MS-E2177 Seminar on Case Studies in Operations Research

# S-Bank: Allocation of the Sales Price of the Credit Collection Portfolio

Final report

Joonatan Honkamaa, Project Manager
Juhana Vehmas
Mattias Österbacka
Petri Koivisto

June 2, 2023

# Contents

# 1 Introduction

S-bank has 3.2 million customers and according to preliminary figures an operating profit of almost 45 million EUR in 2022 [1]. A significant part of S-bank's business is related to lending, having a comprehensive collection of different loans and credits. Examples of these are secured loans such as mortgages, unsecured loans and card credits. Loan products always involve some degree of risk. Thus, in order to make well justified decision it is important to have appropriate tools for measuring and managing risk.

This project seeks to improve S-bank's credit risk modeling by exploring possible solutions for a specific situation, where some information about past loans is lost.

## 1.1 Background

The life cycle of each credit starts from identifying and pre-qualifying new customers. For these potential new customers loan offers are created. If an agreement is made, the loan enters the Account Management phase. If everything goes according to plans and the repayment plan is fulfilled by the customer, the life cycle of a loan comes to an end. However, in the Account Management phase, there might occur some problems with the payments, and the loan enters the pre-delinquency phase where the loans of customers, which have the probability of default in the near future, are proactively managed. In the occurrence of late payments, the bank starts collections without legal action, so called soft collection.

If the payments are late for over 90 days, a default occurs. After a default, the bank can still try to collect recoveries by themselves, move them to collection agency for legal proceedings or sell the defaulted loan to a collection agency.

This project focuses on the phases after the event of default, and more specifically the scope is in the loans that have been sent to a collection agency. The bank can assess the risks of giving loans by calculating expected loss (EL). Expected loss can be calculated as

$$\mathrm{EL} = \mathrm{PD} \cdot \mathrm{EAD} \cdot \mathrm{LGD}, \tag{1}$$

where

- PD is the probability of customer's default,
- EAD is exposure at default, i.e. the balance of the loan in EUR at default,

- LGD is loss given default, i.e. percentage of the loan that bank is not able to collect after default.

Banks are allowed to calculate their own risk parameters, this is known as internal rating-based approach (IRBA). There are two types of IRBAs that can be applied, foundation IRBA, and advanced IRBA. Both allow the banks to calculate their own PD estimate. The advanced IRBA also allows banks to estimate EAD and LGD [2]. This process requires supervisory regulation and approval [3].

## 1.2 Objectives

The main objective of this project is to develop a method to allocate the sales price of a portfolio of defaulted loans in order to make better LGD estimations, i.e. the financial loss a bank ultimately incurs when a borrower stops making loan payments. LGD is in practice estimated based on previous observations of LGD-values of loans. However, data is not easy to collect because the actual costs of a defaulted loan are often scattered over many parts, some of which are not easy to track. On top of that, the bank does not receive any information on loans sold to collection agencies, and thus the estimation has to be made based on total price of the portfolio.

The case team was expected, following the guidance and data provided from S-Bank,

1. to analyze different allocation methods for the portfolio sale price to individual loans for LGD-modelling purposes.

In order to achieve this, the team was required to

2. develop a justified and documented model to estimate a simulated cash flow for loans that have been sent to collection agency,
3. investigate the effect of different characteristics of the loan and lender to the sale price of the collection portfolio,
4. analyze potential effects of asymmetrical information to the sale price of collection portfolios (collection agencies have better picture of the lenders overall financial situation).

# 2 Literature review

There is plenty of scientific literature on credit risk estimation. However, in the expected loss formula (1), the probability of default PD has been studied more extensively than LGD [4]. In the next section a brief literature review on LGD estimation methods is presented. Section 2.2 covers some of the legislation from European Banking Authority [5] about LGD modeling relevant to the scope of this project.

## 2.1 Loss given default

The LGD or loss given default is defined as a percentage of exposure at default that the bank is not able to collect after a loan has defaulted. A bank should calculate an LGD estimate for every defaulted and non-defaulted loan individually. These estimates should be based on bank's own historical data of defaulted loans. [5]

Due to the nature of defaults, the distribution of LGD is often described as bimodal [6]. The person who has defaulted usually either recovers and continues paying off the debt, or stops completely. Previous studies has placed the left mode in the range of 0.00 to 0.30, and the right mode at 0.70 to 1.00. To mimic the bimodal distribution Hlawatsch and Ostrowski [6] propose using a combination of two beta-distributions, one for each mode.

A very popular older LGD model is the LossCalc model introduced in 2002 by Gupton and Stein [7]. It is a statistical model for bonds, preferred stock, and loans. The model is validated for the US market. There are four groups of explanatory variables used in the model: debt-type, firm specific capital structure, industry, and macroeconomic factors.

A model that continues on the LossCalc model is proposed by Hamerle et al. [8]. They estimate recoveries in a similar way to LossCalc. Their model also makes use of a logit transformation of the LGD. The model also has a time component which the previous model lacks.

Without predictive information, LGD can be estimated by just taking mean of the observed values. In practice, however, the bank has a lot of information about every loan which can be used to predict the LGD-value. A simple approach to predict the value is to divide the loans to different groups, and use the mean of the group as an estimate. Another way is to construct regression models. Different types of regression models include linear regression, Tobit regression, beta regression [9], inflated beta regression and censored gamma regression. [10] The two approaches mentioned above can also be combined.

A third approach is to use a two-stage model, where the probability of LGD-value being zero is first estimated, and a regression model is applied to data which only has LGD-values larger than zero. [11]

## 2.2 Legislation concerning LGD models

Due to legislation, a bank should be able to demonstrate that the the model used to estimate LGD and the underlying assumptions of the model are appropriate for the purpose [5]. In the scope of this project, this means that any mathematical methods proposed in the project have to be justifiable and well-documented.

According to [5], any payments to a defaulted loan should be discounted to the moment of default. Moreover, LGD calculation should also take into account all interests and collection fees of the loan both before and after the event of default.

# 3 Data and Methods

This chapter will present the data used, the linear model that is implemented, as well as how the allocation of the sales price is done.

## 3.1 Data

### 3.1.1 Description

In total, we were given data from approximately 30 000 defaulted loans since 2015. The data contains of two different types of loans and in this report these are called *Type 1* and *Type 2* loans. Both loan types are unsecured credits, and the credit limit is higher for *Type 1* loans. The data of a single loan starts when the loan is transferred to collection and ends when no more information is available, for example when the loan is sold to collection agency. The most relevant information in this data is the balance of the loan at the time of default (EAD), the date of default, the current status and balance, and the payment history.

This data does not contain information about the accruing interest and other expenses applied to the loan after the event of default. Thus, the payment history of a loan only contains information about payments that are large enough to cover the ongoing expenses, and are therefore able to amortize the remaining debt. From these 30 000 loans, approximately 36% does not have any payment history after the event of default.

Since the main objective is to allocate the sales price of a portfolio, we were also given information about two loan portfolios. One of these portfolios was sold to the collection agency in the spring of 2022 and the other was not. These datasets were given for S-bank by the collection agency. The dataset for the sold portfolio contains 498 loans of *Type 1* and 1148 loans of *Type 2*. In total approximately 61% of the loans in this portfolio does not have payment-data before it was sold.

This portfolio-data shows the balance of these loans at the time of the sale and the sales price of the portfolio. This total balance at the time of the sale is approximately 10% greater than the total balance calculated from the S-bank's own dataset described earlier. This is due to the fact S-bank does not have complete information about the expenses after the event of default.

### 3.1.2 Preprocessing

Before more comprehensive analysis, the data had to be preprocessed. The data was given in Excel spreadsheets. The data was split into two groups: loans with and without payment history. For each of those loans with payment history, a data point was added to the beginning, showing the balance of the loan at the time of default. This dataset contains information about the loan type, the payments collected, the collection date, the current balance and if it was part of the portfolio sold in spring 2022. It also tells if the loan has been sold, fully paid or if the collection is still ongoing or terminated due to other reasons. The data regarding loans without payment history, contains the same information but of course excluding payment collection information.

We focused on the loans that were sold to the collections agency in the spring of 2022. For some reason, for some loans there were payments recorded after the loan was sold. In order to keep the analysis consistent, these payments were removed from the data such that all of the loans in the portfolio only have payments made before the time of sale.

## 3.2 Linear model

A linear model to predict the future cash flows for loans was developed in the project. When examining the balance for a single loan, an almost linear dependence between the balance and time was noticed. We then determined that a simple linear model that determines how the balance of a single loan will behave in the future will be chosen.

The explanatory variables are the the dates of payment for a single loan. The response variables are the balance of the loan at the date of payment. Using this model makes it easy to extrapolate and predict future cash flows.

In addition we chose a period of 5 years from the point of default, as the time where we predict cash flows. The reason why we stop after 5 years is that people rarely pay after that [12]. Thus, there is a hard cut off after 5 years. Figure 1 shows this concept.

From Figure 1 we notice that the loan marked with yellow is expected to be paid in full within a five year period, while the loan marked in red will not. To use these cash flows to predict LGD for loans, the predicted cash flows have to be discounted on a yearly basis. This means that even though a loan is predicted to be paid in full, the LGD will not be 0, due to the cash flows being discounted.

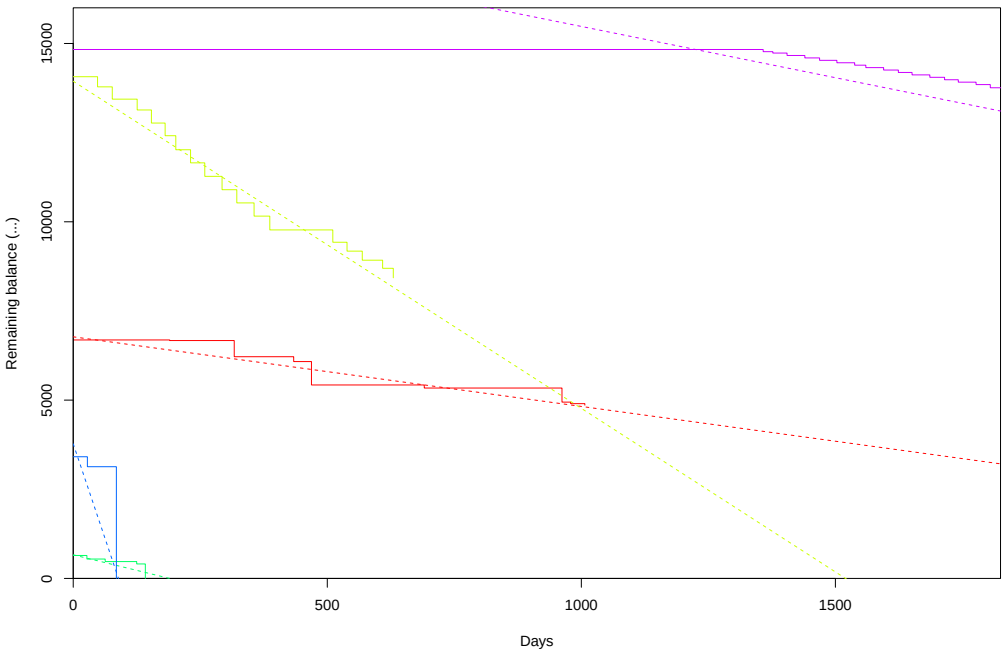


Figure 1: The balance of five randomly chosen loans, and their corresponding linear model, over a five year period.

The future cash flows are discounted using the following formula:

$$DCF = \sum_{i=1}^{5} \frac{CF_i}{(1+r)^i} \tag{2}$$

where $CF_i$ is the cash flow during year $i$, and $r$ is the discount rate, in this model 0.05. Future cash flows are discounted since a given amount of money today is worth more than the same amount tomorrow.

A problem that could arise with this approach is that the debtor pays rapidly, and then stops before the loan is sold. The model could then potentially predict that the loan would be fully paid before the loan actually was sold, which is incorrect. Thus an additional constraint to the linear model is added that states that the regression line has to pass through a data point that is the current balance at the date when it is sold. It might seem like an "artificial" constraint to add, however there is information in the balance amount at the date of transaction. This also makes sure that the loans are never predicted to be fully paid before they are sold.

Figure 2 shows the final linear model with this added constraint, in addition to the loans being shifted to the sell date. The figure clearly shows (from an albeit small sample size) why the linear model works well, and why the line is constrained to pass through the remaining balance amount on the date of transaction. Figure 1 and 2 do not show the discounting.

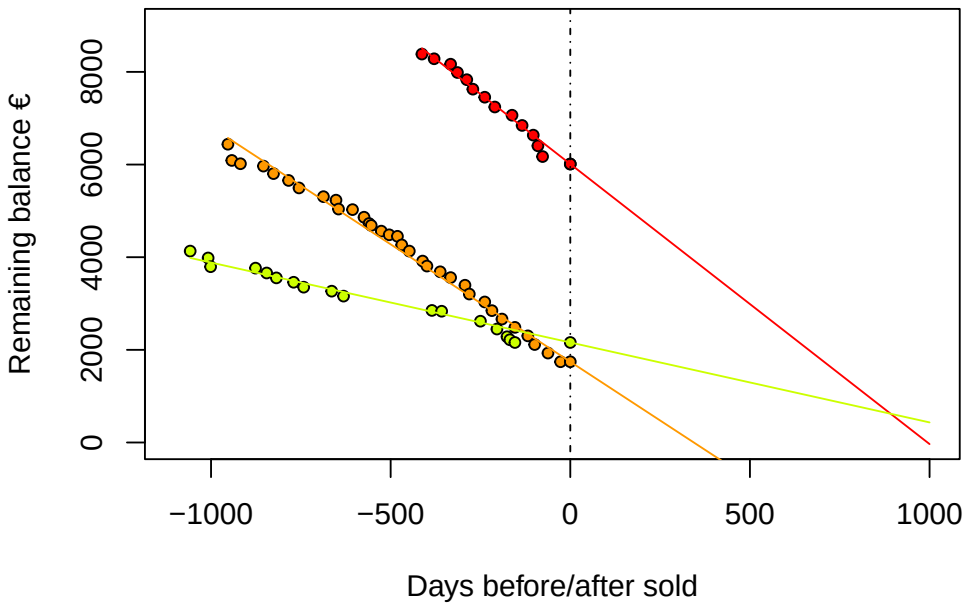


Figure 2: The balance of three randomly chosen loans from the sold portfolio, and their corresponding linear models. Loans sold at Day 0.

A problem with this model is that a payment history is needed for it to be possible top fit a linear model. In the given data set, a large portion of the loans do not have any payment history. Another way of dealing with the loans that lack payments is thus needed. There are several ways of tackling this problem, other information regarding the loan in question could perhaps be used, but the data set provided lacks sufficient information for this approach.

Since the linear model is based on the idea that people will behave in the future as they have done in the past, ie. they will continue to pay at the same rate as they have. Using this way of thinking the future cash flows for the loans without payment history would be 0. There are, however,

some problems with this approach, one is that some loans happen to be sold very quickly after defaulting. It is then reasonable to think that the lack of payment history is simply due to the short time that the defaulted loan was in the control of the bank. It is also unreasonable to think that the collections agency, which has access to much more information regarding the financial state of debtor, would buy a portfolio mostly consisting of loans that would not be paid back at all. Thus our approach is not to directly assume all loans without payment history will not have a single future payment.

A final, more naive approach is to simply use the loans with future cash flows to calculate an average LGD for those loans, and then assign that LGD to the loans without payments. This approach has the advantage that it is not as polarizing as the previous approach, however, the information it gives us is not much. It is also most likely very optimistic towards loans with no payment history. This issue is discussed more in Section 5.

### 3.2.1 Different versions of linear model

There are multiple options how linear model can be implemented in practice, and multiple decisions are to be made. Firstly, we decided to consider only linear models where the prediction line passes through the point at current date and balance. This is justified; if we constructed the model as a linear regression model through the payment data points, the model could in some cases even predict that the loan is completely paid when it is sold, even if there is balance left.

Secondly, it has to be chosen how the model prediction line is forced to pass through current balance and date. In the case of linear regression, the model can simply be built as a linear regression model that passes through that data point. This type of model is referred to in this report as 0-version of linear model. However, especially if there are just few data points, such a model can be significantly more conservative than it should. This is because the last point does not even theoretically fit to the same line with the others, because it is not a balance after the payment but a balance at the moment when the loan is sold, which is a random point from the perspective of payments. This means, for example, that the last two data points always have the same balance. This phenomenon can be taken into account for example by determining the slope of the prediction line based on the actual payment data points, and determining the intercept of the line such that it passes through the current balance and date. This leads to a more optimistic model, which is referred to in this report as 1-model.

Thirdly, linear regression is not the only option. Due to the decreasing nature of debt account balance, a linear model can also be achieved by just drawing a line between the first and last data point. These two options are called in this report as R-version (regression) and L-version (line). Thus we have in total four linear model versions, R0, L0, R1 and L1, which are later compared in this report.

## 3.3 Allocation of sales price

The sales price of the portfolio is to be allocated to the sold loans. This is done using the predicted future cash flows from the linear model. We were given the sales price of a one portfolio sold to the collections agency in the spring of 2022. After the predicted cash flows are calculated and discounted to the time of the sale, we can allocate the sales price for an individual loan $i$ so that

$$S_i = \frac{S_P}{DFC_P} DFC_i, \tag{3}$$

where $DFC_i$ is the discounted future cash flows for loan $i$, $DFC_P$ is the sum of the discounted future cash flows for the whole portfolio, $S_i$ is the sales price allocated for loan $i$ and $S_P$ is the portfolio sales price.

This scaling is done since the sum of all future cash flows for the portfolio will not be exactly the same as the sales price, and it will remove some potential systematic error. It will not however remove any problems related to the relative differences between cash flows for different loans.

If the sum of the discounted future cash flows for the whole portfolio ($DFC_P$) is estimated to be lower than the real portfolio sales price $S_P$, an additional constraint might be needed: The amount allocated cannot exceed the loan balance at the time of sale. This creates a small surplus that should be allocated for the rest of the loans.

# 4 Results

In this section the results are presented. The results consist of the model validation, and how the allocation is done with the help of the model.

## 4.1 Linear model validation

The purpose of validating the linear model is to be able to choose the best among possible linear models, and to get an estimate of how well the model performs. We perform three types of validation analysis for our models.

### 4.1.1 All loans validation

The first validation method is based on dividing the payment history of each loan into two parts: the training set and the test set. This is done in the following way: filtering the loans such that only loans with multiple payments more than 180 days before the last payment are left. Those payments are then used to construct a linear model, and the result is validated at the time of the last payment. This idea is illustrated in Figure 3.

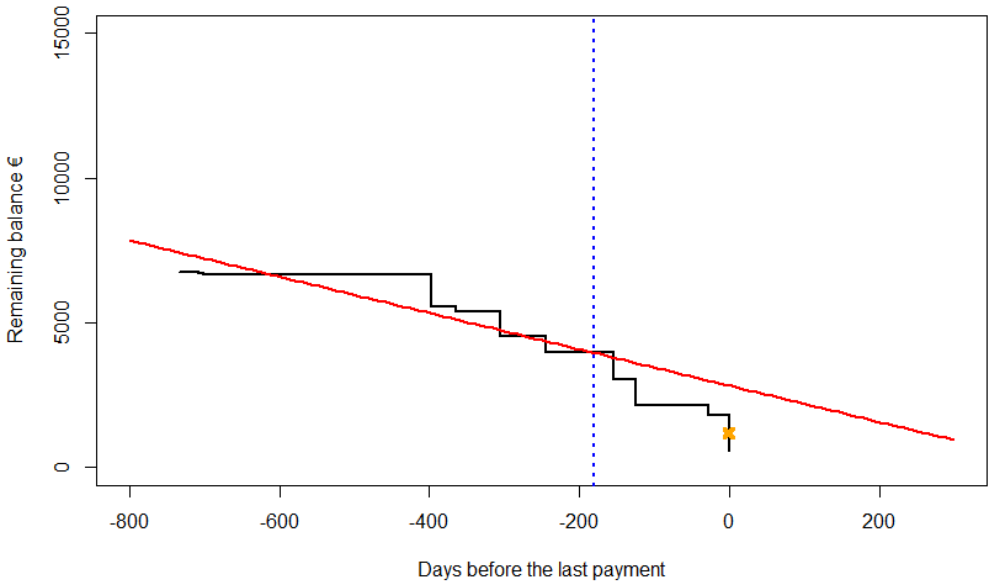


Figure 3: Representation of how a model is constructed when validating the data. Note how the prediction line is based on only information before -180 days. The orange cross represents the point to which the prediction of the model is later compared to.

The accuracy of prediction is then validated by comparing the predicted balance of the model to the midpoint of actual balances before and after the last payment. This is natural comparison balance, because the prediction line passes through the current balance which is in a sense random point with respect to payment points, and thus it can be also thought of as a midpoint. In case of completely linear payment this choice of comparison balance is the correct choice. This is illustrated in Figure 4.

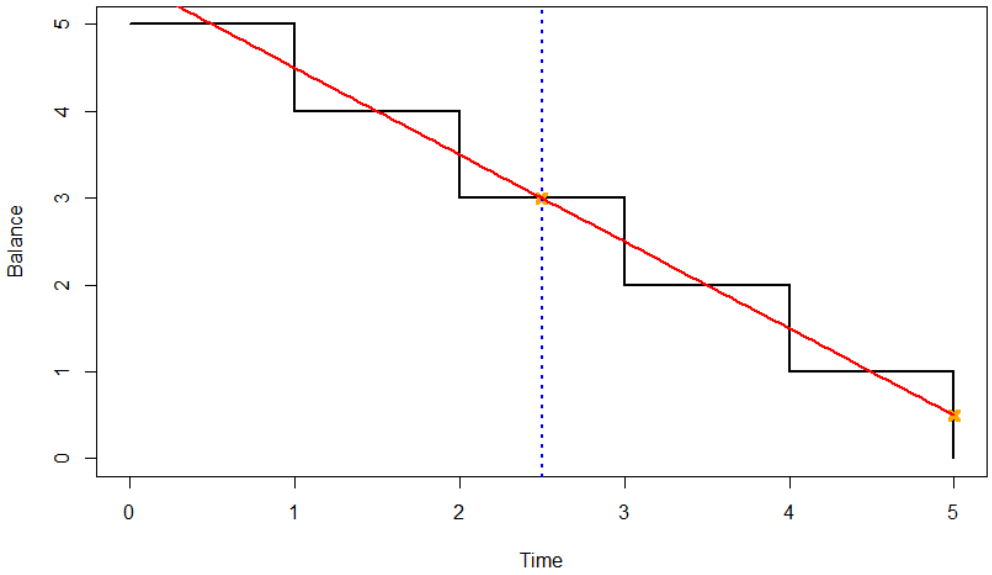


Figure 4: Illustration of why the chosen comparison balance is the correct choice in case of completely linear payment. Note that in the picture the information cut is at the exact midpoint of two payments. If it was earlier or later, it would result to more optimistic or pessimistic predictions, respectively.

The difference between the prediction and comparison balance is then divided by initial balance of the loan at the time of default, resulting in a number between $-1$ and 1. This validation process is then repeated for all loans and all four model versions discussed earlier. The resulting histograms of these validations are presented in Figure 5, and statistical measures of the distributions are presented in Table 1.

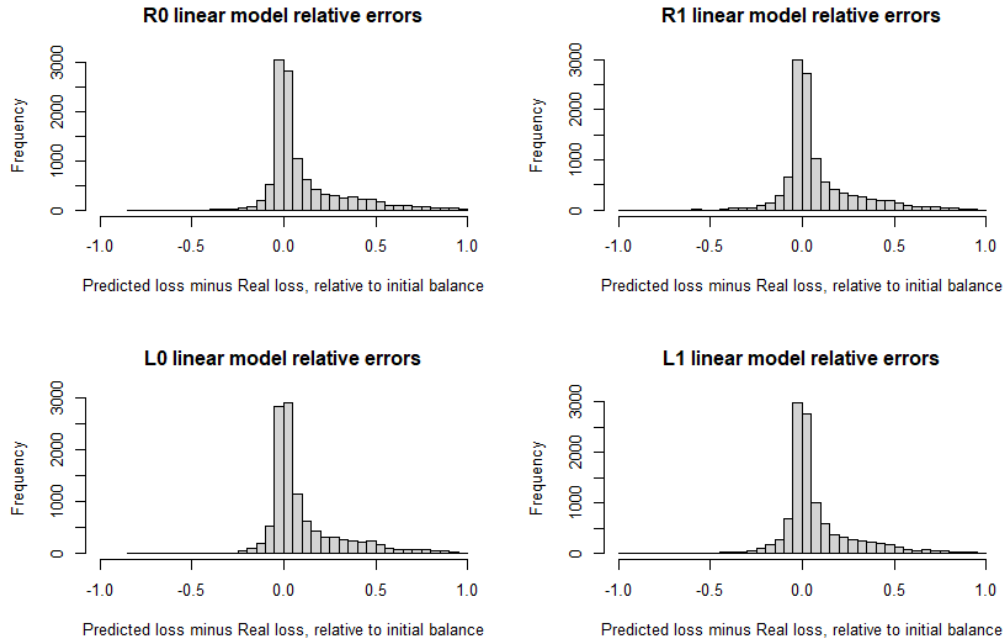


Figure 5: Histograms of relative errors for the four linear models, calculated for all loans.

| | **Mean** | **Median** | **Standard deviation from 0** |
|---|---|---|---|
| **Model R0, all loans** | 0.105 | 0.022 | 0.231 |
| **Model R1, all loans** | 0.079 | 0.015 | 0.219 |
| **Model L0, all loans** | 0.104 | 0.024 | 0.228 |
| **Model L1, all loans** | 0.075 | 0.014 | 0.216 |
| **Model R0, Type 1** | 0.072 | 0.017 | 0.174 |
| **Model R1, Type 1** | 0.065 | 0.017 | 0.174 |
| **Model L0, Type 1** | 0.077 | 0.022 | 0.175 |
| **Model L1, Type 1** | 0.0635 | 0.018 | 0.172 |
| **Model R0, Type 2** | 0.115 | 0.024 | 0.244 |
| **Model R1, Type 2** | 0.083 | 0.014 | 0.230 |
| **Model L0, Type 2** | 0.111 | 0.026 | 0.241 |
| **Model L1, Type 2** | 0.078 | 0.0123 | 0.226 |

Table 1: Statistical measures of the relative error distributions of the four linear model versions described in Section 3.2.1, calculated for different loan types.

Overall, the results suggest that linear models have predictive power, and

there are no significant differences between the model versions. No visible differences can be seen from histograms. However, the statistical measures slightly suggest that 1-versions of linear models have both slightly more realistic mean and are slightly more accurate than 0-versions. There are no significant observable differences between R- and L-versions of models. Despite these observations, we will use the R0-version of the model in the analysis from this point on in this report. More analysis is needed to determine how the model versions perform related to each other.

### 4.1.2 Spring 2022 data set validation

The second validation method uses the same basic idea as the first method, but we use a special data set of a portfolio not sold in spring 2022. We predict the future of the loans based on payment history before the decision in spring 2022, and validate the results based on the actual balance at the time of the end of data. In this dataset, we have a large amount loans with payment information from a time period of over three years, because the portfolio mostly consists of loans defaulted in 2019.

The histogram of relative errors is shown in Figure 6, and statistical measures of the distribution are presented in Table 2. We can see that the model performs better with this dataset. This is most likely because of the good quality and longer payment history of the data.

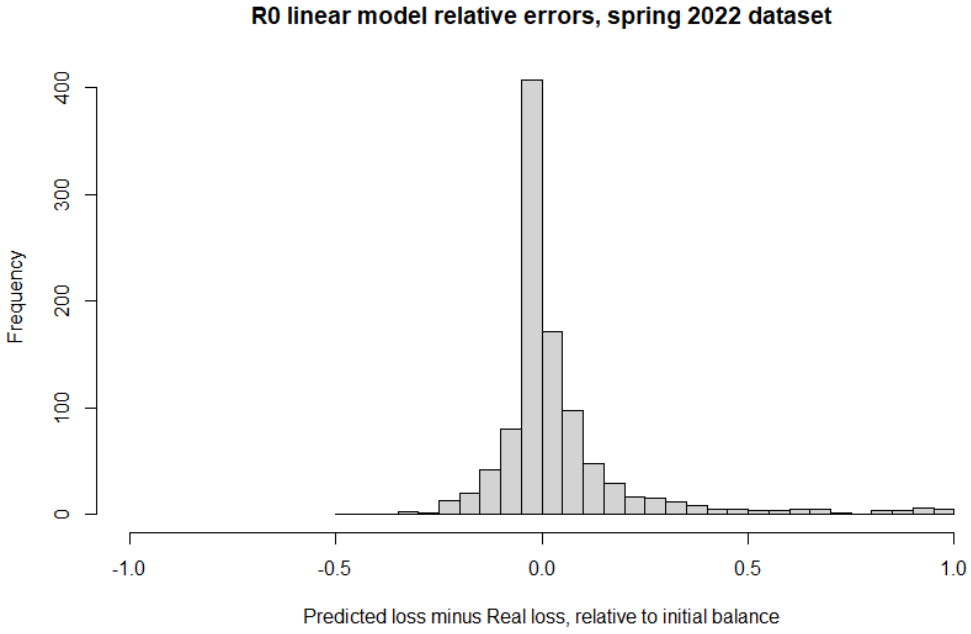


Figure 6: Histogram of relative errors of R0-version of the linear model for the spring 2022 dataset.

| | **Mean** | **Median** | **Standard deviation from 0** |
|---|---|---|---|
| **Model R0, all loans** | 0.047 | -0.000 | 0.188 |
| **Model R0, Type 1** | 0.051 | 0.000 | 0.183 |
| **Model R0, Type 2** | 0.045 | -0.004 | 0.190 |

Table 2: Statistical measures of the relative error distributions of the R0-version of the linear model for different loan types in the spring 2022 dataset.

### 4.1.3 R-squared validation

The third validation method is a simple R-squared analysis of payment history data points. Figure 7 shows the R-squared values for the linear models in the sold portfolio. We see that the values are close to 1 which suggests that approximating the payment data points with lines is justified.

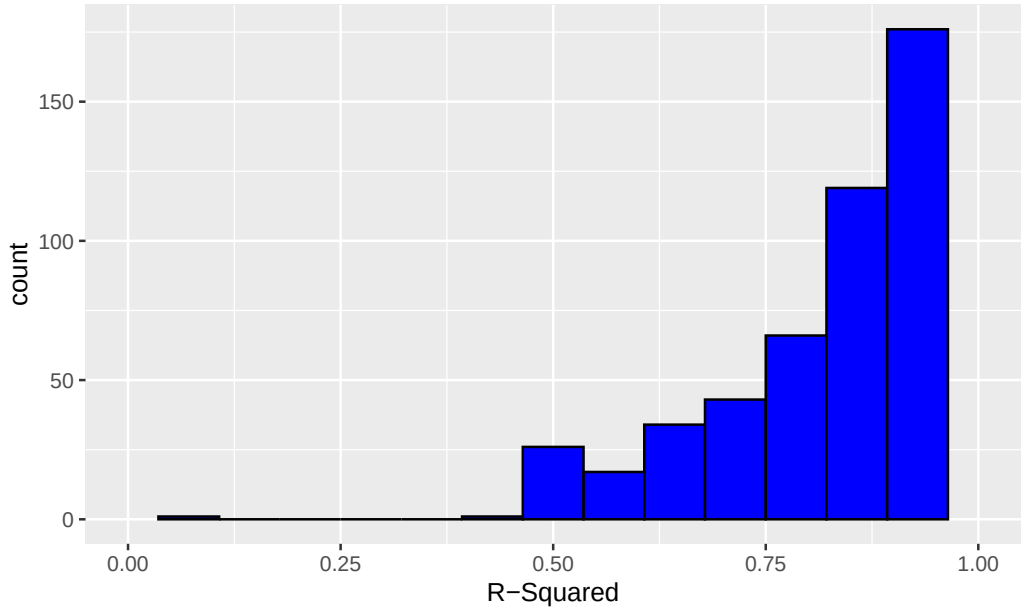


Figure 7: Histogram of R-Squared values for the linear models in the sold portfolio.

## 4.2 Sale price allocation

The sales price of one portfolio was allocated as described in Sections 3.2 and 3.3. The first step was to fit the linear models for each loan with payment history and discount the resulting future cash flows to the time of sale using equation (2). After that, the loss percentage from the balance at time of sale was calculated for each loan $i$. That is

$$Loss\%_i = \frac{B_i - DFC_i}{B_i}, \tag{4}$$

where $B_i$ is the balance of loan $i$ at the time of sale. The resulting distribution is presented in Figure 8. As we can see the distribution is fairly bimodal with peaks close to 0% and 100%. It can be also noticed that the sold portfolio contains fewer number of *Type 1* loans. Also, the distributions are rather similar for both loan types.

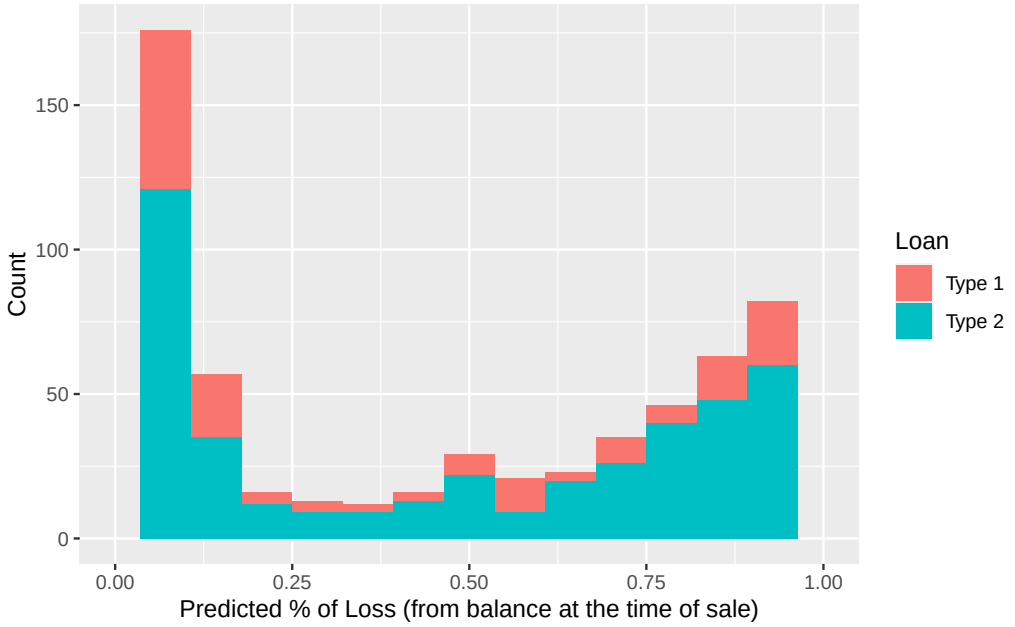


Figure 8: Histogram of predicted *Loss*% for loans with payment history.

Like discussed in Section 3.2, different approach is needed for the loans without any payment history. We assumed that these have the average *Loss*% of those loans with payments. The average *Loss*% for *Type 1* and *Type 2* loans were approximately 45% and 53%, respectively. These averages were used for the rest of the loans in the portfolio and the resulting histogram for all loans is presented in Figure 9. Since large portion of loans in this portfolio were loans without any payment history, relatively high peak in the middle of the distribution can be observed.

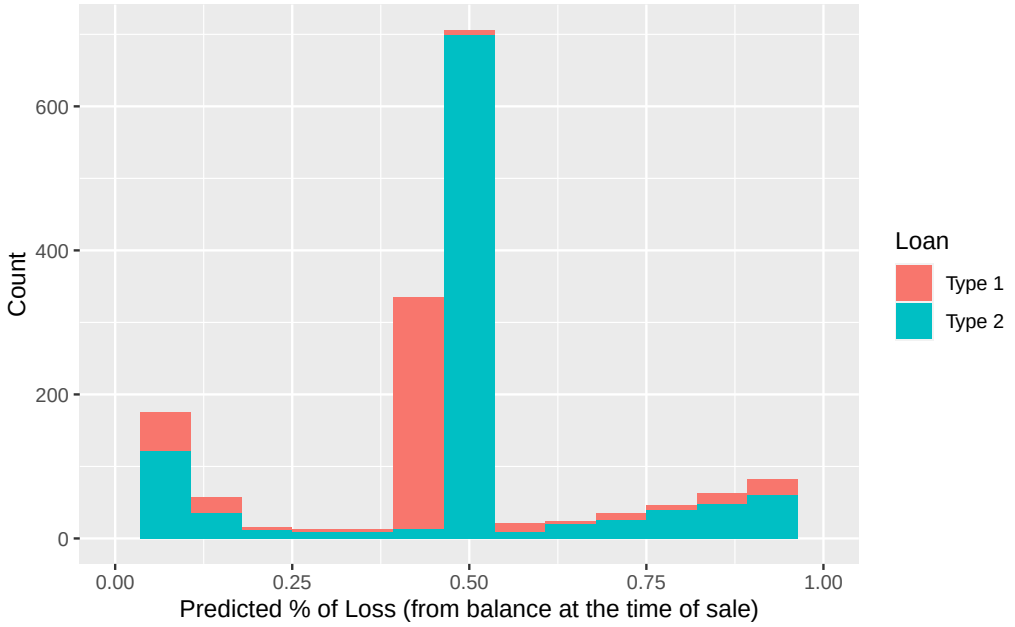


Figure 9: Histogram of predicted *Loss*% for all loans in the sold portfolio.

To investigate the predicted future cash flows from another perspective, for each loan the share from total predicted loss was calculated. The shares were sorted from largest to smallest and summed cumulatively. Figure 10 shows the cumulative loss for loans with payment history and also for all loans. As we can see, using this model, approximately 60% of total portfolio loss is due to 20% of loans. The curve is even steeper when only observing the loans with payment history. This makes sense, since the loss distribution for loans with payment history is more bimodal.

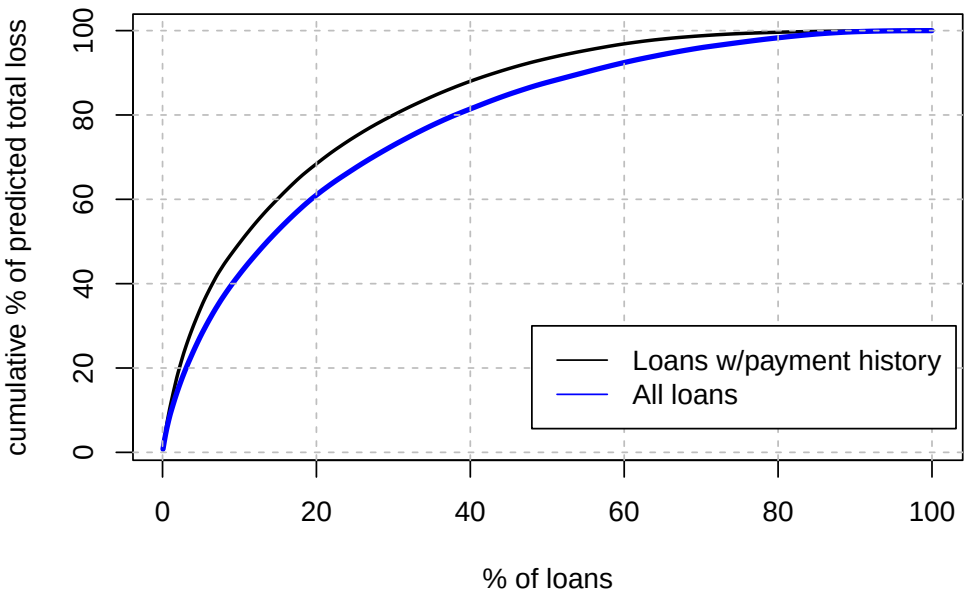


Figure 10: Cumulative % of total predicted loss as a function of % of loans. Loans are sorted from largest to smallest loss.

Since the discounted future cash flows are now determined for each loan, the sales price of the sold portfolio can be allocated using equation (3). The allocation is done separately for both loan types since we have the portfolio sales price by loan type. For both loan types the total received sales price $S_P$ is higher than the total portfolio discounted future cash flow ($DCF_P$). The real sales price for *Type 1* and *Type 2* loans were approximately 15% and 12% higher, respectively.

Because the real sales prices are higher than our model predicts, for some loans the allocated amount $S_i$ might be higher than the balance at the time of sale. To prevent this, the allocated amount is defined as the minimum of $S_i$ and $B_i$. This creates a small surplus, that needs to be allocated among the other loans. For both loan types this surplus is lower than 0.5% of the portfolio sales price. We decided to allocate these small surpluses evenly with the rest of the loans, such that the balance at the time of sale ($B_i$) is never exceeded.

The allocation is visualized in Figure 11. The black line shows if the amount

allocated is the same as the balance at time of sale. These are loans that were predicted to cause small losses in the future. The red dots are loans without payment history. Loans with the largest balance, in this portfolio, seems to be loans without payments after the event default. And because averages were used for this type of loan, even tough the customer has not reduced the amount of debt, a relatively high amount is still allocated to it.

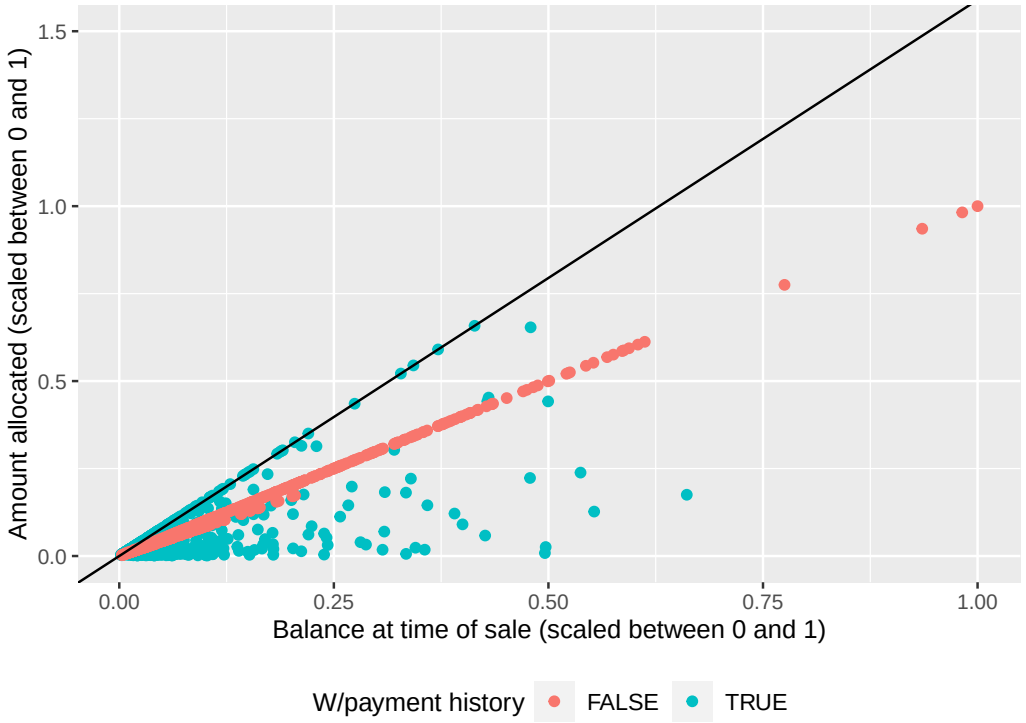


Figure 11: Amount allocated and balance at the time of sale. Loans are coloured based on the existence of payment history. Both axes are scaled separately by diving with their respective maximum observation. Points on the black line are loans where the allocated amount is equal to the balance at the time of sale

# 5 Discussion

There are several reasons to believe the constructed model performs well for loans with payment history. Firstly, it performs very well in the sense that the distribution of the predicted LGD is bimodal. This is how it is often described in literature, discussed in section 2.1. The LGD distribution of the sold loans with payment history is shown in Figure 8. Secondly, the linear model fits well to the payment history of individual loans. It is probable that people who actually want to pay back the loan will try to budget it in a way that there are regular payments which supports the choice of the linear model. Thirdly. the distribution of $R^2$ of the model on the sold portfolio has the mode near 1 and is heavily skewed towards 1 which also indicates that debtors tend to pay in a consistent manner. Fourthly, the results in the validation of the linear model were decent.

Due to the way the portfolios are sold, we only have a single sales price per portfolio. Since the sold portfolio contains both loans with and without payment history, we cannot directly analyse how well only the linear model performs for allocation. Since information is lacking for a large part of the portfolio, we cannot compare the predicted cash flows from the linear model with the sales price. This is unfortunate but a reality of how the data is.

In general, there are a lot of problems with the linear model that was implemented. They way that loans without payment history are handled is not optimal, which is partly attributed to the way the data is. However, it is very probable that when portfolios are sold to collection agencies, a large portion of the loans do not contain payment history. From this perspective some other way of handling these loans should be investigated. Other data regarding the loan or borrower that was not available to us, could perhaps be used by the bank. Though it is a fact that the collections agencies have access to much more information regarding the borrower and regarding the cash flow of Type 1 and Type 2 loans than the bank itself. This is probably one of the reasons why so many loans are bought by the agencies that lack payment information, since they know something else that indicate that the loan might be paid back in some form.

The linear model is based on the assumption that lenders will continue to pay the way that they have already paid. In some sense, this is a reasonable assumption, as discussed before. However, this assumption has problems. Firstly, the assumption implies that debtors who have not paid anything, will not do so in the future either. Using this logic however does not yield feasible results in the sense of the calculated future cash flows from the sold

portfolio would then be substantially lower than the price it sold for. If this prediction was correct, the collections agency would not pay so much for the portfolio. Secondly, there might be a problem with asymmetrical information; the collection agency might collect the debt more actively after it owns the loan completely. We have no way of knowing if this is true or in how large scale this happens.

Instead of using the past to predict the future for loans with no payment history, we assigned the loans future cash flows such that they are the same as the average for the loans with payment history, scaled to the amount left in the loan. This presents another problem as this way of allocation is most likely very optimistic. About half of the debtors who paid slowly, have, with this model, lower predicted future cash flows compared to if they paid nothing at all. This is definitely a flaw with this model, and more pessimistic way of allocation could have been used. However, even under this optimistic assumption the real sales price of the portfolio is greater than the predicted sales price with our model. This is a major problem with the model results, and there are several reasons which could explain it.

First, the interest and other costs are not included in the data. This means that loans without payment history could have the debtor paying interest every month without it showing. It also means that loans with payment history has debtors paying more than what is shown in the data, since interest is paid before the principal amount is decreased. These factors are assumed to have a negligible impact however that is not necessarily true. Since the model is linear, the data points would only be shifted if interest rate would be added.

Second, it would be redundant to discount the future cash flows, since the data used does not include the interest paid. The discounting is done to capture the idea that money today is worth more than the same amount tomorrow. Since we are talking about a timescale of years for the loans in this project, inflation and other factors impact the value of the money in these loans. However, interest rates also exist for the same reason, and since the paid interest rate is already removed in our data, we in a sense discount the cash flows twice. By removing the discounting the model would be more optimistic which would perhaps fit this data set better. Though for LGD modelling purposes, future cash flows have to be discounted according to the Basel Accords [5], and it is unclear whether acknowledging interest rates counts as discounting.

Third, as discussed in Data-section, the total balance of the sales data of the portfolio is approximately 10% greater than the total balance of the payment

history data. The model was created based on the payment history data, so this also explains a part of the difference.

Finally, the linear model itself is a bit pessimistic as discussed in the model validation -section. Due to the fact that only loans with long payment history could be included in the validation, the model might be even more pessimistic than the validation results suggest. The loans with little payment history probably have pessimistic estimates with the linear model, because the debtor has not yet started to pay the debt regularly. A bit more optimistic model could have been achieved with a different choice of the model details.

One of the objectives of the project set by S-bank was to investigate the effect of different characteristics of individual loans and lenders to the sales price of a collection portfolio. However, this was found to be highly difficult due to only having the sales prices of only one collection portfolio. Firstly, since the known sales price of the portfolio were given separately for the two different loan types, it is very probable that the loan type is the most significant explaining variable. Secondly, the data sets contained very little information about individual lenders, as explained in Section 3.1.

Other objective set by S-bank was to analyze potential effects of asymmetrical information to the sale price of collection portfolios. The asymmetric information occurs since collection agencies may have better picture of the lenders overall financial situation: for example, if a lender has defaulted loans from many banks or other institutions, the collection agency is able to estimate the lenders ability to pay to be very low. Since the model developed in this project does not cope well with loans that have no payments, it is not very good at predicting customers of this type. Thus the pricing models used by the collection agencies can probably perform better.

One solution to the problem of pricing in presence of asymmetric information for the bank would be to have it's own collection department or own a collection company. This way the bank would be able to track the cash flows from the customers more accurately. Other method would be to use more personal information of the customer to predict future payments, which understandably was not possible in the settings of a school project.

# 6 Conclusions

The main objective of the project was to analyze different allocation methods for the portfolio sale price to individual loans. Inside the main objective, we mainly focused on developing a model to estimate the cash flow for sold loans. We developed a cash flow model which was divided to four parts: linear model for loans with payment history and a constant estimate for loans without payment history, both separately for Type 1 and Type 2 loans. We also validated the results of the linear models and the linear model is of decent accuracy and it gives reasonable bimodal LGD-distribution. However, we could not validate the results of the model for loans with no payment history.

We combined the predictions of the cash flow models to allocate the portfolio sales price of a portfolio sold in spring 2022. The result of analysis was that the allocation is in many ways justifiable, but there was one big surprise: Even though the validation results for linear model were good and we made a very optimistic assumption for estimating the constant LGD-value of loans with no payment history, the sale price of the portfolio was more than 10% higher than the estimated cash flow of our model. We discussed possible explanations for this apparent paradox, and the conclusion is that it is most likely a sum of several less significant factors. We scaled the allocated values of the loans by a constant factor such that the total sum of the estimated discounted cash flows is equal to the sale price of the portfolio.

We did not completely achieve our initial goal of a justifiable model for portfolio sale price allocation. This was partially because of problems with the data and partially because we encountered surprising issues we did not have enough time to handle. However, we got partly reasonable results and our work is a very good basis for continuing the work with cash flow estimation and LGD-modelling.

# References


[1] S-pankki.fi website. Positive profit warning: S-bank group's operating profit in 2022 exceeds forecast. https://www.s-pankki.fi/fi/tiedotteet/2023/positive-profit-warning-s-bank-groups-operating-profit-in-2022-exceeds-forecast. Retrieved on 14.02.2023.

[2] Risk.net website. Internal ratings-based (IRB) approach. https://www.risk.net/definition/internal-ratings-based-irb-approach. Retrieved on 13.02.2023.

[3] Bundesbank.de website. Internal Ratings-Based Approach. https://www.bundesbank.de/en/tasks/banking-supervision/individual-aspects/own-funds-requirements/credit-risk/internal-ratings-based-approach-622830. Retrieved on 13.02.2023.

[4] E. I. Altman. Default recovery rates and LGD in credit risk modeling and practice: an updated review of the literature and empirical evidence. *New York University, Stern School of Business*, 2006.

[5] European Banking Authority. Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures. Available at 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, 2017.

[6] S. Hlawatsch and S. Ostrowski. Simulation and estimation of loss given default. *The Journal of Credit Risk*, 7(3):39, 2011.

[7] G. M. Gupton, R. M. Stein, A. Salaam, and D. Bren. LossCalcTM: Model for predicting loss given default (LGD). *Moody's KMV, New York*, 2002.

[8] A. Hamerle, M. Knapp, and N. Wildenauer. Modelling loss given default: a "point in time"-approach. *The Basel II Risk Parameters: Estimation, Validation, Stress Testing-with Applications to Loan Risk Management*, pages 137–150, 2011.

[9] X. Huang and C. W. Oosterlee. Generalized beta regression models for random loss-given-default. *Reports of the Department of Applied Mathematical Analysis, 08-10*, 2008.

[10] O. Yashkir and Y. Yashkir. Loss given default modelling: Comparative analysis. *Journal of Risk Model Validation*, 7(1):25–59, 2013.

[11] Mathworks.com website. Model Loss Given Default. https://se.mathworks.com/help/risk/comparing-lgd-models.html. Retrieved on 13.02.2023.

[12] B. Heppe. Valuation of Non-Performing Loans: Calibration of unsecured recovery curves. Available at https://www.nplmarkets.com/es/research/article/valuation-of-non-performing-loans-calibration-of-unsecured-recovery-curves-/", 2022. Retrieved on 2.4.2023.

# Self assessment

- How closely did the actual implementation of the project follow the initial project plan? W ere there any major departures and, if so, what?

  The project followed the initial project plan somewhat closely. We spent more time on the implementation of the future cash flow model than what was initially planned. This left somewhat less time for the analysis of the allocation. However, the the analysis of the allocation methods and what factors impact was not able to be done in a very thorough way due to how the data was faulty.

- In what regard was the project successful?

  The linear model for the loans with payment history seems to work reasonably well as indicated by the model validation and $R^2$ distribution. A big positive in the development process was when the LGD distribution graph was produced, and it resembled the bimodal distribution presented in literature.

  The allocation is in the end not ideal since a vast majority of the loans receive the same amount of money relative to the amount left when the default occurred.

  The project topic initially felt very difficult so when we in the end got some decent results, which is a success.

- In what regard was it less so?

  The model is unsuccessful in the sense that it does not handle the loans without payment history in any meaningful way. We can also not analyze how well the linear model for the loans with payment history performed for the allocation since the sold portfolio contained both. Several factors discussed earlier in section indicate that it might perform decently but with this data set it is not possible to check.

  The sale price allocation results of the model are not credible. Even with very optimistic assumption for loans with no payment history our predicted sales price of the portfolio is less than the actual sales price. In this regard the project was not succesful. We did not manage to create a credible portfolio allocation, and better results would have been achievable, even with the faulty data.

- What could have been done better, in hindsight? (you may analyze

this question from the roles of the project team, the client, and the teacher(s))

The project topic was in a sense not the best. It was very hard to find any relevant articles when doing the literature review which made it very hard to proceed with the project in any way. This is also why our literature review section is rather short, no relevant articles were found.

The data could also have been delivered directly when the course began, we could then have begun earlier. It would also have been very interesting to have a data set without the discussed issues above, might have been better for learning. However, there is an argument to be made that the faulty data is quite realistic.

For the role of the teachers, it felt quite weird that the team sizes were so vastly different. One team with 3 people and another team with 6. This might give vastly different workloads for the different teams depending on the team size.

From our perspective we should have probably tried to be faster when developing and implementing the model. There were (obviously in hindsight) several kinks that appeared one after another, and even during the week before the deadline we were still trying to get rid of the last few. Would we have been a little bit more effective in the beginning we could have these kinks sorted out earlier in the process, and there might have been time to explore the topic more or maybe try to additionally mitigate some of the problems the model suffers from.

For future development, the portfolio could be divided to smaller pieces. Within the linear model the loans with more than just couple data points could be predicted with a model where the first data point, the date and balance of default, is left out of the model, and only payment data points considered for the model. The intuition behind this is that people possibly tend to pay their loans linearly only after they have started paying them, and this is not the case at the moment of default, but at the moment of first payment. This is just an idea, but it could achieve better results for loans with enough payment data. Also some explaining factors, such as balance at the moment of default, initial balance or debtors age, could be added to the model. On top of that, a trial to some way weigh recent data points more than older ones could lead to a more optimistic model and be worth a try.

Within the loans with no payment data we could also divide the port-

folio more. If the explaining factors mentioned above are relevant when considering how well the loan is paid in the case of loans with payment data, they probably are also relevant in no-data case. Based on the payment history we have, we also possibly could perform some kind of analysis of how likely it is that a loan with no payment history starts being paid in the near future.