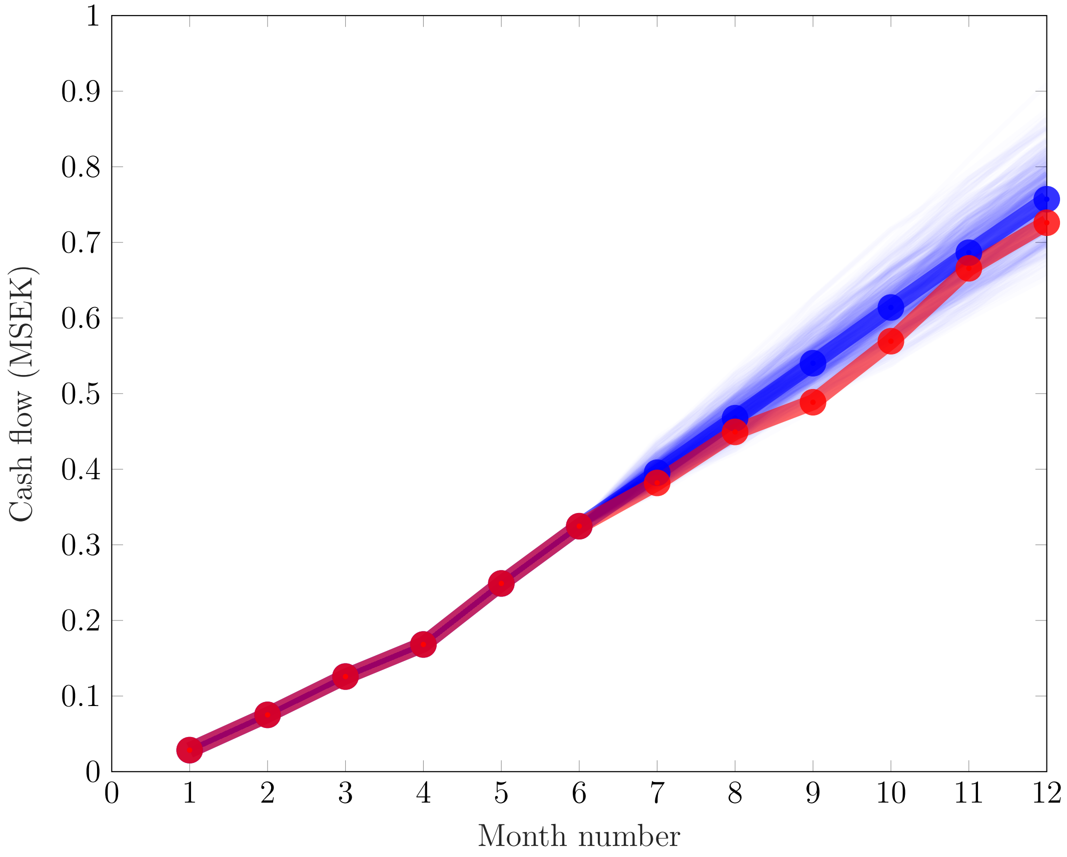
CHALMERS
UNIVERSITY OF TECHNOLOGY



# Valuation of a Non-Performing Loan Portfolio

## Cash flow forecasting using machine learning algorithms and Markov chains

Master's thesis in Engineering Mathematics and Computational Science

MAX PETER BEJMER
LINUS WISKMAN

Department of Mathematical Sciences
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2020

Master's thesis 2020

# Valuation of a Non-Performing Loan Portfolio

Cash flow forecasting using machine learning algorithms and Markov chains

MAX PETER BEJMER
LINUS WISKMAN

Department of Mathematical Sciences
Chalmers University of Technology
Gothenburg, Sweden 2020

Valuation of a Non-Performing Loan Portfolio
Cash flow forecasting using machine learning algorithms and Markov chains
MAX PETER BEJMER
LINUS WISKMAN

Supervisor: Holger Rootzén, Department of Mathematical Sciences
Supervisor: Jörgen Köster, Dignisia AB
Examiner: Serik Sagitov, Department of Mathematical Sciences

Master's Thesis 2020:NN
Department of Mathematical Sciences
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Estimation of the cash flow from an NPL portfolio. See Figure 4.20 for further explanation.

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2020

Valuation of a Non-Performing Loan Portfolio
Cash flow forecasting using machine learning algorithms and Markov chains
MAX PETER BEJMER
LINUS WISKMAN
Department of Mathematical Sciences
Chalmers University of Technology

# Abstract

This master's thesis focuses on valuation of a non performing loans portfolio, provided by partner company Dignisia. Two models are developed; a combined classification-regression model and a Markov chain model. Valuation performances are decent but explanatory power, i.e. $R^2$ values, are lower or on par with similar research.

The two models are tested in two scenarios with the aim of investigating improvement in model performance with knowledge of prior payment history. No clear relation is found between demographic and errand-specific attributes and debt collection rate. The Markov chain model shows similar performance as the more conventional method static pool, in portfolio valuation. However, advantages of the Markov model are the adaptation to new data and the possibility of model extensions which are further discussed. Data quality and quantity are presumed to be the major limiting factors, which is in line with conclusions in the literature.

Keywords: Collection rate, forecasting, prediction, debt collection, non-performing loans, NPL, classification, regression, Markov chain.

# Acknowledgements

First and foremost, the authors wish to express our sincerest gratitude to our supervisor Holger Rootzén at the Department of Mathematical Sciences. Thank you for fruitful discussions and for challenging us to always prioritize the most important questions. We always left your office enlightened and inspired.

Further, we would like to pay our special regards to our supervisor from Dignisia – Jörgen Köster. Thank you for making the project possible and for guiding us through the complexity of the credit market. We value your expertise and are forever grateful for the opportunity you gave us. In addition, we wish to extend our thanks to the entire team at Dignisia.

We also want to thank our extended families for love and support throughout the thesis project and our time at Chalmers.

Last but not least, the authors wish to thank our common friends Jesper, Oscar and Sebastian for making the time at Chalmers special.

Max Peter Bejmer & Linus Wiskman

Gothenburg, March 2020

# Contents

# List of Figures

# List of Tables

# 1

# Introduction

This chapter introduces the reader to the thesis project background and the aim of successfully modelling debt repayments, in Section 1.1 and 1.2 respectively. More specifically, the background introduces important concepts relevant for the thesis context such as a motivation of the increasing importance of credit management, a description of the credit market dynamics and a brief presentation of Dignisia AB. Further, project scope is specified in Section 1.3 and a discussion on ethical considerations and GDPR compliance is presented in Section 1.4.

## 1.1 Background

This section defines the context of the master's thesis. We introduce the field of credit portfolio management and the role and aim of Dignisia AB in the credit market. We also motivate the growing importance of credit portfolio management and portfolio cash flow forecasting.

### 1.1.1 Macroeconomic context and importance of credit portfolio management

After the financial crisis in 2008, credit institutions have been increasingly regulated [1–3]. Regulatory frameworks brought stricter capital and liquidity requirements which in turn put pressure on banks' profit margins [4]. As a consequence, the ability to optimize credit portfolios has become increasingly important among banks [5].
A signal of credit management emerging as a key function in Sweden is the development of the Swedish household debt over the recent years, which has been increasing as a result of a long period of low interest rates [6, 7]. As of August 2019, the total Swedish household debt was 4138 billion SEK which corresponds to 85 percent of GDP [8]. Mortgages account for 82 percent of the debt and the remaining 18 percent are loans for consumption. However, as seen in Figure 1.1 loans for consumption, i.e. credits given without collateral, are growing faster than mortgages [7, 9].
Additionally, loans for consumption account for roughly half of households' monthly debt payments [8]. Swedish financial authorities agree that the increased household indebtedness constitute a significant risk for the financial stability [7, 10] and should hence be considered by credit institutions.

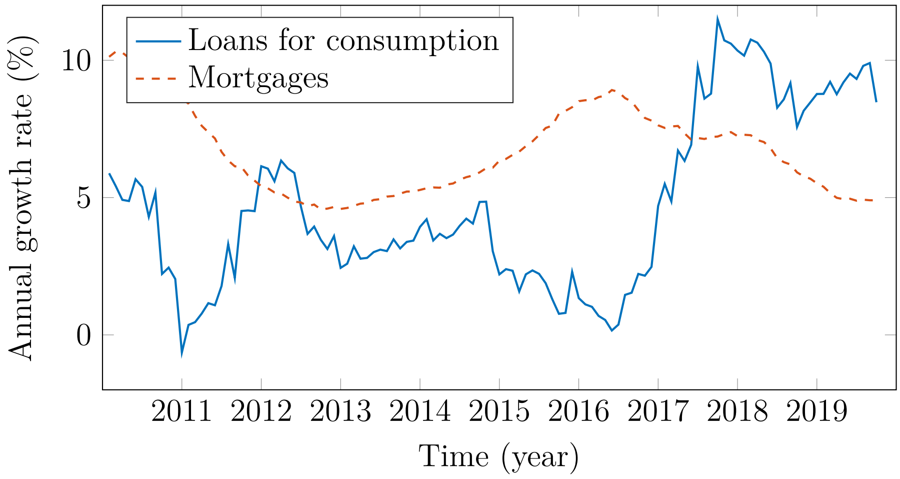


**Figure 1.1:** Annual growth rates of Swedish household mortgages and loans for consumption. Loans with collateral other than houses are excluded. Source: Statistics Sweden [9].

### 1.1.2 Credit market dynamics

The credit market is a complex ecosystem consisting of a variety of players making transactions and granting credits among each other. Players include governments, banks, financial institutions, companies and private investors. On top of these, there are numerous regulatory authorities on national and international level governing the market dynamics. In order to avoid unnecessary complexity we will narrow the scope to credits granted by banks or retailers to a consumer. Typical credits are invoices, mortgages, loans for consumption and credit card products. In essence, banks and retailers gamble on a consumer's ability of repaying a credit. If a consumer fails to repay the debt, the credit is said to *default*. Defaulted credit products are commonly called *non-performing loans*, henceforth denoted NPLs.

Figure 1.2 is an indicative description of a generic payment flow, and depicts different stages in a credit chain ending up in an NPL portfolio. A company, typically retailer or e-commerce player, sells a product or service and offers different payment options. The payment could either be done directly, e.g. via debit card, or via invoice. To be able to grant payment by invoice, the company tests the customer's creditworthiness and makes a credit decision. If the credit is approved, an invoice is sent and is most often paid on time during the payment period. In case of a missing payment, the consumer receives a number of reminders accompanied with fees. If the invoice remains unpaid, the errand is eventually sent to a debt collection agency (DCA) and is considered a non-performing loan. A collection of NPLs are commonly grouped together in an non-performing loan portfolio – or NPL portfolio for short.

NPL portfolios can be traded, which is an important source of financing and risk management for banks and retailers. The Council of the European Union encourage the development of such a secondary market [11] with European financial stability

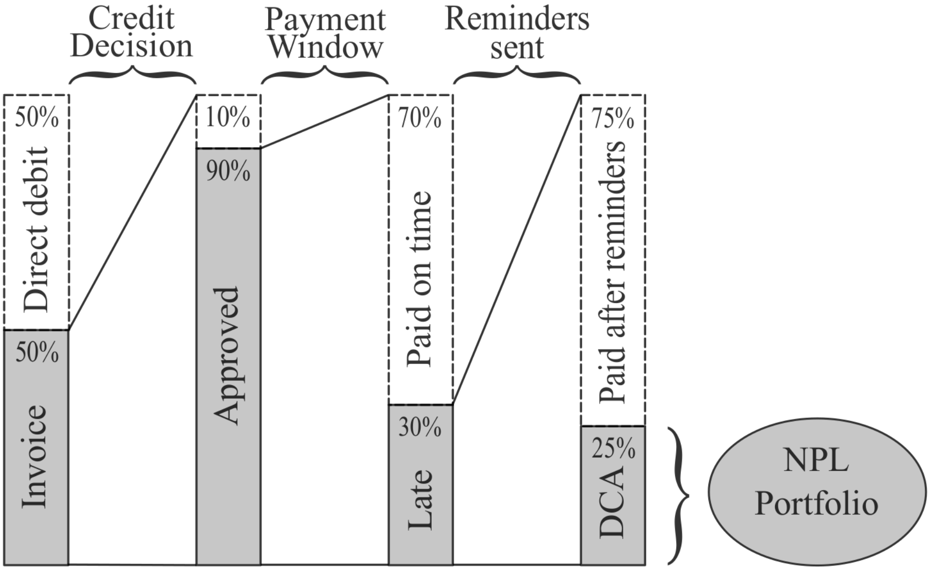


**Figure 1.2:** Illustrates the credit life-cycle of the areas relevant to our analysis. Analysis conducted on the NPL portfolio leads to better portfolio valuation, better case handling, and improved credit decisions. Note: The shares given by the percentages are estimated based on general opinions.

as main motivation. Thus, valuation of NPL portfolios is an increasingly important task given the expected increase in portfolio transactions. Two important questions traders ponder concerning NPL portfolios are:

- How much of the total outstanding debt can the buyer expect to collect?
- How will the cash flow from the portfolio vary over time?

The better the owner of an NPL portfolio can answer these two key questions, the better the opportunity for an accurate valuation of the portfolio.

These questions are the foundation of our thesis project. Our aim and scope are further specified in the following sections, after a short presentation of the partner company Dignisia AB.

### 1.1.3 The role of Dignisia AB

Dignisia AB is a company founded in 2017 by seasoned debt collection and credit management specialists wanting to address an identified need on the client-side for specialized Business Intelligence solutions for credit management. Or, as they call it, Credit Intelligence. The idea is to help companies get a better understanding of their debtors and receivables due with the aim of giving companies better control of the entire credit life-cycle. This is done by analysing customer data to be able to give valuable insights about both the present and the future [12].

As discussed in the previous section, NPL portfolio valuation is a key concern for banks and retailers and is hence important for Dignisia's clients. The relevance of the master's thesis project is motivated on this basis.

## 1.2 Aim

One of the most important aspects of an NPL portfolio is the expected future cash flow. The expectation determines the valuation and gives portfolio managers better control. Therefore, having a well-performing cash flow forecasting model is important to stay competitive and to estimate fair pricing ahead of a portfolio transaction. It is also interesting, from a debt collection operations perspective, to understand if there are attributes explaining and driving loan repayment. This will be the aim of the master's thesis: to build a well-performing model forecasting cash flow from a given NPL portfolio and understanding the drivers of loan repayment.

## 1.3 Thesis Project Scope

While the aim is to understand drivers and build the best possible model, some restrictions have to be made.

In order to avoid seasonality effects and to utilize a sufficient proportion of the available data, the forecasting is considered no further than on a 12 month horizon. Defaulted credits can see collection up to 20 years after being registered at a debt collection agency. Thus, lifetime collection from a NPL portfolio is a complex quantity and the 12 month restriction is necessary. The data availability further limits model development to some extent which is detailed in Section 3.1.

In terms of model selection, we have identified candidate models using insights from our non-exhaustive review of previous research, Section 2.1. Although there are other reasonable approaches, e.g. time series modelling, they are considered out of scope in this thesis.

Our problem solving approach involves two modelling scenarios and two classes of models applied to each scenario. The first scenario is referred to as *blind* and resembles a situation where an analyst wants to predict NPL portfolio cash flow based solely on debt account attributes. That is, no previous cash flow information is available which is the case for an external analyst prior to a portfolio transaction. The second scenario is called *informed* and represents a situation where a portfolio has been monitored for a period of time. The informed scenario could also be seen as a way of having information about prior payment behaviour. This idea is discussed in more detail later. In practice, the blind dataset is complemented with variables explaining initial cash flow.

## 1.4 Ethical Considerations

Ethics is something to always have in mind in every research project. It can take many different shapes and forms. Some ethical aspects are controlled by law, for example processing of personal data, while some are grounded in the personal agenda of the individual conducting the research. This section covers a review of the business landscape in which Dignisia acts and further, a more specific explanation of the setting in which the research for this thesis is performed.

As stated earlier, Dignisia AB acts in the credit business area with the aim of helping companies get a better understanding and control over their debtors and receivables due. Most of the time debts get resolved without complications according to terms agreed upon by all parts involved. However, problems arise when things do not get resolved or when parties do not have mutual agreements. Problems and non-pleasant situations can easily arise and one of these situations is when a debtor is not able or not willing to pay. There exist numerous reasons for this and the origination could be either from the debtor or lendor or somewhere else. Being in debt is not pleasant in several aspects, especially if the debt is overdue. These overdue debts are a big part of what Dignisia works with and is also what this thesis deals with.

With all this said, working at Dignisia, with the assumption that the company itself follows all laws and regulations, one should also make sure that in what way Dignisia contributes to the credit market is something in alignment with ones own personal values.

The aim of this thesis is to create a forecasting model to value and estimate cash flow from an NPL portfolio and also to understand what drives repayment of NPLs. The analysis is performed on a data set containing information about errands of debts which have been sent to debt collection agencies. The data describes a particularly vulnerable segment of people of the society from a financial and economical point of view. The result is a model to valuate and forecast cash flow from a collection of NPLs. Part of the aim is to understand the drivers behind repayment of loans. When the results are presented and analyzed it is of importance to be aware of what is done and what affects it might have. A result that estimates repayment ability based on attributes could become discriminating if the attribute is of a certain character, for example gender or race. One could discuss the ethical concern about including results like these in the report. One question to ask is how much impact, direct or indirect, the results can have. In our thesis, there is not much direct impact of our results. For example, we do not build and implement a credit model based on our findings. We analyse data and present what it says, nothing less, nothing more.

Another aspect to think about when dealing with personal data is what kind of regulations and laws apply. In this case, where the work will be published, GDPR apply and must be followed. The data must be anonymised.

# 2

# Theory

This chapter is a deep-dive into relevant models and theoretical frameworks relevant for the thesis project. A review of previous results on debt modelling is presented in Section 2.1. Further, the static pool method, which is used for benchmarking purposes, is breifly described in Section 2.2. Sections 2.3-2.6 are used for in depth presentation of regression models, decision trees, artificial neural networks and Markov chains respectively.

## 2.1 Previous Research

There have been plenty of attempts to model and predict collection rates for debt collection agencies. Overall, the prediction performance has been poor with typical $R^2$-values ranging from 10 to 25 percent. Several researchers highlight bad data quality and availability as main obstacles for better modelling, which in turn is due to low degree of transparency in debt collection operations and lack of reporting requirements. However, in cases with highly detailed data $R^2$-values of close to 70 percent are reachable. A non-exhaustive review of previous research on collection modelling and prediction is presented in the remainder of this section.

As previously mentioned, data availability is a major limiting factor for debt collection forecasting performance. The relation between availability and performance has been studied. Kribel and Yam [13, 14] address forecasting of insurance collectables from a German DCA on two granularity levels: at first using data available to a third-party DCA, and secondly complementing with insurance agency in-house collection data. They find that in-house data, particularly through credit bureau scores and previous repaid accounts, dramatically improves prediction performance. The adjusted $R^2$ is tripled on inclusion of in-house data; increasing from 10-15 percent to 43 percent. A similar comparison was conducted by Thomas et al. [15] for a UK-based credit institute. The authors propose a two-stage model for each information set, i.e. in-house or third party collection data. It is deemed necessary to separate the analysis since accounts sold to third parties have passed through in-house collection systems unsuccessfully, and are thus harder to collect. The idea of using a multiple-stage model is seen elsewhere in the literature. Belotti and Ye [16] observe a strong trimodal distribution of collection rates when analyzing a portfolio of NPLs originating from a European bank. No collection and full collection are frequent occurring events, while accounts paid in part are smoothly distributed between collection rate zero and one. The authors use classification methods to distinguish accounts paid in full, and a Beta Mixture Regression model to produce a collection rate in the interval $[0, 1)$ for remaining accounts. Their two-stage Beta Regression model is one of the most successful approaches encountered with an adjusted $R^2$-value of 15-20 percent and rising to 69 percent on inclusion of detailed in-house data. In-house data included information about payment frequency, number of DCA interactions and multiple credit scores which is not seen elsewhere in the literature. Albeit the unique dataset, their results further stress the importance of data quality and availability.

Papke and Woolridge [17] discuss the problem of modelling a bounded fractional dependent variable. The authors propose functional forms extending a generalized linear modelling framework, which is used and further developed by several researchers in their attempts to model collection rates [13, 14, 18]. Belotti et al. [19] introduce non-linear and machine learning approaches in a comparative study from 2019. Out of the 19 models applied, random forest emerges as the top-performing option. A different approach to a slight variation of the debt collection prediction problem is to use survival analysis techniques. Cox's proportional hazard models [20] was used to predict recovery times of delinquent credit cards by Ha and Krishnan [21] and Boutachaktchiev [22] used a Markov chain model to predict cure rates of NPLs.

Although prediction performances of aforementioned models are not very promising, the techniques suggested offer a wide range of modelling possibilities and are all more advanced than the most popular framework – vintage analysis. While vintage analysis is not a model in itself, it offers a way of structuring data based on a constant factor and a risk factor. The most common selection is to use the registration month as constant factor and days past due as an indicator of risk [23]. In this way, different averaging techniques can be used to predict future cash flow.

## 2.2 Static Pool

One of the most widely used methods today for valuating and estimating cash flow from an NPL portfolio is called *static pool* and is a special case of vintage analysis. Accounts with common characteristics, such as time of registration, is grouped together in a pool or sub-portfolio.

To valuate and estimate cash flow from an NPL-portfolio using the static pool method, a common way of conducting the analysis is according to the following [12]. The NPL portfolio to be valuated is analysed with the objective of finding what type of accounts it consists of. The next step is to find an old NPL portfolio, with known outcome, with similar attributes as the one to be analysed. The idea is that the two portfolios will share enough common attributes that they can be expected to behave the same way. For example, if 5% of the total debt was collected during the first month for the old portfolio, then it is expected that around 5% will be collected from the new portfolio. The time frame for which the comparison is made could either be one month as in the example or one year or ten years or anything else that is suitable.

## 2.3 Regression Models

Linear regression is a widely used modelling technique and can be formulated in a statistical, probabilistic or machine learning framework. The concept, however, is the same and its straight-forward implementation and simplicity of interpretation makes the method a good choice in many disciplines. Medicine, physiotherapy, social sciences and economics are examples where regression is the major modelling technique. While the idea in linear regression modelling is simple, there are multiple extensions of higher complexity — two of which are treated in more detail in Section 2.3.2, namely generalised linear models (GLMs) and the special case logistic regression. Section 2.3.1 deals with the basic concepts of linear regression. The theory is considered common knowledge and hence not referenced thoroughly, although notation and structure are inspired by Rogers and Girolami [24] and Bishop [25].

### 2.3.1 Linear regression

Some notation is introduced in order to treat the concept of linear regression. Let $x^{(k)}$ be an observation of a $D$-dimensional vector of variables, $x$, and let $t^{(k)}$ be a corresponding target observation. In linear regression, the goal is to fit the best model to a sequence of outcomes, or targets, $\mathbf{t} = \{t^{(i)}\}_{i=1}^{n}$ using a linear combination of an observed sequence of data, $\{x^{(i)}\}_{i=1}^{n}$. Call the coefficients, or weights $\mathbf{w}$, and let $f$ be a linear mapping from $\mathbb{R}^D$ to $\mathbb{R}$ representing the model for the target observation $t^{(k)}$. The linear model can be written as

$$f(x^{(k)};\mathbf{w}) = f(x_1^{(k)}, x_2^{(k)}, ..., x_D^{(k)}; w_0, w_1, \dots, w_D) = \mathbf{w}^T\mathbf{x}^{(k)} \ ,$$

where

$$\mathbf{w} = \Big[w_0, w_1, \dots, w_D\Big]^T \ ,$$

$$x^{(k)} = \Big[x_1^{(k)}, x_2^{(k)}, \dots, x_D^{(k)}\Big]^T \ ,$$

$$\mathbf{x}^{(k)} = \Big[1, x^{(k)}\Big]^T = \Big[1, x_1^{(k)}, x_2^{(k)}, \dots, x_D^{(k)}\Big]^T \ .$$

The problem of interest is to find a set of weights which makes the model $f$ fit the data in the best possible way. To accomplish this a loss function $\mathcal{L}$ is introduced. The loss function is scalar and measures the distance between two functions in some sense. The quadratic loss is given by

$$\mathcal{L}(u, v) = (u - v)^2 \ ,$$

which is by far the most commonly used. Here, $u$ and $v$ are arbitrary vectors defined on the same space. Having defined such a loss function, the problem of finding the best weights can be formulated as the minimization problem

$$\underset{\mathbf{w}\in\mathbb{R}^{D+1}}{\arg\min} \quad \frac{1}{N}\sum_{i=1}^{n}\mathcal{L}(f(x^{(i)};\mathbf{w}), t^{(i)}) = \underset{\mathbf{w}\in\mathbb{R}^{D+1}}{\arg\min} \quad \frac{1}{N}\sum_{i=1}^{n}(f(x^{(i)};\mathbf{w}) - t^{(i)})^2 \ .$$

In order to solve the problem, it is convenient to introduce a so called design matrix

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}^{T(1)} \\ \mathbf{x}^{T(2)} \\ \vdots \\ \mathbf{x}^{T(n)} \end{bmatrix} = \begin{bmatrix} 1 & x^{T(1)} \\ 1 & x^{T(2)} \\ \vdots & \vdots \\ 1 & x^{T(n)} \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \dots & x_D^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \dots & x_D^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(n)} & x_2^{(n)} & \dots & x_D^{(n)} \end{bmatrix} .$$

The minimization problem becomes

$$\underset{\mathbf{w} \in \mathbb{R}^{D+1}}{\arg\min} \quad \frac{1}{N}(\mathbf{t} - \mathbf{X}\mathbf{w})^T(\mathbf{t} - \mathbf{X}\mathbf{w}) \ . \tag{2.2}$$

Assuming $\mathbf{X}$ is invertible, the optimal weight vector $\hat{\mathbf{w}}$ is given by expanding the multiplication above and setting the partial derivative to zero. The resulting weight given by linear regression is expressed as

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t} \ .$$

### 2.3.2 Generalized linear regression

There are numerous ways of extending the framework offered by linear regression. A set of models created from such an extension is the *generalized linear model* (GLM). In general, the design matrix is modified to allow arbitrary functions of each input variable according to

$$\mathbf{X} = \begin{bmatrix} \mathbf{h}(\mathbf{x}^{T(1)}) \\ \mathbf{h}(\mathbf{x}^{T(2)}) \\ \vdots \\ \mathbf{h}(\mathbf{x}^{T(n)}) \end{bmatrix} = \begin{bmatrix} h_0 & h_1(x_1^{(1)}) & h_2(x_2^{(1)}) & \dots & h_D(x_D^{(1)}) \\ h_0 & h_1(x_1^{(2)}) & h_2(x_2^{(2)}) & \dots & h_D(x_D^{(2)}) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ h_0 & h_1(x_1^{(n)}) & h_2(x_2^{(n)}) & \dots & h_D(x_D^{(n)}) \end{bmatrix} .$$

An important special case of a GLM is *logistic regression*, commonly used to model binary response variables. In a probabilistic framework, logistic regression can be formulated by mapping the output of linear model to the unit interval according to

$$f(\mathbf{x}, \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{x}^T\mathbf{w})} \ .$$

Clearly the model $f$ is restricted to values between zero and one, and can hence be considered to be a predicted probability of a binary target variable $t$:

$$P(t = 1|\mathbf{x}, \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{x}^T\mathbf{w})} \ .$$

Such a probability can easily be used in binary classification. The most natural option is to assign class 1 to all instances where the probability is greater than one half, and class 0 otherwise. There is, however, no restrictions on choosing the cut-off probability differently.

## 2.4 Classification and Regression Trees

Classification and regression trees are methods for constructing prediction models from data. The models are obtained by partitioning the data and fitting a simple prediction model within each partition. Different regions are created where the data can be organized in. Starting from the top of the tree, the data is split into groups based on their attributes. The splits are constructed with the intention to obtain the greatest possible separation of the data [26].

For a classification tree, obtaining a measure for the maximum separation is done by calculating the Gini Impurity, $I_G$, for each node. $I_G$, for a node $n$ is a measure of how often, a randomly chosen data point, would be incorrectly labeled. It is given by

$$I_G(n) = 1 - \sum_{i=1}^{J} p_i^2 \Bigg|_{\text{node}=n} , \tag{2.3}$$

where $J$ is the number of classes, and $p_i$ is the probability of randomly choosing class $i$. A split in a node will always decrease the total Gini Impurity. Another concept that is used instead of Gini Impurity is the Information Gain Entropy which works in a similar fashion [26].

For a regression tree the split is based on the maximal reduction in the standard deviation of the target value. For a node the sample standard deviation, $S$, of the set of target values, $T$, is calculated according to

$$S(T) = \sqrt{\frac{\sum_{x_i \in T}(x_i - \overline{x})^2}{|T| - 1}} . \tag{2.4}$$

This is then compared to the total standard deviation after a split given by

$$S(T, X) = P(\text{left node})S_{\text{left}}(T) + P(\text{right node})S_{\text{right}}(T) , \tag{2.5}$$

where X indicates on what attribute and value to make the split. The reduction in the standard deviation

$$SDR(T, X) = S(T) - S(T, X) , \tag{2.6}$$

is what decides on what attribute to perform the split.

Following is a brief and conceptual description of how to generate a decision tree. For a more through explanation the reader is encouraged to read ”Classification and Regression Trees” by W. Loh [26].

1. Start at the root node
2. For each attribute, partition the data at the node for the different values for the full range of values of the attribute. Compute the separation, the Gini Impurity for a classification tree, see (2.3), or the reduction in the standard deviation for a regression tree according to (2.4)-(2.6). The attribute that results in the greatest separation will make up the splitting criteria.

3. Continue splitting the tree recursively according to step 2. If no type of splitting will result in a separation, the node will only contain instances from one class and become a leaf node. Another way to stop the process is if a stopping criteria is met, for example reaching a predefined maximum dept or maximum number of splits.

The tree is constructed with the help of training data. It is possible to construct a tree classifying all training data points correctly. However, this would not give optimal results when testing the tree on a test set, the model is over-fitted and does not generalize well on new data. The generalisation ability of the tree is in general related with the dept of the tree. The deeper the tree, the worse the generalisation. Usually, one defines maximum number of splits or maximum depth of the tree. These are also parameters that can be optimized.

### 2.4.1 Random forest

A random forest is an ensemble of decision trees. The trees are constructed as described above. However, instead of using the whole set of data the trees are grown based on a random subset of the data. The subset is constructed by randomly sampling data, with replacement, from the entire set. This is called bootstrapping. When constructing the splits in a random forest, the construction is based only on a random subset of the features. Typically this amounts to the square root of the total number of features.

Predictions are made by averaging the predictions among the different trees. Random trees are robust against over-fitting [27].

## 2.5 Artificial Neural Networks as Classification and Regression Models

Artificial neural networks, ANN, have grown popular over the last decade in the application of solving complex problems. The strengths of ANNs lies in the information processing. Among these abilities are the notion of finding complex and non-linear relations, having characteristic of high parallelism, be robust against faulty and noisy data, and be able to generalize well. This can allow for a better fit, fast processing at the same time as having a high hardware failure-tolerance, and be applicable to unlearned data [28].

An ANN has an architecture inspired by the biological neural network in the human brain. It consists of connected nodes called neurons. Each neuron can receive a signal, process it, and signal other neurons connected to it. The architecture of the network consists of an input layer, usually the size of the feature space, a number of hidden layers, and an output layer. The input layer usually has some kind of activation function which processes the incoming signal. The neurons in the hidden layer have individual biases and a separate set of weights. The neurons in the output layer also have some kind of activation function, usually the same for all neurons in the layer [29].

In supervised learning the ANN is trained with training data. The weights and biases are updated according to the difference in the network output and the target output. The aim is to minimize the overall network error. The most widely used networks are the ones where back propagation can be used. Back propagation is a way of updating the weights and biases based on training data. In each iteration the data is propagated forward to produce a solution. In the next step the error between the solution and the target is propagated backward through the network to update weights and biases [28].

ANN is a versatile algorithm and can be used for both classification and regression tasks.

## 2.6 Discrete Time Markov Chains

Andrei A. Markov (1856-1922) developed the concept of Markov chains. He studied sequences of random variables with a particular dependence property now known as the Markov property. His research launched a new branch of probability theory – stochastic processes – with applications in financial modelling, signal processing and cell biology to mention a few [30]. The remainder of this section treats basic concepts of discrete time Markov chains. For a more thorough theoretical review the reader is referred to the work of G. Grimmett and D. Stirzaker [31]. The theory presented below is inspired by the same book, if nothing else is noted.

Let $X = \{X_n,\ n = 0, 1, 2, \dots\}$ be a sequence of random variables taking values on some discrete countable state space[1] $S$. The process $X$ is said to be a *Markov chain* if it fulfills the *Markov property.*

**Definition 1 (Markov property)** *Assume that $s, x_0, x_1, \dots, x_{n-1} \in S$ are known states, and let $X$ be a sequence of random variables. The sequence $X$ fulfills the Markov property if*

$$\mathbb{P}(X_n = s | X_0 = x_0, X_1 = x_1, \dots, X_{n-1} = x_{n-1}) = \mathbb{P}(X_n = s | X_{n-1} = x_{n-1}) \quad (2.7)$$

The property can be understood as a lack of memory of the process. The distribution of variable $X_n$ conditional on the previous state $X_{n-1}$ is independent of the history $\{X_0, X_1, \dots, X_{n-2}\}$, i.e. the only information governing the future values of the process is the previous value. With this property, it is possible to prove useful results for Markov chains. An important concept relevant to the thesis project is a *transition probability matrix.*

**Definition 2 (Transition probability matrix)** *Let $i$ and $j$ be two states in $S$. The probability of a Markov chain, $X$, moving from $i$ to $j$ at time step $n + 1$ is*

[1] Implicitly assumed that there exists an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and that $X_n$ is $\mathcal{F}$-measurable and maps $\Omega$ into $S$ for each $n = 0, 1, 2, \dots$.

*denoted*

$$p_{ij} = \mathbb{P}(X_{n+1} = j | X_n = i)$$

*Collecting all possible transitions we get a* $|S| \times |S|$*-matrix of probabilities,* $\mathbf{P}$*, called transition probability matrix*

$$\mathbf{P} = \left(p_{ij}\right)_{i,j=0}^{|S|} = \begin{bmatrix} p_{00} & p_{01} & \cdots & p_{0|S|} \\ p_{10} & p_{11} & \cdots & p_{1|S|} \\ \vdots & \vdots & \ddots & \vdots \\ p_{|S|0} & p_{|S|1} & \cdots & p_{|S||S|} \end{bmatrix}$$

*Here,* $|\cdot|$ *denotes the norm.*

**Definition 3 (Homogeneity)** *A Markov chain is called homogeneous, or time homogeneous if*

$$\mathbb{P}(X_{n+1} = j | X_n = i) = \mathbb{P}(X_1 = j | X_0 = i)$$

*for all* $i, j$ *and* $n$*.*
Markov chains encountered in this thesis project are time homogeneous.

**Definition 4 (Recurrent Markov state)** *State* $i$ *is called* ***recurrent****, or* ***persistent****, if*

$$P(X_n = i \text{ for some } n \geq 1 | X_0 = i) = 1$$

*which is to say that the probability of eventual return to* $i$*, having started from* $i$*, is 1. If this probability is strictly less than 1, the state* $i$ *is called* ***transient****.*

**Definition 5 (Absorbing Markov state)** *A state is called* ***absorbing*** *if the probability of leaving the state is zero. An absorbing state is a special case of a reccurent state.*

**Definition 6 (Absorbing Markov chain)** *A Markov chain is an absorbing chain if*

1. *There is at least one absorbing state and*
2. *It is possible to go from any state to at least one absorbing state in a finite number of steps.*

The transition matrix $\mathbf{P}$ for an absorbing Markov chain, with the transition states coming first, can be written on the canonical form [32]

$$\mathbf{P} = \begin{bmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{0} & \mathbf{I}_r \end{bmatrix}$$

where the dimensions of the matrices $\mathbf{Q},\ \mathbf{R},\ \mathbf{0},\ \mathbf{I}_r$ are $[t \times t],\ [t \times r],\ [r \times t],\ [r \times r]$ respectively where $t$ is number of transient states and $r$ is number of absorbing(also called recurrent) states.

The probability of transitioning from $i$ to $j$ in exactly $k$ steps is the $i, j$:th entry of $\mathbf{Q}^k$. Summing these up for a certain number of steps $K$, results in the matrix $\mathbf{N}_K$ in which the $i, j$:th entry is the expected number of visits of state $j$ given starting in state $i$:

$$\mathbf{N}_K = \sum_{k=0}^{K} \mathbf{Q}^k \ . \tag{2.8}$$

Inspired by theorem 11.6 in [32]. The probability that an absorbing chain will be absorbed in the absorbing state $j$, given start in state $i$, in $K$ steps, is given by the $i, j$:th entry of the matrix $\mathbf{B}_K$:

$$\mathbf{B}_{K_{i,j}} = \sum_n \sum_l q_{il} r_{lj} = \sum_l \sum_n q_{il} r_{lj} = \sum_l n_{il} r_{lj} = (\mathbf{N}_K \mathbf{R})_{ij} \ , \tag{2.9}$$

where $\mathbf{R}$ is the $[t \times r]$-matrix containing transition probabilities to the absorbing states.

# 3 Methods

This chapter deals with the problem solving strategy based upon the questions and statements written in the aim of the thesis, see Section 1.2.

First, the data set is presented, with information about origin, context, attributes, limitations and how it has been used.

Two different models are built and tested: the classification-regression model and the Markov chain model. In the end the models aim to do the same thing, namely to value and forecast an NPL portfolio. The two models use different approaches, the models are based upon different assumptions and ideas which results in different conclusions. The classification-regression models are based on the assumption that there exist a relation between attributes of an account and collection rate. It uses a bottom-up approach where it analyses each account by itself and sums the parts to build a portfolio valuation. The Markov chain model, on the other hand, uses a top-down approach and uses the assumption of patterns in repayment. The simulation of the Markov chain is done on the portfolio as a whole.

Further, to investigate the increase of predictive capacity under information gain the two models are tested in two different scenarios; blind and informed. In the blind scenario, the analysis and prediction is only based on the account attributes. In the informed scenario the analysis and prediction is done with additional information about payment history. The way this is done differs between the two models. In the sections which describe the models, the application of the two scenarios, blind and informed, is discussed in more detail.

Matlab has been used for all programming tasks.

## 3.1 Data

This section describes data origin, business context, data attributes and high-level statistics of the dataset. Limitations and data pre-processing decisions are also discussed.

### 3.1.1 Data origin

The available data originates from a transaction of an NPL portfolio consisting of five types of defaulted banking credit services – invoices, account credits, loans, credit cards and mortgages from several European banks. The portfolios were presented with cash flow from active accounts in the portfolio for three years, accompanied with several attributes for each account, and investors were invited to value the portfolio and place bids. Eventually the portfolios were partially sold at undisclosed terms.

The data set at disposal for this Master's thesis project is an anonymized (no personal data and no possibility to relate any item to any actual person or debt) and time-translated version of the very same data set presented to investors before the auctions.

### 3.1.2 Dataset construction and pre-processing

The NPL portfolio data is structured on account level. Each row represents an account, or errand, registered at a DCA with information about the outstanding debt and the debtor. Rows are equivalently referred to as accounts or errands in this report.

Numerous data constrictions have to be made in order to obtain a useful dataset. This section presents and motivates each decision on data limitations. In total, 88 % percent of the available data is considered of no use for the purpose of the thesis project. Figure 3.1 summarizes limitations and indicates relative data loss in each step. The steps are motivated below.

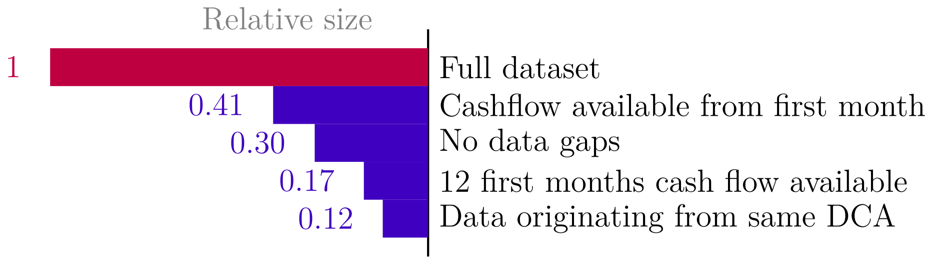
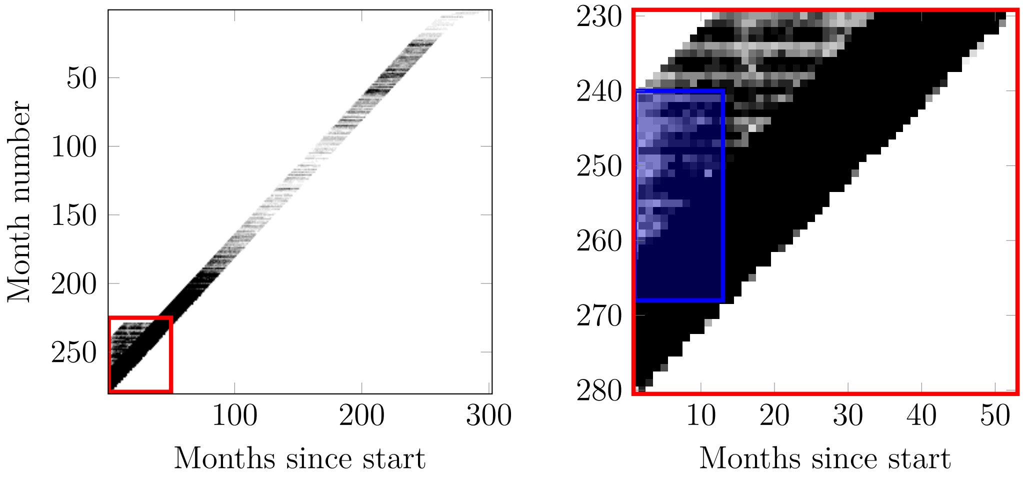


**Figure 3.1:** Restrictions imposed on the dataset and corresponding size reduction.

The most important limitations is the one imposed to deal with the availability of cash flow from the first month. It is not trivial to understand why this is necessary, which is why it is explained in detail below. The other limitations are straightforward: accounts with data gaps are removed. Reporting structure differs between DCA's, and hence accounts from the largest and best documenting DCA are kept. Since the valuation is done on a 12 month horizon, accounts younger than 12 months are removed.

Figure 3.2a is an illustration of the cash flow from the available NPL portfolio. Registration month, which is the date given by month and year at which the account was registered at the DCA, is shown on the vertical axis, indexed from the oldest account in the data set. The horizontal axis values are number of months since the registration month. High "heat", i.e. dark pixels, represent a relatively high cash flow and low heat, i.e. white pixels, means no cash flow. The cash flow matrix has an upper triangular shape, since no cash flow can exist before an account is registered. There is also a clearly visible lower triangular structure, producing a ribbon of cash flow across the diagonal in Figure 3.2. The meaning of this is that data was only recorded in a fixed time period, which is the height of the ribbon. No information about earlier activity of the accounts is available. This is a major limitation, since the data is implicitly conditioned on existence of cash flow. Closed accounts or accounts with zero cash flow in the period is not featured in the data set. Therefore, a large proportion of the data is chosen not to be used because the intention is to investigate cash flow the first months after account registration.



**(a)** Full data set. Cash flow from 300 months old accounts

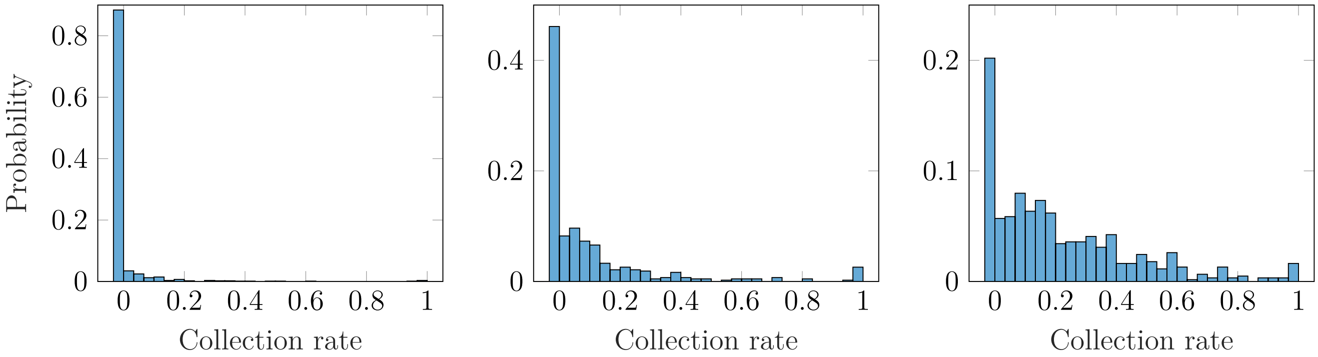
**(b)** Highlighted area. The blue proportion indicates time period for which the 12 first months of cash flow is available.

**Figure 3.2:** Heatmap of portfolio cash flow. y-axis represents DCA registration month for an account, while x-axis value describes months since registration. Hence, no "heat" i.e. cash flow, can exist below the diagonal.

Further limitations shown in Figure 3.1 are

### 3.1.3 Description of data attributes

After the pre-processing is done, as described in Section 3.1.2, the set contains three different types of debts, namely, invoices, loans, and credit card debts. The types of debt are believed to, in general, behave differently. By looking at Figure 3.3, which shows the distribution of collection rates, which is the share of the debt that has been repaid, one can see big differences. Most of the invoices have no repayment. About half of the loans have no repayment. Credit card debts are repaid to a greater extent compared to the other two debt types. In the classification-regression model all errands that are invoices are removed. Such a skewed distribution could easily give the notion of the model being better than what it actually is. Also, low collection in combination with low debt size, results in a low contribution of invoices to the whole portfolio which makes it less interesting to model. In the Markov model the debt types are modelled separately and invoices are included in the analysis.



**(a)** Invoices.

**(b)** Loans.

**(c)** Credit card debts.

**Figure 3.3:** Distribution of collection rate, which is the share of the debt that has been repaid, after 12 months for different debt types. The leftmost bin in every figure contains only CR=0.

Further attributes of the data are described in Tables 3.1 and 3.2. The variables used in the blind scenario are listed in Table 3.1. In the informed scenario, in addition to the variables used in the blind scenario, the variables listed in Table 3.2 are used. These are all variables describing payment behaviour during the first six months.

Descriptive statistics for all the variables are summarized in Table 3.3. More information about variable distributions are found in appendix A.1.

**Table 3.1:** Variable names and their explanations for attributes.

| Variable name | Description |
|---|---|
| `debtType` | Categorical variable describing the type of debt (loan or credit card) |
| `regDateQuarter` | Quarter of the year in wich the debt was registered at the DCA |
| `age` | Age of the debtor |
| `debt` | Size of the debt registered by the DCA |
| `income` | Yearly income of the debtor at the time of granting of the credit. Data from the Swedish tax agency |
| `sumADebt` | Sum of the total debt owed to the government by the debtor registered at the Swedish Enforcement Authority (Kronofogdsmyndigheten) at the time the case was presented in portfolio to investors |
| `sumEDebt` | Sum of the total debt owed to a private company or other person by the debtor registered at the Swedish Enforcement Authority(Kronofogdsmyndigheten) at the time the case was presented in portfolio to investors |
| `numEDebt` | Number of debts to a private company or other person by the debtor registered at the Swedish Enforcement Authority(Kronofogdsmyndigheten) at the time the case was presented in portfolio to investors |

**Table 3.2:** Variable names and their explanations only used in the informed scenario.

| Variable name | Description |
|---|---|
| `collectionRate_1` | Total amount collected for the first month since registration at DCA |
| `collectionRate_1to3` | Total amount collected for the first three months since registration at DCA |
| `collectionRate_1to6` | Total amount collected for the first six months since registration at DCA |

**Table 3.3:** Descriptive statistics of the variables in the data set and the constructed variables used in the informed scenario. Categorical variables are presented with frequency and relative frequency(%) for each category. Numeric variables are presented according to; min, mean(std), max. For more detailed statistics, see appendix A.1.

| Variable | Type | Statistics |
|---|---|---|
| `debtType` | categorical | loan: 425 (41%), credit card: 614 (59%) |
| `regDateQuarter` | categorical | 1$^{st}$: 153 (15%), 2$^{nd}$: 319 (31%), 3$^{rd}$: 347 (33%), 4$^{th}$: 220 (21%) |
| `age` | numeric | 22, 45.3(12.4), 84 (years) |
| `debt` | numeric | 2.71, 57.9(64.7), 550 (SEK, thousands) |
| `income` | numeric | 0, 210(139), 988 (SEK, thousands) |
| `sumADebt` | numeric | 0, 25.2(106), 1690 (SEK, thousands) |
| `sumEDebt` | numeric | 0, 171(259), 4510 (SEK, thousands) |
| `numEDebt` | numeric | 0, 5.53(6.39), 42 |
| `collectionRate_1` | numeric | 0, 0.0031(0.0216), 0.5288 |
| `collectionRate_3` | numeric | 0, 0.0136(0.0478), 0.8329 |
| `collectionRate_6` | numeric | 0, 0.0387(0.0977), 1 |

### 3.1.4 Splitting data into training- and test set

The data set is divided in training and test sets which is standard procedure in model evaluation. The split is done differently for the two different models, combined classification-regression model and Markov chain model. The reason behind this is mainly that the models partly aim to do different things.

The primary objective for the combined model is to investigate the existence of a relation between attributes of a debtor and collection rate. And if there exists a relation, find the driving factors. The portfolio valuation is secondary. The model is trained on the train set and applied to the test set. During the evaluation of the model, this is done several times. Qualitative measures are taken each time and averaged over the rounds. This is done to avoid an unlucky or lucky split of data potentially deciding the result. The division of training and test set is done solemnly at random at a 75/25 ratio for train and test.

The Markov model is constructed based on the training set in order to value the test set. The split is done at a 75/25 ratio. However, now the split is done based on registration date where the training set consist of the 75% of accounts that are earliest in time. The test set is then the 25% latest in time. This means that in Figure 3.2, the top 75% is the training set and the bottom 25% is the test set.

## 3.2 Classification and Regression Model

Our model is inspired by logic developed by Kriebel & Yam and Belotti et al. who use different regression models to forecast collection rate [14, 19]. L.C. Thomas et al. and Belotti & Ye proposes models which are combination of classification and regression models [15,16]. The combined classification-regression model is used when the distribution of collection rate is bi- or trimodal. Models like these are based on assumptions that one can group together accounts with similar collection rate based on their characteristics. It seems logical that accounts that are fully repaid have more in common with each other than with accounts with low repayment. If the distributions are bi- or trimodal there will exist clearer segments, which make for using a classifier more appropriate.

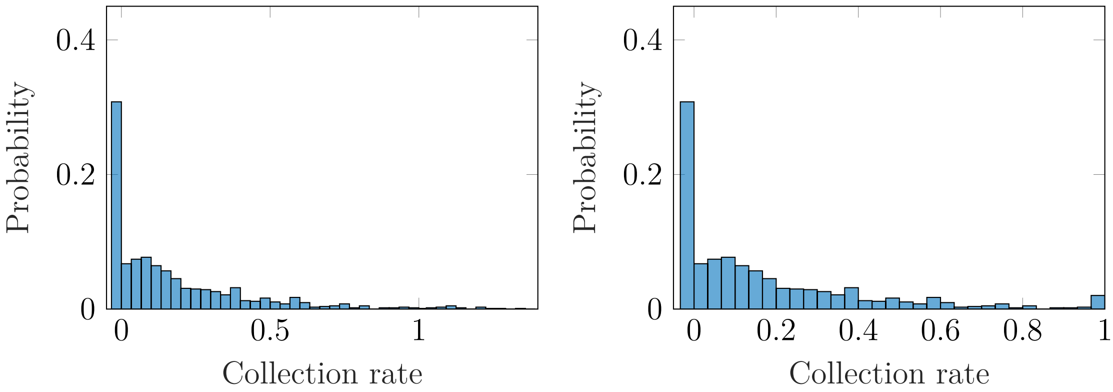
### 3.2.1 Dependent variable – Collection Rate (CR)

Collection rate is a measure of repayment of loans. To put it simply it is the proportion of the debt that has been repaid. Usually one wants to sum repayments done in a fixed time span, the first 12 months in our study. There are discussions in the literature whether the collection rate should include delayed payments fees and interest or not. Beck et al. and Belotti & Ye do adjust the collection rate for costs associated with damages caused by the delay [16, 18]. Kriebel and Yam, means that the collected money first serve as repayment of the initial debt and secondly covers costs associated with the delay [13].

Our data set contains no information about fees or interest caused by the delay. Hence, when taking the ratio between total amount collected and initial debt, one obtains values on an unbounded interval, see Figure 3.4a. Values above one imply a collection more than the initial debt, obviously the payments then also include fees and interest associated with the delay. The extra fees and interest make up $0-40\%$ of the total debt, which is reasonable. We define collection rate as

$$CR = \min\left\{1,\ \frac{\text{amount collected first 12 months}}{\text{initial principal}}\right\}.$$

To model collection rate, with for instance a regression model, it is generally more convenient to have it on a bounded interval. We have decided to truncate the collection rate within the interval $[0, 1]$ as shown in the definition, which is a standard approach [16]. The assumption that accounts with collected amount more than or equal to the initial debt are fully paid is presumably often satisfied. Especially if one treats the debt and other fees separately, like Kriebel and Yam, where the initial collection service the debt. The truncated collection rate is seen in Figure 3.4b.



**(a)** Non-truncated CR.

**(b)** Truncated CR.

**Figure 3.4:** The distribution of the CR when the CR is truncated and when it is not.

### 3.2.2 Model architecture

We use a combined classification-regression model similar to the one used by Bellotti and Ye [16].

In the classification step accounts are classified in two classes: CR=0 and CR $> 0$. After the classifier has been applied, some accounts will be given a definite value of the collection rate CR=0 and some will be further estimated using a regression model. In the regression step the CR will be estimated in the interval $0 \leq \text{CR} \leq 1$. In the end, each errand will have an estimation of the expected value of the CR after 12 months. By multiplying each accounts initial debt with its estimated CR one obtains the estimated amount to be repaid. Summing up all the accounts estimated

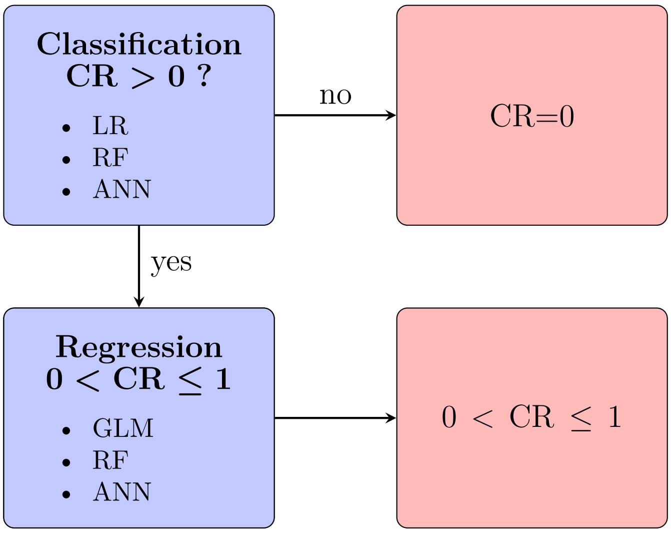


**Figure 3.5:** Illustration of the combined classification regression model to estimate the collection rate for each account based on its attributes.

amount to be repaid will give an indication of the amount of money that one could expect to retrieve from the portfolio as a whole. This translates to the value of the portfolio. The model is illustrated in Figure 3.5.

#### 3.2.2.1 Classification models

Three different classification models are evaluated. These are:

- Logistic regression, LR
- Random forest classification model, RF
- Artificial neural network classification model, ANN

See appendix A.2 for more details. LR is a good base line classifier. It is generally faster and gives more interpretable results. This is the reason we choose to include it. RF and ANN regression models have performed well in similar setups [16]. Hence we were interested in using them also as classification models. The random forest is trained using an ensemble of 30 decision trees. The artificial neural network consists of one input layer, one hidden layer with 10 neurons, and one output layer. The network is trained by propagating training data forward and the error backward, as described in Section 2.5.

#### 3.2.2.2 Regression models

Three different regression models are used:

- Generalized linear model, GLM
- Random forest regression model, RF
- Artificial neural network regression model, ANN

See appendix A.2 for more details. As GLM is used extensively in the literature [14, 18] we choose to include it. As noted earlier RF and ANN regression models have performed well in similar studies [16]. RF and ANN regression models are trained in the same way as their classification model counterparts. The regression models are evaluated by testing the different algorithms on an isolated regression scenario. This is done by training and testing the models on accounts where the CR is known to be greater than zero. Consequently, all the accounts with CR=0 are filtered out for this evaluation.

#### 3.2.2.3 Combined model

The combined model consists of a combination of one classification and one regression model of the ones listed in Sections 3.2.2.1 and 3.2.2.2. The models are trained on a training set and then applied to a test set. The classification models are trained on all the data in the training set. The regression models are trained only on accounts with CR $>$ 0 from the training set.

#### 3.2.2.4 Performance assessment of models

The models are evaluated and compared with each other by qualitative measures of the performances. It is important to choose the measures relevant to the work. As the classification and regression models can be treated as separate models the assessment of them is also done separately and with different metrics.

For evaluation of the classification models the confusion matrix is studied. The confusion matrix illustrates the distribution between true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). This is useful when looking for a better understanding of the model classifications which is valuable information when adjusting the model to avoid certain miss-classifications more than others, i.e. avoid FPs or FNs. With the information given by the confusion matrix it is easy to calculate the accuracy, which is a common evaluation metric for a classification model. The accuracy is given by

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \ .$$

To put it simply it is a measure of how many data points are correctly classified.

Another evaluation method, specific to our problem, is to plot the real distributions of the model classified accounts. This gives a feeling for what type of accounts are classified correctly and not correctly.

The regression models and the combined classification-regression models are also evaluated by comparing the distributions of the estimations and the target values. In addition, some other measures will be taken. In similar studies the mean absolute error (MAE), mean squared error (MSE), and the coefficient of determination ($R^2$) have been used [16, 19]. We choose to use the MAE and $R^2$ as performance metrics.

MAE is calculated according to

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \ ,$$

where $n$ is the total number of data points, $x_i$ is the predicted value and $y_i$ is the target value.

$R^2$ is a measure of the proportion of variance, of the dependent variable, predicted by the independent variables. It is given by

$$R^2 = 1 - \frac{SS_{reg}}{SS_{tot}} \ ,$$

where $SS_{reg}$ is the squared error for the estimates and $SS_{tot}$ is the squared difference between the mean and the data according to

$$SS_{reg} = \sum_i (f_i - \bar{y})^2 \ ,$$
$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \ ,$$

where $\bar{y}$ is the mean of the data, and $f_i$ the estimations.

Another interesting evaluation metric, specific for our case, is the estimated value for the whole portfolio. To increase comparability we choose to study the relative portfolio value, RPV, and define it as the ratio between the estimated portfolio value and the real portfolio value according to

$$\text{RPV} = \frac{\text{estimated portfolio value}}{\text{actual portfolio value}} \ . \tag{3.1}$$

Here, the actual portfolio value is defined as the sum of all cash flow in the portfolio.

It could also be valuable to estimate the variable importance of a model. The variable importance is a measure of the relative impact the different variables or predictors have on the estimation of the CR. This is something that is done for the classification and regression step. For more details see appendix A.3.

### 3.2.3 Applying the model to the blind and informed scenarios

To investigate the concept of predictive capacity in relation to information availability, the model is applied to the two different scenarios; blind and informed. More specifically, for the classification-regression model it is about what variables are available to the model in the train and test phase. Similar to the blind scenario, CR is to be estimated also in the informed scenario. In the blind scenario CR for the period of month 1-12 is estimated. However, in the informed scenario, collection rate for the time period of month 7-12 is estimated. Using this setup, it is possible to use information given by payments during month 1-6, in the estimation of payments to be done during month 7-12. Information about payments done during month 1-6 is given as three different variables: CR for month 1, cumulative CR for month 1-3, and cumulative CR for month 1-6. These variables are used in addition to the regular attributes. See also Table 3.2.

## 3.3 Markov Chain Model

To complement the combined classification and regression model, a Markov chain inspired model is implemented. However, there is a slight difference in the scope and anticipated output from the two approaches. While the combined model uses errand-level attributes to predict an errand-specific collection rate, the Markov model is more of a tool to analyze payment behaviour on a portfolio level as well as a tool to produce accurate portfolio valuations. The approach is to use the predictive power of human behaviour. By focusing on the observed structures of debt repayment, rather than total volumes or shares of debt, prediction could be more accurate. The combined model is, as discussed in Section 3.2.2, developed through a comparative study of candidate algorithms. The Markov model, in contrast, makes no comparison claims and should be thought of as a model framework with potential for further development.

The Markov model is tested in the blind and informed scenarios, and bench-marked against a static pool implementation.

### 3.3.1 Markov model motivation and design

The conceptual motivation for using a Markov approach comes from industry expertise. Debt collection experts at Dignisia acknowledge that debtors tend to continue repaying their debts once they have entered a repayment plan. It is reasonable to assume that payments do not occur at random, but rather follow some sort of structure. Inspection of cash flow data also supports existence of a payment structure. The structure can be captured by defining a set of payment states adopted by each debtor, and letting transitions between states adhere to conditions of a Markov chain, see Definitions 1-3 in Section 2.6.

The most important aspect of designing the model is state and time unit definitions. Since the available cash flow data are recorded on a monthly basis it is natural to select the time unit for a step in the chain to be one month. Months are standard units of time in financial applications and has been used in previous Markov modelling attempts [22]. The state definitions, however, are non-trivial. The easiest state definition is to look at each month separately and label it with one of three states: a paying, non-paying or fully paid state. Suggested states are presented in Table 3.4 and the corresponding chain is illustrated in Figure 3.6. Note that the suggested Markov chain is time homogeneous.

**Table 3.4:** Summary of Markov model states

| State name | Symbol | Description |
|---|---|---|
| State 0 | $s_0$ | No payment |
| State 1 | $s_1$ | Payment |
| State 2 | $s_2$ | Debt fully paid |

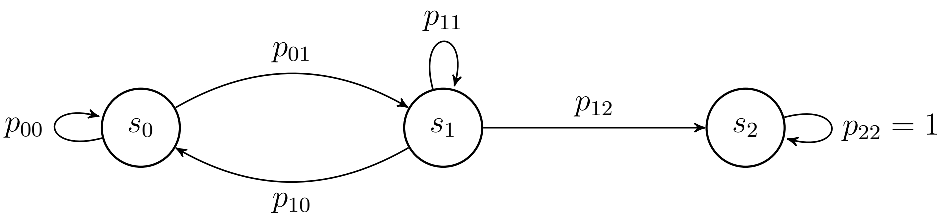


**Figure 3.6:** Time homogeneous Markov model with possible transitions and corresponding probabilities illustrated

An illustrative example of a 6 month cash flow matrix in SEK transformed into a state matrix is shown in (3.2).

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3500 & 0 & 0 & 0 \\ 0 & 0 & 1000 & 1000 & 1000 & 1000 \\ 0 & 0 & 1000 & 0 & 2000 & 1000 \\ 500 & 500 & 517 & 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} s_0 & s_0 & s_0 & s_0 & s_0 & s_0 \\ s_0 & s_0 & s_1 & s_0 & s_0 & s_0 \\ s_0 & s_0 & s_1 & s_1 & s_1 & s_1 \\ s_0 & s_0 & s_1 & s_0 & s_1 & s_1 \\ s_1 & s_1 & s_1 & s_2 & s_2 & s_2 \end{bmatrix} \tag{3.2}$$

Notice how the final row ends in state 2, i.e. full repayment. Full repayment is not distinguishable solely from the cash flow matrix but needs complementary information about original debt. Another observation to make is the similarities between row three and four in the cash flow matrix. It is reasonable to assume that the intention in row four was to follow a structured payment plan of 1000 SEK per month, but that one payment was missed. Therefore, the amount is doubled the following month. This is observed frequently in the real dataset as well.

When the state matrix has been constructed based on the data, it is easy to estimate the transition probabilities $p_{ij},\ i, j = 0, 1, 2$. One simply counts number of occurrences of a specific transition $s_i \rightarrow s_j$ and divides with the total number of occurrences of the originating state $s_i$. By doing this for each transition the transition probability matrix $\mathbf{P}$ is obtained, for definition see Definition 2 in Section 2.6.

The model contains three states, see Table 3.4. The transition matrix is of the form

$$\mathbf{P} = \begin{bmatrix} p_{00} & p_{01} & p_{02} \\ p_{10} & p_{11} & p_{12} \\ p_{20} & p_{21} & p_{22} \end{bmatrix}.$$

As a consequence of the problem design some of the $p_{ij}$ have a restriction on them. Because it is not possible to go from the no payment state $s_0$ to the fully paid state $s_2$, it follows that $p_{02} = 0$. The fully paid state is an absorbing state, hence

$$p_{2j} = \begin{cases} 0, & \text{if } j = 0, 1 \\ 1, & \text{if } j = 2 \end{cases},$$

$\mathbf{P}$ then becomes

$$\mathbf{P} = \begin{bmatrix} p_{00} & p_{01} & 0 \\ p_{10} & p_{11} & p_{12} \\ 0 & 0 & 1 \end{bmatrix}. \tag{3.3}$$

The rest of the transition probabilities have to be calculated based on the state matrix.

### 3.3.2 Markov model as an analysis tool

As the transition matrix $\mathbf{P}$ defines the Markov chain, conclusions can be drawn by analysing the matrix. Given $\mathbf{P}$ and starting in state $s_i$, the expected number of visits to a specific state $s_j$ for a specific time $K$ is given by (2.8) in Section 2.6. Multiplying this with the initial state vector $\mathbf{p}_0$, which is of dimensions $[t \times t]$ and indicates starting distribution between the two transient states, gives the expected number of visits for the different states during the given time period $K$ according to

$$\mathbf{v}_K = \mathbf{p}_0 \mathbf{N}_K \ . \tag{3.4}$$

Especially interesting is to know the expected number of visits to the paying state.

Another interesting quantity is fraction of accounts expected to be absorbed $A_K$, i.e. fully paid, within the $K$ first months. An expression for this is

$$A_K = \mathbf{p}_0 \mathbf{B}_K \mathbf{R} \ . \tag{3.5}$$

Here, $\mathbf{B}_K$ is a $[t \times r]$-matrix where the $i, j$:th entry is the probability that the chain will be absorbed in the absorbing state $s_j$ for a time $K$, given starting in state $s_i$, see (2.9) in Section 2.6.

### 3.3.3 Markov model as a forecasting model

Aside from functioning as an analysis tool, another application for the Markov chain model is portfolio valuation. The valuation is obtained through forecasting cash flow the coming 12 months.

The forecast of the NPL portfolio is conducted through the creation of a Markov chain, which is defined by the empirical transition matrix in the training set. Each account is represented by one Markov chain, and is propagated 11 steps forward to produce a 12 month state matrix. The state matrix is then converted into a cash flow matrix by generating a random payment for each time a Markov chain has been in the payment state $s_1$. The payment is drawn from the empirical distribution of payments in the training set. This is to avoid modelling payments separately.
Number of errands, or accounts, in the portfolio to be modelled decides the number of chains to be simulated. The initial state distribution, $\mathbf{p_0}$, is chosen to be the empirical distribution in the training set.

The distribution from which payments are drawn is a previously known distribution of payments from NPLs-portfolios of similar debt type. One thing to note is that no consideration is taken to the potential difference in debt size of the errands between the portfolio on which the forecast is intended and the portfolio from which the payment distribution is based upon. This is a simplification that has been made.

However, it has not been made without consideration. It is based on the assumption that payments and size of debt are not strongly correlated. For an investigation of this, see Appendix A.4.

Summing the payments from each chain results in a total sum for the whole portfolio. This is the estimated portfolio value. To reduce the effect of stochasticity and obtain as fair a valuation as possible, several portfolios are simulated and the average and standard deviation is calculated.

As well as an end value of the portfolio, the simulations also give an indication of the progress of the portfolio over time. This can be plotted and compared against the real cash flow curve. In the same way as for the classification-regression model, the portfolio valuation can be expressed as a fraction of the actual value, see (3.1) for definition of relative portfolio valuation, called RPV.

For comparison and benchmark in portfolio valuation, the static pool method is used. In our setting the static pool method valuates the test set based on information given in the training set.

### 3.3.4 Interpretation of blind and informed scenarios

In the blind scenario an NPLs-portfolio is valuated based on no prior knowledge of the errands. The data set is split up in two, one training set and one test set, see Section 3.1.4. The transition matrix is constructed based on the information given in the training set. A number of chains, equivalent to number of accounts in the test set is simulated. The payments are drawn from the empirical distribution of payments from the training set. 500 portfolios are simulated and an average is taken.

In the informed case, the first six months cash flow is known in the test set. This is to simulate a scenario where a portfolio analyst wants to update a blind portfolio valuation for instance. The aim is to forecast cash flow the coming half year, i.e. cash flow month 7 to 12. Three changes are made compared to the blind simulations; changing transition matrix, starting distribution and number of steps taken in the chain. Starting distribution $\mathbf{p}_0$ is set to the empirical state distribution in month 6 and the the number of propagation steps is changed from 11 to 6. The changes to the transition matrix are a bit more complicated. By merging the training set transition matrix $\mathbf{P}_{\text{train}}$ with the empirical transition matrix in the informed period $\mathbf{P}_{\text{1-6}}$, the informed transition matrix

$$\mathbf{P}^{\text{informed}} = w_{\text{bias}}\mathbf{P}_{\text{1-6}} + (1 - w_{\text{bias}})\mathbf{P}_{\text{train}}$$

is obtained. The parameter $w_{\text{bias}} \in [0, 1]$ governs the how much weight is put on the new information, i.e. cash flow from month 1 through 6. The parameter is important when the two merged matrices differ. The value

$$w_{\text{bias}} = 0.75$$

is used for all debt types in the simulations. A range of values where tested, but the variance in the results was negligible.

# 4
# Results

In this chapter, results are presented from the classification-regression model and the Markov chain model. The results are presented independently of each other in different sections. Respective section further describes the way in which results are presented.

## 4.1 Combined Classification-Regression Model

The combined model is made up of a classification step and a regression step. The two steps can be isolated and analysed separately. This is also what has been done in this study. The results from a classification setting is presented, where models are compared against each other. Afterwards, results from a regression setting is presented where models are compared to each other. Finally, different combined models are tested and evaluated.
Qualitative measures, see Section 3.2.2.4, are taken and presented in tables. To get a more fair evaluation, the measures are taken 30 times and an average are presented together with the sample standard deviation, $\hat{\sigma}$.

### 4.1.1 Blind scenario

The results of the separate evaluation of the classification and regression models are presented in Sections 4.1.1.1 and 4.1.1.2. The evaluation of the combined model, consisting of different combinations of classification and regression models, are presented in Section 4.1.1.3.

#### 4.1.1.1 Classification models

Table 4.1 shows that three models are similar in their performances with accuracies in the interval 68-69%. LR and ANN classifies around 16% of the accounts as CR=0 while RF classifies 22% of the accounts as CR=0.

By analysing the confusion matrix, a more nuanced picture of the performance is obtained, Figure 4.1. The matrices, where the numbers corresponds to percentages, show similar characteristics. The majority of the errands with CR $> 0$ are classified correctly, which is because most errands are classified as CR $> 0$. On the other hand this results in a lot of FP, which in other words means that many accounts with actual class CR=0 are classified as CR $> 0$.Further, by looking at the actual distributions for the different classes of the classified accounts valuable information is obtained, see Figure 4.2. The different classification models show similar results. The distributions for the classes, CR = 0 and CR > 0, seem to be different. The models have been able to do some kind of distinction.

**Table 4.1:** Comparison of different classification models for the blind scenario through qualitative measures. The true distribution of accounts between CR=0 and CR > 0 are ≈ 31%/69%(depends on the split in train and test set).

| Class. model | Accuracy ( ± $\hat{\sigma}$) | Prop. classified CR=0/CR > 0 |
|---|---|---|
| LR | 68.6% ± 2.8% units | 15.8% / 84.2% |
| RF | 68.7% ± 2.1% units | 22.1% / 77.9% |
| ANN | 68.4% ± 2.1% units | 17.3% / 82.7% |

| Actual | | Predicted | |
|---|---|---|---|
| | | CR=0 | CR > 0 |
| Actual | CR=0 | 7.8 | 22.7 |
| | CR > 0 | 7.9 | 61.6 |

**(a)** Logistic regression.

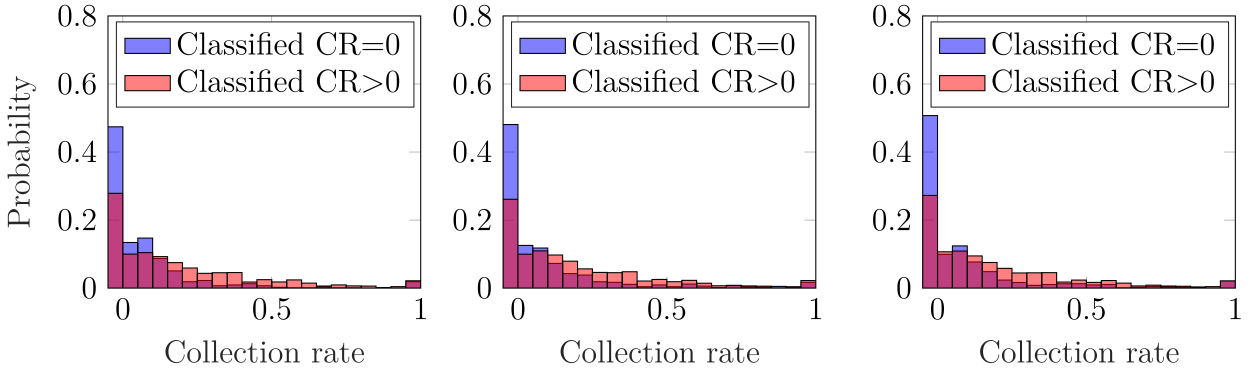
| | Predicted | |
|---|---|---|
| | CR=0 | CR > 0 |
| CR=0 | 10.7 | 19.9 |
| CR > 0 | 11.3 | 58.1 |

**(b)** Random forest.

| | Predicted | |
|---|---|---|
| | CR=0 | CR > 0 |
| CR=0 | 9.3 | 21.3 |
| CR > 0 | 8.6 | 60.8 |

**(c)** Artificial neural network.

**Figure 4.1:** Confusion matrices for the respective classification model on a test set in the blind scenario. Numbers are given in percentages.



**(a)** Logistic regression.

**(b)** Random forest classification model.

**(c)** Artificial neural network classification model.

**Figure 4.2:** The trained classification models applied on a test set in the blind scenario. The actual distribution of the two different classes, CR = 0 and CR > 0, classified by the different models are shown. The leftmost bin only contains CR=0. Optimal is to have blue only at zero and red only at bins greater than zero.

#### 4.1.1.2 Regression models

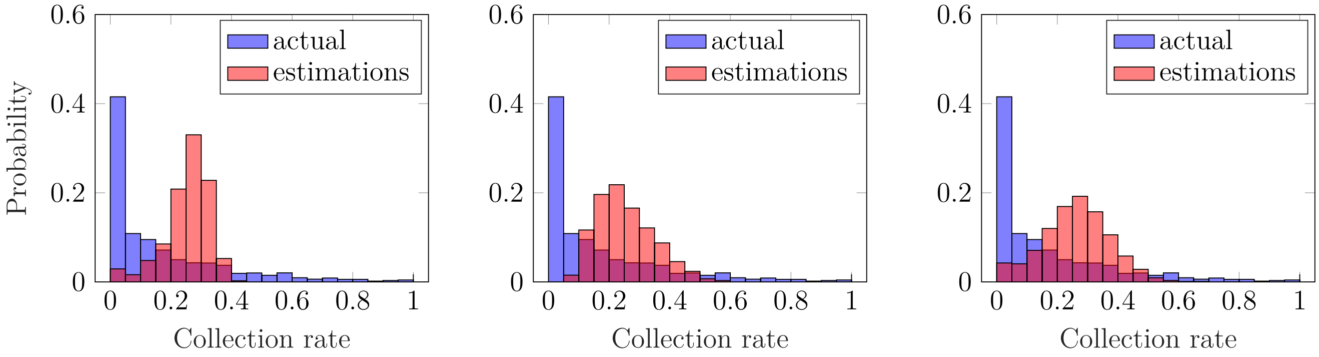
In Table 4.2 the results for the regression step is presented for the different models. Note that for the isolated regression step, only accounts with CR > 0 are used.

Judging from the MAE, RF performs best, GLM is the second best and the ANN performs worst. Similar observations are made when judging from the $R^2$. RF is the better one with a value of 0.14. GLM has an $R^2$ of 0.09 while ANN's $R^2$ is zero, which means that the explanatory strength of the model is zero. For all models the the variance of the $R^2$ is high. Surprisingly, there does not seem to be any correlation between $R^2$ and portfolio valuation.

The relative portfolio valuation is presented in the third column in Table 4.2. One could expect that the better performing model, with the lowest MAE and highest $R^2$ would have a portfolio valuation closest to the actual one but that is not the case as can be seen. By looking at the distribution of the estimations, further insights of how the different regression models behave are obtained, see Figure 4.3. All three distributions share a common characteristics in the sense that they tend to have most of their mass close to the mean of the actual CR.

**Table 4.2:** Comparison of the different regression models in the blind scenario through qualitative measures. The results are averages and standard deviations of 30 measurements.

| Regression model | MAE $\pm \hat{\sigma}$ | $R^2 \pm \hat{\sigma}$ | RPV $\pm \hat{\sigma}$ |
|---|---|---|---|
| GLM | $0.168 \pm 0.012$ | $0.089 \pm 0.038$ | $1.07 \pm 0.13$ |
| RF | $0.156 \pm 0.010$ | $0.143 \pm 0.05$ | $1.12 \pm 0.12$ |
| ANN | $0.176 \pm 0.015$ | $-0.002 \pm 0.08$ | $1.12 \pm 0.21$ |



**(a)** Generalized linear model.

**(b)** Random forest regression model.

**(c)** Artificial neural network regression model.

**Figure 4.3:** The trained regression models applied on a test set for the blind scenario. The distribution of the model estimations together with the distribution of the actual estimations are visualized in the same plot. Only accounts with CR > 0 are used when evaluating the regression models separately.

#### 4.1.1.3 Combined model

For the combined model, one classification model and one regression model is chosen. Because RF regression model perform notably better than the other regression models it is chosen to be the regression model for all the combined models. For the classification step, the three different classification models perform equally well more or less. All three is then tested for different combined models. Results are presented in Table 4.3.

Judging from the MAE the models are similar. RF-RF performs slightly better than ANN-RF which performs slightly better than LR-RF. For all the combined models the $R^2$ is around zero with a high variance.

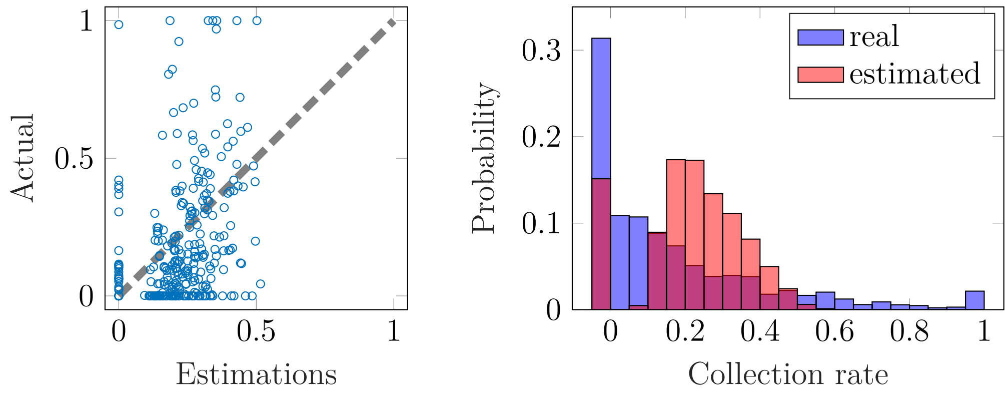
The degree of explanation can also be investigated by looking at a correlation plot between the estimated value of the CR and the actual value of the CR, Figures 4.4a, 4.5a, and 4.6a. The data points deviate from the diagonal line in general. In other words there is no clear relation between estimated CR and actual CR. Additionally the deviations does not appear in a structured manner, which indicates no relation rather than systematic model error.

The distribution of the estimated CRs for the accounts are compared with the actual distribution of CRs in Figures 4.4b, 4.5b, and 4.6b. The models do not estimate CR to be above about 0.6 for any accounts even though there exist a number of accounts with actual CR above that. Estimates are concentrated around the mean.

Finally, an analysis is conducted on the notion of predictor importance, see Figure 4.7. This is done for the RF classification model and the RF regression model for the isolated classification and regression step. As the $R^2$ is as low as it is, the results should not be taken too seriously upon. However, the indications are that `debtType`, `numEDebt`, and `debt` are more important than the other predictors.

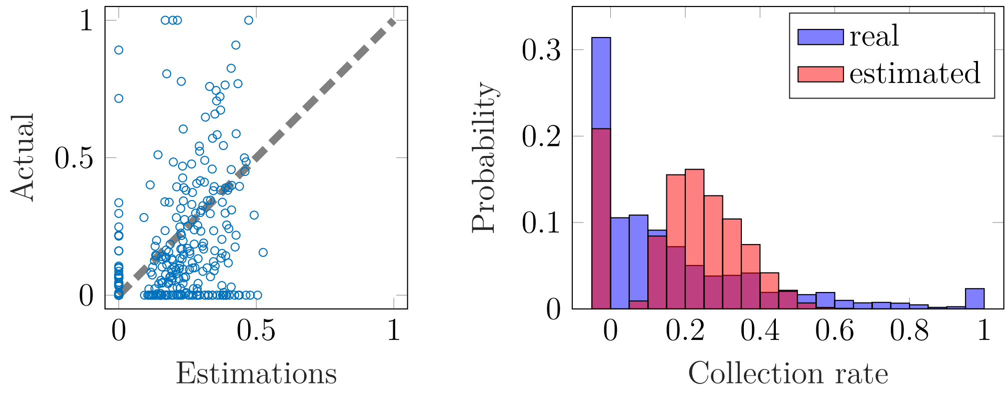
**Table 4.3:** Comparison of the different combined models in the blind scenario, through qualitative measures. The qualitative measures are an average over 30 measurements. The standard deviation is also included.

| Combined Model | MAE $\pm\ \hat{\sigma}$ | $R^2 \pm\ \hat{\sigma}$ | RPV $\pm\ \hat{\sigma}$ |
|---|---|---|---|
| LR + RF | $0.172 \pm 0.007$ | $-0.005 \pm 0.063$ | $1.22 \pm 0.16$ |
| RF + RF | $0.163 \pm 0.007$ | $0.010 \pm 0.077$ | $1.21 \pm 0.18$ |
| ANN + RF | $0.168 \pm 0.009$ | $-0.029 \pm 0.080$ | $1.25 \pm 0.23$ |



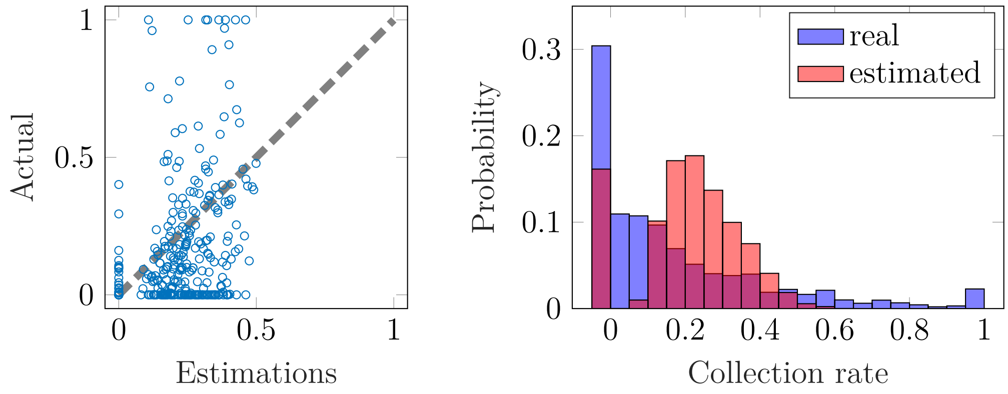
**(a)** Account-wise comparison. **(b)** Comparison of distributions.

**Figure 4.4:** Combined model consisting of LR classifier and RF regression model applied on a test set in the blind scenario.



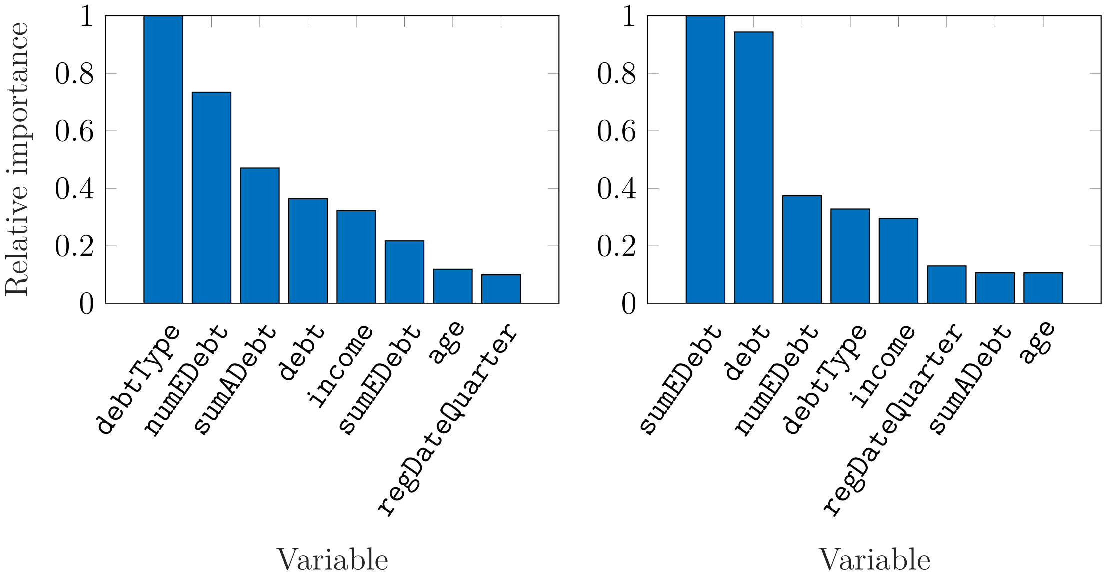
**(a)** Account-wise comparison. **(b)** Comparison of distributions.

**Figure 4.5:** Combined model consisting of RF classifier and RF regression model applied on a test set in the blind scenario.



**(a)** Account-wise comparison. **(b)** Comparison of distributions.

**Figure 4.6:** Combined model consisting of ANN classifier and RF regression model applied on a test set in the blind scenario.



**(a)** Classification step. All errands.

**(b)** Regression step. Only errands with CR > 0

**Figure 4.7:** Estimations of the predictive capacities for the different variables for the RF model in the blind scenario. The error is a measure of the out of bag permuted predictor delta error which has been normalized. The predictive capacity is calculated for the isolated classification step and the isolated regression step.

### 4.1.2 Informed scenario

In this subsection, the results of applying a classification-regression model in an informed scenario are presented. The variable to be estimated is CR for month 7-12. The model performances in the informed scenario will be compared with the model performances in the blind scenario in Section 4.1.1.

Results from classification-regression models applied to an informed scenario are displayed in Table 4.4.

**Table 4.4:** Comparison of the different combined models in the informed scenario through qualitative measures. Qualitative measures are averages over 30 measurements. The standard deviation is also presented.

| Model | Accuracy class. step | $R^2$ regress. step $\pm$ $\sigma$ | MAE full model $\pm$ $\sigma$ | $R^2$ full model $\pm$ $\sigma$ | RPV $\pm$ $\sigma$ |
|---|---|---|---|---|---|
| LR+GLM | 0.728 | 0.142 ± 0.041 | 0.089 ± 0.006 | 0.051 ± 0.067 | 0.910 ± 0.157 |
| RF+RF | 0.773 | 0.125 ± 0.064 | 0.087 ± 0.005 | 0.048 ± 0.077 | 1.088 ± 0.150 |

In the classification step the accuracies for the two different models are 72.8% and 77.3% where RF is the one with the highest value. This is about 9% units more than the blind scenario, see Section 4.1.1. 63.2% and 36.8% of the accounts are classified as CR=0 and CR > 0 respectively, for both LR and RF in the informed scenario, see Figure 4.8.

| | | Predicted | |
|---|---|---|---|
| | | CR=0 | CR > 0 |
| Actual | CR=0 | 50.0 | 14.0 |
| | CR > 0 | 13.2 | 22.9 |

**(a)** Logistic regression.

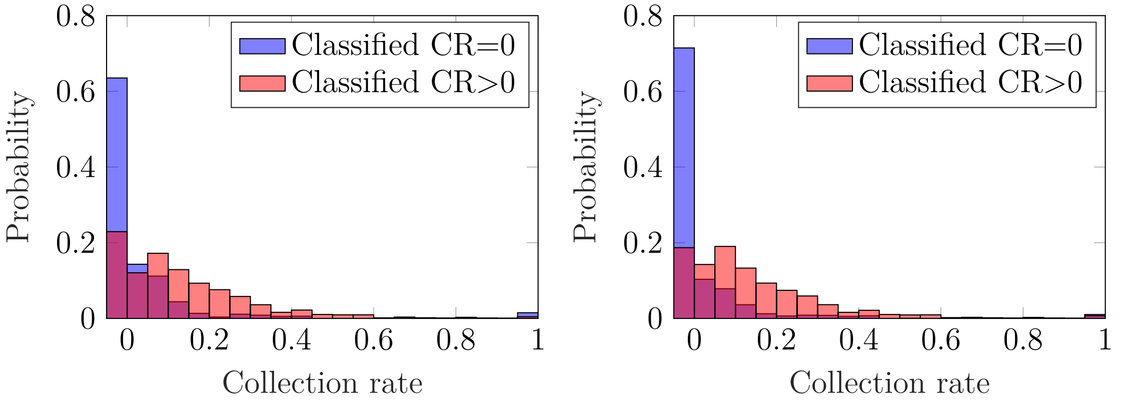
| | | Predicted | |
|---|---|---|---|
| | | CR=0 | CR > 0 |
| Actual | CR=0 | 52.0 | 11.5 |
| | CR > 0 | 11.2 | 25.4 |

**(b)** Random forest.

**Figure 4.8:** Confusion matrices for the respective classification models on a test set for the informed scenario. Numbers are given in percentages.

The real distributions of the model classified accounts are seen in Figure 4.9. Both the models share similar results. The majority of the accounts classified as CR=0 is in reality 0 or close to 0 with the exception of a number of accounts with actual CR of 1 or close to 1.

Analyzing the regression step separately, GLM performs better than RF by both having a higher value and a lower variance. The $R^2$ is 2-6% units higher in the informed scenario compared to the blind scenario. However, for all the models in both scenarios the variance of the $R^2$ is high.



**(a)** Logistic regression.

**(b)** Random forest classification model.

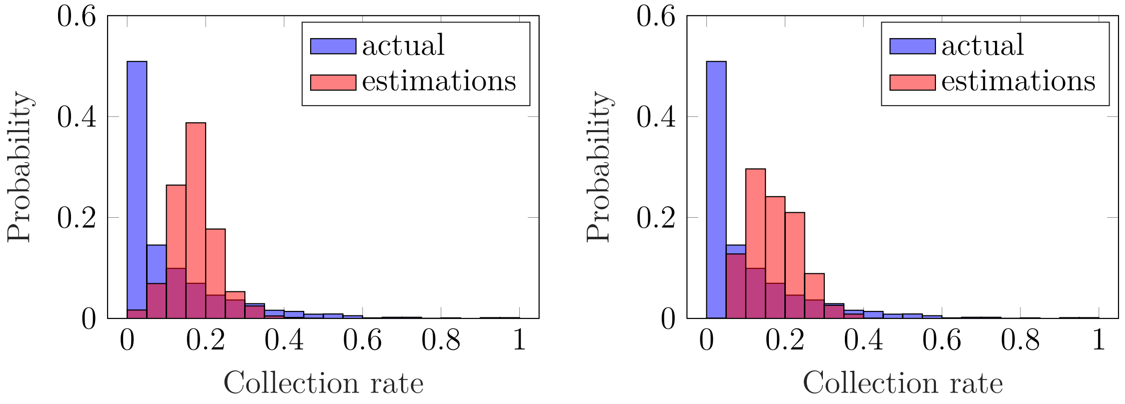
**Figure 4.9:** The trained classification models applied on a test set in the informed scenario where knowledge about prior payments is available. The actual distribution of the two different classes, CR = 0 and CR > 0, classified by the different models are shown. The leftmost bin only contains CR=0. Optimal is to have blue only at zero and red only at bins greater than zero.

Similarly as in Section 4.1.1.2, analysis is conducted separately on the regression step, which means that training and testing is only performed on data with CR $> 0$. The distribution of the estimated values are compared with the distribution of the actual values for the two different models in Figure 4.10. Comparing the models with each other in the informed scenario, GLM has a higher concentration closer to the mean than RF. Comparing the different scenarios blind and informed, in other words comparing the Figures 4.3 and 4.10 with each other, big similarities are seen for the respective models. The way GLM estimates the CR in a blind scenario is very similar to the way it estimates the CR in an informed scenario. The same is true for RF.

For the full model, judging from the MAE RF+RF is the slightly better model with a lower error and a lower variance. The MAE is significantly lower in the informed scenario than in the blind scenario. Judging from the $R^2$ LR+GLM perform slightly better. As expected, both of the models perform better in the informed scenario. Still, the $R^2$ is low and the variance is high. Even though the explanatory strengths of the models are better in the informed scenario, the portfolio valuation does not differ much.

The performance of the combined models are further investigated by comparing estimations of CR with actual values of CR, see Figures 4.11 and 4.12.

Estimations of CR are compared with actual values of CR on an account-wise level, see Figures 4.11a and 4.12a. There exists some visible correlation between estimations and actual values. The exceptions are the accounts with actual CR close to 1, as noted earlier. The distributions of the estimated values of CR are compared to the actual distribution of the values of CR, see Figures 4.11b and 4.12b. They both have an equal amount of accounts classified as 0, which is almost the exact same as the actual distributions. A note to be made is that the distribution does



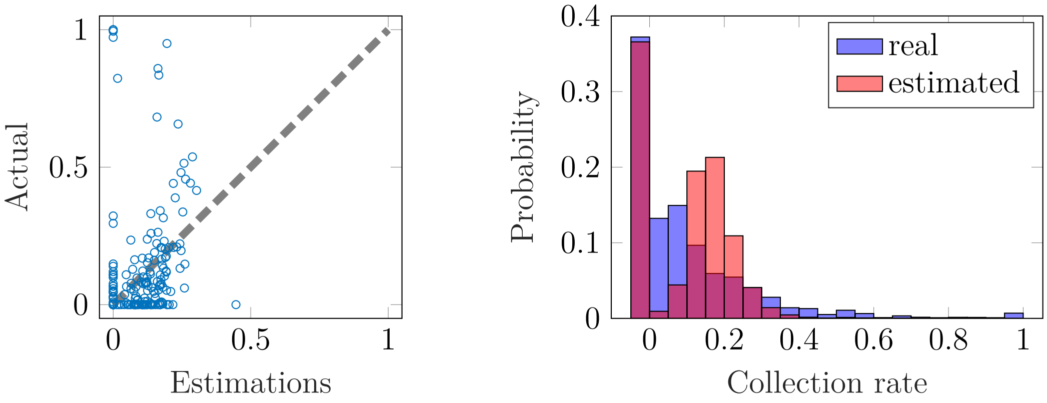
**(a)** Generalized linear model.

**(b)** Random forest regression model.

**Figure 4.10:** The GLM and RF regression models applied on a test set for the informed scenario to evaluate the regression step. The distribution of the model estimations together with the distribution of the actual estimations are visualized in the same plot.

not say anything about the estimations of CR on an account level, two identical distributions could have completely different estimations of CRs on an account level. The distributions is rather a visualization of how the model estimates CR as a whole.

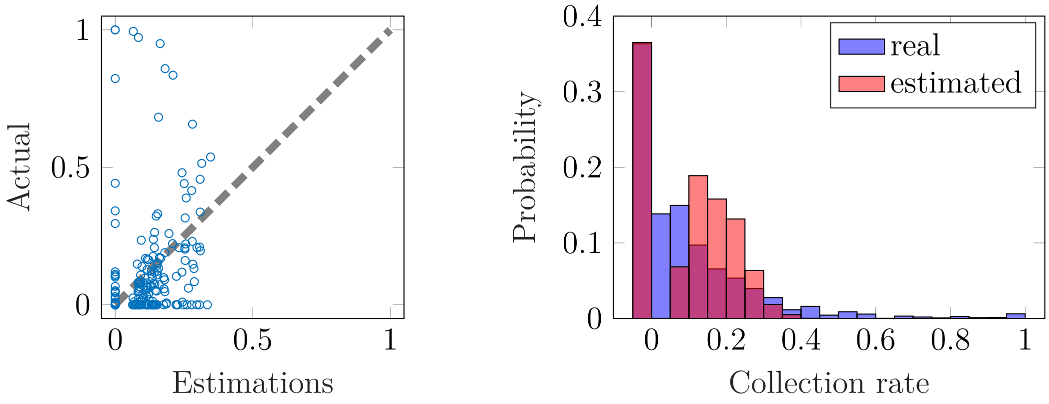
Looking at the predictor importance estimation, see Figure 4.13, not surprisingly `collectionRate_6` is topping for both the classification and regression step. In the regression step the importance relative to the other variables, especially `sumEDebt` and `debt`, is not that significant however. `age` and `regDateQuarter` are still bad predictors which is consistent with findings in the blind scenario. `collectionRate_1` is interestingly not at all a decisive variable.



**(a)** Account-wise comparison.

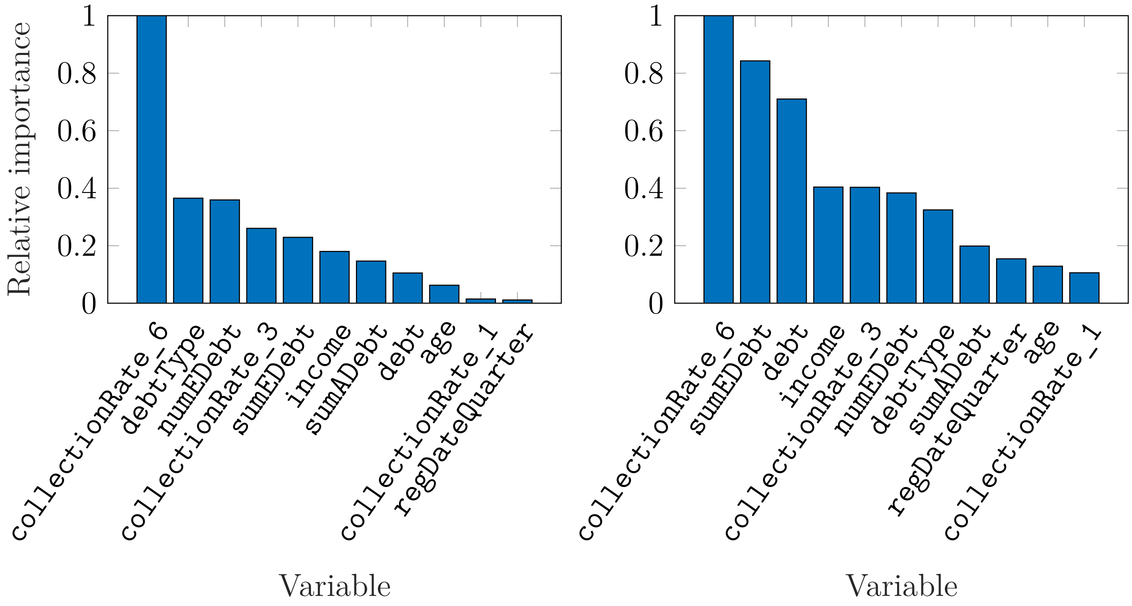
**(b)** Comparison of distributions.

**Figure 4.11:** Combined classification-regression model consisting of an LR classification model and an GLM regression model applied on a test set for the informed scenario.



**(a)** Account-wise comparison. **(b)** Comparison of distributions.

**Figure 4.12:** Combined classification-regression model consisting of an RF classification model and an RF regression model applied on a test set for the informed scenario.



**(a)** Classification step. All errands. **(b)** Regression step. Only errands with CR > 0

**Figure 4.13:** Estimations of the predictive capacities for the different variables for the RF model in the informed scenario. The error is a measure of the out of bag permuted predictor delta error which has been normalized. The predictive capacity is calculated for the isolated classification step and the isolated regression step.

## 4.2 Markov Chain Model

The Markov section of the results are made up of four subsections. In the first subsection, a summary of the most important results is presented. The following subsections are considered as deep-dives. The summary is followed by results from using the Markov model as an analysis tool as described in method Section 3.3.2. In the forecasting subsections the model is used as a forecasting tool applied to a blind and an informed scenario respectively.

### 4.2.1 Summary of results

There are differences in debtors' payment patterns across debt types. An invoice debtor who did not produce any cash flow the previous month has 98.5% probability of not paying the coming month either. The corresponding figure for loans and credit card debtors are 89.2% and 81.2% respectively. Conversely, non-paying invoice debtors have a probability of converting to payment of 1.5%. For loans and credit card debtors the conversion probability is 10.8% and 18.8%. The overall probability of an account being fully repaid after 12 months is roughly 1%, and the expected number of months with payment ranges from 0.5 for invoices to 3 and 4 for loans and credit cards.

Regarding forecasting, it is clear that the informed scenario is more successful. This is no revolutionary result, but still speaks for the importance of information availability. The forecasting results are summarized in Table 4.5 and are visualized in Figure 4.14. All results are produced with a chronological training proportion of 75% of available accounts and a recency bias of 75%. In the informed scenario, the first six months of cash flow is given for the test set. See Section 3.3.1 for details.

**Table 4.5:** Portfolio valuation for the Markov model and the static pool method in the scenarios blind and informed. Portfolio valuation numbers are given in million SEK.

| | Debt type | Actual | Markov model | | Static pool | |
|---|---|---|---|---|---|---|
| | | | Port. est. | RPV ($\pm\sigma$) | Port. est. | RPV |
| Blind | Invoice | 0.135 | 0.0426 | $0.315 \pm 0.082$ | 0.038 | 0.279 |
| | Loans | 0.971 | 0.646 | $0.665 \pm 0.101$ | 0.666 | 0.686 |
| | Credit card | 0.726 | 0.924 | $1.273 \pm 0.079$ | 0.935 | 1.288 |
| Informed | Invoice | 0.135 | 0.061 | $0.451 \pm 0.060$ | 0.061 | 0.450 |
| | Loans | 0.971 | 0.625 | $0.644 \pm 0.057$ | 0.688 | 0.708 |
| | Credit card | 0.726 | 0.757 | $1.043 \pm 0.056$ | 0.893 | 1.230 |

Judging from the RPV, credit card debtors behave in the most predictable way for the given NPL portfolio. It is also for credit cards that the improvement in the informed scenario is the greatest. However, as seen in Figure 4.14, the actual portfolio cash flow increases dramatically in month 8 and 12 for invoices and credit cards. Since that behaviour is not observed in the training data, it is impossible to capture with the modelling framework at hand. Another interesting observation to

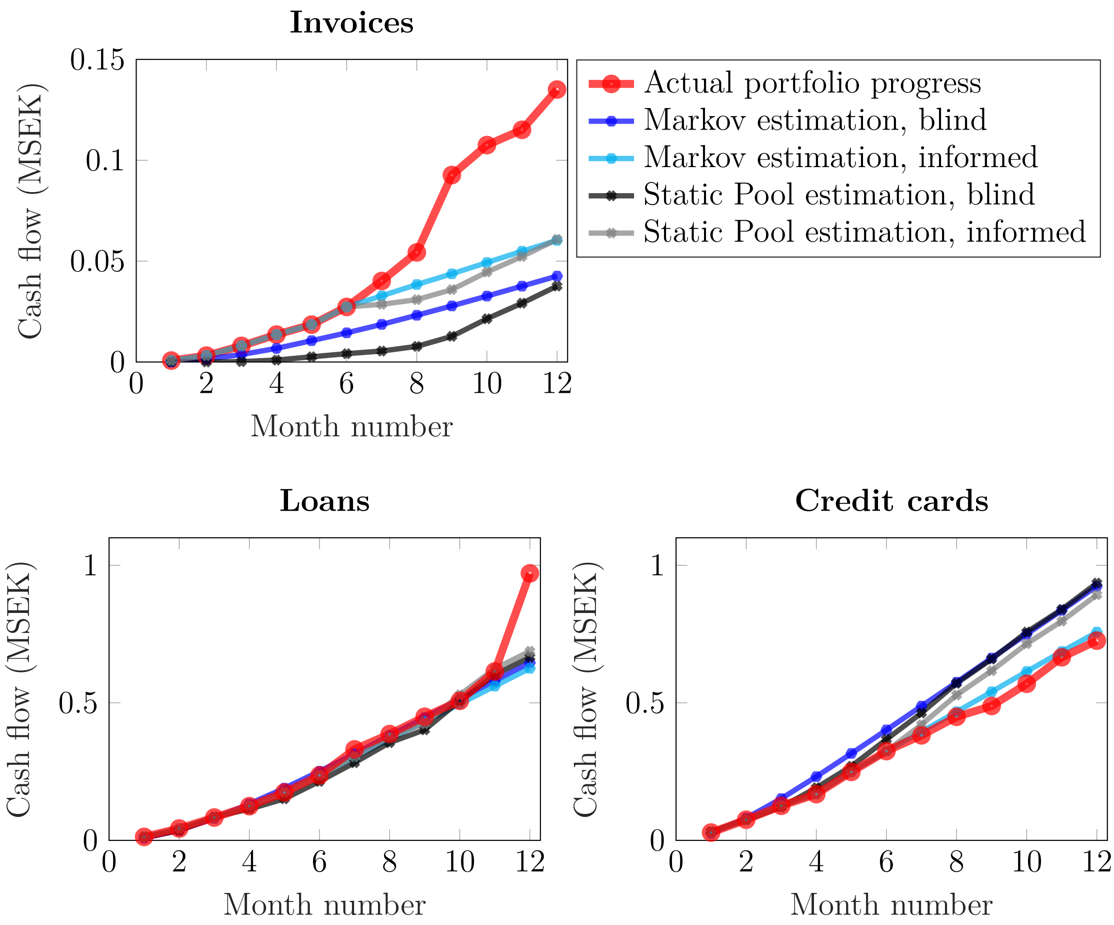


**Figure 4.14:** Markov model cash flow forecast in blind and informed scenarios, compared to portfolio progress and static pool forecast.

make is that the Markov model is on par or better than the static pool method in every scenario.

### 4.2.2 Model insights

Recall from Section 2.6 the notation

$$p_{ij} = \mathbb{P}(X_{n+1} = j | X_n = i), \quad i, j \in \{0, 1, 2\}$$

indicating the probability of transitioning from state $i$ to $j$. Also recall from Table 3.4 in Section 3.3.1 the definition of the three states, $s_0$, $s_1$, and $s_2$, and the construction of the transition matrix

$$\mathbf{P} = \begin{bmatrix} p_{00} & p_{01} & 0 \\ p_{10} & p_{11} & p_{12} \\ 0 & 0 & 1 \end{bmatrix}.$$

The transition probabilities $p_{i,j}$ are calculated by counting the states and transitions between them for the cash flow matrix. Initially, this is done for the whole data set, as the intention is to confirm the validity of the model, analyse the behaviour and draw conclusions from all the accounts.

The transition matrices for the different debt types becomes:

$$\mathbf{P}^{\text{invoice}} = \begin{bmatrix} 0.985 & 0.015 & 0 \\ 0.307 & 0.685 & 0.008 \\ 0 & 0 & 1 \end{bmatrix} \tag{4.1a}$$

$$\mathbf{P}^{\text{loans}} = \begin{bmatrix} 0.892 & 0.108 & 0 \\ 0.246 & 0.747 & 0.007 \\ 0 & 0 & 1 \end{bmatrix} \tag{4.1b}$$

$$\mathbf{P}^{\text{cc}} = \begin{bmatrix} 0.812 & 0.188 & 0 \\ 0.268 & 0.729 & 0.002 \\ 0 & 0 & 1 \end{bmatrix}. \tag{4.1c}$$

There are several interesting conclusions to draw from the construction and structure of the transition matrices.

First, the probabilities confirms the initial assumption of the payments being ordered. The assumption is that a debtor is more likely to remain in the same state. Expressed in a more mathematical way, $p_{ii} > p_{ji}$, $\forall i, j$ where $i \neq j$. Had payments been randomly distributed, the following would have been valid, $p_{ii} = p_{ji}$, $\forall i, j$ where $i \neq j$, with exceptions of small perturbations because of stochasticity. Observing (4.1) we can tell this is not the case. Note: Because $p_{13} = 0$ and $p_{23} \geq 0$ by design, the condition for complete random distribution of payments cannot be fully met in general. However, because $p_{23} << 1$ this does not have to be accounted for when investigating the question of randomness of payments. The bottom line is that the assumption that payments are ordered, more specifically that it is more likely to remain in a state than to change state, are supported by the data.

As noted before the the transition matrices given in (4.1) are all absorbing Markov chains, see Definitions 5 and 6 in Section 2.6.

The expected number of visits to the transient states $s_0$ and $s_1$, given the initial distribution $\mathbf{p}_0$ is given by the entries in vector $\mathbf{v}_{12}$, for formula see (3.4) in Section 3.3.1.

$$\mathbf{v}_{12}^{\text{invoice}} = \mathbf{p}_0^{\text{invoice}}\mathbf{N}_{12}^{\text{invoice}} = \begin{bmatrix}0.997 & 0.0034\end{bmatrix}\begin{bmatrix}11.574 & 0.415\\ 8.366 & 3.408\end{bmatrix} = \begin{bmatrix}11.56 & 0.422\end{bmatrix}$$

$$\mathbf{v}_{12}^{\text{loan}} = \mathbf{p}_0^{\text{loan}}\mathbf{N}_{12}^{\text{loan}} = \begin{bmatrix}0.969 & 0.031\end{bmatrix}\begin{bmatrix}9.173 & 2.746\\ 6.277 & 5.473\end{bmatrix} = \begin{bmatrix}9.09 & 2.83\end{bmatrix}$$

$$\mathbf{v}_{12}^{\text{cc}} = \mathbf{p}_0^{\text{cc}}\mathbf{N}_{12}^{\text{cc}} = \begin{bmatrix}0.873 & 0.127\end{bmatrix}\begin{bmatrix}7.945 & 4.015\\ 5.734 & 6.180\end{bmatrix} = \begin{bmatrix}7.66 & 4.29\end{bmatrix}$$

The expected proportion of accounts to be fully paid during the first 12 months, $A_{12}$, is given by

$$A_{12}^{\text{invoice}} = \mathbf{p}_0^{\text{invoice}}\mathbf{B}_{12}^{\text{invoice}} = \begin{bmatrix}0.9966 & 0.0034\end{bmatrix}\begin{bmatrix}0.0032\\ 0.0265\end{bmatrix} = 0.33\ \%$$

$$A_{12}^{\text{loan}} = \mathbf{p}_0^{\text{loan}}\mathbf{B}_{12}^{\text{loan}} = \begin{bmatrix}0.969 & 0.031\end{bmatrix}\begin{bmatrix}0.0183\\ 0.0365\end{bmatrix} = 1.89\ \%$$

$$A_{12}^{\text{cc}} = \mathbf{p}_0^{\text{cc}}\mathbf{B}_{12}^{\text{cc}} = \begin{bmatrix}0.873 & 0.127\end{bmatrix}\begin{bmatrix}0.0088\\ 0.0135\end{bmatrix} = 0.94\ \%.$$

For formula see (3.5) in Section 3.3.1.

The above metrics, $\mathbf{v}_{12}$ and $A_{12}$, are compared to the metrics from the actual data set. They are easily found by just counting the average number of occurrences of each state to find $\mathbf{v}_{12}$ or the proportion of fully paid accounts to find $A_{12}$. The estimated metrics based on the transition matrices in the Markov model and the actual metrics found from the empirical data set are displayed in Table 4.6.

**Table 4.6:** Metrics from the system expressed as a Markov chain transition matrix compared to the metrics of the actual empirical data set.

| | Debt type | $\mathbf{v}_{12}$: # visits to $s_0$ or $s_1$ | $A_{12}$ |
|---|---|---|---|
| Analysis of trans. | loan | $s_0$: 9.085, $s_1$: 2.830 | 1.89 |
| matrix | credit card | $s_0$: 7.664, $s_1$: 4.290 | 0.94 |
| Actual metrics | loan | $s_0$: 9.186, $s_1$: 2.793 | 1.65 |
| from data | credit card | $s_0$: 7.782, $s_1$: 4.209 | 0.81 |

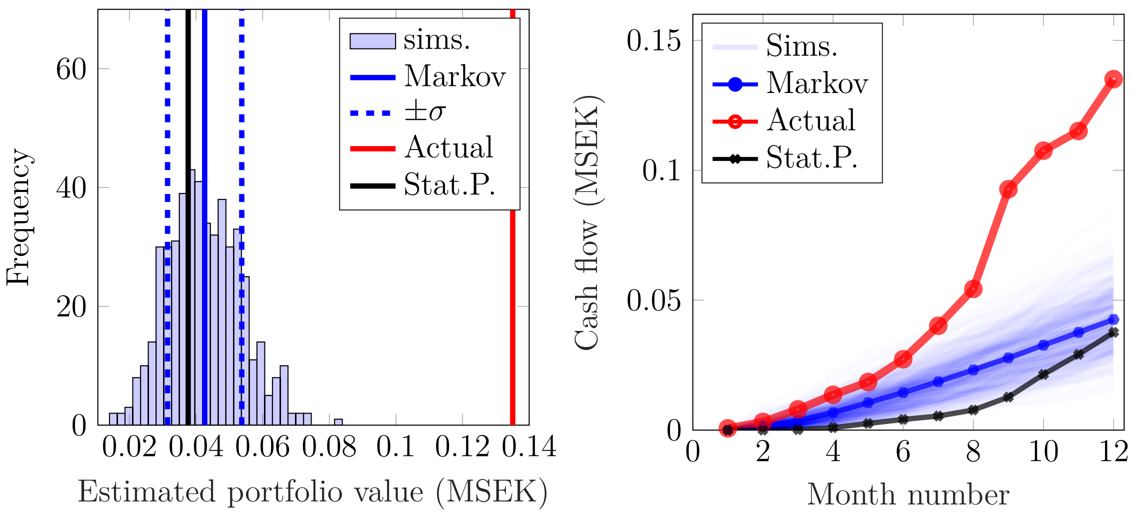
**Table 4.7:** Empirical transition probabilities for three debt types. Numbers annotated with a dagger ($^\dagger$) are not results, but consequences of the state design. See method Section 3.3.1 for details.

| Transition | Invoice | Loan | Credit card |
|---|---|---|---|
| $p_{00}$ | 0.9849 | 0.8924 | 0.8120 |
| $p_{01}$ | 0.0151 | 0.1076 | 0.1880 |
| $p_{02}$ | $0^\dagger$ | $0^\dagger$ | $0^\dagger$ |
| $p_{10}$ | 0.3074 | 0.2459 | 0.2684 |
| $p_{11}$ | 0.6848 | 0.7474 | 0.7294 |
| $p_{12}$ | 0.0078 | 0.0067 | 0.0022 |
| $p_{20}$ | $0^\dagger$ | $0^\dagger$ | $0^\dagger$ |
| $p_{21}$ | $0^\dagger$ | $0^\dagger$ | $0^\dagger$ |
| $p_{22}$ | $1^\dagger$ | $1^\dagger$ | $1^\dagger$ |

### 4.2.3 Forecasting - Blind scenario

The interpretation of forecasting in the blind scenario for the Markov model is straightforward: compute a transition matrix based on the training set and extrapolate to the test set. This is not very different from the static pool method, and hence the results are similar. Empirical transition probabilities for the training set for the different debt types is found in appendix A.5. Note how transitions from state $s_0$ to itself or $s_1$ differ substantially, while the transitions from $s_1$ to itself or $s_0$ are similar for the different debt types. This indicates that once a debtor enters a repayment plan, they are equally likely to stay in state $s_1$ across debt types. On the other hand, credit card debtors are almost 20 times more likely than invoice debtors to enter a structured payment plan. For loans, the likelihood is roughly half of the figure for credit cards. This is in line with what is to be expected considering the differences in collection rate distributions, see Figure 3.3.

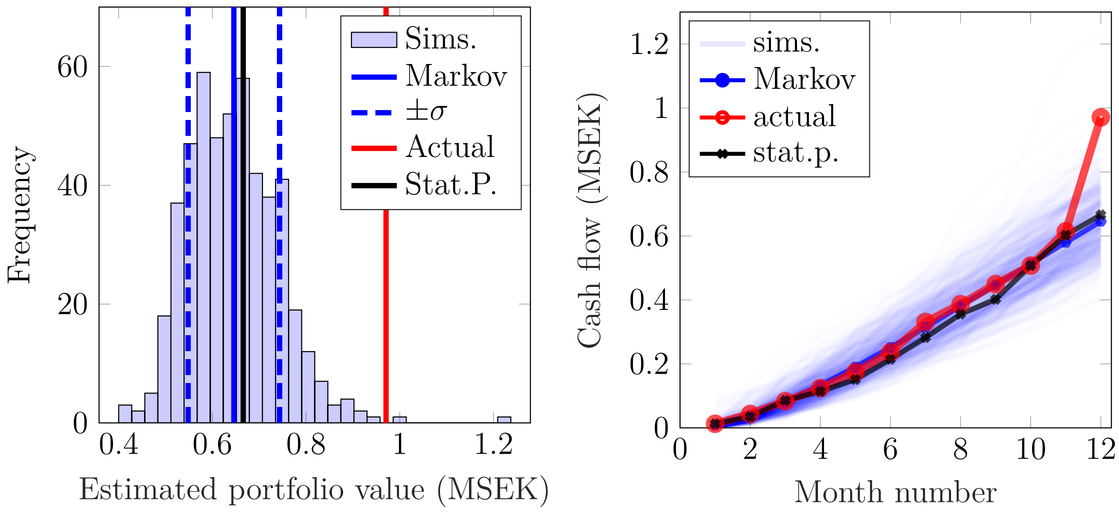
Figures 4.15, 4.16, and 4.17 show simulation results and prediction performance for each debt type in the NPL portfolio. Predictions are fairly stable, as indicated by the low sample standard deviation. The simulations do a good job capturing the overall behaviour in the training set, which is seen by the similarities with the static pool estimate. However, the test set behaves somewhat differently and the blind predictions are not accurate in this case. The RPV and corresponding uncertainties are shown in Table 4.5.



**(a)** Portfolio valuation distribution.

**(b)** Cash flow simulations.

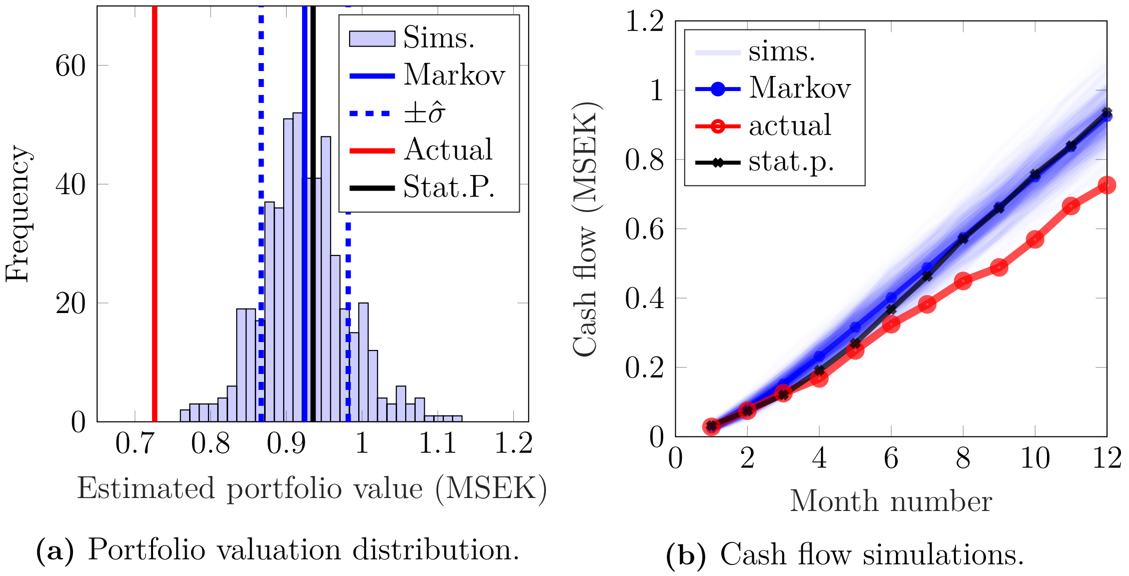
**Figure 4.15:** Simulation of 500 portfolios of **invoices** in the blind scenario.



**(a)** Portfolio valuation distribution.

**(b)** Cash flow simulations.

**Figure 4.16:** Simulation of 500 portfolios of **loans** in the blind scenario.



**(a)** Portfolio valuation distribution.

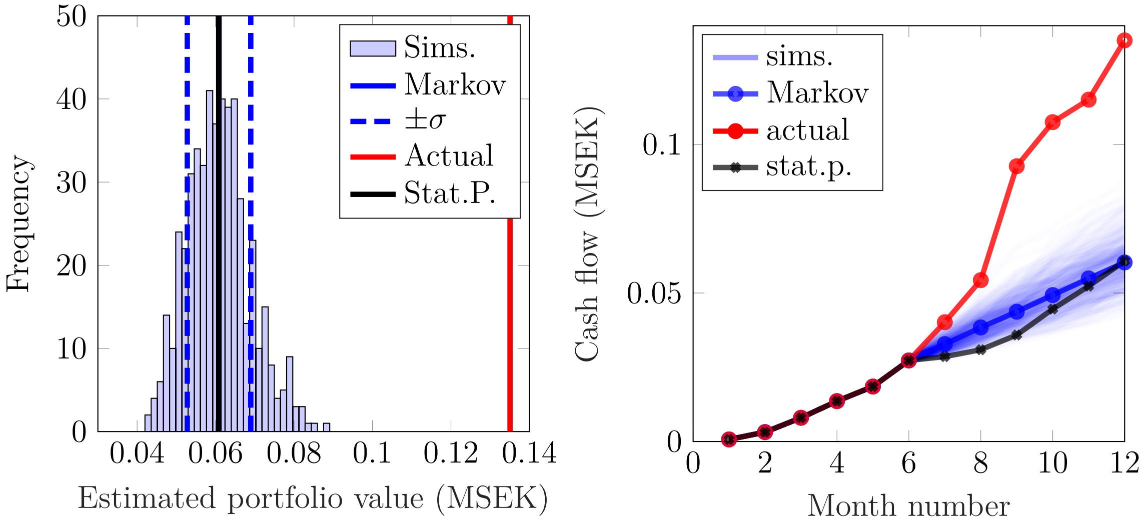
**(b)** Cash flow simulations.

**Figure 4.17:** Simulation of 500 portfolios of **credit cards** in the blind scenario.

### 4.2.4 Forecasting - Informed scenario

From a portfolio point of view, the Markov model applied to the informed scenario performs better for invoice and credit card debts. For invoices the forecast is still very off compared to the actual value. This is simply due to the training set behaving differently from the test set, and that the rapid cash flow increase during month number nine is unseen in training. The improvement in the informed scenario is small but noticeable. For credit card debts, the already good valuation is improved upon. For loans, the informed scenario produces a slightly worse valuation than the blind case, see Table 4.5.

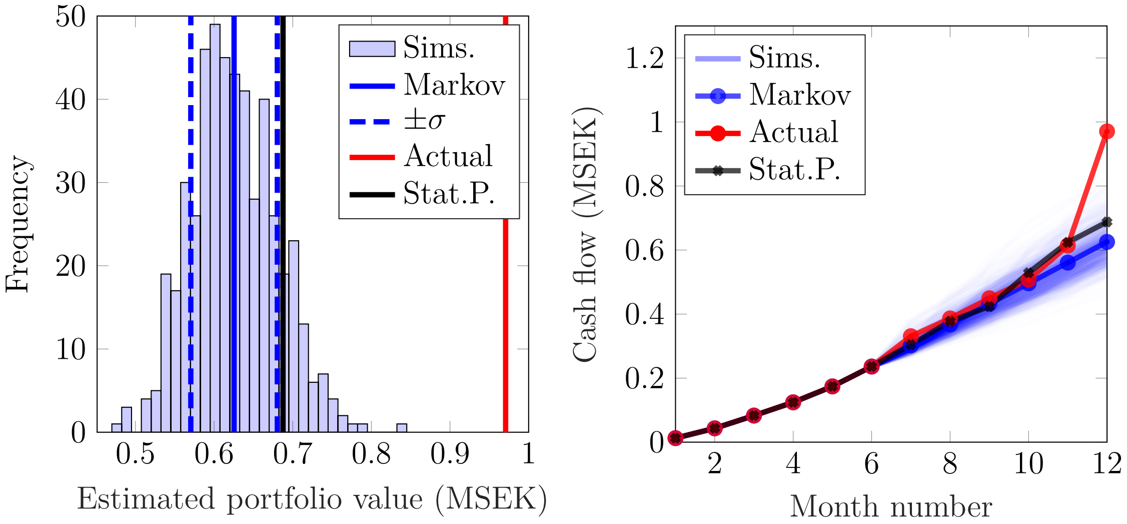
Compared to the static pool method, the informed Markov model performs better for invoice and credit card debts. For loans the static pool performs better, see Table 4.5.



**(a)** Portfolio valuation distribution.

**(b)** Cash flow simulations.

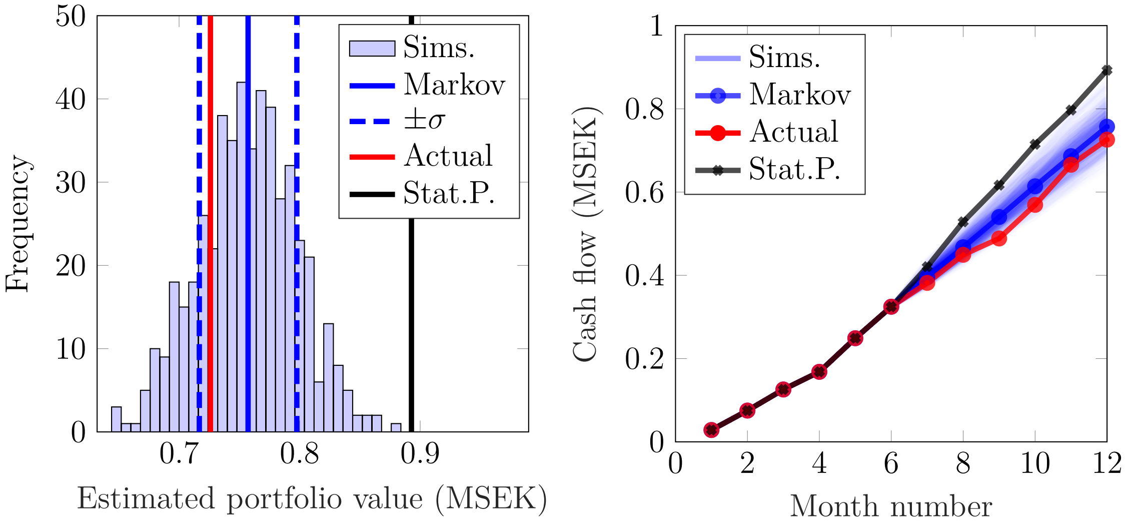
**Figure 4.18:** Simulation of 500 portfolios of **invoices** in the informed scenario.



**(a)** Portfolio valuation distribution.

**(b)** Cash flow simulations.

**Figure 4.19:** Simulation of 500 portfolios of **loans** in the informed scenario.



**(a)** Portfolio valuation distribution.

**(b)** Cash flow simulations.

**Figure 4.20:** Simulation of 500 portfolios of **credit cards** in the blind scenario.

# 5

# Discussion

The discussion is initiated by an analysis of the results of the classification and regression model followed up by an analysis of the results of the Markov chain model. In the literature there is a general consensus that modelling CR is difficult. One of the reasons for this is challenges connected with data availability which affect quantity and quality. Therefore this subject is discussed further. Finally, a section is added that deals with possible extension of the Markov model.

## 5.1 Model Performance Assessment

In this section the performance of the classification-regression model is discussed, based on the results in Section 4.1, and the performance of the Markov model is discussed, based on the results in Section 4.2.

### 5.1.1 Classification-regression model

The combined classification-regression model captures some interesting features, but does not perform well-enough to be useful in practice. Generally, and as expected, the model performs better when applied to the informed scenario. The scenarios are initially discussed separately. For the blind scenario, the classification step and regression step are treated separately, before a concluding discussion on the combined model viability ends the section.

The respective classification models LR and RF seem to perform equally well based on the results in Table 4.1 and Figure 4.2. The accuracies are low as the models only succeed to predict the right class in about 69% of the cases. However, one needs to investigate the classified elements in order to do a full evaluation of the model performance. More information of the classified data points can be obtained from Figure 4.2, which shows the respective distribution of the different classes classified by the model. A fair random classifier would produce two almost identical distributions and achieve a classification accuracy of 50%. Figure 4.2 display a discrepancy, for all models, in actual collection rate distribution of accounts classified as CR $= 0$ and CR $> 0$. The CR $= 0$ class distribution has more mass at collection rate zero than the class CR $> 0$. Additionally, the class mean is indeed lower for class CR $= 0$, even though almost 40% of accounts with actual collection rate greater than zero are classified as CR $= 0$.

When it comes to the regression step, RF performs best based on the qualitative measures in Table 4.2. The GLM is not too far behind with a slightly higher MAE and an $R^2$ about half as big. The ANN regression model has an $R^2$ close to zero with comparatively high estimation error. There are several remarks to make about these results. Firstly, observations based on Figure 4.3 show that all the models tend to estimate the CR close to the actual mean. Secondly, even though there are some differences in $R^2$ between models, the differences are not as apparent in MAE. CR is estimated on a bounded interval [0,1], which means that the MAE is bounded too. Hence, bad model performance is not reflected well in the MAE. Thirdly, it is notable that GLM produces the most accurate portfolio valuation, with a 7% overestimation, while RF and ANN both produce an overestimation of 12% despite $R^2$ differences. Even though the $R^2$ is low the portfolio valuation is still decent. This is probably due to the law of large numbers which comes into play when account collections are summed in portfolio valuation. Over- and underestimations on account level due to wrong classification or bad regression results cancel out in summation, producing a prediction close to the actual portfolio value. This effect is stronger if the predicted collection rates are concentrated around the true mean, which is the case for the regression models as seen in Figures 4.3.

The three different combined models have MAE of about 0.160-0.171 and an $R^2$ close to zero with high variance, as presented in Table 4.3. The high variance in $R^2$ implies that the split of training and test set have a big effect. The distributions of estimated CRs compared to the distribution of real CRs can be seen in Figures 4.4b, 4.5b, and 4.6b. There is little difference between the different combined models. The performances are equally bad.

One thing to notice about the $R^2$ is significantly higher for the regression step than for the combined model, especially for the RF model. One reason for this could be difference in the scenario, the combined model models all the errands but for the isolated regression step only errands where CR $> 0$ are included.

When the classification-regression model is applied to an informed scenario, the performance is clearly better. Observing the classification and regression steps by themselves, the improvement is obvious in both cases.

For the classification step the improvement in accuracy is 9% units. Additionally, in contrast to the blind scenario where the majority of the accounts are classified as CR $> 0$, in the informed scenario the majority of the errands are classified as CR=0.

In the regression step the performance is 2-6 % units better in the informed scenario. Similarly between the two scenarios, the models estimate the CR close to the mean.

For the combined model, the $R^2$ is 5% units higher when applied to the informed scenario with a considerable lower MAE. An important thing to notice is that, because of only looking at half a year instead of a full year, the actual collection rates in the informed scenario are lower by nature. If the collections in general are lower, they tend to be more grouped together. Estimating CR in this area results automatically in lower errors.

Looking at the predictor importance estimates, it is more interesting to focus on the predictor importance estimates of the informed model where there actually exist some degree of explanation, see Figure 4.13. The variable `collectionRate_6` is the one with the most impact. Interestingly between the classification step and regression step, the importance relative to the other variables are different. A considerable difference between the two steps is that in the classification step, all the errands are considered while in the regression step, only errands with CR $> 0$ is considered. Earlier payment behaviour tells more about whether there will be any collection at all rather than the extent of the collection. Whether there exist collection the first month or not seem to have low impact in both the classification and regression step. It could be that in the few cases, more precisely 8.8%, where there exist a payment the first month, it can tell quite a bit, but if it does not exist a payment the first month it does not convey any information.

To summarize, the type of scenario(blind or informed) have a far bigger impact on the predictive power than the actual models used(LR, RF, ANN, or GLM etc.). With more relevant information, a better estimate of CR can be made.

### 5.1.2 Markov chain model

Judging from the results in Section 4.2.2 the idea of representing the NPL portfolio as a collection of Markov chains is a fair approximation. The Markov representation seems to give a slightly more optimistic view on the portfolio where the expected number of visits to state $s_1$ and the expected proportion to be fully paid $A_{12}$ are slightly overestimated compared to the actual case, see Table 4.6.

One of the main purposes for the Markov chain model is portfolio valuation. The model captures the big features of the portfolio and gives a good indication of average behaviour. The result of the forecast is similar to the one of the static pool method, which estimates average behaviour for each month. One important thing to note is that the division of train and test set can give rise to differences in the respective portfolios that are very difficult to predict. An example is the forecast of loans in the blind scenario, see Figure 4.16. The forecast seem to do well up until month 11, but because of some unusually large payments in the actual case, the forecast falls short of the actual value. These notions are difficult to predict, also the Markov chain model is not designed to capture these. In this setup, the results are very dependent on the split of train and test set. This became most apparent in the modelling of invoices. The overall collection is very low because only a few errands repay their debts. The total collection then becomes very dependent on these payments. A few big payments in the last months have a huge impact on the total collection which makes the forecasts to be quite off, see Figure 4.18.

The forecasting results produced by the Markov chain model are not groundbreaking. In fact, they are similar to the results produced by the conventionally used method, static pool. However, there are other aspects that are in favour of the Markov chain model, the easy adaptation to new data, and the possibility of model extensions. Having a portfolio under investigation, it is easy to update the forecast of it, as new information about payments are available. It is easy to perform new simulations with current states on the start vector. The informed scenario is an example of this. Further adaptions concern the payment distribution from which the payments are drawn. The potential extensions of the model are discussed in a later section.

Another strength of the Markov chain model is, as we see it, the idea that the focus is on number of paying accounts, and not on total amount collected. This is contrary to the static pool method which only focus on the total amount collected on a portfolio level. Let us investigate an example. Suppose there is a situation were there are a few accounts where the collection is very high but where the rest of the accounts have a very low or no collection. By only looking at total amount collected, it is easy to trick oneself to believe that the performance is pretty good on a portfolio level. But, as the well performing accounts are paid off in full and as the rest of the accounts are performing bad, the portfolio will quickly stop generating cash. As the Markov chain model takes into consideration number of paying accounts the model can expect a lower collection when the well performing accounts are paid off. Had the situation been the contrary, with many accounts doing small payments in a structured manner, the Markov model could detect this notion and expect a steady collection for the upcoming time.

## 5.2 Data Availability and Quality

It is interesting to compare our results with results obtained in other studies with similar settings. This comparison applies to the results produced by the classification-regression model.

Belotti et al. receives an $R^2$ of about 14% when modelling CR with the RF regression model [19]. Kriebel & Yam obtains an $R^2$ of 9-14%, for different data sets, in similar settings [13]. Thomas et al. obtains an $R^2$ of 15% for their model applied in a similar setting [15]. One of the most probable explanations of our low model performances are aspects concerning the quality and quantity of data. As noted in Section 2.1, data availability is a big factor in model performance.

For Belotti et al. the available data consisted of socio-demographic and loan file data as well as information from the bank recovery history. In their model, the four most important attributes where the original principal amount, total principal of all debts of the person, credit limit and debt interest. Original principal amount alone was as important as the other three combined [19]. Something to note is that the original principal amount differs from the `debt` attribute in our data set. Original principal is the original amount of the debt given by the bank, while `debt` is the size of the remaining debt when the errand reaches the DCA. In the study done by Kriebel and Yam, two of the variables with highest importance are the age of the account and the existence of phone contact details to the debtor. Further, Kriebel and Yam, introduces credit bureau score as a variable. This results in a significant increase in $R^2$ [13]. The data set at our disposal does not contain any of these attributes. Had we had access to more of the stated important variables, our belief is that the model performance for the classification-regression model had been improved. However, we cannot exclude that there are approaches different from ours that would do a better job in capturing features in the available data set.

Data- and information gathering, organisation and processing is an issue in the whole credit industry [12] and there are no industry-wide reporting standards. Additionally, in general, the information sharing between credit businesses and DCAs is limited. For example, DCAs can have poor insight in the banks or business's own in-house collection. On the other hand, banks or businesses have limited knowledge concerning the working processes of the DCAs. When performing data analysis, this becomes an issue. This idea is backed up by an example from the data set available to us. See Section 3.1 for a detailed explanation of the data set. The data set contains old errands as well as newly registered ones. However, the errands that were registered before the ”window of observation” are conditioned on existence of cash flow. This means that older errands that have been closed or older errands with no cash flow during the observation time are not included in the portfolio. This implicit a heavy bias on the old errands in the portfolio. This resulted in the decision of not including old errands in the analysis.

As these data- and information issues exist, it is interesting to model CR based on what data is available. Belotti et al. sees an increase in predictive capacity, about 5% units rise in $R^2$, by including variables referring to a borrowers past repayments and/or the bank recovery process, i.e. in-house collection [19]. Thomas et al. obtains an $R^2$ of 15% when applying a model to a 3$^{\text{rd}}$ party data set and an $R^2$ of 23% when applying a model to in-house data, the increase is 8% units [15]. Our attempt to model information gain is through the informed scenario. The introduction of information about payments in an earlier time gave rise to an increase in $R^2$ with 5%. The introduction of this type of additional information could somewhat be compared to information gain through acquisition of information about the in-house process.

Being in the position of a DCA, if no information is obtained about the in-house process, the initial information is often sparse [14]. Kriebel and Yam then investigates what type of information could be gathered by a DCA and how this effects predictive performance. They see increases in $R^2$ of 9-31% units, for different data sets, by including DCA gathered information. This is a significant increase which supports their statement about the importance of the DCA information gathering process.

### 5.2.1 Limitations imposed by time-framing in CR modelling

One of the simplifications we have undertaken is to only model the CR for the first 12 months. Had we not been limited by the data and instead looking to model the CR for a longer time span, for example 24 or 36 months or up to as much as 10 years, there may have been a clearer relation between the attributes and CR. Bellotti et al. models the CR by using a data set where the whole lifetime of an account, from the granting of debt until closure, is available [16, 19]. Kriebel and Yam models the CR over a four year period [14]. Thomas et al. models recovery rate, or CR as we call it, for 24 and 36 months [15]. All three studies show better model performances.

In the data, there are interesting patterns relevant to the discussion about the time frames limiting effect on modelling CR. Many of the accounts that in the end repays most of their debts, do not start their structured repayments until several months after getting registered at the DCA. For illustration, see appendix A.6. The reason for this could be the limitation of the DCA, it takes time to agree on payment plans and some errands are not contacted until several months after the errand has been registered. This idea further strengthens the limitation a 12 month time frame imposes.

## 5.3 Markov Model Extensions

There are plenty of interesting extension options to the basic Markov model presented in the report. Detailing state definitions, incorporating account attributes, parameterization of payment distributions and trying Bayesian updating techniques are all viable options. Suggested model extensions are, however, not possible to test in the context of the thesis project, due to data limitations. Although they are not tested, the ideas are considered valuable for the development of Dignisia's system and are hence discussed further in the coming subsections. Additionally, the extensions to the Markov modelling framework constructively contribute to the discussion on modeling NPLs in the literature by offering a tool for flexible portfolio valuation and, in the long run, better understanding loan repayment drivers.

### 5.3.1 Redefining debtor states

The debtor states in the report are chosen in a simplistic manner as presented in Section 3.3.1. Debtors are either in a paying or non-paying state. It is easy to construct more complex state definitions, which might prove more useful in specific modelling applications. For instance, it might be interesting to distinguish structured payments from single payments. The hypothesis to test would be that there is a difference in expected collection from accounts where there are several consecutive payments and accounts with only one payment. If there is a difference, it would be interesting to analyze payment behaviours even further. A suggestion of such an analysis is presented in Appendix A.7. Note that the more advanced state definition is only one of many possible extensions. The best state definition is probably data and context dependent.

Another note to make on state definitions is geographical differences and local debt collection systems. If the dataset reaches over several geographies, it might be a good idea to consider separate state definitions for different locations.

### 5.3.2 Introduce data dependencies in the transition matrix

The results obtained in the report are produced based on empirical transition matrices for each debt type, but there are no further data dependencies. This constitutes a major opportunity of improvement. A natural problem formulation is to extend the concept of the transition matrix to be a function of the account data[1]

$$\mathbf{P} = \mathbf{P}(\mathbf{x}) = \begin{bmatrix} p_{00}(\mathbf{x}) & p_{01}(\mathbf{x}) & \dots & p_{0n}(\mathbf{x}) \\ p_{10}(\mathbf{x}) & p_{11}(\mathbf{x}) & \dots & p_{1n}(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ p_{n0}(\mathbf{x}) & p_{n1}(\mathbf{x}) & \dots & p_{nn}(\mathbf{x}) \end{bmatrix}.$$

The meaning of the above is that the constant transition probabilities presented in Definition 2 are extended to be functions of debtor attributes. A natural first attempt would be to model each transition probability as a linear model of the data

$$p_{ij} = p_{ij}(\mathbf{x}) = w_0^{(ij)} + w_1^{(ij)} x_1 + \dots + w_n^{(ij)} x_n = \mathbf{w}^{(ij)}\mathbf{x}$$

while reassuring $p_{ij} \in [0, 1]$ in a convenient way. For instance using a logistic regression approach would guarantee parameters taking values on the unit interval.

There are arguments to be made that this regression would fail in the application in this thesis project. The regression modelling of collection rates did not succeed, which indicates that it is hard to find separation in the data. However, it is not clear that attributes with no or low correlation with collection rates does not have an impact on transition probabilities. The likelihood of staying or leaving a state might be a less difficult problem.

### 5.3.3 Parameterization of payment distributions

When using the Markov model as a forecasting method, there is an inherent subproblem regarding modelling of amount repayed for each debtor in state $s_1$. One train of thought is to try parameterization, instead of using the empirical payment distribution in the training set. There are several reasons to try parameterization. Firstly, it is interesting to compare parameters between different portfolios to grasp whether or not there are differences in payment structure. Fitting a distribution and comparing parameters is an easy method of comparison and could be useful for analysts. Secondly, parameters are easy to update as more information becomes available. For instance, a portfolio analyst who receives new cash flow data every month can update model parameters with a specific weight on. If necessary, the expectation maximization algorithm [33] can be used to update parameter estimates,

[1] Notation is borrowed from Chapter 2.

and a Bayesian modelling framework might prove successful. A global prior payment distribution can be obtained by studying payments across different portfolios. From what the authors have gathered, the Gamma and Weibull distributions are promising candidates. The probability densities have two parameters respectively and are given by

$$f_{\text{Gamma}}(z;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} z^{\alpha-1} e^{-\beta z}$$

$$f_{\text{Weibull}}(z;\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{z}{\lambda}\right)^{k-1} e^{-(z/\lambda)^{k}}, & z \geq 0 \\ 0, \quad z < 0 \end{cases} .$$

### 5.3.4 Suggested usage of the Markov model

Even if the predictive performance of the Markov model has not proved to be much stronger than the static pool method, it still offers plenty of improvements for portfolio analysts.

Firstly, the Markov model introduces a number of useful metrics. The transition probabilities can be considered as debt collection performance metrics and could be interesting to monitor over time and across different portfolios. For instance, the probability of starting a payment, $p_{10}$, can be considered a measure of the DCA's ability to convert non-paying debtors. The payment retention $p_{11}$, i.e. probability of continued payment, indicates a DCA's ability to follow through on agreed payment plans with debtors. As a credit institution, these metrics can be used to compare how different DCA's perform on similar portfolios. In the long run, credit institutions could see higher collection by efficient hiring of the best DCA for the specific portfolio at hand.

Secondly, the Markov model can work as a strategic decision-making tool. By constructing artificial transition matrices or payment distributions, managers can understand how operational changes could affect the portfolio valuation. For instance, a manager might wonder how the portfolio valuation would change if the DCA increased their conversion of non-paying debtors, i.e. $p_{01}$, by 10%. This increase in portfolio value could then be compared to a situation where the average payment amount is 10% higher than in the original portfolio. Simulation and analysis of the two artificial portfolios could help management find the best strategy, and hence optimize investments with respect to expected increase in portfolio value.

Finally, the Markov model could be used in operational excellence purposes at DCA's. Trying new collection strategies at a small number of debtors and using the Markov framework to produce portfolio valuations could be an efficient way of evaluating and comparing debt collection processes.

# 6
# Conclusion

This master's thesis aims to investigate the possibility of building a well-performing model to forecast cash flow from a given portfolio of non-performing loans. The dataset at hand consists of defaulted credits, with socio-demographic as well as debt-specific attributes and information about monthly payments for each account. Two different methods are used: a combined classification-regression model which aims to find a relation between the attributes and collection rate and a Markov chain model which aims to estimate total cash flow on a portfolio level based on payment behavioural characteristics. The two models are also tested in two different scenarios, blind and informed, which aims to investigate improvement in model performances with knowledge of prior payment history.

With an $R^2$ of close to zero, the classification-regression model does not succeed in finding a relation between attributes and collection rate in the blind scenario. This could be compared to results in other studies in similar settings where $R^2$ values of 10-15 % are obtained [13, 15, 16, 19]. Even though the degree of explanation is low, the portfolio valuation is not completely off. This probably has to do with the law of large numbers where under- and over-estimations cancel out on average. The area of NPL forecasting is heavily influenced by the way information and data is managed in the credit industry. It seems that it does not exist standardized ways of dealing with information which have the effect that different institutions sit on different data. This influence the predictive ability of models estimating CR. Including information of the in-house process have shown to improve predictions with 5-8% units [15, 19]. With the intention of investigating information gain through knowledge of in-house collection, a scenario was set up where collection rate for month 7-12 is estimated based on the initially available attributes as well as payment behaviour for month 1-6. Applying the model to this scenario results in an increase of 5% units for the $R^2$ value.

The Markov chain model is based on the assumption that payments do not occur at random, but rather follow some sort of structure. The conclusion is that it is a reasonable way of describing the system. The model is used both as a tool to analyze payment behaviour on a portfolio level and as a tool to produce portfolio valuations. The forecasting performance is equal to the more conventionally used static pool method. It captures the general behaviour of the portfolio but fails to model the more stochastic elements. The advantage of the Markov chain model is the adaptation to new data and the possibility of some interesting model extensions that are further discussed.

The conclusion to be drawn is that we did not succeed in building a well-performing forecast model to estimate repayment in an NPL portfolio for the given data set. The perception is that the biggest reason for this lies in the quantity and quality of data rather than in model design. There is a general consensus in the literature that information availability is a deciding factor when modelling CR of NPLs [13, 14, 19].

With that said, a new approach to model structured payments has been developed in the form of the Markov chain model. There is potential in further extending it and continue using it as a forecasting model as well as a strategic decision-making support tool.

# Bibliography

[1] Basel Committe on Banking Supervision, “Basel III: Finalizing post-crisis reforms,” *Bank for International Settlements*, December 2017.

[2] European Union, Single Resolution Borad, “Minimum Requirement for Own Funds and Eligible Liabilities (MREL) – 2018 SRB Policy for the first wave of resolution plans,” November 2018.

[3] Sveriges Riksbank, “Minimum Requirement for Own Funds and Eligible Liabilities (MREL) – 2018 SRB Policy for the first wave of resolution plans,” *Sveriges Riksbank: Economic Commantaries*, no. 1, 2016.

[4] A. Sironi, “The evolution of banking regulation since the financial crisis: a critical assessment,” *BAFFI CAREFIN Centre Research Paper*, no. 103, 2018.

[5] L. Nario, T. Pfister, T. Poppensieker, and U. Stegemann, “The evolving role of credit portfolio management,” *McKinsey & Company - Risk*, July 2016.

[6] Statistics Sweden *SCB Economic Indicators*, no. 9, p. 18, 2019.

[7] Sveriges Riksbank, *Financial Stability Report,*, no. 1, pp. 9–13, 2019.

[8] Finansinspektionen, “Swedish Consumption Loans,” June 2019.

[9] Statistics Sweden, Enheten för Finansmarknadsstatistik, “Finansmarknadsstatistik 2019-10-25,” 2019.

[10] Finansinspektionen, *Stability in the Financial System,*, no. 1, 2019.

[11] Council of the EU, “Non-performing loans: Council adopts position on secondary markets for bad loans,” Mar 2019.

[12] Jörgen Köster, Dignisia AB. `https://dignisia.com/en`, Oct 2019. Accessed on 2019-10-09.

[13] J. Kriebel and K. Yam, “Gathering information: Forecasting recoveries in debt collection,” 2018.

[14] J. Kriebel and K. Yam, “Forecasting recoveries in debt collection and information production,” 2019.

[15] L. Thomas, A. Matuszyk, and A. Moore, “Comparing debt characteristics and LGD models for different collection policies,” *International Joural of Forecasting*, vol. 12, pp. 196–203, 2012.

[16] H. Ye and A. Bellotti, “Modelling recovery rates for non-performing loans,” *Risks*, vol. 7, no. 1, p. 19, 2019.

[17] L. E. Papke and J. M. Wooldridge, “Econometric methods for fractional response variables with an application to 401 (k) plan participation rates,” *Journal of applied econometrics*, vol. 11, no. 6, pp. 619–632, 1996.

[18] Beck, Timo and Grunert, Jens and Neus, Werner and Walter, Andreas, “What Determines Collection Rates of Debt Collection Agencies?,” *Financial Review*, vol. 52, no. 2, pp. 259–279, 2017.

[19] A. Bellotti, D. Brigo, P. Gambetti, and F. D. Vrins, "Forecasting recovery rates on non-performing loans with machine learning," *Credit Scoring and Credit Control XVI*, 2019.

[20] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[21] S. H. Ha and R. Krishnan, "Predicting repayment of the credit card debt," *Computers & Operations Research*, vol. 39, pp. 765–773, 2012.

[22] V. Boutchaktchiev, "A Markov-Chain Model for the Cure Rate of Non-Performing Loans," *SSRN Electronic Journal*, 2018.

[23] P. Siarka, "Vintage analysis as a basic tool for monitoring credit risk," *Mathematical Economics*, no. 7 (14), pp. 213–228, 2011.

[24] S. Rogers and M. Girolami, *A first course in machine learning.* Chapman and Hall/CRC, 2016.

[25] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics).* Berlin, Heidelberg: Springer-Verlag, 2006.

[26] W. Loh, "Classification and regression trees," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 14–23, 2011.

[27] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[28] I. A. Basheer and M. Hajmeer, "Artificial neural networks: fundamentals, computing, design, and application," *Journal of microbiological methods*, vol. 43, no. 1, pp. 3–31, 2000.

[29] W. S. Sarle, "Neural networks and statistical models," 1994.

[30] J. J. O'Connor and E. F. Robertson, "Andrei A. Markov." `http://mathshistory.st-andrews.ac.uk/Biographies/Markov.html`, Aug 2006. Accessed on 2019-12-11.

[31] G. R. Grimmett and D. R. Stirzaker, *Probability and Random Processes.* Oxford university press, 3 ed., 2001.

[32] C. M. Grinstead and J. L. Snell, "Markov chains," *Introduction to probability*, pp. 405–470, 1997.

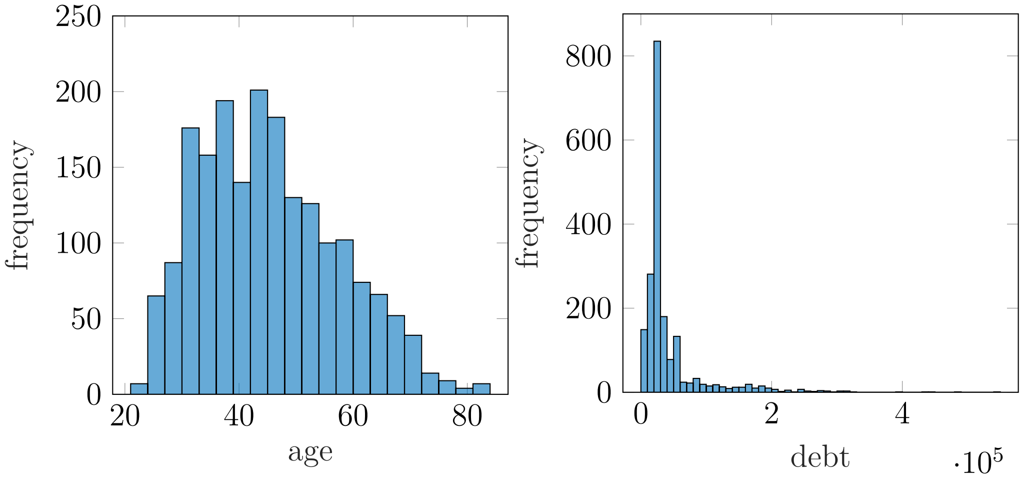
[33] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal processing magazine*, vol. 13, no. 6, pp. 47–60, 1996.

# A

# Appendix

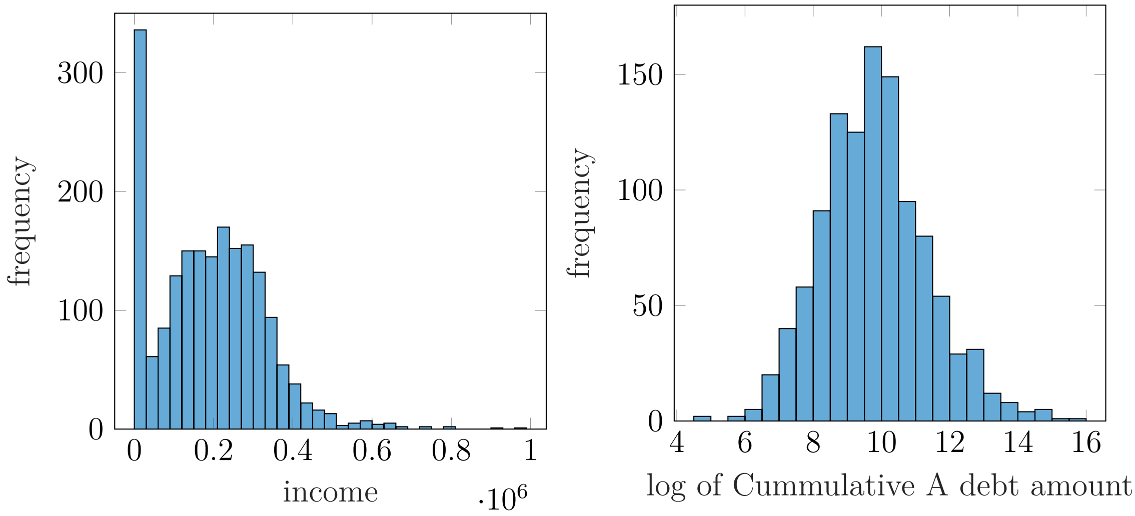
## A.1 Histograms of Variables in the Dataset

To complement to the information of the dataset given in Section 3.1, more specifically to give more information about the value distributions of the variables in Table 3.1.



**(a)** Distribution of `age`. **(b)** Distribution of `debt`.

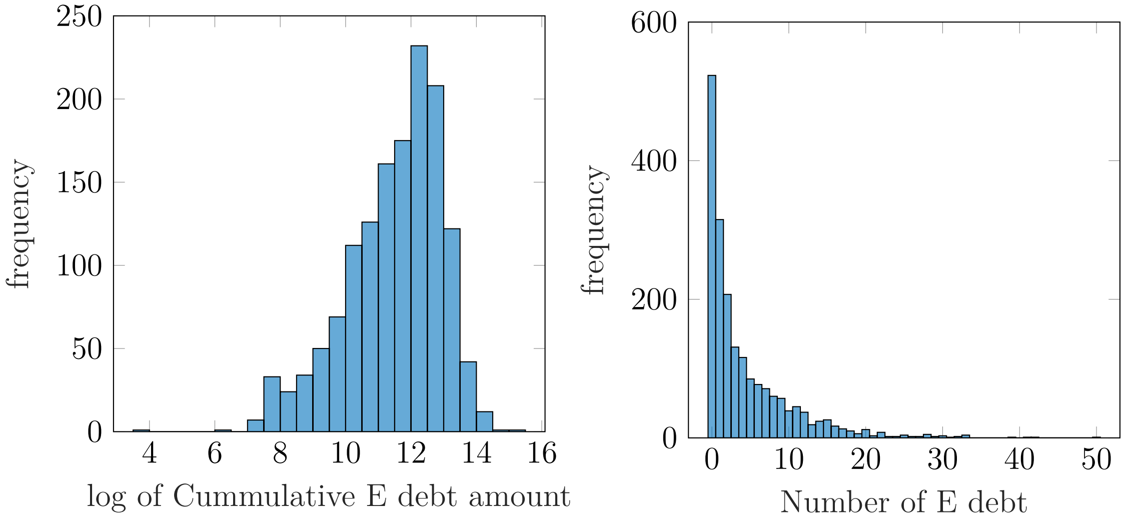
**Figure A.1:** Distribution of variables in the data set. For more information, see Tables 3.1 and 3.3 in Section 3.1.



**(a)** Distribution of `income`.

**(b)** Distribution of the logarithm of `sumADebt` for the accounts where the sum of A-debts are greater than zero. 43% does not have an A-debt, i.e. it is zero.

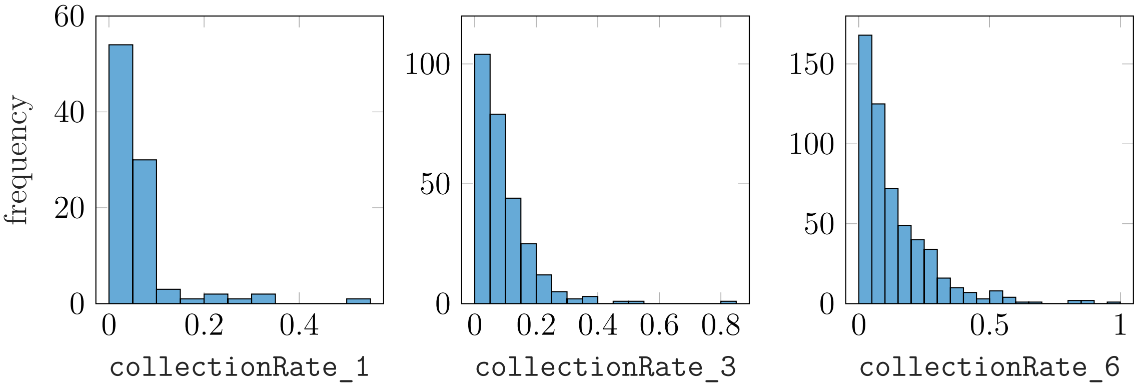
**Figure A.2:** Distribution of variables in the data set. For more information, see Tables 3.1 and 3.3 in Section 3.1.



**(a)** Distribution of the logarithm of `sumEDebt` for the accounts where the sum of E-debts are greater than zero. 27% does not have an E-debt, i.e. it is zero.

**(b)** Distribution of `numEDebt`.

**Figure A.3:** Distribution of variables in the data set. For more information, see Tables 3.1 and 3.3 in Section 3.1.



**(a)** Distribution of `collectionRate_1` where CR is greater than 0. 95% of the accounts have a CR=0 for the first month.

**(b)** Distribution of `collectionRate_3` where the CR is greater than 0. 86% of the accounts have a CR=0 for the first three months.

**(c)** Distribution of `collectionRate_6` where the CR is greater than 0. 72% of the accounts have a CR=0 for the first six months.

**Figure A.4:** Distribution of the informed variables in the data set. For more information, see Tables 3.2 and 3.3 in Section 3.1.

## A.2 Implementation of Classification- and Regression Models (reference packages)

### A.2.1 Classification models

The classification models have been implemented according to the following:

- Logistic regression: Function generated using the machine learning application in MATLAB.
- Random forest: `TreeBagger`-function.
- Artificial neural network: Function generated using the machine learning application in MATLAB.

### A.2.2 Regression models

The regression models have been implemented according to the following:

- Generalized linear model: `fitglm`.
- Random forest: `TreeBagger`.
- Artificial neural network: `feedforwardnet` and `train`.

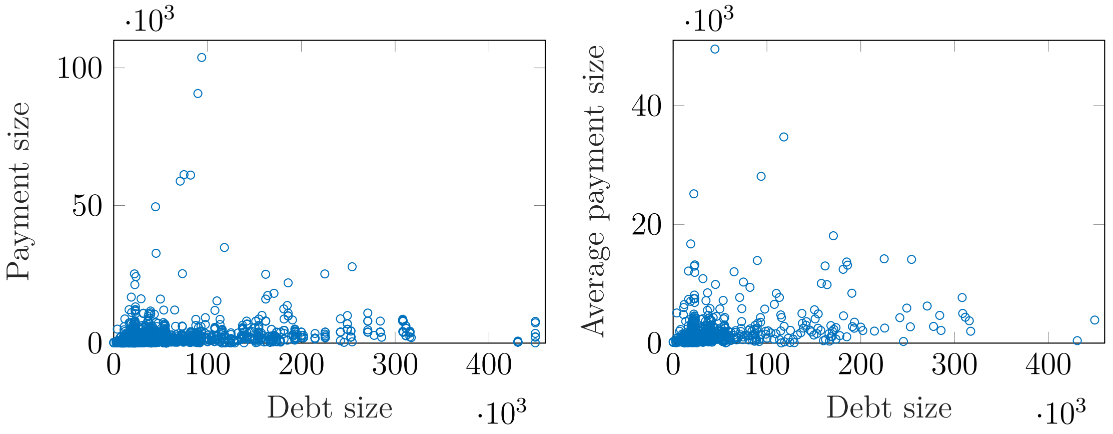
## A.3 Estimation of Predictor Importance

For the RF classification and regression model the variable importance is estimated through the out of bag permuted predicted delta error. An explanation to this follows: During the training phase of the model, the model is trained on a set of data and validated on another set of the data. To estimate a variables effect on the dependent variable, i.e. its importance, the total error of a model created with the specific variable removed is compared with the total error of a model created with the complete set of variables. The difference in error gives an indication of the impact that specific variable has. This is done for all the variables in the data set. To more easily compare the variables the measures could be normalized and sorted in descending order and represented in a bar chart, which is what we choose to do.

## A.4 Distributions for Drawing Payments in the Markov Model

The distribution from which payments are drawn is a previously known distribution of payments from NPL portfolios of similar debt type. One thing to note is that no consideration is taken to the potential difference in debt size of the errands between the portfolio on which the forecast is intended and the portfolio from which the payment distribution is based upon. This is a simplification that has been made. This section aims to motivate this simplification.

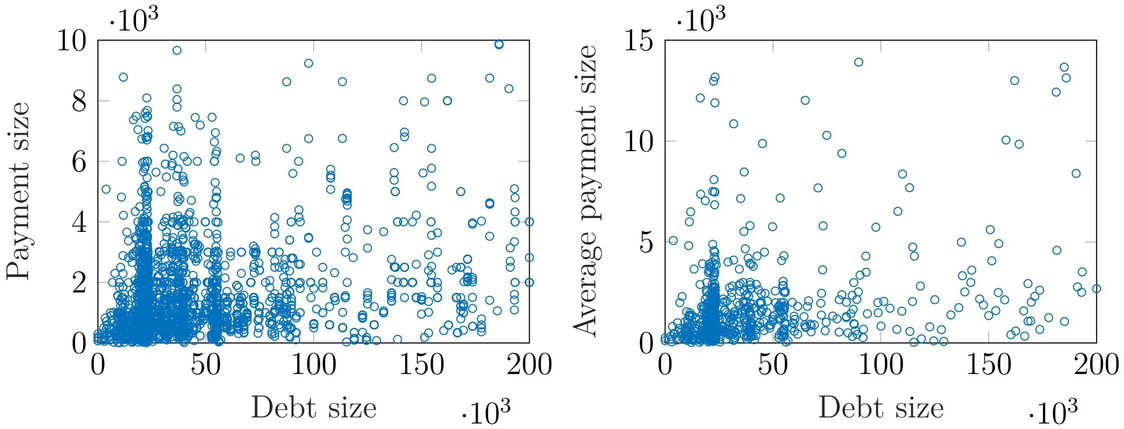
The assumption is that average payment size does not scale with debt size. To investigate this matter, payment size and average payment size per errand are looked closer upon with respect to debt size. Payment size is the size of a monthly payment for a specific account. Average payment size per errand is the average taken over all monthly payments, conditional that the payment is greater than 0, for a specific account. Payment size and average payment size are plotted against the account specific debt size, see Figure A.5. All available debt types, invoice, loans, and credit card debts, are plotted in the same plot and no distinction is made. To more easily analyse the relation, the graphs are zoomed in by setting appropriate axis limits, see Figure A.6. There is no clear relation between payment size and debt size. A bigger debt does not mean that a debtor is able to repay in a bigger extent. One has to remember that we are talking about the ability to pay of people in economically pressured situations. It is reasonable to believe that these people does not have more than a couple of thousand available for repayment each month.



**(a)** Payment size plotted against debt size.

**(b)** Average payment size for an errand plotted against debt size.

**Figure A.5:** Scatter plot of debt size and payment size. One outlier at $(x, y) = [243 \cdot 10^3, 240 \cdot 10^3]$ is omitted in both figures due to readability reasons.



**(a)** Payment size plotted against debt size. A zoomed in view of Figure A.5a.

**(b)** Average payment size for an errand plotted against debt size. A zoomed in view of Figure A.5b.

**Figure A.6:** Zoomed in view of Figure A.5.

## A.5 Markov Chain Model Applied to a Blind Scenario

The transition matrices for each debt type is constructed based on the training set. For the different debt types the matrices are:

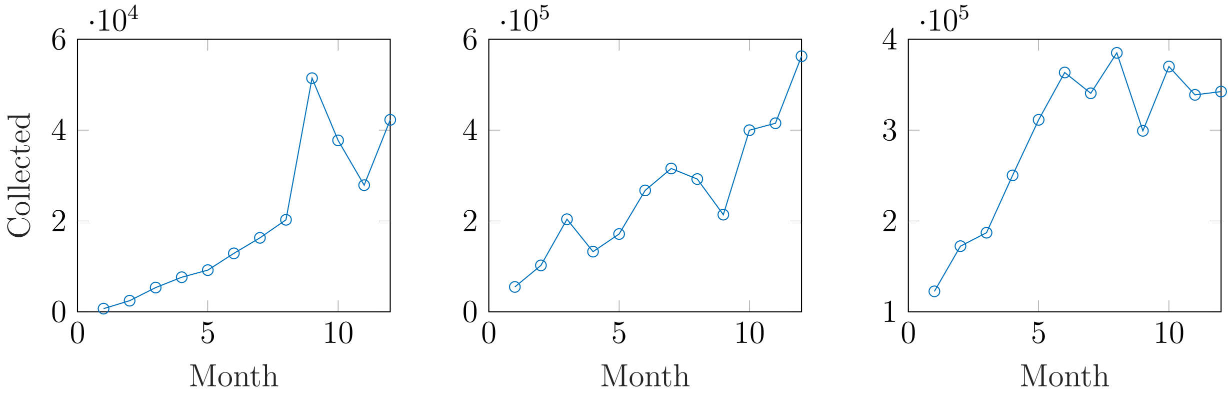
$$\mathbf{P}^{\text{invoice}}_{\text{train}} = \begin{bmatrix} 0.989 & 0.011 & 0 \\ 0.250 & 0.743 & 0.007 \\ 0 & 0 & 1.000 \end{bmatrix}$$

$$\mathbf{P}^{\text{loans}}_{\text{train}} = \begin{bmatrix} 0.894 & 0.106 & 0 \\ 0.251 & 0.743 & 0.006 \\ 0 & 0 & 1.000 \end{bmatrix}$$

$$\mathbf{P}^{\text{cc}}_{\text{train}} = \begin{bmatrix} 0.789 & 0.211 & 0 \\ 0.274 & 0.726 & 0.001 \\ 0 & 0 & 1.000 \end{bmatrix} .$$

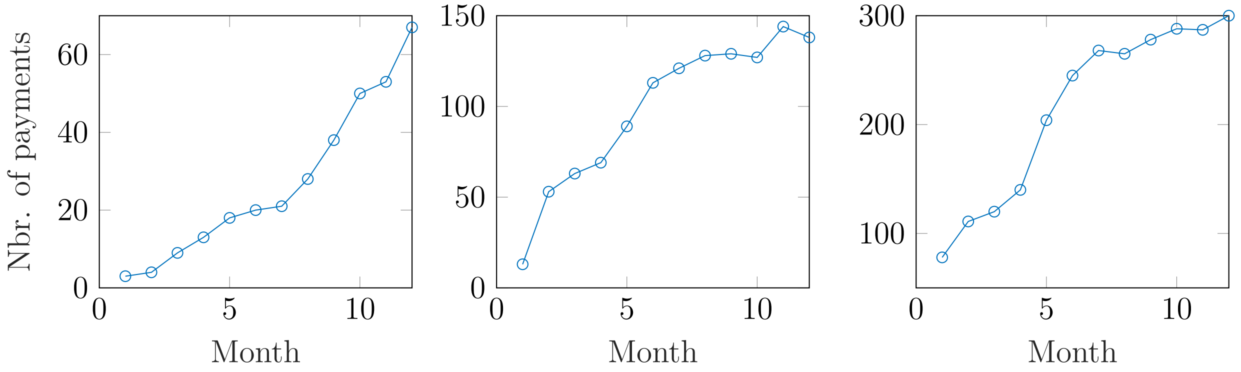
## A.6 Collections on a Portfolio Level Over Time

A simple analysis conducted on total collection per month for a portfolio, shows a steady increase in collection per month, see Figure A.7. The reason for this is either an increase in payment amount on an account level, or that more accounts enter in structured payment. By looking at the total number of payments per month for a portfolio, see Figure A.8, the matter is decided upon, the reason is more accounts doing payments.



**(a)** Invoices. **(b)** Loans. **(c)** Credit card debts.

**Figure A.7:** Total amount collected per month on an portfolio level for the different debt types, invoices, loans, and credit card debts.



**(a)** Invoices. **(b)** Loans. **(c)** Credit card debts.

**Figure A.8:** Number of payments per month on a portfolio level for the different debt types, invoices, loans, and credit card debts.

## A.7 Markov Model Potential Extension

The following is a possible extension on the Markov chain model. It divides the paying state in two where the distinction is done based on the characteristic of the payment. It uses the following definition of states:

- State 0. This state indicates that there is no cash flow from a given account at a particular month.
- State 1. A debtor is in a structured payment. This is defined as having at least two payments in a three month period.
- State 2. Single payment, i.e., all payments that does not fulfill conditions of being State 1.
- State 3. Debt fully paid.

To highlight the difference between the advanced and simple state definitions, the same sample cash flow matrix from Section 3.3.1 is reused.

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3500 & 0 & 0 & 0 \\ 0 & 0 & 1000 & 1000 & 1000 & 1000 \\ 0 & 0 & 1000 & 0 & 2000 & 1000 \\ 500 & 500 & 517 & 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 3 & 3 & 3 \end{bmatrix} \tag{A.1}$$

The fourth month in row four is still considered a structured payment state, since the month of no payment is encapsulated by two paying months. The payment in row two is considered a single payment.

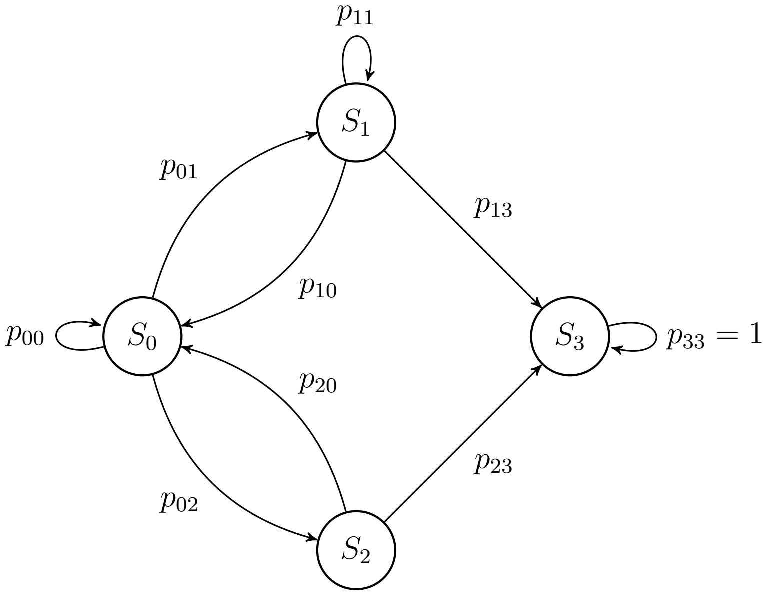


**Figure A.9:** Markov model with possible transitions and corresponding probabilities illustrated