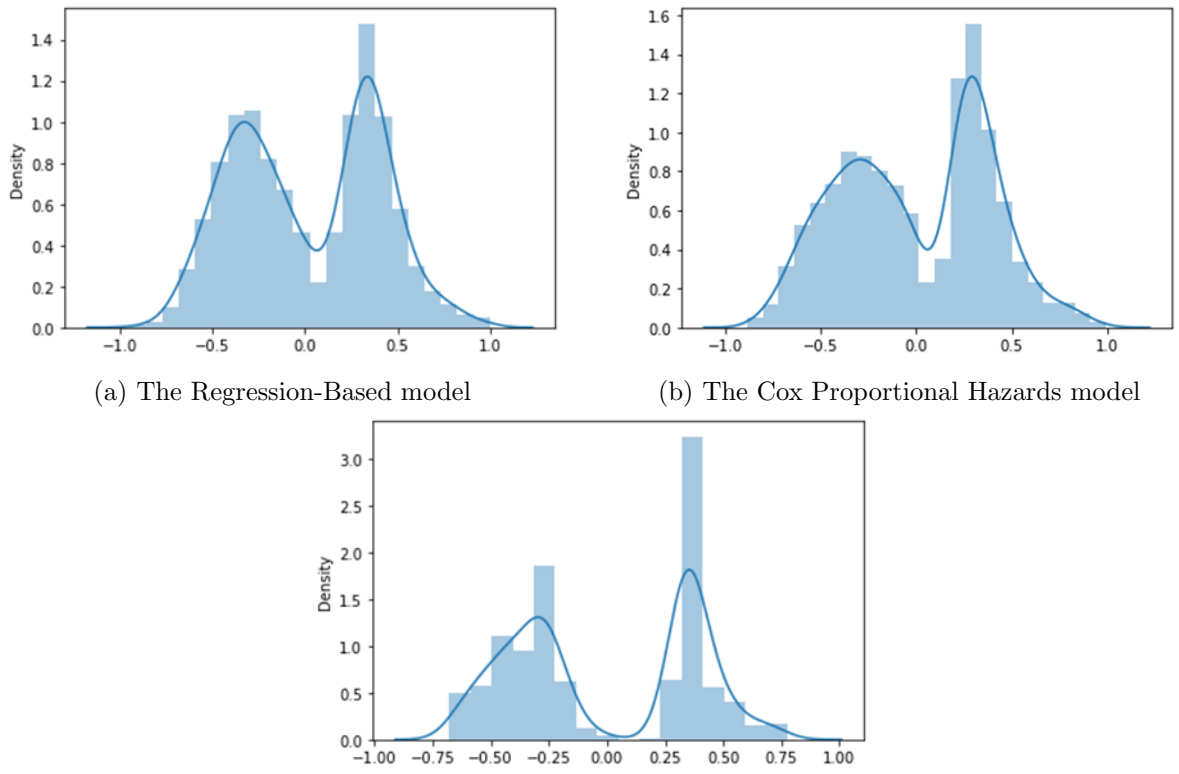


Figure 7: Loss capture curves for resolved cases

that the predictions are closer to the realized values, however, the predictions are consistently over- or underpredicting. This improves the MAE and MSE, but the LS can worsen. Overall, all models perform similarly for both methods, so we can conclude that all models can handle unresolved cases well.

Furthermore, backtesting can only be done on resolved cases. We cannot compare the model's predictions with the realized LGDs, as the realized LGDs are not the actual final observed LGDs of the loans. These are the final observed LGDs at the performance cutoff date, whereas the model predicts the final observed LGDs till the case is closed. Backtesting unresolved cases would require an estimation of the final realized LGDs, however, this would add extra uncertainty to



(a) The Regression-Based model

(b) The Cox Proportional Hazards model

(c) The Random Survival Forest model

Figure 8: Distribution of the differences between realized and predicted LGDs for resolved cases

the validation process. We considered a binning method in which we binned final realized LGDs for resolved cases based on the most important risk drivers in Table 6 and the number of months the observation was currently in default. After this a scale factor, defined as the ratio between the average final LGD and the average current LGD of each bin was computed. Then, we binned the unresolved cases in a similar way and used the scale factor to calculate the final estimated LGD. These final estimated LGDs were then compared to the model-predicted LGDs. Although this method works, it still brings too much uncertainty to the model performances. Backtesting on unresolved cases should, therefore, be left for further research.

Then, we also compare the predictions at time $t = 12, 24, 36, 48, 60$, to see whether the models can also predict on a short-term basis. This is particularly useful when one wants to know the total loss at a specific date, rather than the total final loss. Knowing the total loss at a specific date allows lenders to have a better and timely understanding of the financial performance and risk associated with their loans, enabling better decision-making and risk management. Table 10 shows the performance results of the three models when we trained on resolved and unresolved cases but tested on resolved cases only. As with the final LGD, we find that the RSF outperforms the Regression-Based and Cox PH model based on the LCR, however, it performs worse based on the MAE and MSE. Interesting to see, is that all models have a negative LS, i.e. they overpredict the LGDs, for $t = 12, 24, 36, 48$, and even $t = 60$ for the RSF model. This may be favourable for risk-averse lenders who may prefer a more conservative prediction to help protect their financial stability. Furthermore, we find that the Regression-Based model has the

best LS for each twelve-monthly prediction until $t = 60$. Overall, it seems like the Regression-Based model either performs the best or second-best based on the four performance measures. Nevertheless, computing these twelve-monthly predictions is a lot more cumbersome compared to the two SA techniques. Calculating the twelve-monthly predictions for the Regression-Based model requires fitting the model for each t. For the Cox PH and RSF model, on the other hand, the model has to be fitted only once, and the LGD for each observation is predicted for each t directly.

Table 10: Model performances of twelve-monthly LGD predictions

	Model \Time	12	24	36	48	60
LCR[1]	Regression-Based	0.3407	0.4547	0.4728	0.4786	0.4828
	Cox Proportional Hazards	0.5070	0.5812	0.5833	0.5822	0.5833
	Random Survival Forest	**0.6924**	**0.7520**	**0.7463**	**0.7437**	**0.7466**
	Calibrated Regression-Based	0.4906	0.5768	0.5975	0.6073	0.6229
	Calibrated Cox Proportional Hazards	0.5101	0.5854	0.5850	0.5840	0.5848
	Calibrated Random Survival Forest	**0.6924**	**0.7520**	**0.7463**	**0.7437**	**0.7466**
MAE[2]	Regression-Based	0.3132	0.3297	0.3380	0.3446	0.3470
	Cox Proportional Hazards	0.3032	0.3317	0.3398	0.3445	0.3450
	Random Survival Forest	0.3183	0.3477	0.3593	0.3667	0.3702
	Calibrated Regression-Based	0.3034	0.3167	0.3238	0.3313	0.3329
	Calibrated Cox Proportional Hazards	0.3026	0.3151	0.3206	0.3279	0.3302
	Calibrated Random Survival Forest	**0.2967**	**0.3097**	**0.3147**	**0.3196**	**0.3223**
MSE[3]	Regression-Based	0.1476	0.1526	0.1538	0.1551	0.1543
	Cox Proportional Hazards	0.1505	0.1546	0.1541	0.1543	0.1533
	Random Survival Forest	0.1486	0.1551	0.1561	0.1570	0.1568
	Calibrated Regression-Based	0.1457	0.1499	0.1513	0.1528	0.1522
	Calibrated Cox Proportional Hazards	0.1463	0.1491	0.1497	0.1511	0.1507
	Calibrated Random Survival Forest	**0.1424**	**0.1460**	**0.1465**	**0.1469**	**0.1468**
LS[4]	Regression-Based	-0.0064	-0.0095	-0.0034	-0.0033	0.0020
	Cox Proportional Hazards	-0.0428	-0.0380	-0.0147	-0.0041	0.0055
	Random Survival Forest	-0.0247	-0.0292	-0.0166	-0.0130	-0.0071
	Calibrated Regression-Based	0.0062	0.0116	0.0148	0.0152	0.0122
	Calibrated Cox Proportional Hazards	0.0022	0.0094	0.0175	0.0223	0.0248
	Calibrated Random Survival Forest	**-0.0020**	**-0.0017**	**0.0024**	**0.0023**	**-0.0003**

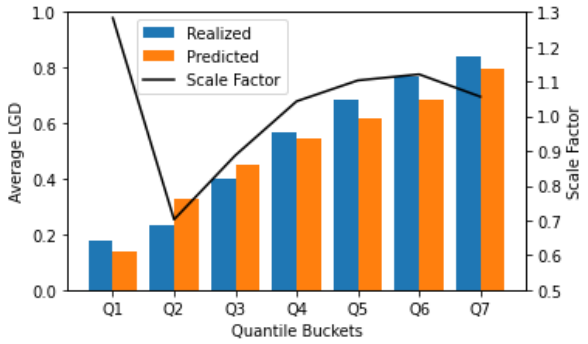
[1] LCR = Loss Capture Ratio
[2] MAE = Mean Absolute Error
[3] MSE = Mean Squared Error
[4] LS = Loss Shortfall
Note: Bold values indicate the corresponding model performs the best

Lastly, we observe that when predicting the final and twelve-monthly LGDs, the RSF model performs poorly based on the MAE and MSE, but very well based on the LCR, and LS for the final predictions. Furthermore, we found that the RSF was the only model that was able to capture the bimodal distribution of the LGD. This implies that the RSF model has high discriminatory power, but low calibration power. Nevertheless, this is a matter of calibrating the model correctly. To calibrate the model correctly, we propose the following method. We bin the predicted final LGDs of the RSF model in seven equal-sized bins. This is the minimum number of buckets required in a credit risk context required by the European Banking Authority (European Parliament, Council of the European Union, 2013). We compute a scaling factor

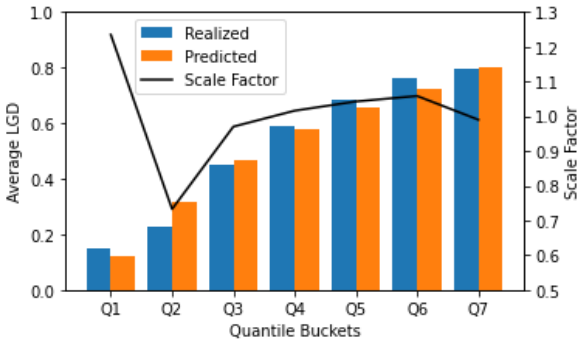
based on the ratio between the average realized final LGD and the average predicted final LGD per bucket. The predicted final LGDs are multiplied by this scaling factor. Finally, we calculate the performance measures of the adjusted final LGDs. For a fair comparison, we also calibrate the Regression-Based model and Cox PH model in the same way. Note that the calibration may be specific to the data set considered. Applying the model to a different data set may require a recalibration of the model.

Figure 9 shows the average realized and average predicted LGD per bucket with the corresponding scale factor for all three models. Here, we clearly see that the scale factors of the Regression-Based and Cox PH model are not monotonically increasing. The scale factors of the RSF model are almost monotonically increasing, with just a slight decline from the fourth to the fifth quantile bucket. Therefore, we get that the smallest LGD predictions become even smaller, the largest LGD predictions become even larger, and similarly for the other bins. Furthermore, as the observations within each bin are multiplied with the same scaling factor, we get that the rank order of the LGDs remains the same within the bins. Therefore, overall, the order of the predictions does not change, but the predictions are spread over a wider range. For this reason, we improve the calibration power, while the discriminatory power remains the same. Furthermore, we find that the Regression-Based and Cox PH model follow a similar pattern, the average predicted LGD is slightly below the average realized LGD in the first, and fourth till sixth quantile bucket, while it is slightly above in the second and third bucket. As their scale factor pattern is similar as well, the effect of calibrating the models will likely be the same for these two models. We will support this later in Tables 9 and 10. The RSF model, on the other hand, consistently overpredicts for the first few buckets, and thereafter consistently underpredicts. This characteristic (two monotonic relations) is favourable for improving the calibration power. Monotonic relations are easier to calibrate when one uses a single calibration function rather than a calibration factor per bucket. Lastly, we find that the difference between the average realized and predicted LGD increases per bucket further away from the median bucket for the RSF model. In the fourth quantile bucket, the average realized and predicted are almost the same. This result will positively influence the calibration performance of the RSF model. We will support this later in Tables 9 and 10.

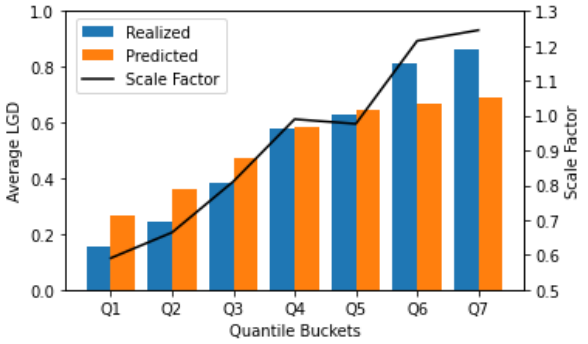
We find that the Calibrated RSF model outperforms the (Calibrated) Regression-Based model and (Calibrated) Cox PH model on the LCR and MSE (Resolved Test Set, Table 9). When a lender prioritizes the identification and reduction of large errors in LGD prediction, a lower MSE is desired, and the Calibrated RSF would be a good model. The Calibrated Regression-Based model performs slightly better than the (Calibrated) Cox PH and (Calibrated) RSF model based on the MAE and LS, though the difference with the Calibrated RSF model is minimal. Overall, the calibration power of the RSF model improved substantially, while the discriminatory power remained the same. The discriminatory power of the Regression-Based and Cox PH model, on the other hand, decreased. Figure 10 shows the distributions of the adjusted final LGDs. We observe that the predicted LGDs for all three models are slightly more spread over the range of 0 to 1 when we calibrate the LGDs. We also find that the Regression-Based and Cox PH



(a) The Regression-Based model



(b) The Cox Proportional Hazards model



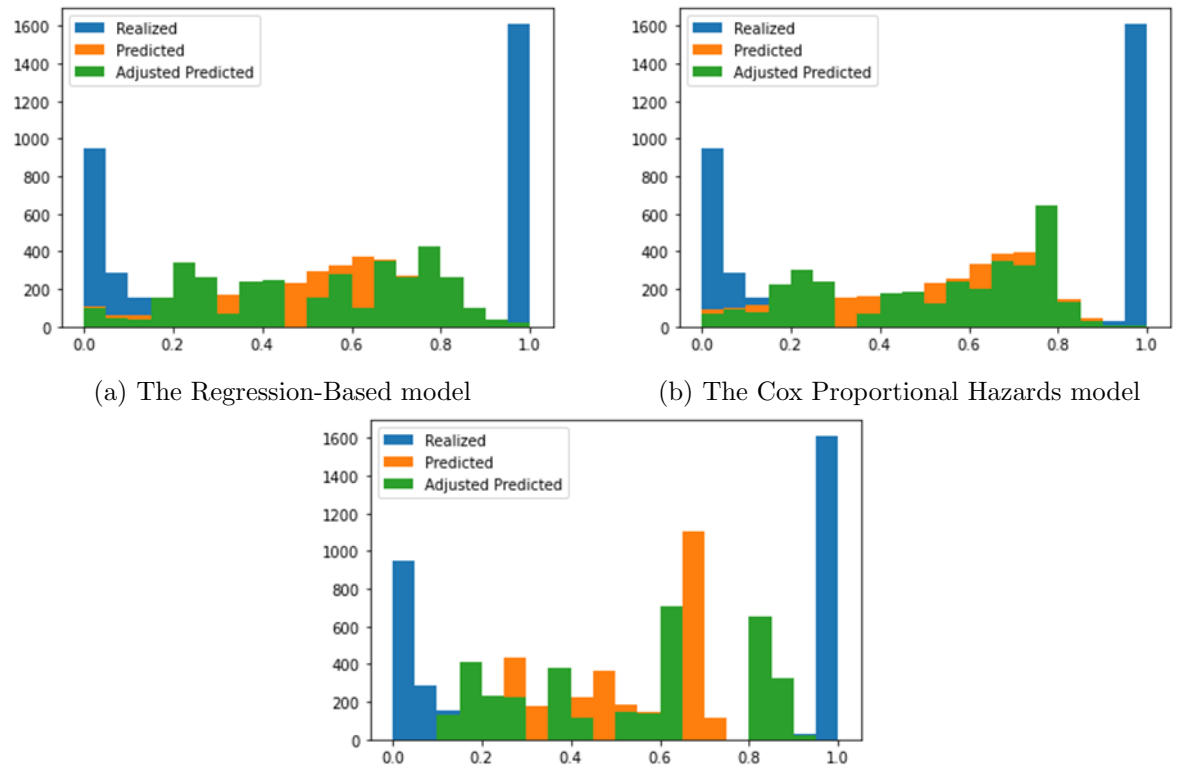
(c) The Random Survival Forest model

Figure 9: Average realized vs average predicted LGDs for resolved cases of the calibrated models and their scale factors per bucket

models become slightly more bimodal after calibration (Figures 10a and 10b). Again, as an extra check, we also train and test the calibrated models on resolved cases only. We find that the Calibrated RSF model outperforms all other models and that the performances of the Calibrated RSF and Calibrated Cox PH model are similar to when we trained on both resolved and unresolved cases. This means that the Calibrated RSF and Cox PH models are also able to differentiate between resolved and unresolved cases. However, the Calibrated Regression-Based model performs worse on all four measures when trained on resolved cases only. This could be an indication that the Calibrated Regression-Based model performs better when trained on more dissimilar data. When trained on resolved cases only, it may become overly fit to the specific patterns and characteristics of those cases. It may not be able to perform well on new or unseen resolved cases. Lastly, we also calibrate the LGDs for the twelve-monthly predictions. The results can be found in Table 10. We find that the Calibrated RSF model outperforms the (Calibrated) Regression-Based model and the (Calibrated) Cox PH model based on all four performance measures for all months $t = 12, 24, 36, 48, 60$.

5.6 Discovery: LGD Prediction with Survival Analysis

During the process of predicting LGD with SA, we discovered an interesting finding. SA is able to predict the final LGD well for a period $t_1 = 0$ till another period t_2. However, we found that SA is not able to predict the expected future recoveries between a certain period t_1 and t_2, $0 < t_1 < t_2$. In this case, the model considers unrecovered parts of resolved cases as still



(a) The Regression-Based model

(b) The Cox Proportional Hazards model

(c) The Random Survival Forest model

Figure 10: Histogram of the realized vs predicted LGDs for resolved cases of the calibrated models

performing at time t_1 and predictions become inaccurate. The expected LGD at time t_2 is therefore incorrect. In Table 11, we give a simple example, where we have six loans, all with outstanding amounts of 100 at $t_1 = 0$, time of default. We indicate the moment a loan resolves with underlined values. The aim is to predict the LGD between $t_1 = 0, 1, 2$ and $t_2 = 3$. For illustration purposes, we use a simplified SA model, namely, the Kaplan-Meier model. This model works similarly to the Cox PH model, however, it does not include explanatory variables in the estimation (Witzany et al., 2012).

The SA predicted recovery rate between time t_1 and t_2 is calculated by

$$RR_{t_1,t_2} = \frac{\text{Survival Function}_{t_1} - \text{Survival Function}_{t_2}}{\text{Survival Function}_{t_1}}, \tag{39}$$

where *Survival Function*$_{t_1}$ is the predicted LGD based on the survival curve at time t_1, similarly for *Survival Function*$_{t_2}$. Note that the survival curve is the same for each observation as we use the Kaplan-Meier model, that is, no risk drivers are taken into account. In this example, the values of the survival curve are the same as the average Realized Predicted LGDs in Table 11b as we only have resolved cases. We divide by *Survival Function*$_{t_1}$ in equation (39) to adjust for the LGD already known at time t_1 (the recovery between time t_1 and t_2 can be seen as a conditional probability). In Table 11b, we see that the predicted LGD at $t_2 = 3$ for $t_1 = 0, 1, 2$, $LGD_{t_1,t_2} = LGD_{t_1} - RR_{t_1,t_2}$, is only correct for $t_1 = 0$, where LGD_{t_1} is the average Realized

Table 11: Example of LGD prediction with Survival Analysis

(a) Outstanding amounts

Loan	$t_1 = 0$	$t_1 = 1$	$t_1 = 2$	$t_2 = 3$
1	100	100	100	100
2	100	100	100	0
3	100	100	100	100
4	100	100	0	0
5	100	100	100	100
6	100	0	0	0

(b) Current Survival Analysis realized predicted LGD

Loan	$t_1 = 0$	$t_1 = 1$	$t_1 = 2$	$t_2 = 3$	Workout Length[1]
1	1	1	1	1	3
2	1	1	1	0	3
3	1	1	1	1	2
4	1	1	0	0	2
5	1	1	1	1	1
6	1	0	0	0	1
Average Realized Predicted LGD	1	0.83	0.67	0.50	
SA Predicted RR_{t_1,t_2}[2]	0.50	0.40	0.25	-	
SA Predicted LGD_{t_1,t_2}[3]	0.50	0.43	0.42	-	
Scaling Factor	1	1.25	2	-	

(c) Survival Analysis adjusted predicted recovery[4]

Loan	$t_1 = 0$	$t_1 = 1$	$t_1 = 2$	$t_2 = 3$
1	0.50	0.50	0.50	0
2	0.50	0.50	0.50	0
3	0.50	0.50	0	0
4	0.50	0.50	0	0
5	0.50	0	0	0
6	0.50	0	0	0
Average	0.50	0.33	0.17	0

(d) Final adjusted predicted LGD[5]

Loan	$t_1 = 0$	$t_1 = 1$	$t_1 = 2$	$t_2 = 3$
1	0.50	0.50	0.50	1
2	0.50	0.50	0.50	0
3	0.50	0.50	1	1
4	0.50	0.50	0	0
5	0.50	1	1	1
6	0.50	0	0	0
Average	0.50	0.50	0.50	0.50

[1] Workout Length: the number of months the loan was in default
[2] SA Predicted RR_{t_1,t_2}: Unadjusted predicted recovery rate between time t_1 and t_2
[3] SA Predicted LGD_{t_1,t_2}: Predicted LGD at $t_2 = 3$ when we look at current time t_1, LGD_{t_1,t_2} = LGD_{t1} - RR_{t_1,t_2}
[4] Survival Analysis adjusted predicted recovery: SA Predicted RR_{t_1,t_2} × Scaling Factor, or 0 if loan is already resolved
[5] Final adjusted predicted LGD: Current Survival Analysis realized predicted LGD - Survival Analysis adjusted predicted recovery, i.e. adjusted predicted LGD at $t_2 = 3$ when we look at current time t_1
Note: Underlined values indicate the case is resolved

Predicted LGD at time t_1. The issue of incorrect recovery predictions is only present for predictions between t_1 and t_2, $0 < t_1 < t_2$ because at time $t_1 = 0$, there are no closed cases with unrecovered exposure yet. To the best of our knowledge, this has never been discovered before.

The problem arises due to the way the data is used in the model. In binary scenarios, such as in medical cases, we can say with certainty that at time $t = \infty$ a client has "died" and no more data is observed. In an LGD context, on the other hand, the remaining exposure of resolved cases is set to an artificial observation at time t_{max}, as explained in Section 4.2.3. This, however, implies that we still expect recoveries even though the observation is already resolved.

We call this the Naive Approach and also show it in Table 12. Note that the Naive Approach does not regard the recovery between time t_1 and t_2 as a conditional probability, therefore, we do not divide by *Survival Function*$_{t_1}$ as in equation (39). We find that using the Naive Approach also leads to the correct final LGDs. Nevertheless, SA assumes that once a case is closed, there cannot be any recoveries anymore. For this reason, we need to use a scaling factor.

Table 12: The Naive Approach

(a) Naive recovery rates

Loan	$t_1 = 0$	$t_1 = 1$	$t_1 = 2$	$t_2 = 3$
1	0	0	0	$\underline{0}$
2	1	1	1	$\underline{0}$
3	0	0	$\underline{0}$	$\underline{0}$
4	1	1	$\underline{0}$	$\underline{0}$
5	0	$\underline{0}$	$\underline{0}$	$\underline{0}$
6	1	$\underline{0}$	$\underline{0}$	$\underline{0}$
Average	0.50	0.33	0.17	0

(b) Naive predicted LGDs

Loan	$t_1 = 0$	$t_1 = 1$	$t_1 = 2$	$t_2 = 3$
1	1	1	1	$\underline{1}$
2	0	0	0	$\underline{0}$
3	1	1	$\underline{1}$	$\underline{1}$
4	0	0	$\underline{0}$	$\underline{0}$
5	1	$\underline{1}$	$\underline{1}$	$\underline{1}$
6	0	$\underline{0}$	$\underline{0}$	$\underline{0}$
Average	0.50	0.50	0.50	0.50

If we first multiply the recovery rate by a scaling factor

$$\text{Scaling Factor}_{t_1} = \frac{\sum_{i=1}^{N} LGD_{i,t_1}}{\text{Number of observations with workout length larger than } t_1}, \tag{40}$$

where LGD_{i,t_1} is the LGD of loan i at time t_1, and set the recovery to 0 if the case is already resolved (Table 11c), we do get the correct predicted LGDs (0.50 in this example, Table 11d). To show that this also works for more complex examples of resolved cases, we provide an example in Tables B2 and B3.

This method of scaling the LGDs works for resolved cases, however, further research is required about the appropriateness of this approach for unresolved cases. For unresolved cases, the "desired" final LGD is not as straightforward as for resolved cases. When there are also unresolved cases, the predicted LGD will be lower for the SA model since it assumes a certain extent of recoveries to come in for the unresolved cases, whereas the average LGD calculation simply looks at realized cashflows. An additional scale factor for unresolved cases may be needed. This, nevertheless, goes beyond the topic of this research and will, therefore, be left for further research.

6 Conclusion

Lenders are required to estimate the Loss Given Default (LGD) accurately to adhere to the Basel Accord. An accurate LGD estimation can help determine a more precise capital buffer size, which in turn can absorb losses and free up capital for investments, for example. Traditional LGD forecasting techniques, however, can only use resolved and adjusted unresolved cases in the modelling process, while resolved and unresolved cases may exhibit different recovery behaviours. Therefore, these techniques may result in biased estimates. The purpose of this research

was to accurately predict the expected future recovery cash flows of defaulted loans and investigate which method performs best. We proposed a machine learning-based Survival Analysis (SA) model; the Random Survival Forest (RSF), and compared this model to the traditional Regression-Based model and the semi-parametric Cox Proportional Hazards (Cox PH) model. We tested this on a set of American mortgages from Freddie Mac.

The results showed that before calibrating the RSF model, the RSF was the only model that could capture the bimodality of the LGD. Moreover, it had the best Loss Capture Ratio (LCR) and Loss Shortfall (LS) implying that it was able to differentiate between the severity of losses better and it was able to capture the total loss better. However, the Cox PH model slightly outperformed the RSF model and the Regression-Based model based on the Mean Absolute Error (MAE) and Mean Squared Error (MSE), implying that the Cox PH model was able to make more accurate LGD predictions. Furthermore, the built-in RSF function in Python was not able to correctly implement the model yet, nevertheless, via a workaround we were able to implement the RSF model. This, however, affected the computation speed and the obtained performance of the RSF model. Due to these modelling limitations, the obtained RSF performance was not optimal. We also performed an extra check by training and testing on resolved loans only. The results showed that all models performed similarly to when we trained on the full training set. This implies that all models were able to differentiate well between resolved and unresolved cases. Furthermore, we predicted the LGD at twelve-month intervals. The RSF model again outperformed the other two models based on the LCR, but the Regression-Based model outperformed the Cox PH and RSF model for most intervals based on the MAE, MSE and LS. However, calculating the twelve-monthly predictions for the Regression-Based model is more cumbersome as it requires fitting the model for each time period, rather than only once for the Cox PH and RSF model.

Lastly, we observed that when predicting the final and twelve-monthly LGDs, the RSF model had high discriminatory power, but low calibration power. Nevertheless, this was a matter of calibrating the model correctly. After calibrating the model via a binning method, we found that the calibration power improved, while the discriminatory power remained the same. We found that the Calibrated RSF model outperformed the (Calibrated) Regression-Based model and (Calibrated) Cox PH model based on the LCR and MSE. When a lender prioritizes the identification and reduction of large errors in LGD prediction, a lower MSE is desired, and the Calibrated RSF would be a good model. Furthermore, the Calibrated RSF model outperformed all models based on all four measures for the twelve-monthly predictions. This implies that the Calibrated RSF model was the best in short-term LGD predictions.

Overall, based on these findings, we can conclude that the RSF model is the only model that can capture the bimodal distribution of the LGD and after calibrating the RSF model it often outperforms the (Calibrated) Regression-Based model and (Calibrated) Cox PH model on all four measures. Specifically, when one prioritizes high discriminatory power and the reduction of large errors in LGD prediction, or when one wants to perform short-term LGD predictions,

the Calibrated RSF model is an appropriate model.

One of the potential limitations of this study is that the RSF model was not optimally implemented in Python. For a more credible conclusion, the RSF model must first be implemented correctly in Python's RSF package. Furthermore, the RSF model could not be optimally trained with the best hyperparameters due to computation constraints. The obtained performance was therefore affected, and we most likely did not obtain the optimal results. Lastly, all three models used the risk drivers that were optimal for the Regression-Based model. However, we saw that eight out of 19 risk drivers did not have any predictive power in the RSF model. Therefore, the risk drivers could probably have been selected more appropriately. It would then be interesting to see whether and how the results would change if we use the optimal risk drivers corresponding to the model.

We have several suggestions for potential further research. First, during the process of predicting LGD with SA, we discovered that SA is not able to predict the expected future recoveries between a certain time period t_1 and t_2, $0 < t_1 < t_2$, when the data set has resolved and unresolved cases. It would be valuable to find a way to be able to predict between any two time periods, rather than only between the start and another period. This is particularly useful when we have cases for which the information at the month of default is unavailable.

Furthermore, backtesting was only done on resolved cases. We found a way to backtest on unresolved cases, however, this method still brought too much uncertainty to the model performances. It would be interesting to find a way to backtest unresolved cases without adding extra uncertainty to the validation process. This way, one can validate whether the models are able to capture the different recovery behaviour of resolved and unresolved cases.

In this research, only mortgage-specific risk drivers were used, however, it has been shown that macroeconomic variables have a positive effect on the prediction performance of LGD. It would be useful to see how the results would change, and whether the performance of one model would be affected more by the macroeconomic variables than the performance of the other models.

Lastly, there are several machine learning-based SA models, including the Survival Support Vector Machine, Gradient Boosted Survival and Survival Extreme Gradient Boosting models. Other research papers found that these models could also outperform the Regression-Based model and Cox PH model in binary-event prediction. It would be interesting to see if the same holds in the case of LGD prediction and whether these models could outperform the RSF model.

References

Araj, V. (2023). *A guide to first-time home buyer programs, loans and grants.* Retrieved from https://www.rocketmortgage.com/learn/first-time-home-buyer

Bhakta, A., Kim, Y. & Cole, P. (2021). Comparing machine learning-centered approaches for forecasting language patterns during frustration in early childhood. *arXiv*.

Breslow, N. E. (1974). Covariance analysis of censored survival data. *Biometrics*, *30*(1), 89–99.

Brock, M. (2023). *Is a mortgage secured or unsecured debt?* Retrieved from https://www.rocketmortgage.com/learn/is-a-mortgage-secured-or-unsecured#:~:text=Mortgages%20are%20%22secured%20loans%22%20because,higher%20risk%20to%20the%20lender.

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L. & Lopez, A. (2020). A comprehensive survey on support vector machine classification: applications, challenges and trends. *Neurocomputing*, *408*, 189–215.

de Amorim, L. B., Cavalcanti, G. D. & Cruz, R. M. (2023). The choice of scaling technique matters for classification performance. *Applied Soft Computing*, *133*.

Doan, L. M. T., Angione, C. & Occhipinti, A. (2022). Machine learning methods for survival analysis with clinical and transcriptomics data of breast cancer. In *Computational biology and machine learning for metabolic engineering and synthetic biology* (Vol. 2553, pp. 325–292). Humana, New York, NY.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, *72*(359), 557–565.

European Banking Athority. (2018). *Capital requirements regulation (CRR) Article 178.* Retrieved from https://www.eba.europa.eu/regulation-and-policy/single-rulebook/interactive-single-rulebook/1738

European Banking Authority. (2017). *Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures.* Retrieved from https://www.eba.europa.eu/sites/default/documents/files/documents/10180/2033363/6b062012-45d6-4655-af04-801d26493ed0/Guidelines%20on%20PD%20and%20LGD%20estimation%20%28EBA-GL-2017-16%29.pdf?retry=1

European Banking Authority. (2018). *Capital requirements regulation (CRR) Article 92.* Retrieved from https://www.eba.europa.eu/regulation-and-policy/single-rulebook/interactive-single-rulebook/536

European Central Bank. (2019). Instructions for reporting the validation results of internal models: IRB Pillar I models for credit risk. In *Banking supervision.* European Central Bank.

European Parliament, Council of the European Union. (2013). *Regulation (EU) No 575/2013 of the European parliament and of the council - Article 170.* Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32013R0575

Freddie Mac. (2023a). *Single family loan-level dataset.* Retrieved from https://www.freddiemac.com/research/datasets/sf-loanlevel-dataset

Freddie Mac. (2023b). *Single family loan-level dataset general user guide.* Retrieved from https://www.freddiemac.com/fmac-resources/research/pdf/user_guide.pdf

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, *29*(5), 1159–1232.

George, B., Seals, S. & Aban, I. (2014). Survival analysis and regression models. *Journal of Nuclear Cardiology*, *21*(4), 686–696.

Gujarati, D. N. & Porter, D. C. (2009). Dummy variable regression models. In *Basic econometrics* (pp. 178–215).

Herzog, M. H., Francis, G. & Clarke, A. (2019). The multiple testing problem. In *Understanding statistics and experimental design* (pp. 63–66).

Hirschberg, J. & Lye, J. (2001). The interpretation of multiple dummy variable coefficients: an application to industry evects in wage equations. *Applied Economics Letters*, *8*, 701–707.

Hosmer, D. W., Lemeshow, S. & May, S. (2008). Regression models for survival data. In *Applied survival analysis: Regression modeling of time-to-event data* (pp. 67–91). John Wiley Sons.

Hurlin, C., Leymarie, J. & Patin, A. (2018). Loss functions for loss given default model comparison. *European Journal of Operational Research*, *268*(1), 348–360.

Hussain, A. (2022). *Mortgage putback: What it is, how it works, history.* Retrieved from `https://www.investopedia.com/terms/m/mortgage-putback.asp`

Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, *2*(3), 841–860.

Kim, T. K. & Park, J. H. (2019). More about the basic assumptions of t-test: normality and sample size. *Korean J Anesthesiol*, *72*(4), 331–335.

Li, D., Bhariok, R., Keenan, S. & Santilli, S. (2009). Validation techniques and performance metrics for loss given default models. *The Journal of Risk Model Validation*, *3*(3), 3–26.

Liu, X. (2012). *Survival analysis - model and applications.* Higher Education Press.

Miller, P. & Töws, E. (2018). Loss given default adjusted workout processes for leases. *Journal of Banking and Finance*, *91*, 189–202.

Moerbeek, M. & van Schie, S. (2016). How large are the consequences of covariate imbalance in cluster randomized trials: a simulation study with a continuous outcome and a binary covariate at the cluster level. *BMC Medical Research Methodology*, *16*(79).

Nasejje, J. B. & Mwambi, H. (2017). Application of random survival forests in understanding the determinants of under-five child mortality in Uganda in the presence of covariates that satisfy the proportional and non-proportional hazards assumption. *BMC Research Notes volume*, *10*(459).

Oleszak, M. (2023). *Bad machine learning models can still be well-calibrated.* Retrieved from `https://www.nannyml.com/blog/probability-calibration`

Peng, Y. & Dear, K. (2000). A nonparametric mixture model for cure rate estimation. *Journal of the International Biometric Society*, *56*(1), 237–243.

Prívara, S., Kolman, M. & Witzany, J. (2013). Recovery rates in consumer lending: empirical evidence and model comparison. *Bulletin of the Czech Econometric Society*, *21*(32).

Pölsterl, S. (2023a). *Ensemble forest.* Retrieved from `https://github.com/sebp/scikit-survival/blob/v0.22.2/sksurv/ensemble/forest.py#L73-L169`

Pölsterl, S. (2023b). *Evaluating survival models.* Retrieved from `https://github.com/sebp/`

scikit-survival/blob/v0.22.2/doc/user_guide/evaluating-survival-models.ipynb

Qu, Z., Li, H., Wang, Y., Wang, Y., Zhang, J., Abu-Siada, A. & Yao, Y. (2020). Detection of electricity theft behavior based on improved synthetic minority oversampling technique and random forest classifier. *Energies*, *13*(8).

Ratner, B. (2009). The correlation coefficient: its values range between +1/1, or do they? *Journal of Targeting, Measurement and Analysis for Marketing*, *17*, 139–142.

Slud, E. V. (1982). Consistency and efficiency of inferences with the partial likelihood. *Biometrika Trust*, *69*(3), 547–552.

Song, Y. (2020). *The AIC-BIC dilemma: an in-depth look.* Retrieved from https://repository.tudelft.nl/islandora/object/uuid%3A66216cab-2669-45a1-a054-c7ccf73e7112

Trion Properties. (2023). *Pros and cons of investing in multi-family properties.* Retrieved from https://trionproperties.com/news-and-articles/pros-and-cons-of-investing-in-multi-family-properties/

Vidiyala, R. (2020). *Normalization vs standardization.* Retrieved from https://towardsdatascience.com/normalization-vs-standardization-cb8fe15082eb

Wade, C. & Glynn, K. (2020). XGBoost unveiled. In *Hands-on gradient boosting with XGBoost and scikit-learn.* Packt Publishing.

Winnett, A. & Sasieni, P. (2001). A note on scaled Schoenfeld residuals for the proportional hazards model. *Bimoetrika*, *88*(2), 565–571.

Witzany, J., Rychnovský, M. & Charamza, P. (2012). Survival analysis in LGD modeling. *European Financial and Accounting Journnal*, *7*(1), 6–27.

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

Yingchun, L. (2014). Random forest algorithm in big data environment. *Computer Modelling New Technologies*, *18*(12A), 147–151.

Zhang, J. & Thomas, L. C. (2012). Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD. *International Journal of Forecasting*, *28*(1), 204–215.

A Data

A.1 Variables

Table A1: Variables and their description

Risk Driver	Description	Requirement Violation[7]
Borrower Assistance Status Code	The type of assistance plan that the borrower is enrolled in that provides temporary mortgage payment relief.	Missing Values
Channel	Denotes whether a Broker or Correspondent originated or was involved in the origination of the mortgage loan.	Loan Characteristic
Credit Score	Number that represents the borrower's creditworthiness at the time of origination.	
Current Actual UPB	The Current Unpaid Principal Balance (UPB): the mortgage ending balance for the corresponding monthly reporting period.	Correlation
Current Interest Rate	The current interest rate on the mortgage note.	
Current Loan Delinquency Status	A value corresponding to the number of days the borrower is delinquent.	Correlation
DDLPI [1]	Due date of the loan's scheduled principal and interest is paid through.	Missing Values
Deferred Payment Plan	Indicates whether the loan follows a deferred payment plan.	Normal Loans Characteristic
Delinquency due to Disaster	Denotes whether a disaster-related hardship took place.	Missing Values
Estimated LTV[2] (ELTV)	The current LTV based on the current value of the property.	Correlation & Missing Values
Exposure at Default	The Unpaid Principal Balance (UPB) at the time the loan went into default.	
First Time Homebuyer Flag	Indicator that denotes whether the borrower had no ownership in a residential property in the three years preceding the origination date.	
Interest-Bearing UPB	Portion of the UPB that will accrue interest.	Correlation
Loan Age	The number of scheduled payments from the time the loan was originated.	Correlation
Loan Purpose	Indicates whether the mortgage loan is a Cash-out Refinance, No Cash-out Refinance or a Purchase mortgage.	
Loan Sequence Number	Unique identifier assigned to each loan.	
Loss Given Default (LGD)	The current loss rate of the loan.	
MI[3] Percentage (%)	The percentage of loss coverage on the loan, at the time of Freddie Mac's purchase of the loan, that a mortgage insurer is providing to cover losses incurred as a result of a default on a loan.	
Modification Flag	Indicates whether the loan has been modified.	Normal Loans Characteristic
Months in Default	The number of months the loan has been in default.	
MSA[4]	The five-digit value for the MSA or Metropolitan Division of the mortgaged property.	Correlation
Non-Interest-Bearing UPB	Portion of the UPB that will not accrue interest.	Correlation
Number of Borrowers	The number of borrowers who are obligated to repay the mortgage note secured by the mortgaged property.	
Number of Units	Denotes whether the mortgage is a one-, two-, three-, or four-unit property.	
Occupancy Status	Denotes whether the mortgage type is owner occupied, second home, or investment property.	
Original Combined LTV[2] (CLTV)	The ratio between all outstanding loans and the value of the property at origination.	

Original DTI[5] Ratio	The ratio between the total monthly debt expense and the total monthly income of the borrower at origination.	
Original Loan Term	The number of scheduled monthly payments of the mortgage.	
Original LTV[2]	The ratio between the original mortgage loan amount and the value of the property at origination.	Correlation
Original UPB	The Unpaid Principal Balance (UPB) of the mortgage on the note date.	Correlation
Postal Code	The three-digit postal code for the location of the mortgaged property.	
PPM[6] Flag	Denotes whether the mortgage is a PPM, that is, a mortgage where the borrower is, or at any time has been, obligated to pay a penalty in the event of certain repayments of principal.	
Property State	The two-letter abbreviation indicating the state or territory within which the mortgaged property is located.	Correlation
Product Type	Denotes whether the product is a fixed-rate or adjustable-rate mortgage.	More than One Value
Property Type	Denotes whether the property type secured by the mortgage is a condominium, leasehold, planned unit development, cooperative share, manufactured home, or Single-Family home.	
Remaining Months to Legal Maturity	The remaining number of months to the mortgage maturity date.	Correlation
Zero Balance Code (ZBC)	A code indicating the reason the loan's balance was reduced to zero.	
Zero Balance Effective Date	Date on which the event triggering the ZBC took place.	Missing Values
Zero Balance Removal UPB	The amount of total UPB remaining on the loan immediately before the application of the ZBC.	Correlation

[1] DDLPI = Due Date of Last Paid Installment
[2] LTV = Loan-to-Value
[3] MI = Mortgage Insurance
[4] MSA = Metropolitan Statistical Area
[5] DTI = Debt-to-Income
[6] PPM = Prepayment Penalty Mortgage
[7] Requirement definitions as stated in Table 1

A.2 Data Removal

We exclude observations from the data set that were repurchased, i.e. mortgages that have been repurchased from Freddie Mac by the original seller. These mortgages contain faulty origination documents in which the creditworthiness of the mortgagor or value of the property was misrepresented (Hussain, 2022). Repurchased mortgages, therefore, do not represent a mortgage under normal circumstances. Moreover, loans that have been modified are removed from the data set as well. Modifications include changes of a loan such that these loans follow different patterns, and are, thus, a misrepresentation of a loan under normal circumstances. The data set also contains over 9,000 defaulted loans with a deferred payment plan. We see that these loans hardly have any cash flows before the start of the deferred payment plan (Figure A.1), however, after the start of the plan, these loans directly go back to performing loans. We consider the case that these loans are treated as unresolved, that is, data until the month of deferral is included in the data set, and data from the month of deferral onwards is removed from the data set. We

then compare the average cash flow of the unresolved deferred loans to the average cash flow of normal unresolved loans. The average cash flow of the first group is approximately 90, whereas it is approximately 342 for the latter group. After performing a t-test, we can conclude that these two groups are significantly different from each other (p-value of <0.0001). Therefore, these deferred payment plans follow a different pattern than normal loans, including these loans in the data set could lead to biases. We exclude all loans with a deferred payment plan.

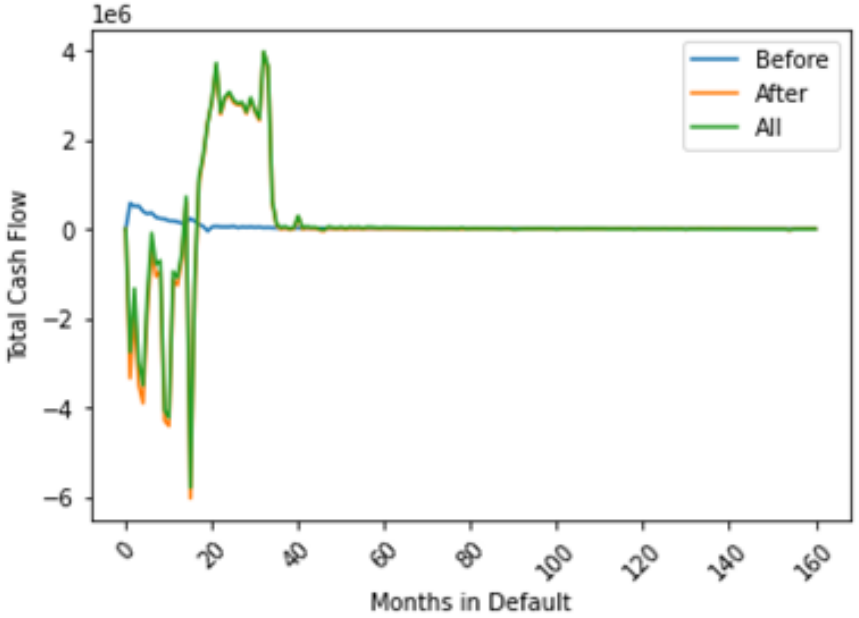


Figure A.1: Total cash flow against months in default for loans before and after deferred payment plan, and the full loan

Table A2: Data removal

Data Set	N	%
Full	52.4M	100
Sub-sample	899,968	1.72
Non-defaulted loans removed	40,600	4.51
Repurchased loans removed	39,447	97.16
Modified loans removed	27,528	69.78
Deferred loans removed	18,119	65.82
Loans for which termination is the same as default month removed	17,439	96.25

Furthermore, we consider the following procedure for removing certain variables. First, we look at correlated variables. We have the variables *Original Loan-to-Value (LTV)* and *Original Combined Loan-to-Value (CLTV)*. The first is the ratio between the loan at origination and the value of the property. The latter indicates the ratio between all outstanding loans and the property's value. The *Original CLTV*, therefore, captures at least the same amount of risk as the *Original LTV* and is highly correlated with the *Original LTV*. For this reason, we exclude the variable *Original LTV*. The *Estimated LTV (ELTV)* is the monthly LTV, as 65% of the data is missing and the *ELTV* is also slightly contained in the *Original CLTV*, we delete this variable. Next,

the *Current Actual Unpaid Principal Balance (UPB)* is calculated by the sum of the *Interest-bearing UPB* and *Non-interest-bearing UPB*, due to the correlation we remove the *Interest-* and *Non-interest-bearing UPB*. The *Current Actual UPB* is itself highly correlated to the *Exposure at Default*. We disregard the first as a risk driver, however, do still need it to calculate the LGD, explained in Subsection 3.3.

Then, variables with missing values are treated. The variable *Number of Units* has one missing value, we set this value to the mode of similar loans, based on *Exposure at Default*, *Postal Code* and *Property Type*. Furthermore, variables with more than 70% missing data or variables for which only one unique value exists, are removed. Moreover, the *Channel* variable is not a property of the loan itself, and is, therefore, excluded. Lastly, variables related to the deferred payment plan and modified loans are removed as well, as these are irrelevant due to the construction of our data set. There are three variables that describe the location of the mortgaged property. Since the three-digit *Postal Code* has no missing values, we only include this variable.

It is important that the remaining risk drivers are not highly correlated. A strong correlation can have a negative effect on the predictor standard error. Typically, a correlation coefficient with an absolute value of > 0.7 is considered to be a strong correlation between risk drivers (Ratner, 2009). When deriving the correlations between risk drivers, the strongest correlation is 0.54, which is between *Loan Purpose - No Cash Out* and *Original Debt-to-Income (DTI) Ratio*. Based on this, multicollinearity should not be a problem in the models. The full correlation matrix can be found in Figure A.2.

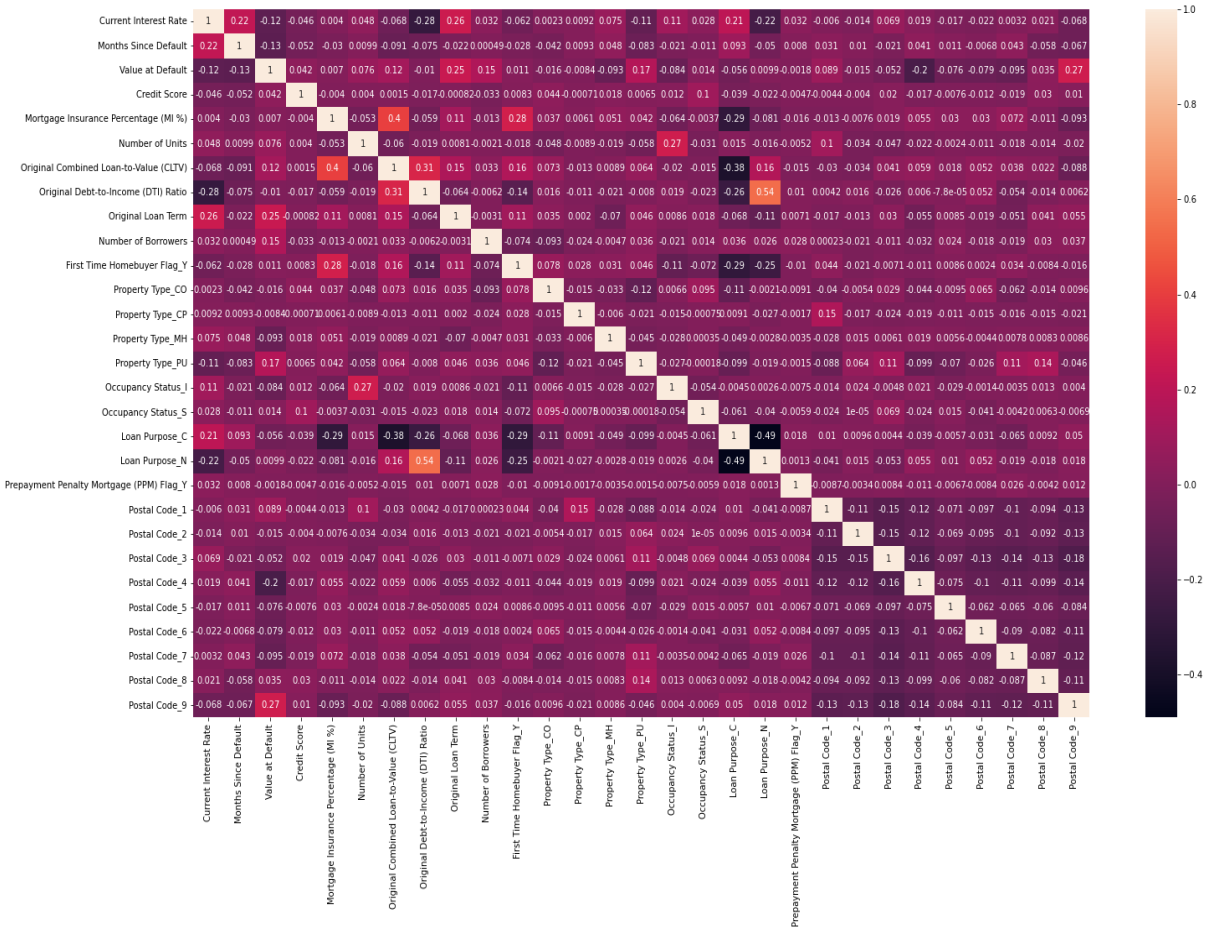


Figure A.2: Correlation matrix of risk drivers

B Results

Table B1: Grid search Random Survival Forest

Hyperparameter	Description	Values
n_estimators	The number of trees in the forest	[**100**, 200, 300]
min_samples_split	The minimum number of samples required to split an internal node	[4, **6**, 8]
min_samples_leaf	The minimum number of samples required to be at a leaf node	[**1**, 3, 5]
max_depth	The maximum depth of the tree	[6, **10**, 12]
max_leaf_nodes	The maximum number of leaf nodes of the tree	[10, **12**, 14]

Note: Bold values indicate that value was chosen by the grid search

Table B2: More complex example of LGD prediction with Survival Analysis

(a) Outstanding amounts

Loan	$t_1 = 0$	$t_1 = 1$	$t_1 = 2$	$t_2 = 3$
1	100	100	100	<u>100</u>
2	100	80	50	<u>40</u>
3	50	50	<u>50</u>	<u>50</u>
4	50	40	<u>0</u>	<u>0</u>
5	30	<u>10</u>	<u>10</u>	<u>10</u>
6	20	<u>0</u>	<u>0</u>	<u>0</u>

(b) Current Survival Analysis realized predicted LGD

Loan	$t_1 = 0$	$t_1 = 1$	$t_1 = 2$	$t_2 = 3$	Workout Length[1]
1	1	1	1	<u>1</u>	3
2	1	0.80	0.50	<u>0.40</u>	3
3	1	1	<u>1</u>	<u>1</u>	2
4	1	0.80	<u>0</u>	<u>0</u>	2
5	1	<u>0.33</u>	<u>0.33</u>	<u>0.33</u>	1
6	1	<u>0</u>	<u>0</u>	<u>0</u>	1
Average Realized Predicted LGD	1	0.66	0.47	0.46	
SA Predicted RR_{t_1,t_2}[2]	0.54	0.31	0.04	-	
SA Predicted LGD_{t_1,t_2}[3]	0.46	0.35	0.43	-	
Scaling Factor	1	0.98	1.42	-	

(c) Survival Analysis adjusted predicted recovery[4]

Loan	$t_1 = 0$	$t_1 = 1$	$t_1 = 2$	$t_2 = 3$
1	0.54	0.30	0.05	<u>0</u>
2	0.54	0.30	0.05	<u>0</u>
3	0.54	0.30	<u>0</u>	<u>0</u>
4	0.54	0.30	<u>0</u>	<u>0</u>
5	0.54	<u>0</u>	<u>0</u>	<u>0</u>
6	0.54	<u>0</u>	<u>0</u>	<u>0</u>
Average	0.54	0.20	0.02	0

(d) Final adjusted predicted LGD[5]

Loan	$t_1 = 0$	$t_1 = 1$	$t_1 = 2$	$t_2 = 3$
1	0.46	0.70	0.95	<u>1</u>
2	0.46	0.50	0.45	<u>0.40</u>
3	0.46	0.70	<u>1</u>	<u>1</u>
4	0.46	0.50	<u>0</u>	<u>0</u>
5	0.46	<u>0.33</u>	<u>0.33</u>	<u>0.33</u>
6	0.46	<u>0</u>	<u>0</u>	<u>0</u>
Average	0.46	0.46	0.46	0.46

[1] Workout Length: the number of months the loan was in default
[2] SA Predicted RR_{t_1,t_2}: Unadjusted predicted recovery rate between time t_1 and t_2
[3] SA Predicted LGD_{t_1,t_2}: Predicted LGD at $t_2 = 3$ when we look at current time t_1, $LGD_{t_1,t_2} = LGD_{t1}$ - RR_{t_1,t_2}
[4] Survival Analysis adjusted predicted recovery: SA Predicted RR_{t_1,t_2} × Scaling Factor, or 0 if loan is already resolved
[5] Final adjusted predicted LGD: Current Survival Analysis realized predicted LGD - Survival Analysis adjusted predicted recovery, i.e. adjusted predicted LGD at $t_2 = 3$ when we look at current time t_1
Note: Underlined values indicate the case is resolved

Table B3: The Naive Approach

(a) Naive recovery rates

Loan	$t_1 = 0$	$t_1 = 1$	$t_1 = 2$	$t_2 = 3$
1	0	0	0	$\underline{0}$
2	0.60	0.40	0.10	$\underline{0}$
3	0	0	$\underline{0}$	$\underline{0}$
4	1	0.80	$\underline{0}$	$\underline{0}$
5	0.67	$\underline{0}$	$\underline{0}$	$\underline{0}$
6	1	$\underline{0}$	$\underline{0}$	$\underline{0}$
Average	0.54	0.20	0.02	0

(b) Naive predicted LGDs

Loan	$t_1 = 0$	$t_1 = 1$	$t_1 = 2$	$t_2 = 3$
1	1	1	1	$\underline{1}$
2	0.40	0.40	0.40	$\underline{0.40}$
3	1	1	$\underline{1}$	$\underline{1}$
4	0	0	$\underline{0}$	$\underline{0}$
5	0.33	$\underline{0.33}$	$\underline{0.33}$	$\underline{0.33}$
6	0	$\underline{0}$	$\underline{0}$	$\underline{0}$
Average	0.46	0.46	0.46	0.46

C Programming Code

Sample_svcg_data: This code loads the monthly performance data and cleans it as explained in Table A2.
Sample_orig_data: This code loads the origination data and cleans it by only keeping the same loans as in the monthly performances data set.
nonTimeVaryingDataDummies: This code converts the categorical variables to dummy variables and concatenates the monthly performances data set with the origination data set.
nonTimeVaryingDataDummies Statistics: This code looks at the summary statistics of the data set.
Lin Regr with SA data: This code implements the Regression-Based model and the Stepwise Selection procedure.
Lin Regr Calibrated: This code implements the Calibrated Regression-Based model.
Cox PH Updated: This code implements the Cox Proportional Hazards model.
Cox PH Calibrated: This code implements the Calibrated Cox Proportional Hazards model.
RSF: This code implements the Random Survival Forest model.
RSF calibrated: This code implements the Calibrated Random Survival Forest model.