UNIVERSITY OF ECONOMICS, PRAGUE
FACULTY OF INFORMATICS AND
STATISTICS

New Methods in Credit Underwriting

Doctoral Thesis

Author: Ing. Mgr. Michal Rychnovský, MSc

Supervisor: prof. Ing. Josef Arlt, CSc.
Study program: Quantitative Methods in Economics
Field of study: Statistics

Prague, September 2017

Declaration

I declare that I carried out this doctoral thesis independently and cited all used sources and literature.

Brno, 1 June 2017

Michal Rychnovský

Acknowledgment

I would like to thank to my supervisor prof. Ing. Josef Arlt, CSc. for his support during my doctoral studies. My thanks belong also to prof. RNDr. Jiří Witzany, Ph.D. and RNDr. Pavel Charamza, CSc. for their help with my research and for co-authoring our articles that have been used in this thesis. Finally, I would like to thank my family and friends for their long-lasting support and encouragement.

Michal Rychnovský

Abstract

This thesis contributes to the field of applied statistics and financial modeling by analyzing mathematical models used in retail credit underwriting processes. Specifically, it has three goals. First, the thesis aims to challenge the performance criteria used by established statistical approaches and propose focusing on predictive power instead. Secondly, it compares the analytical leverage of the established and other suggested methods according to the newly proposed criteria. Third, the thesis seeks to develop and specify a new comprehensive profitability-based underwriting model and critically reflect on its strengths and weaknesses.

In the first chapter I look into the area of probability of default modeling and argue for comparing the predictive power of the models in time rather than focusing on the random testing sample only, as typically suggested in the scholarly literature. For this purpose I use the concept of survival analysis and the Cox model in particular, and apply it to a real Czech banking data sample alongside the commonly used logistic regression model to compare the results using the Gini coefficient and lift characteristics. The Cox model performs comparably on the randomly chosen validation sample and clearly outperforms the logistic regression approach in the predictive power.

In the second chapter, in the area of loss given default modeling I introduce two Cox-based models, and compare their predictive power with the standard approaches using the linear and logistic regression on a real data sample. Based on the modified coefficient of determination, the Cox model shows better predictions.

Third chapter focuses on estimating the expected profit as an alternative to the risk estimation itself and building on the probability of default and loss given default models, I construct a comprehensive profitability model for fix-term retail loans underwriting. The model also incorporates various related risk-adjusted revenues and costs, allowing more precise results. Moreover, I propose four measures of profitability, including the risk-adjusted expected internal rate of return and return on equity and simulate the impact of the model on each of the measures.

Finally, I discuss some weaknesses of these approaches and solve the problem of finding default or fraud concentrations in the portfolio. For this purpose, I introduce a new statistical measure based on a pre-defined expert critical default rate and compare the GUHA method with the classification tree method on a real data sample.

While drawing on the comparison of different methods, this work contributes to the debates about survival analysis models used in financial modeling and profitability models used in credit underwriting.

Keywords: Probability of default, loss given default, profitability model, survival analysis, Cox model

Abstrakt

Tato práce přispívá do oblasti aplikované statistiky a finančního modelování analýzou matematických modelů používaných v procesech schvalování retailových úvěrů. Konkrétně má tři cíle. Za prvé, diskutuje vhodnost výkonostních kritérií užívaných zavedenými statistickými postupy a navrhuje zaměřit se místo toho na sílu predikce. Za druhé, porovnává analytickou přidanou hodnotu stávajících a nově navrhovaných metod podle navržených kritérií. A třetím cílem práce je potom výstavba a detailní specifikace rozsáhlého modelu pro odhad profitability včetně kritické reflexe jeho silných a slabých stránek.

V první kapitole pracuji v oblasti modelování pravděpodobnosti defaultu (selhání dlužníka) a navrhuji srovnání predikční síly modelů v čase, místo v akademické literatuře běžně používaného srovnání na náhodném testovacím vzorku. K tomuto účelu používám koncept analýzy přežití a Coxův model, který společně s běžně používanou logistickou regresí aplikuji na vzorek reálných českých bankovních dat a porovnávám výsledky pomocí Giniho koeficientu a charakteristiky lift. Na náhodném validačním vzorku vykazuje Coxův model podobnou přesnost jako logistická regrese, zatímco při porovnání predikčních schopností v čase vychází Coxův model znatelně lépe.

Ve druhé kapitole, zaměřené na modelování ztráty při defaultu (LGD), představuji dva modely založené na Coxově regresi a na reálných datech srovnávám jejich predikční sílu se standardními přístupy lineární a logistické regrese. Ve srovnání pomocí modifikovaného koeficientu determinace vykazuje Coxův model lepší predikce.

Třetí kapitola se zaměřuje na odhad očekávané profitability jako alternativy k odhadům rizika jako takového a staví na modelech pravděpodobnosti defaultu a ztráty při dafaultu. Zde konstruuji rozsáhlý a detailní model profitability pro schvalování retailových úvěrů s fixní dobou spláceni. Do modelu vstupují také další související výnosy a náklady očištěné o riziko plynoucí z defaultu dlužníka, což vede k přesnějším výsledkům. Dále navrhuji čtyři charakteristiky profitability, včetně rizikově očištěného očekávaného vnitřního výnosového procenta a rentability vlastního kapitálu, a simuluji vliv tohoto modelu na každou z těchto měr.

Nakonec poukazuji na některé slabiny těchto přístupů a řeším problém nalezení koncentrací defaultů či podvodů v portfoliu. Proto také představuji novou statistickou míru založenou na předem stanované expertní hodnotě kritické míry defaultu a srovnávám GUHA metodu s použitím klasifikačních stromů na reálném datovém vzorku.

Pomocí srovnání různých metod tato práce přispívá k debatám ohledně použití modelů analýzy přežití ve finančním modelování a modelů profitability používaných pro schvalování úvěrů.

Klíčová slova: Pravděpodobnost defaultu, ztráta při defaultu, model profitability, analýza přežití, Coxův model

Contents

Introduction

In this thesis I deal with the field of consumer credit underwriting and I aim to propose new mathematical and statistical methods to enhance the standard credit underwriting automated scoring. Particularly, I aim to challenge the performance criteria based on ex-post random testing samples,[1] which is often suggested by various researches in the literature when comparing the credit risk related models. Instead I propose comparing the predictive power of the models on an ex-ante sample of the most recent data. Then I seek to use this new criteria and a real Czech banking data sample to compare the standard models performance with some suggested alternatives. Finally, I aim to construct a new comprehensive underwriting model that would be based on an estimation of loan profitability instead of the standard evaluation of the riskiness of the client. Such model should be described in detail, the results simulated and compared with the standard approach and its weaknesses treated by proposed alternative methods.

Theoretically, the thesis contributes to the scholarly literature on mathematical modeling in finance by showing how different performance criteria can change the outcomes of the credit risk models comparison. This opens up prospects for the so far overlooked models to be further studied and considered as relevant alternatives in financial modeling.

At the empirical level, the thesis explores ways how to improve the precision of credit underwriting models in consumer finance. Moreover, it promotes and explores the concept of profitability models used in a loan approval process that can potentially increase the companies' profit. Finally, it solves the fraud concentrations discovery problem, and thus helps to secure the underwriting models against fraud attacks and other risky segments.

Before going into details about the mathematical models analyzed in this thesis I briefly outline the main principles and terms of the credit underwriting. In reality the loan approval process is usually very complicated and contains a great amount of specific conditions, calculations and sub-processes. However, with major simplification, I could say that the main parts of the process could be the evaluation

[1] The testing samples are also called validation or comparison samples. In this thesis I will treat these expressions as synonymous.

of client's ability to repay the loan and verification of income and other provided information. The repayment ability is then studied from the perspective of checking stability and sufficiency of income to cover all the expenses, and from the perspective of evaluation of the riskiness of the client – and it is the riskiness of the client that is the key topic for this thesis.

The riskiness of the client is usually based on estimation of the *probability of default (PD)* based on the client's characteristics. *Default* is usually defined as a violation of debt contract conditions, such as a lack of will or a disability to pay a loan back. In the case of default, the creditor (e.g. a bank or other financial institution) suffers a loss. The probability of default is then usually estimated using the logistic regression models. The regression model, also called scoring model, assigns a score to each client, which is then used as a key factor for automated approval or rejection of the loan application in the process, or as one of the main inputs for the following manual underwriting.

Even though the probability of default estimation is a simple and widely used concept for credit underwriting, recently the more attention is paid to the models, where not only risk, but also the whole *expected loan profitability* is considered for the loan approval. I take inspiration from these concepts and aim to propose a comprehensive scheme consisting of several models to calculate the expected profitability of each specific loan application in the consumer loan business.

In the modeling I often experiment with the *survival analysis* models. Survival analysis is a common approach to model the time to death of biological organisms or failure in mechanical systems – generally a time to some defined event for some subjects. However, the time-to-event variables can be easily applied to the credit risk modeling as well. Using these models one can estimate the time to default or the time to recovery based on the client's characteristics. Also using the resent censored observations can bring some additional value in the model performance.

Throughout all the thesis and in particular for the proposition of the new methods, I build on my master theses (Rychnovský, 2009) and (Rychnovský, 2011) and combine the research findings and theory from the cited sources with the practical experience I gained from my professional career working in the consumer finance business. The thesis is structured in four chapters covering four topics – the probability of default modeling, the loss given default modeling, the profitability model concept and the fraud detection problem. Generally, in this thesis I focus on the comparison of methods.

The first topic is the area of the probability of default modeling. In this area I build upon our past project (Pazdera et al., 2009) that deals with the application of the survival analysis theory to the probability of default estimation,[2] and my master thesis (Rychnovský, 2011) that describes the development process of a standard

[2] The project (Pazdera et al., 2009) was made on request of one of the biggest Czech banks, and

logistic regression based scoring model. The idea of using the survival analysis models for the probability of default estimation is not new and has been published before by several researchers, including (Banasik et al., 1999), (Glennon and Nigro, 2005) or already (Narain, 1992), however it is still very popular among researchers and professionals. When comparing the performance of the survival analysis models with the logistic regression, it is usually concluded, see e.g. (Stepanova and Thomas, 2002), (Cao et al., 2009) or (Bellotti and Crook, 2009), that the survival analysis models perform similarly to the logistic regression on a random testing sample.

However, it is rather the predictive power of the models in time that could be more relevant for the modelers in the financial practice. Therefore, in this thesis I introduce an alternative performance indicator based on the predictive power of the models and compare two methods on the real financial set of data. I aim to compare not only the precision but also the predictive power of the standard logistic regression model with the Cox model alternative. It is the Cox model's ability to work with the recent time-censored observations that gives us the motivation that such a model could bring potential added value for prediction. Empirically, I build both the logistic regression scoring model and the Cox regression survival model on a set of real Czech banking data, taken from the past project (Pazdera et al., 2009) and further adjusted by omitting the data vintages containing small number of observations and selecting the fix-term loans only, and compare their performance on the standard random validating sample as well as its prediction power on a specially constructed ex-ante data sample.

The second topic deals with the loss given default modeling, where I apply similar logic and aims as in the first topic. The *loss given default (LGD)* is the part of the credit exposure that has not been recovered after the client defaulted. This characteristic has been studied as one of the expected loss components even earlier, e.g. (Asarnow and Edwards, 1995) or (Gupton et al., 2000), but it gained its importance mainly after the new Basel II Capital Accord, see (Basel II, 2001), was signed. After that a new wave of explanatory notes such as (Schuermann, 2004) and models proposals like (Gupton, 2005), (Huang and Oosterlee, 2011) or (Loterman et al., 2012) arrived.

In my master thesis (Rychnovský, 2009) I first combined the survival analysis methodology with the LGD and introduced some new survival analysis models for its estimation. I compared those models with other approaches using the linear regression, logistic regression, two-step beta regression and regression trees. In that thesis I introduced a modified weighed coefficient of determination and compared the models on their development sample. There the survival analysis models performed worse.

Having worked with the models further, I realized that the added value of the

even though the original seminar paper was never published, it was mentioned in several theses and papers including (Nehrebecka et al., 2016) from the National Bank of Poland.

proposed survival-based Cox regression models should be in the predictive power – again due to the fact that it uses the time censored recent data. Therefore, I continued working on this topic and introduced a way to compare the predictive power of the models by developing it on a censored data sample and comparing it on the ex-ante sample that was not used for development. In this case I figured that the survival analysis models outperform the standard models in the predictive power, which has been published as (Witzany et al., 2010) and later as (Witzany et al., 2012). Following our publications, this idea has been further dealt with by (Bonini and Caivano, 2013) in the Journal of Credit Risk, (Louzada et al., 2014) in the Journal of Statistics Applications & Probability, (Belyaev et al., 2012) in the Working Paper Series of the Czech National Bank, or (Thomas et al., 2016) in the European Journal of Operational Research. This topic became further studied also by (Bonini and Caivano, 2012), (Zhang and Thomas, 2012) or (Prívara et al., 2014), who come with new methods for LGD modeling.

Drawing on my previous research of (Witzany et al., 2012) and the recent academic discussions of (Thomas et al., 2016) and (Zhang and Thomas, 2012), I incorporated one new survival analysis model, based on the event of a full or partial recovery, and applied it on the data set from (Witzany et al., 2012) to compare the predictive power of the adjusted set of models. These results have been presented as (Rychnovský, 2015).

The second chapter of this thesis therefore presents two contributions made during my doctoral studies, which are the change of the performance criteria focusing on the predictive power of the models rather then their comparison on the development sample and the development of a new survival-based model to be compared with the previous models.

The third topic promotes the importance of the profitability modeling in the underwriting process and combines the probability of default modeling with the loss given default modeling and various risk-adjusted revenues and costs into the above mentioned comprehensive *profitability model* specifically designed for fix-term retail loans. Here I get the theoretical inspiration from (Allen et al., 2004) and (Stein, 2005) combined with my professional experience with consumer finance products, to build the profitability model concept. This concept uses the outcomes of the survival analysis models and incorporates various sources of revenues and costs to offer four profitability measures that can be used in financial practice.

The idea of a probability model is a reaction to my professional experience from the consumer credit business, where I realized that not only the risk management, but mainly the profit management are the key factors for a credit company. Therefore, I take the inspiration from a simple profitability model used in practice and various literature to create a comprehensive profitability model fitting the retail credit management needs. This model combines the estimation of the probability of default with several techniques to extrapolate it to all the instalments of the whole loan existence, and with the estimation of the loss given default to get the expected loss of the loan.

I enrich this concept with a variety of potential cost and revenue streams coming from this loan leading to the risk-adjusted expected profit from providing this loan. Moreover, I calculate four alternative profitability measures coming from this model, that can each support the priorities of the company. Especially, the risk-adjusted expected return on equity from a specific loan could be a really beneficial measure on some markets – this I have not seen it implemented in practice nor suggested in literature.

Furthermore, I run a data simulation where the profitability model with all four suggested measures is implemented, to understand the correlation between the measures and evaluate the impact of the whole profitability model under various simulated data and market situations. This simulation aims to suggest that in practice such models not only expand the horizons of the loan providing companies, but with a proper timing and good implementation they can significantly increase the profitability of the business.

The last topic reacts on the underwriting model as such, points to some of their weaknesses and provide a possible solution for searching for segments with high *concentration of default or fraud* within the portfolio. In this task I aim to find a proper method for identification of risky segments based on the default rate, sample size and some expert evaluation of the default severity, as well as proposing some methods for finding these segments within big data structures. Here I combine the usual statistics and the GUHA method of (Hájek et al., 1966) with my experience from the credit fraud detection. Up to my knowledge, the solution of such task has not been published to this moment.

Further on, I run the GUHA method alongside the classification tree on a real loan portfolio to compare the results and discus the advantages of both methods for practical use. The chapter provides a statistical tool combined with data mining techniques to solve a practical problem of banks and other companies from various industries that need to identify and find some concentrations in their portfolios od data. In my opinion and experience, for credit companies it is the timely identification of fraud segments that enables to successfully deal with the fraud, together with adjusting the underwriting system to prevent any further losses coming from such pattern.

Chapter 1

Probability of Default Modeling

In this chapter I deal with the most common task in the field of retail credit risk – probability of default (PD) modeling – and aim to set the new performance criteria focusing on the predictive power of the models and compare the standard logistic regression model with the alternative of the survival-based Cox model on the real sample of Czech banking data.

Probability of default modeling is a frequently discussed topic with various applications and comparisons of methods, including the comparison of the standard logistic regression model with the survival analysis alternatives. It has been shown, e.g. in (Stepanova and Thomas, 2002), that the survival analysis models have a similar performance to the logistic regression model in the terms of its precision.

Since this topic is still up-to-date and relevant to banks and other financial institutions managing credit risk, I decided to extend the research by comparing the standard logistic regression model with its survival analysis alternative also on a regionally specific Czech banking data sample. Moreover, I aim to focus on the predictive power and compare the two models' performance on an ex-ante validation sample. The motivation for choosing the Cox model is mainly the fact that the survival analysis method incorporates also the recent censored observations, and thus can enhance the predictions.

For this chapter I use the logic of the logistic regression model and diversification power measures described in my master thesis (Rychnovský, 2011), and the idea of the survival analysis model together with the data for comparison from our previous seminar project (Pazdera et al., 2009). Then I adjust the data sample by excluding the vintages with few observations and select the fix-term loans only for the analysis. Finally, I create two sets of development and validation samples (one for the random testing and one for the prediction testing), implement both the models on these adjusted samples and compare the results by the Gini and lift characteristics.

As the research result for this thesis I present the new setup of data samples for model predictive performance comparison, the practical application of these models on the adjusted data and the comparison of the results in two qualitatively varying angles – on both the standard random testing sample as well as on the ex-ante validation sample.

1.1 Probability of Default Models

In banks and other financial companies providing retail loans, one of the basics of the retail risk management is the risk assessment of applicants.[1] This is usually done by mathematical models developed on the company's historical data.

The company can take a history of applicants that have an approved loan and can observe their repayment history in time. Then they use some definition of default – e.g. a client is called defaulted if he or she was more than 90 days past due (DPD) on at least one of the first 12 monthly payments – to create a development sample consisting of individual clients, their potential explanatory variables[2] and a binary target variable indicating observed defaults. Such sample is then used to develop a precise and stable model to estimate the probability of default for new clients.

1.1.1 Standard Approaches

According to my professional experience the probability of default modeling is most often done using the logistic regression models. In that case we need a historical sample that is mature enough in order to observe a given repayment period after the loan was issued. In the above mentioned example default definition of 90 DPD on one of the first 12 payments, we are talking about loan vintages that are at least 15 months old.

Then we can take the development data sample consisting of vectors $(\boldsymbol{x}_k, y_k)$, where $\boldsymbol{x}_k$ is the vector of potential explanatory variables and y_k, where $y_k = 1$ in the case of default and $y_k = 0$ otherwise, is the target variable. Then using the logistic regression model we can estimate the probability of default $\pi(\boldsymbol{x})$ as

$$\pi(\boldsymbol{x}) = \frac{e^{\boldsymbol{\beta}'\boldsymbol{x}}}{1 + e^{\boldsymbol{\beta}'\boldsymbol{x}}}. \tag{1.1}$$

[1] Among others I can mention also methods of verification of the information provided by the client, an economical model considering applicant's income, costs and minimal living standard, as well as calculation of provisions or collection strategies.

[2] Usually hundreds to thousands of categorized or standardized characteristics – data from application form, credit bureaus, behavioral and transactional data within the bank and other available external data.

The parameters $\boldsymbol{\beta}$ are then estimated using the maximum likelihood method, see (Lehmann and Casella, 1998) or (Van der Vaart, 2000), and tested for significance.

Here the standard modeling usually consists of several rounds of variable categorizations, correlation adjustments and model building using the standard automated selection methods (such as forward, backward or stepwise).[3] The final model is then tested on precision, stability and logic.

For more information about the logistic regression model, its parameter estimation and significance testing I refer to (Agresti, 1990) and (Hosmer and Lemeshow, 2000). More practical recommendations for probability of default model building can be then found e.g. in (Witzany, 2010).

1.1.2 Repayment Survival Model

Before defining the repayment survival model, I shortly outline the basic definitions and principles of the survival analysis and the Cox model used for this task. This summary is mainly taken from (Pazdera et al., 2009) and (Rychnovský, 2011) and the original sources (Reisnerová, 2004), (Kalbfleisch et al., 1980), (Peto, 1972) and (Breslow, 1974).

Survival analysis deals with modeling of the time elapsed until some particular event occurs (it is called *exit* or *end-point*), conditional on the specific characteristics of the subject.

Assume that X is an absolutely continuous nonnegative random variable representing the time to exit of a subject. Denote F the distribution function and f the density of X. Then we define a *hazard function* (or *intensity*) of the subject as

$$\lambda(t) = \lim_{h \to 0+} \frac{1}{h} \mathbb{P}(t \leq X < t + h | X \geq t). \tag{1.2}$$

By a *survivor function* $S(t)$ (also called *survival function*) we denote the probability that the subject will not exit until time t (will survive), i.e. $S(t) = 1 - F(t)$. Using this relation we can rewrite the hazard function (1.2) into the form

$$\lambda(t) = \lim_{h \to 0+} \frac{F(t+h) - F(t)}{h} \frac{1}{S(t)} = \frac{f(t)}{S(t)} = -\frac{\mathrm{d}}{\mathrm{d}t} \ln \big(S(t)\big) \tag{1.3}$$

and a converse relation

$$S(t) = \exp\left[-\int_0^t \lambda(u) \mathrm{d}u \right]. \tag{1.4}$$

[3] A comparison of these methods on a simulated data can be found in (Derksen and Keselman, 1992).

Finally, we define a *cumulative hazard function* as

$$\Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u = -\ln\big(S(t)\big). \tag{1.5}$$

Typically, the survival analysis models work with *censoring*, i.e. the fact that we do not usually have complete information about our subject – whether it had exited or not – simply because we can only observe it during a fixed time interval of length T. During this interval there are three possibilities of a subject status to be observed: exit at time X, no exit until time T or the subject leaving the survey at time C before the final status could have been obtained. For more information about censoring and parameter estimation, see (Reisnerová, 2004), (Kalbfleisch et al., 1980), (Peto, 1972) and (Breslow, 1974),

Finally, there are several parametric and non-parametric alternative models based on various formulas for the hazard function $\lambda(t)$. In this task I use the non-parametric Cox model as follows.

D. R. Cox in (Cox, 1972) assumed the hazard function $\lambda(t; \boldsymbol{x}_i)$ of subject i at time t in the form

$$\lambda(t; \boldsymbol{x}_i) = \lambda_0(t)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}), \tag{1.6}$$

where $\boldsymbol{x}_i$ is the vector of characteristics of subject i and $\boldsymbol{\beta}$ is a vector of parameters. The function $\lambda_0(t)$ is then called a *baseline hazard function*, independent of the subject's characteristics. Due to the fact that the relation

$$\frac{\lambda(t; \boldsymbol{x}_i)}{\lambda(t; \boldsymbol{x}_j)} = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{\exp(\boldsymbol{x}_j'\boldsymbol{\beta})},$$

depends only on subjects' characteristics, the Cox model is often called the *proportional hazards model.* For more information about the Cox model and modeling, see also (Therneau and Grambsch, 2000) or (Persson, 2002).

Now for the repayment survival model in this chapter, I assume that subjects are our loan clients and exits are defaults (e.g. 90 DPD after some of their instalments). Then we assume that every client would default at least once in a lifetime (either before the end of the repayment schedule – this would be a real default – or after the end of the repayment schedule – this would be a virtual default) and that the baseline hazard function is the same for all clients, i.e. that the probabilities of default of any two clients are proportional for all time intervals. This is a basic assumption that is in practice usually accepted. In reality there are often patterns present in the loan life-cycle (e.g. higher probability of default at the first payments followed by better repayment moral) that are common for the loan portfolio and the information about the individual applicant is usually not strong enough to aim for modeling individual shapes of the hazard functions.

Then I can use the full set of observations and for each observation define one of the following outcomes:

- Default occurred in time t, for the case when the client was more than 90 DPD after the instalment scheduled at time t – this is an observation with exit.
- Observation censored in time t, for the case when the client did not default until time t (so called right censoring).

In this case the censoring can be called non-informative (i.e. there is no relation with the default event), because it is only caused by the fact that the loan was issued later, and therefore the observation window is shorter.

Under these assumptions we can use the Cox model to estimate the hazard function, survival function and the vector of parameters to get the probability of default of a client until time t as

$$\pi_t(\boldsymbol{x}) = 1 - S(t, \boldsymbol{x}), \tag{1.7}$$

where $S(t, \boldsymbol{x})$ is the survival function of a subject with a vector of characteristics $\boldsymbol{x}$ at time t. Then the model can be examined for precision, logic and stability in a very similar manner as the common scoring models.

One big advantage of the survival analysis models is the fact that it can incorporate the censored observations, and thus extend the development data sample for the most recent observations. Secondly, the survival function can give the probability of default for all observed times (contrary to the logistic regression approach where only one time horizon is used for modeling).

1.2 Real Data for Modeling

For this task I use the data and its initial transformations that have been used before for the project (Pazdera et al., 2009). It is a real data sample provided for research purposes by one of the biggest Czech banks.

1.2.1 Data Overview

The provided data sample consists of several data files with information about the clients (see the provided categorization in tables 1.1, 1.2 and 1.3), application date and loan maturity, as well as the date of default according to several definitions of default. In total there are 19,139 clients with the following default rates:

- client defaulted on any of his/her loans – 10.6% of the original sample,
- default 90 days past due on this loan – 5.5% of the original sample,
- default 180 days past due on this loan – 4.3% of the original sample.

Table 1.1: Full list of variables and categories – part 1

Variable	Categories
Sex	Female
	Male
Marital status	Single
	Married female
	Married male
	Divorced
	Yokefellow
	Widowed
Education	Basic educ.
	Skilled
	Voc. educ.
	Voc. grad. educ.
	Full sec. educ
	Full sec. gen. educ.
	High voc. educ.
	University
Employment status (empl.since)	Unemployed
	Household
	Retaired
	Student
	Employed <3M
	Employed <1Y
	Employed <5Y
	Employed <10Y
	Employed >10Y
	Entrepreneur
	Civil servant
Employer	State
	SHC state
	SHC non state
	Foreign corporate
	Cooperative
	LTC
	Entrepr.himself
	Entrepreneur
	Other Employer

Table 1.2: Full list of variables and categories – part 2

Variable	Categories
Housing status	House
	Rent estate
	Rent indiv.
	Housing by parents
	Hostel
	Housing other
Repayment type	Cash
	Transf. same bank
	Transf.
	Drawback
	Overdraft
Credit card	Yes
	No
Kind of employment	Empl. non rank
	Middle manager
	Employed enterpr.
	Employed
	Manual
	Freelance occup.
	Student
	Pensioner
	Entrepreneur
	Household
Telephone private	Tel. priv. fix
	Mobile phone priv.
	No tel. priv.
Telephone at work	Tel. work fix
	Mobile phone work
	No tel. work
	Tel. does not work
Number of dependent persons	No dependent person
	1 dependent person
	2 dependent persons
	3 dependent persons
	4 dependent persons
	5 or more dependent persons

Table 1.3: Full list of variables and categories – part 3

Variable	Categories
Monthly income	Numeric
Other income	Numeric
Credit limit	Numeric
Loan distribution channel	3 categories
Age	15 categories

1.2.2 Data Structure for Modeling and Comparison

Further on, I use some of the earlier performed transformations in (Pazdera et al., 2009):

1. Cleaning the data in the sense of handling outliers.
2. Running univariate statistics on each variable in order to find out and solve possible inconsistencies.
3. Re-categorizing nominal variables having too many categories.
4. Omitting the correlated variables, e.g. number of dependent persons or other income.
5. Finally, calculated the variables necessary for the model (time in months, indicators of default, etc.).

For the purpose of this task I additionally adjusted the sample as follows. Because of the practical differences in the risk management between fix-term and revolving loans I proceed with the fix-term products only.[4] Furthermore, we can see from tables 1.4 and 1.5 that the number of cases for the fix-term loans fluctuates in the early samples and after January 2006. Therefore, to achieve a robust number of observations for each month I cleared the data sample to contain only vintages from the period of January 2002 to December 2005. As the sample was originally provided in 2008, the latest observations from December 2005 are mature enough to allow the 24 month default measuring.

Thus I get a new and more homogenous sample of data that has 9,835 observations with measurable default of 90 DPD on one of the first 24 payments.[5]

[4] In my professional experience the fix-term loans (i.e. the loans with the pre-defined installment structure, e.g. fixed monthly payments) are having a different repayment behavior and default times than the revolving loans (such as credit cards or overdrafts).

[5] By the term measurable default I mean the fact that we can see whether the case defaulted within 24 months after issuing the loan or whether the default did not occur within this time frame.

Table 1.4: Date of loan issuing for the whole sample (till December 2003)

Date of loan issuing	Number of revolving loans	Number of fix term loans
Before January 2001	6	17
January 2001	0	1
February 2001	0	1
March 2001	0	2
April 2001	0	4
May 2001	0	1
June 2001	0	4
July 2001	0	4
August 2001	0	5
September 2001	0	1
October 2001	0	4
November 2001	1	14
December 2001	1	9
January 2002	0	126
February 2002	0	150
March 2002	0	160
April 2002	0	162
May 2002	0	174
June 2002	0	136
July 2002	0	157
August 2002	4	125
September 2002	30	191
October 2002	15	205
November 2002	47	334
December 2002	18	132
January 2003	37	132
February 2003	63	145
March 2003	82	156
April 2003	81	162
May 2003	200	157
June 2003	227	151
July 2003	237	221
August 2003	169	162
September 2003	228	163
October 2003	247	207
November 2003	259	192
December 2003	177	155
Total before January 2004	2,129	4,122

Table 1.5: Date of loan issuing for the whole sample (after January 2004)

Date of loan issuing	Number of revolving loans	Number of fix term loans
Before January 2004	2,129	4,122
January 2004	170	120
February 2004	203	146
March 2004	210	125
April 2004	226	195
May 2004	316	153
June 2004	255	159
July 2004	270	140
August 2004	251	120
September 2004	347	167
October 2004	304	211
November 2004	407	214
December 2004	291	168
January 2005	253	122
February 2005	317	165
March 2005	323	211
April 2005	322	490
May 2005	328	753
June 2005	286	263
July 2005	208	187
August 2005	264	300
September 2005	413	175
October 2005	211	336
November 2005	648	584
December 2005	280	276
January 2006	3	1
February 2006	1	0
Total sample	9,236	9,903

Altogether, I get 448 defaults and, thus 4.6% default rate. Now I prepare two sets of data samples – one for the random testing sample comparison and one for the predictive performance comparison.

Random Sample

In the first task I divide the data into a development sample and a comparison sample randomly in order to develop both models on the development sample and compare their diversification power on the independent comparison (validation) sample. Since for the repayment survival model we use all observed defaults with the time of default (even though the default occurred later after 24 months), I denote it as exit in the following text. See the sample overview in table 1.6. As we can see from this table, there are additional exits in the development sample that can be used for the survival model building. In the comparison sample I compare the performance on the defaults only, therefore the exits are not relevant here – thus not shown in the table.

Table 1.6: Random sample overview

Sample	Clients	Defaults	Default rate	Exits
Development	7000	319	4.6%	500
Comparison	2835	129	4.6%	—
Total	9835	448	4.6%	—

Progressive Time Sample

In the second task I divide the sample by time. Imagine it is beginning of 2006 and we are developing a scoring model. Then using the standard logistic regression model we can only use the clients from January 2002 to December 2003 as a development sample, whereas for the survival analysis model we can use all the time-censored observations from 2002–2005, including the partially observed vintages 2004–2005.

This is illustrated in figure 1.1, where the area A is the development sample for the logistic regression model, areas B and C are the additional exit observations that can be used for the repayment survival model and area D contains the information that is censored for both models and used for final comparison only.

This is why I divide the sample into the full vintages of 2002–2003 as the development sample for the logistic regression model, the vintages from 2002–2005 time-censored to the date of 1 January 2006 as an development sample for the repayment survival model and the full sample of 2004–2005 as the sample for comparison. For details see table 1.7. Again, in the comparison sample I compare the

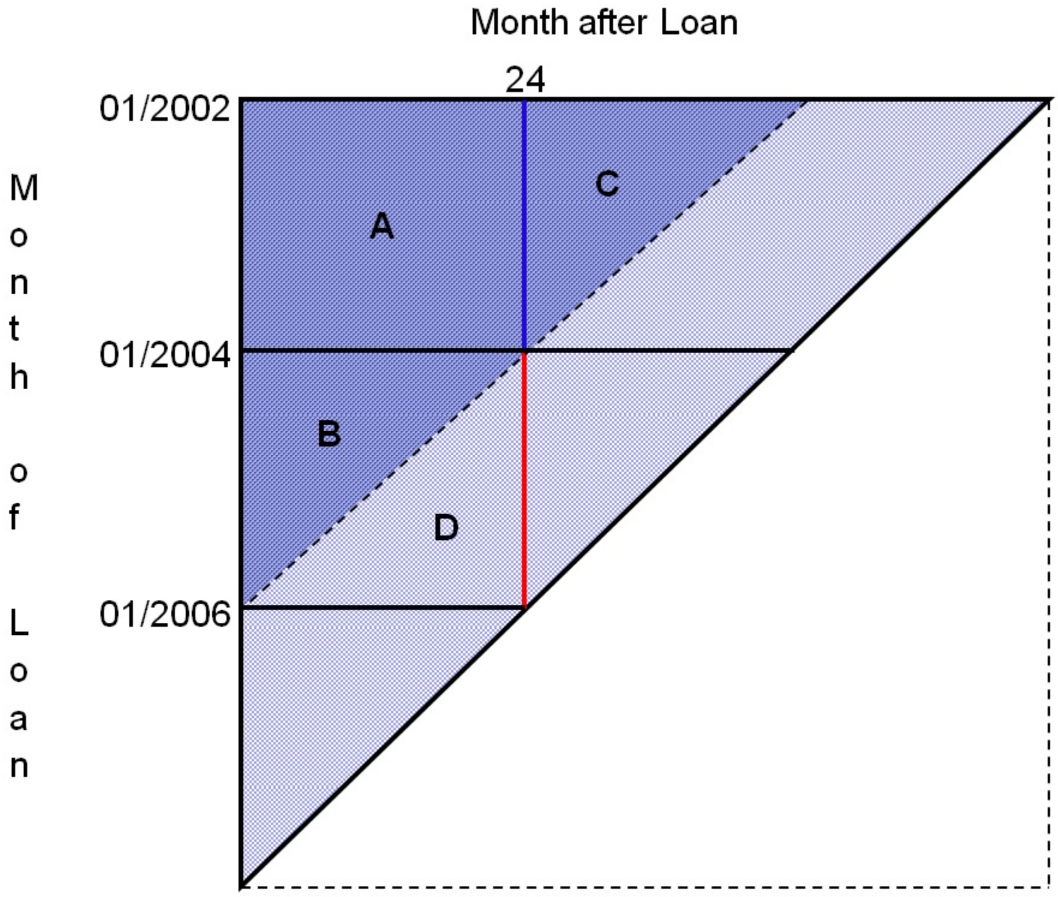

Figure 1.1: Illustration of the progressive time sample structure

performance on the defaults only, therefore the exits are not relevant here – thus not shown in the table.

Table 1.7: Progressive time sample overview

Sample	Clients	Defaults	Default rate	Exits
Development 2002–2003	4055	215	5.3%	279
Development 2004–2005	5780	—	—	79
Comparison 2004–2005	5780	233	4.0%	—
Total	9835	448	4.6%	—

1.3 Goodness of Fit Definition

In this section I briefly outline the basic information about the Gini coefficient with the Somers' d calculation method and lift characteristic, that are used for comparison of the rival models. The information in this section is taken from (Rychnovský, 2011), (Somers, 1962) and (Witzany, 2017).

The idea of model *diversification power* is based on the ability of a scoring model to distinguish bad clients (i.e. the clients who will default) from good clients (i.e. the clients who will not default). Every scoring model assigns to each client a score value (e.g. the estimated probability of default). If we then order the clients according to their scores we get an ordering of clients from which we can see how powerful the model really is.

1.3.1 Gini Coefficient

The Gini coefficient is usually defined using the distribution curve (also Lorenz curve or ROC curve – from Receiver Operating Characteristic). For more information see also (Hanley et al., 1983), (Řezáč et al., 2011) or (Witzany, 2010).

First denote $\mathbb{S} = \big\{\mathbb{S}(\boldsymbol{x}), \boldsymbol{x} \in \boldsymbol{X}\big\}$ the set of all values of a scoring function $\mathbb{S}(\boldsymbol{x})$. Then for every value of score $s \in \mathbb{S}$ define the *distribution function of bad clients* $\mathrm{F}^B(s)$ as the probability that a randomly chosen bad client will have a score lower then s; and analogically, the *distribution function of good clients* $\mathrm{F}^G(s)$ as the probability that a randomly chosen good client will have a score lower then s.

The explicit distribution functions $\mathrm{F}^G(s)$ and $\mathrm{F}^B(s)$ are in practice not known; and therefore, they are usually replaced by their consistent estimates. The function $\mathrm{F}^B(s)$ is estimated as the ratio of bad clients with scores lower than s and all bad clients, and the function $\mathrm{F}^G(s)$ is estimated as the ratio of good clients with scores lower than s and all good clients.

Then we can define the *distribution curve* as the connection of the set

$$L = \Big\{\big[\,\mathrm{F}^B(s), \mathrm{F}^G(s)\big] \in \mathbb{R}^2 : s \in S\Big\}, \tag{1.8}$$

with the points $[0, 0]$ and $[1, 1]$ (see an illustration in figure 1.2).

Now the *Gini coefficient* can be defined as the ratio of the oriented area between the distribution curve and the diagonal of the square (A) and the total area above the diagonal ($A + B$), thus $GC = \frac{A}{A+B}$ (see again figure 1.2).

To compute the Gini coefficient of a model, the *Somers' d statistic* from (Somers, 1962) is often used. Then

$$d = \frac{a - b}{a + b + c}, \tag{1.9}$$

where for $\mathbb{B}$ the index set of all bad clients and $\mathbb{G}$ the index set of all good clients and s_k the score of the k-th client, we have $a = \sum_{l \in \mathbb{B}} \left|\{k : k \in \mathbb{G}, s_k < s_l\}\right|$ is the

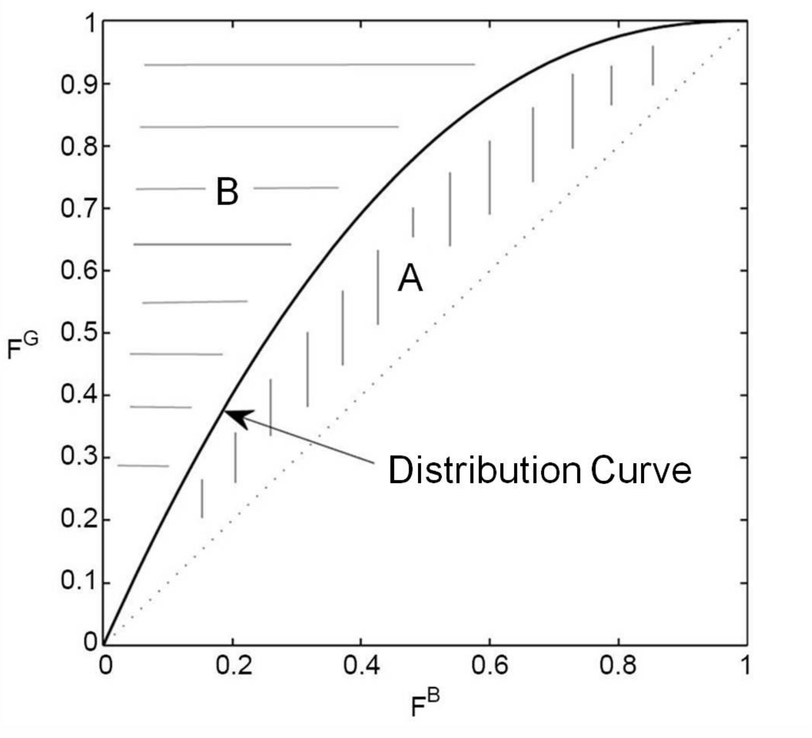

Figure 1.2: Distribution curve from (Rychnovský, 2011)

number of all pairs of a good and a bad client where the good client has lower score than the bad client (i.e. number of pairs in a correct order – also called *concordant*); $b = \sum_{l \in \mathbb{B}} |\{k : k \in \mathbb{G}, s_k > s_l\}|$ is the number of all pairs of a good and a bad client where the good client has higher score than the bad client (i.e. number of pairs in an incorrect order – also called *discordant*), and $c = \sum_{l \in \mathbb{B}} |\{k : k \in \mathbb{G}, s_k = s_l\}|$ is the number of all pairs of a good and a bad client where the good client has the same score as the bad client (also called *irrelevant*). This statistics is then used in practice to estimate the Gini coefficient of scoring models. For more information about the Somers' d in categorical data analysis I refer to (Somers, 1962).

The value of the Gini coefficient is then in the interval $[-1, 1]$, where

- $GC = 1$ for an ideal diversification power (i.e. a model, where all good clients have scores lower than all bad clients),
- GC close to zero for a random model, and
- $GC < 0$ for a reversal model (i.e. with a contradictory classification).

Even though from the economic theory the Gini coefficient is usually in the interval $[0, 1]$, in the risk management the complete interval of $[-1, 1]$ is usually used to evaluate both good and bad models.

1.3.2 Lift

Another characteristic used to compare the rival models is *lift*. For the purpose of scoring modeling, we define the $P\%$ value of lift as the ratio of the default rate for the $P\%$ worst cases divided by the default rate for the whole population.

Compared to the Gini coefficient, the lift is a characteristic evaluating the model performance with the reference to one relative point only (e.g. a decile lift for $P = 10$). This can be used for example when we know that those $P\%$ of the worst clients should be rejected by the model and we want to see the direct impact of such model on the rejected population. Therefore, to evaluate the whole model, more information is usually taken from the complete lift curve (e.g. the values for all $P \in [5, 100]$).

For more information about the distribution power measures I refer to (Řezáč et al., 2011) or (Witzany, 2009).

1.4 Results

In this section I present the key results of the modeling and comparison of the standard approach using the logistic regression model with the repayment survival model on the real banking data. As mentioned earlier, I compare the results from two points of view – on the random testing sample and on the progressive time testing sample. All the calculations are made using SAS 9.4 and MS Excel.

Before running the models, I check all the used predictors, their categorization and performance on my sample. This is performed on a standard set of charts with the number of cases and default rate in all the categories of all predictors. From figures 1.3 to 1.17 we can see the performance of individual variables.[6] From these figures we can understand the the risk of individual categories together with their share in the portfolio, i.e. showing us the logic and giving us some indication of the relevance to the model.

From my professional experience, I can say that from most of the predictors we can see the standard risk behavior of the credit portfolio similar on various markets (e.g. men are riskier then women, clients living in their own house are less risky than the others, higher education suggests less risk and so on), some predictors are very sensitive to the bank's limit calculation and verification process (e.g. Credit limit or Monthly income).

[6] In figure 1.7 we can see that the category Unemployed is merged with other middle risk categories. From the professional point of view I would personally prefer the logical approach in this case and rather put the Unemployed category together with the high risk categories, or reject those cases at all due to the income instability.

Here we can see an interesting phenomena that some of the predictors (e.g. existence of a fix phone line at home or work) are quite strong in this sample, however in reality they are loosing their significance in time and could be very weak nowadays. Therefore, whenever there is a longer sample available, it is always important to test the stability of the predictors in time before the real implementation.

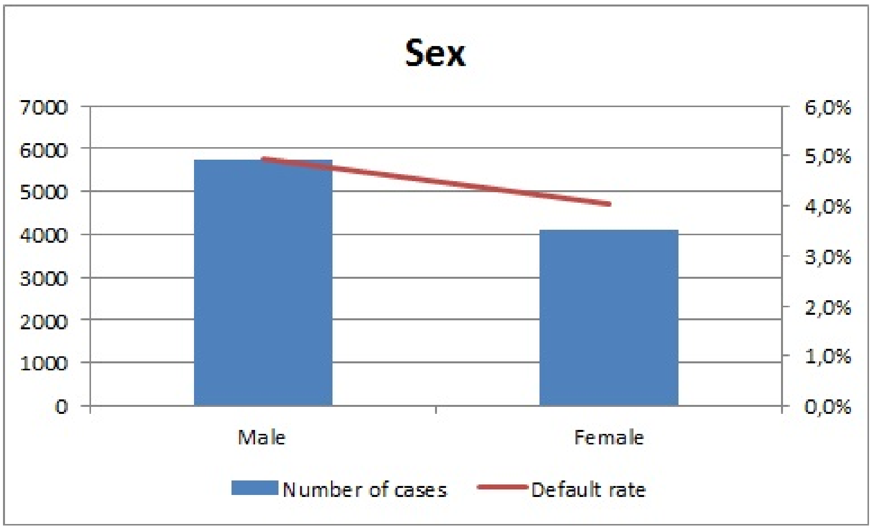

Figure 1.3: Variable Sex – number of cases (L) and default rate (R)

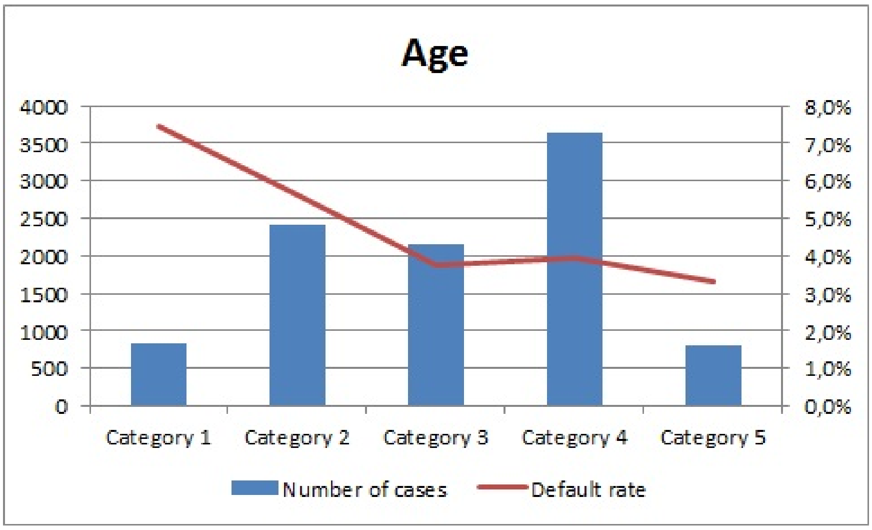

Figure 1.4: Variable Age (with undisclosed categories) – number of cases (L) and default rate (R)

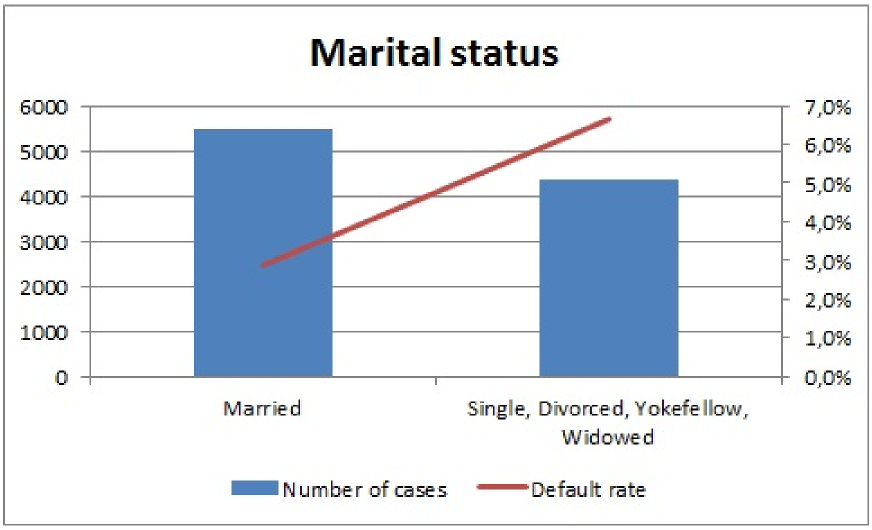

Figure 1.5: Variable Marital status – number of cases (L) and default rate (R)

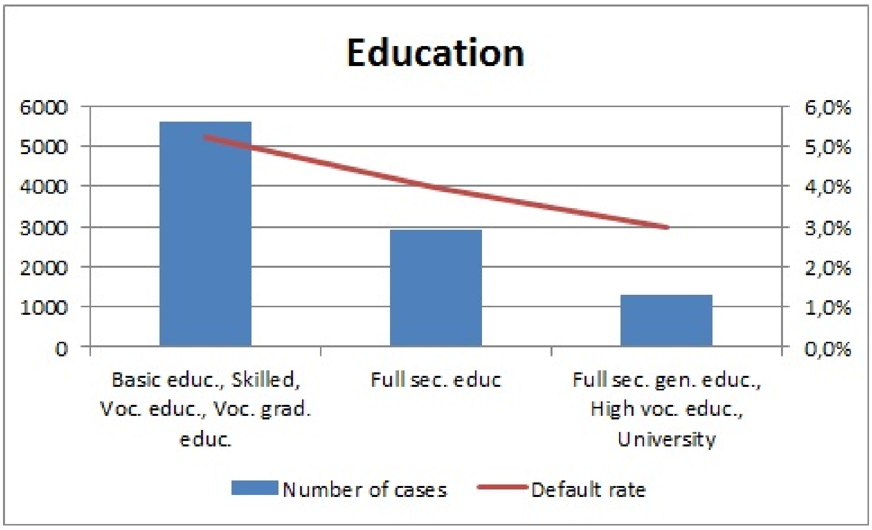

Figure 1.6: Variable Education – number of cases (L) and default rate (R)

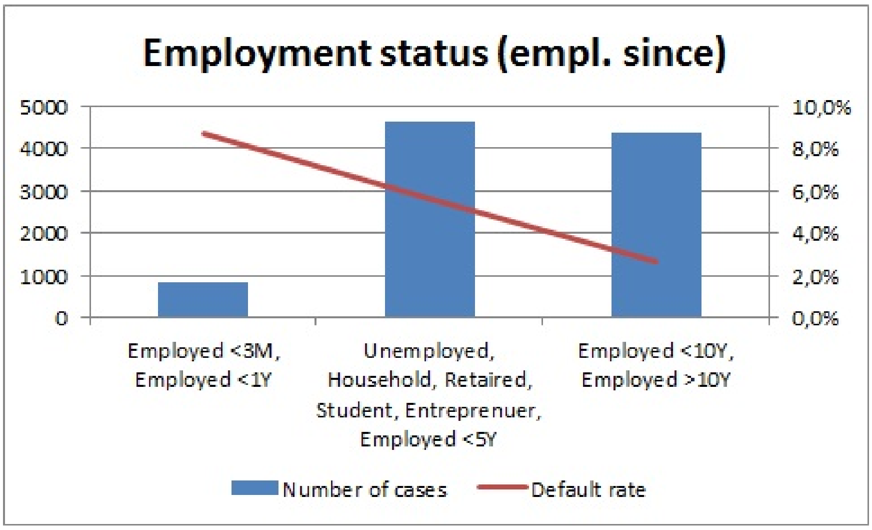

Figure 1.7: Variable Employment status – number of cases (L) and default rate (R)

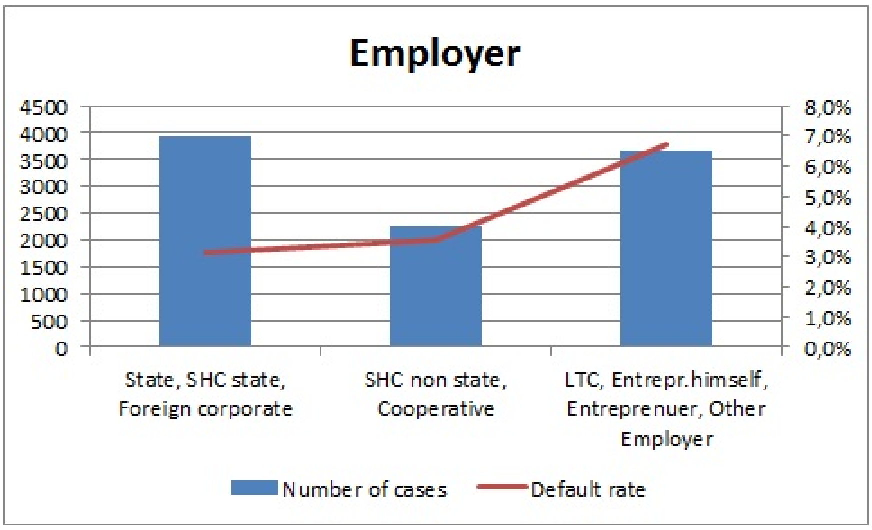

Figure 1.8: Variable Employer – number of cases (L) and default rate (R)

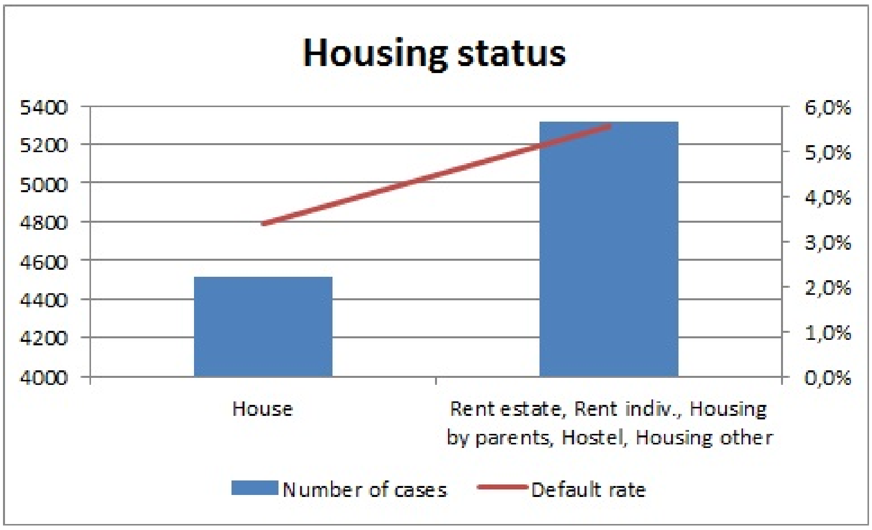

Figure 1.9: Variable Housing status – number of cases (L) and default rate (R)

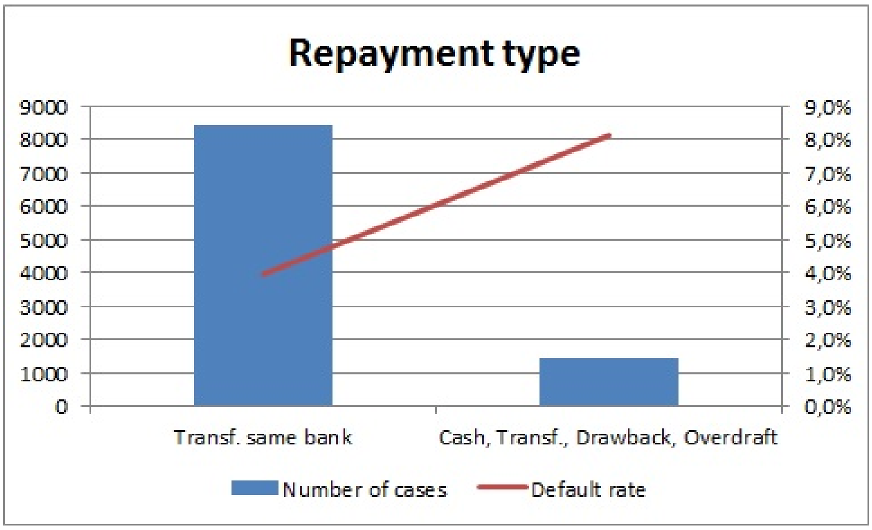

Figure 1.10: Variable Repayment type – number of cases (L) and default rate (R)

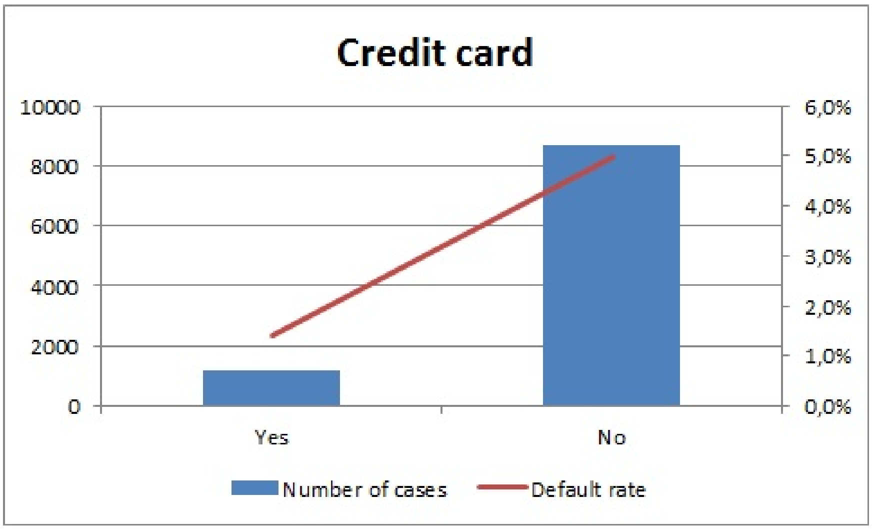

Figure 1.11: Variable Credit card – number of cases (L) and default rate (R)

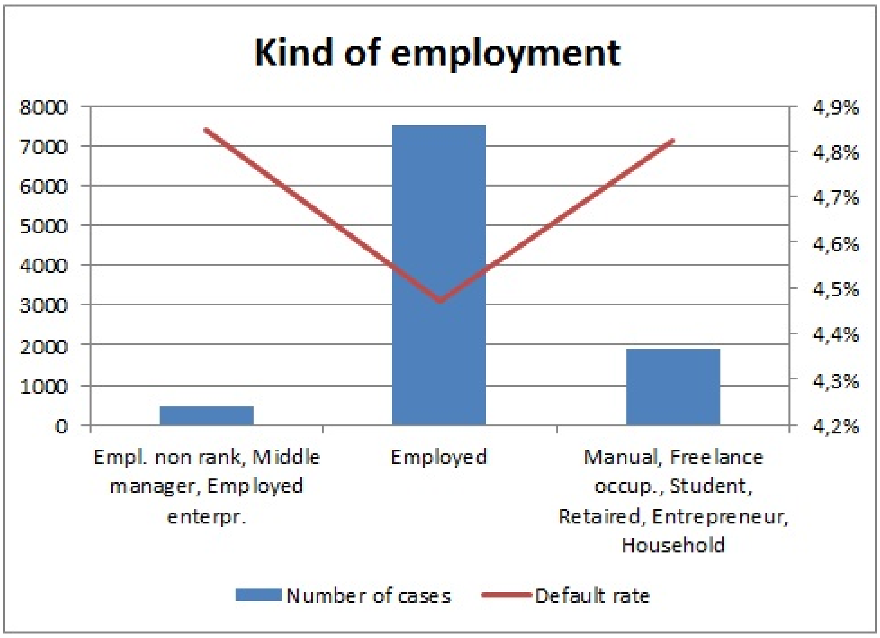

Figure 1.12: Variable Kind of employment – number of cases (L) and default rate (R)

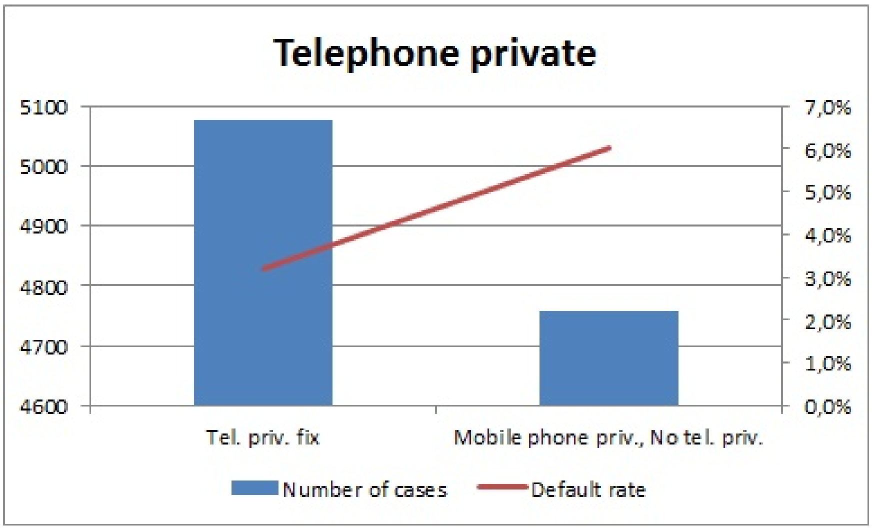

Figure 1.13: Variable Telephone private – number of cases (L) and default rate (R)

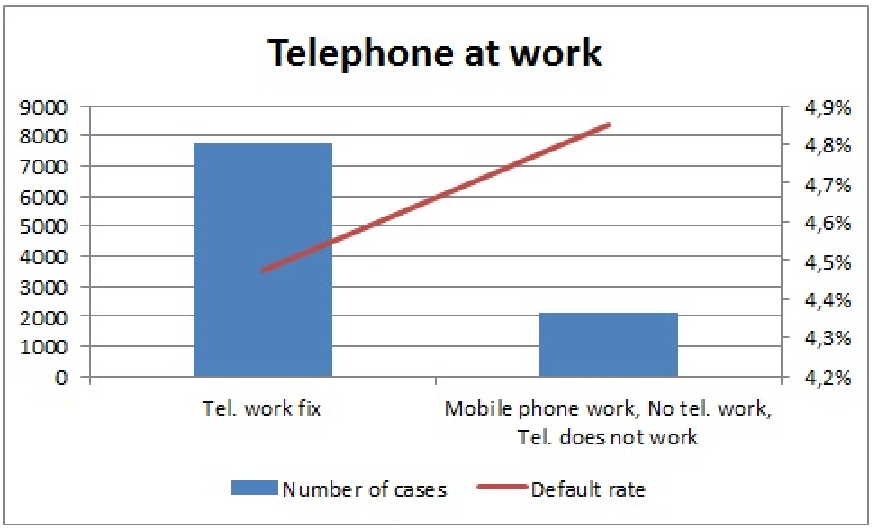

Figure 1.14: Variable Telephone at work – number of cases (L) and default rate (R)

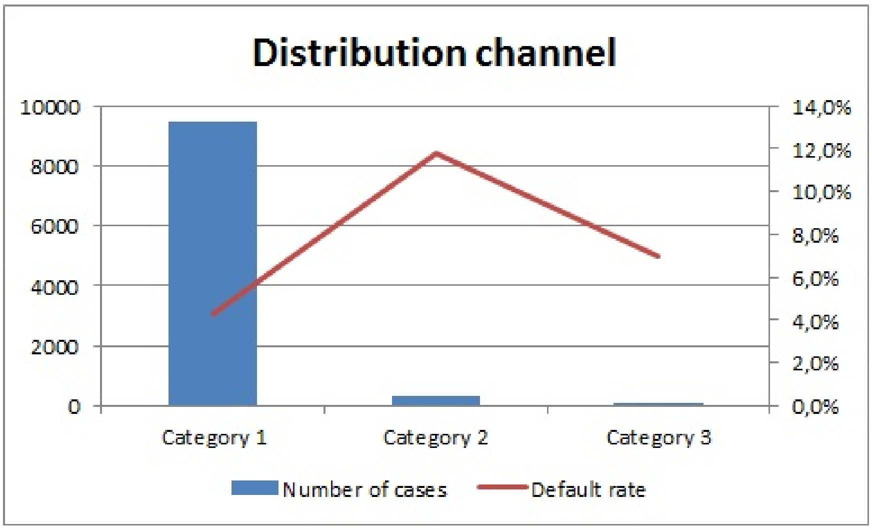

Figure 1.15: Variable Loan distribution channel (with undisclosed categories) – number of cases (L) and default rate (R)

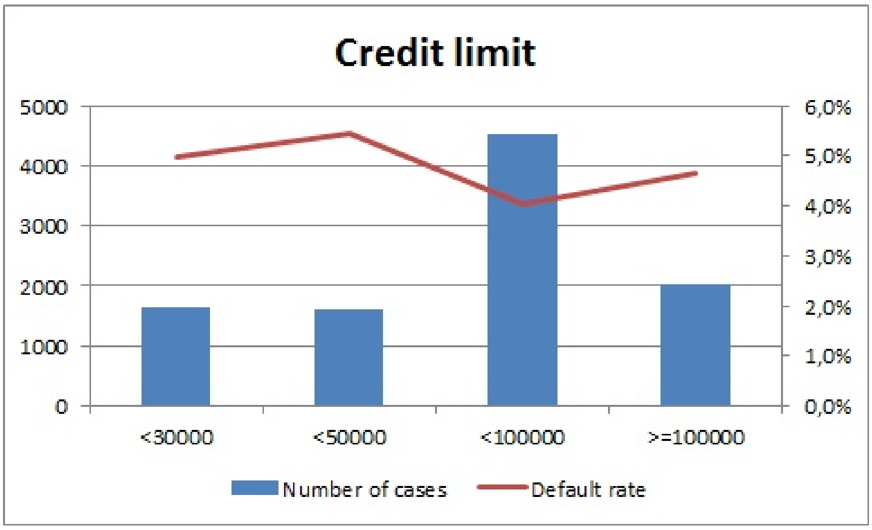

Figure 1.16: Variable Credit limit – number of cases (L) and default rate (R)

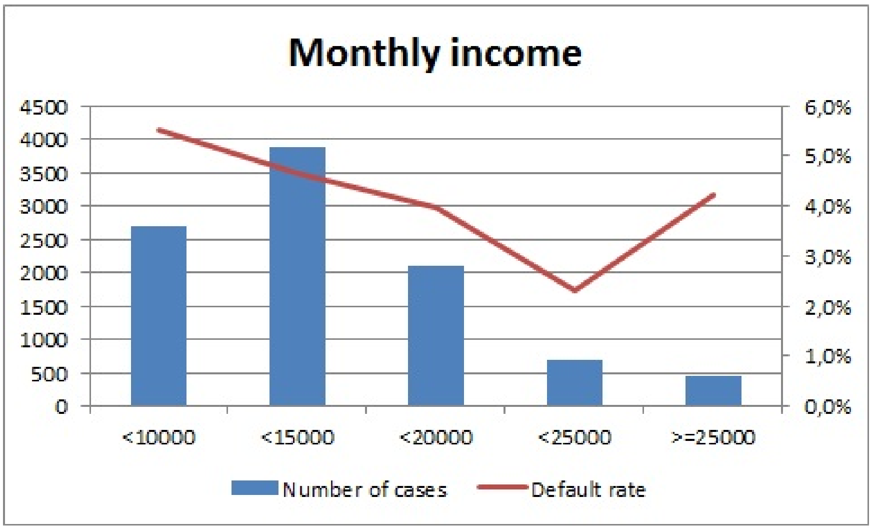

Figure 1.17: Variable Monthly income – number of cases (L) and default rate (R)

1.4.1 Standard Approach

In this section I present the results of the standard approach using the logistic regression model on both sets of samples.

Random Sample

First I run the logistic regression procedure with the stepwise predictors selection method on the significance level 0.05 for both entry and exit on the development data sample with 7000 observations. Then I check the results shown in tables 1.8 and 1.9, and the correlation matrix in table 1.11. As we can see from table 1.9, the p-value for both categories of predictor distribution channel are above 0.05, however the p-value for the whole predictor is below 0.05 as we can see from table 1.8.

Since all the correlations between different non-intercept predictors (not just categories) are in absolute values less than 0.4 (that is generally recommended as the acceptability threshold),[7] I accept this model for comparison and calculate the score for all clients in both samples. Finally, I evaluate the Gini coefficient and lift for both samples.

As we can see from table 1.10, the Gini coefficient on the comparison sample is much lower than on the development sample. This suggests that the model is not very stable and even with higher lift I would not accept such model for real probability of default estimation. However, for the comparison purposes I prefer the models to be developed in a similar way and not manually adjusted.

Table 1.8: Analysis of the selected predictors of the standard approach on the random sample

Predictor	Degrees of Freedom	Wald Chi-Square	Pr > ChiSq
Sex	1	4.8023	0.0284
Marital status	1	31.6047	< .0001
Employment status	2	34.3346	< .0001
Employer	2	30.0245	< .0001
Housing status	1	5.5632	0.0183
Repayment type	1	27.0599	< .0001
Credit card	1	15.8031	< .0001
Telephone private	1	17.3618	< .0001
Distribution channel	2	21.7976	< .0001

[7] Some more information about how to work with correlated data in the credit risk modeling can be found in (Rychnovský, 2011). For an article about modeling correlated data, see e.g. (Le Cessie and Van Houwelingen, 1994).

Table 1.9: Analysis of maximum likelihood estimates of the standard approach on the random sample

Parameter	Category	Estimate	Standard Error	Pr > ChiSq
Intercept		2.0074	0.5605	0.0003
Sex	1	-0.2724	0.1243	0.0284
Marital status	1	0.7096	0.1262	$< .0001$
Employment status	1	-1.0609	0.1856	$< .0001$
Employment status	2	-0.5872	0.1414	$< .0001$
Employer	1	0.6176	0.1377	$< .0001$
Employer	2	0.7260	0.1683	$< .0001$
Housing status	1	0.2977	0.1262	0.0183
Repayment type	1	0.7166	0.1378	$< .0001$
Credit card	1	641368	0.3435	$< .0001$
Telephone private	1	0.5153	0.1237	$< .0001$
Distribution channel	1	0.0264	0.5318	0.9604
Distribution channel	2	-0.9814	0.5658	0.0828

Table 1.10: Summary of the standard approach on the random sample

Summary	Development	Comparison
Gini	0.51	0.39
Lift 10%	2.73	2.78

Progressive Time Sample

Secondly, I run the logistic regression procedure on the progressive time sample. I put the stepwise selection method again on the significance level 0.05 for both entry and exit and run it on the development data sample with 4,055 observations. Then again I check the results shown in tables 1.12 and 1.13, and the correlation matrix in table 1.15.

Even here all the correlations between different non-intercept predictors (not just categories) are in absolute values less than 0.4 and I accept this model for comparison and calculate the score for all clients in both samples. Finally, I evaluate the Gini coefficient and lift for both samples.

As we can see from table 1.14, there is again a minor drop in the Gini coefficient and a major drop in lift.

Table 1.11: Analysis of correlations of the standard approach on the random sample

Par.	Cat.	Int.	Sex	M.st.	E.st.	E.st.	Emp.	Emp.	H.st.	Rep.	C.c.	T.p.	D.ch.	D.ch.
Int.		1.00	−0.17	−0.08	−0.16	−0.21	−0.10	−0.10	−0.06	−0.16	−0.02	−0.03	−0.93	−0.88
Sex	1	−0.17	1.00	−0.10	0.02	0.04	0.11	−0.01	0.04	0.04	−0.02	0.01	0.01	0.01
M.st.	1	−0.08	−0.10	1.00	0.07	0.08	0.03	0.04	−0.18	0.03	0.02	−0.10	0.01	0.01
E.st.	1	−0.16	0.02	0.07	1.00	0.54	0.15	0.12	0.01	−0.06	0.04	−0.01	0.02	0.01
E.st.	2	−0.21	0.04	0.08	0.54	1.00	0.05	0.13	0.00	−0.02	0.03	−0.01	0.02	0.01
Emp.	1	−0.10	0.11	0.03	0.15	0.05	1.00	0.29	0.00	0.04	−0.01	−0.03	−0.03	−0.03
Emp.	2	−0.10	−0.01	0.04	0.12	0.13	0.29	1.00	0.00	0.07	−0.01	0.01	−0.01	−0.01
H.st.	1	−0.06	0.04	−0.18	0.01	0.00	0.00	0.00	1.00	0.01	0.02	−0.10	0.00	−0.02
Rep.	1	−0.16	0.04	0.03	−0.06	−0.02	0.04	0.07	0.01	1.00	−0.08	−0.01	−0.04	0.03
C.c.	1	−0.02	−0.02	0.02	0.04	0.03	−0.01	−0.01	0.02	−0.08	1.00	0.00	0.01	0.01
T.p.	1	−0.03	0.01	−0.10	−0.01	−0.01	−0.03	0.01	−0.10	−0.01	0.00	1.00	−0.03	−0.02
D.ch.	1	−0.93	0.01	0.01	0.02	0.02	−0.03	−0.01	0.00	−0.04	0.01	−0.03	1.00	0.92
D.ch.	2	−0.88	0.01	0.01	0.01	0.01	−0.03	−0.01	−0.02	0.03	0.01	−0.02	0.92	1.00

Table 1.12: Analysis of the selected predictors of the standard approach on the progressive time sample

Predictor	Degrees of Freedom	Wald Chi-Square	Pr > ChiSq
Sex	1	4.1815	0.0409
Marital status	1	18.6966	< .0001
Employment status	2	28.9289	< .0001
Employer	2	23.7860	< .0001
Housing status	1	6.5490	0.0105
Telephone private	1	11.9081	0.0006
Telephone at work	1	6.9916	0.0082

Table 1.13: Analysis of maximum likelihood estimates of the standard approach on the progressive time sample

Parameter	Category	Estimate	Standard Error	Pr > ChiSq
Intercept		2.0662	0.2868	< .0001
Sex	1	−0.3141	0.1536	0.0409
Marital status	1	0.6559	0.1517	< .0001
Employment status	1	−1.1975	0.2235	< .0001
Employment status	2	−0.5900	0.1778	0.0009
Employer	1	0.7007	0.1705	< .0001
Employer	2	0.7590	0.2080	0.0003
Housing status	1	0.3949	0.1543	0.0105
Telephone private	1	0.5033	0.1459	0.0006
Telephone at work	1	0.4720	0.1785	0.0082

Table 1.14: Summary of the standard approach on the progressive time sample

Summary	2002–2003	2004–2005
Gini	0.45	0.38
Lift 10%	3.15	1.83

Table 1.15: Analysis of correlations of the standard approach on the progressive time sample

Par.	Cat.	Int.	Sex	M.st.	E.st.	E.st.	Emp.	Emp.	H.st.	T.p.	T.w.
Int.		1.00	−0.36	−0.23	−0.43	−0.61	−0.42	−0.23	−0.19	−0.26	−0.63
Sex	1	−0.36	1.00	−0.09	0.03	0.05	0.10	0.01	0.02	0.03	−0.05
M.st.	1	−0.23	−0.09	1.00	0.09	0.12	0.02	0.03	−0.14	−0.11	0.09
E.st.	1	−0.43	0.03	0.09	1.00	0.54	0.16	0.11	−0.01	−0.01	0.06
E.st.	2	−0.61	0.05	0.12	0.54	1.00	0.10	0.13	0.02	0.02	0.22
Emp.	1	−0.42	0.10	0.02	0.16	0.10	1.00	0.27	0.05	0.00	0.22
Emp.	2	−0.23	0.01	0.03	0.11	0.13	0.27	1.00	0.03	0.02	−0.01
H.st.	1	−0.19	0.02	−0.14	−0.01	0.02	0.05	0.03	1.00	−0.07	0.06
T.p.	1	−0.26	0.03	−0.11	−0.01	0.02	0.00	0.02	−0.07	1.00	0.10
T.w.	1	−0.63	−0.05	0.09	0.06	0.22	0.22	−0.01	0.06	0.10	1.00

1.4.2 Repayment Survival Model

In this section I present the results of the repayment survival model using the nonparametric Cox model described in section 1.1.2.

Random Sample

This time I run the Cox procedure on the random sample, put the stepwise selection method again on the significance level 0.05 for both entry and exit and run it on the development data sample with 7000 observations and 500 exits.

In figure 1.18 we can see the estimated baseline function for the model and in figure 1.19 the estimated survival function for the first contract in the sample as an example. Both functions are plotted with their 90% confidence limits.

Then similarly to the logistic regression method I check the results shown in tables 1.16 and 1.17,[8] and the correlation matrix in table 1.19. Even here all the correlations between different predictors (not just categories) are in absolute values less than 0.4 and I accept this model for comparison and calculate the score for all clients in both samples as described in 1.1.2. Finally, I evaluate the Gini coefficient and lift for both samples.

As we can see from table 1.18, there is also a drop in the Gini coefficient and lift.

[8] As the baseline level is given by the baseline hazard function, there is no intercept in the model.

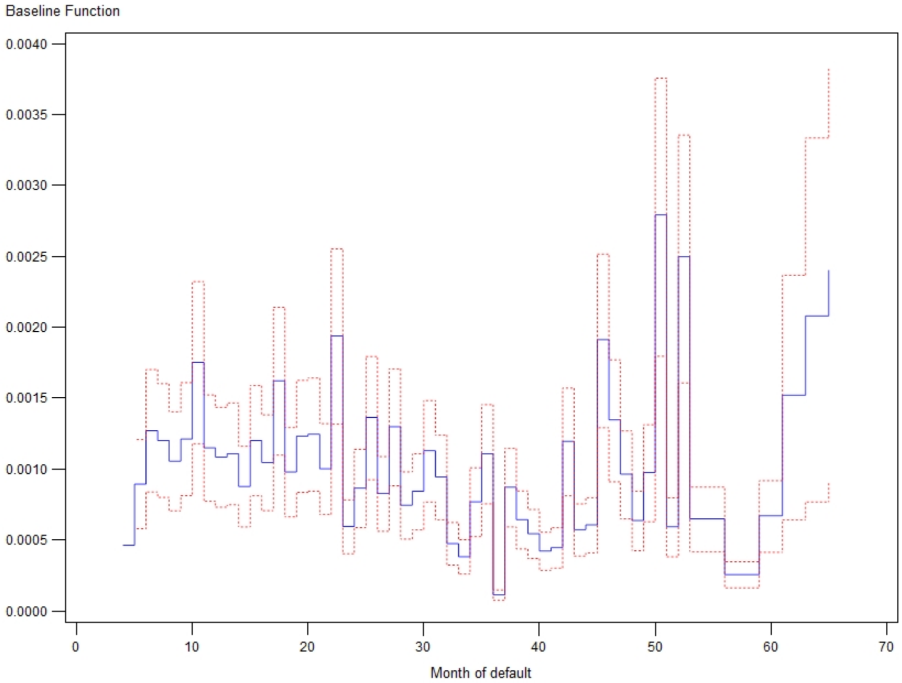

Figure 1.18: Baseline function for the Cox model on the random sample

Progressive Time Sample

Finally, I run the Cox procedure on the progressive time sample, put the stepwise selection method again on the significance level 0.05 for both entry and exit and run it on the development data sample with 9835 observations and 358 exits.

Again, in figure 1.20 we can see the estimated baseline function for the model and in figure 1.21 the estimated survival function for the first contract in the sample as an example. Both functions are plotted with their 90% confidence limits.

I check the results shown in tables 1.20 and 1.21, and the correlation matrix in table 1.23. All the correlations between different predictors (not just categories) are in absolute values less than 0.4. I accept this model for comparison and calculate the score for all clients as described in 1.1.2. Finally, I evaluate the Gini coefficient and lift for both vintages 2002–2003 and 2004–2005. From table 1.22, we can see the Gini coefficient remains stable and there is a minor drop in lift.

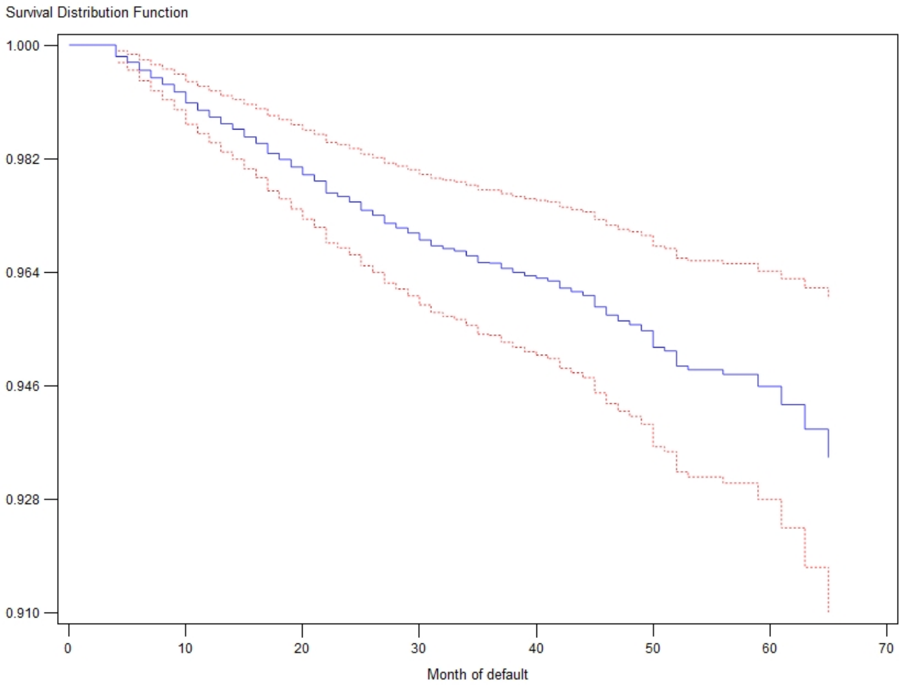

Figure 1.19: Survival function of the first contract on the random sample

Table 1.16: Analysis of the selected predictors of the repayment survival model on the random sample

Predictor	Degrees of Freedom	Wald Chi-Square	Pr > ChiSq
Sex	1	4.8385	0.0278
Marital status	1	32.7541	< .0001
Education	2	18.6069	< .0001
Employment status	2	43.7940	< .0001
Employer	2	27.2749	< .0001
Housing status	1	10.8025	0.0010
Repayment type	1	29.9899	< .0001
Credit card	1	13.3729	0.0003
Telephone private	1	14.7220	0.0001
Distribution channel	2	15.0739	0.0005

Table 1.17: Analysis of maximum likelihood estimates of the repayment survival model on the random sample

Parameter	Category	Estimate	Standard Error	Pr > ChiSq
Sex	1	0.20854	0.09481	0.0278
Marital status	1	−0.54307	0.09489	< .0001
Education	1	0.64284	0.17682	0.0003
Education	2	0.32907	0.19070	0.0844
Employment status	1	0.91189	0.14007	< .0001
Employment status	2	0.46189	0.10568	< .0001
Employer	1	−0.38070	0.10484	0.0003
Employer	2	−0.59314	0.12669	< .0001
Housing status	1	−0.31685	0.09640	0.0010
Repayment type	1	−0.58169	0.10622	< .0001
Credit card	1	−0.78838	0.21559	0.0003
Telephone private	1	−0.35800	0.09330	0.0001
Distribution channel	1	−0.01917	0.38219	0.9600
Distribution channel	2	0.67663	0.41522	0.1032

Table 1.18: Summary of the repayment survival model on the random sample

Summary	Development	Comparison
Gini	0.50	0.40
Lift 10%	2.96	2.32

Table 1.19: Analysis of correlations of the repayment survival model on the random sample

Par.	Cat.	Sex	M.st.	Educ.	Educ.	E.st.	E.st.	Emp.	Emp.	H.st.	Rep.	C.c.	T.p.	D.ch.	D.ch.
Sex	1	1.00	−0.10	−0.03	0.02	0.03	0.04	0.11	−0.01	0.03	0.03	−0.04	0.01	0.00	0.00
M.st.	1	−0.10	1.00	−0.01	0.02	0.08	0.09	0.01	0.04	−0.18	0.02	0.02	−0.10	0.00	0.01
Educ.	1	−0.03	−0.01	1.00	0.83	0.02	0.00	0.09	−0.02	−0.01	0.02	0.05	0.05	0.01	0.02
Educ.	2	0.02	0.02	0.83	1.00	−0.01	0.01	0.04	−0.02	−0.02	0.00	0.02	0.01	0.01	0.02
E.st.	1	0.03	0.08	0.02	−0.01	1.00	0.51	0.17	0.13	0.01	−0.04	0.05	0.00	0.03	0.01
E.st.	2	0.04	0.09	0.00	0.01	0.51	1.00	0.05	0.13	0.00	−0.02	0.04	−0.01	0.02	0.00
Emp.	1	0.11	0.01	0.09	0.04	0.17	0.05	1.00	0.30	0.00	0.03	0.00	−0.01	−0.03	−0.02
Emp.	2	−0.01	0.04	−0.02	−0.02	0.13	0.13	0.30	1.00	0.00	0.05	−0.02	0.01	−0.03	−0.02
H.st.	1	0.03	−0.18	−0.01	−0.02	0.01	0.00	0.00	0.00	1.00	0.00	0.02	−0.11	0.00	−0.02
Rep.	1	0.03	0.02	0.02	0.00	−0.04	−0.02	0.03	0.05	0.00	1.00	−0.10	−0.01	−0.04	0.04
C.c.	1	−0.04	0.02	0.05	0.02	0.05	0.04	0.00	−0.02	0.02	−0.10	1.00	0.01	0.01	0.01
T.p.	1	0.01	−0.10	0.05	0.01	0.00	−0.01	−0.01	0.01	−0.11	−0.01	0.01	1.00	−0.04	−0.02
D.ch.	1	0.00	0.00	0.01	0.01	0.03	0.02	−0.03	−0.03	0.00	−0.04	0.01	−0.04	1.00	0.90
D.ch.	2	0.00	0.01	0.02	0.02	0.01	0.00	−0.02	−0.02	−0.02	0.04	0.01	−0.02	0.90	1.00

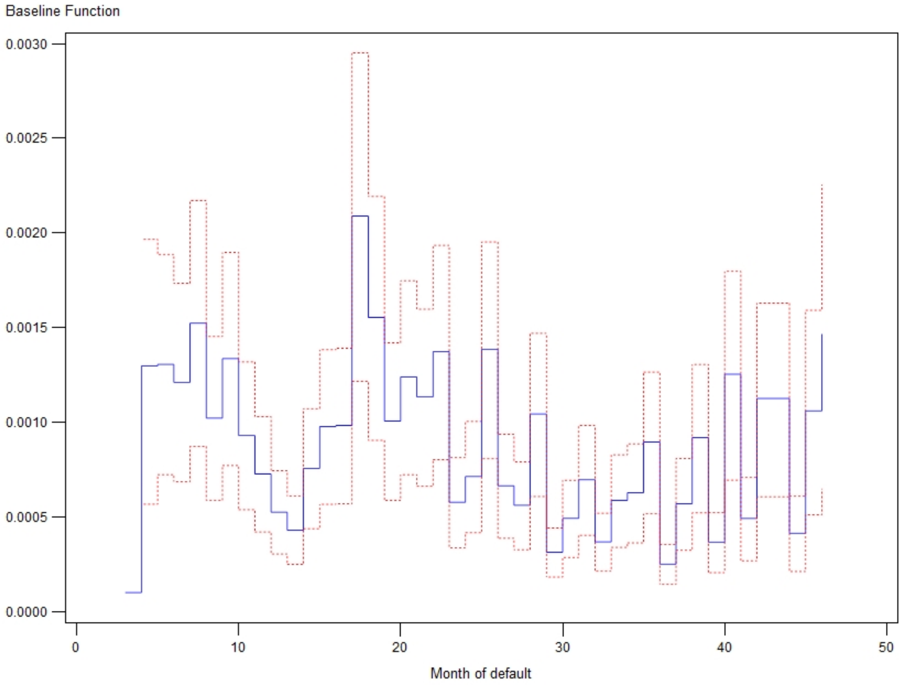

Figure 1.20: Baseline function for the Cox model on the progressive time sample

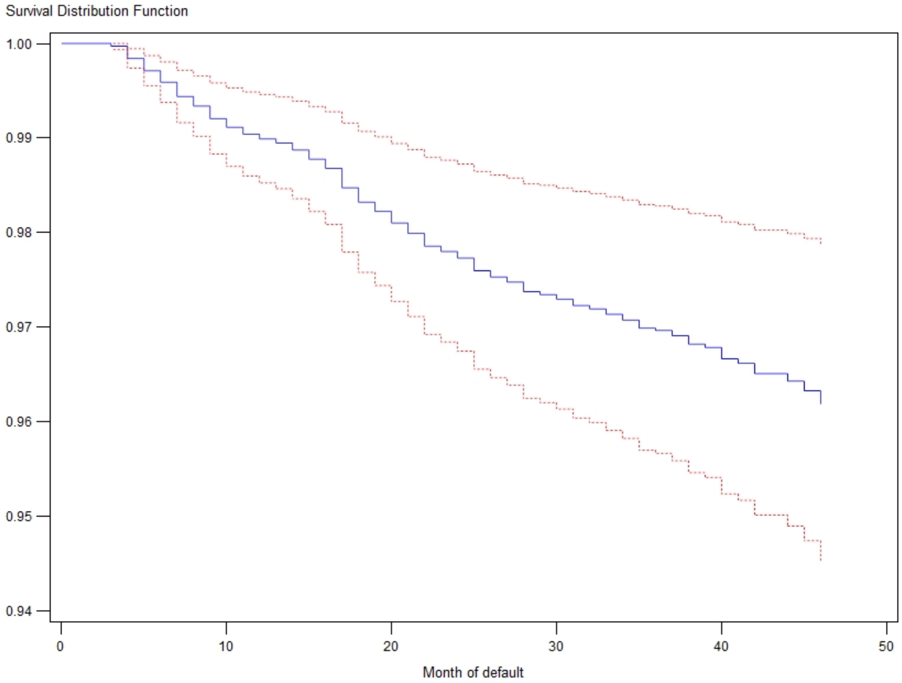

Figure 1.21: Survival function for the first contract in the sample as an example on the progressive time sample

Table 1.20: Analysis of the selected predictors of the repayment survival model on the progressive time sample

Predictor	Degrees of Freedom	Wald Chi-Square	Pr > ChiSq
Sex	1	8.0746	0.0045
Marital status	1	38.7425	< .0001
Employment status	2	31.1130	< .0001
Employer	2	25.2562	< .0001
Repayment type	1	16.0285	< .0001
Credit card	1	6.9649	0.0083
Telephone private	1	21.9056	< .0001
Distribution channel	2	6.1591	0.0460

Table 1.21: Analysis of maximum likelihood estimates of the repayment survival model on the progressive time sample

Parameter	Category	Estimate	Standard Error	Pr > ChiSq
Sex	1	0.32320	0.11374	0.0045
Marital status	1	−0.69728	0.11202	< .0001
Employment status	1	0.90381	0.16398	< .0001
Employment status	2	0.44249	0.12432	0.0004
Employer	1	−0.47280	0.12316	0.0001
Employer	2	−0.64566	0.15196	< .0001
Repayment type	1	−0.51152	0.12777	< .0001
Credit card	1	−0.94965	0.35984	0.0083
Telephone private	1	−0.51026	0.10902	< .0001
Distribution channel	1	0.05390	0.58230	0.9262
Distribution channel	2	0.72095	0.63303	0.2547

Table 1.22: Summary of the repayment survival model on the progressive time sample

Summary	2002–2003	2004–2005
Gini	0.44	0.43
Lift 10%	2.71	2.42

Table 1.23: Analysis of correlations of the repayment survival model on the progressive time sample

Par.	Cat.	Sex	M.st.	E.st.	E.st.	Emp.	Emp.	Rep.	C.c.	T.p.	D.ch.	D.ch.
Sex	1	1.00	−0.08	0.04	0.04	0.12	0.00	0.04	−0.01	0.04	−0.02	−0.01
M.st.	1	−0.08	1.00	0.08	0.10	0.01	0.03	0.04	0.02	−0.15	0.00	0.01
E.st.	1	0.04	0.08	1.00	0.51	0.18	0.11	−0.04	0.04	0.00	0.04	0.02
E.st.	2	0.04	0.10	0.51	1.00	0.09	0.13	−0.03	0.03	−0.01	0.02	0.01
Emp.	1	0.12	0.01	0.18	0.09	1.00	0.30	0.02	−0.01	−0.02	−0.02	−0.01
Emp.	2	0.00	0.03	0.11	0.13	0.30	1.00	0.06	−0.02	0.01	−0.03	−0.01
Rep.	1	0.04	0.04	−0.04	−0.03	0.02	0.06	1.00	−0.06	−0.01	−0.05	0.02
C.c.	1	−0.01	0.02	0.04	0.03	−0.01	−0.02	−0.06	1.00	−0.01	0.01	0.01
T.p.	1	0.04	−0.15	0.00	−0.01	−0.02	0.01	−0.01	−0.01	1.00	−0.03	−0.02
D.ch.	1	−0.02	0.00	0.04	0.02	−0.02	−0.03	−0.05	0.01	−0.03	1.00	0.91
D.ch.	2	−0.01	0.01	0.02	0.01	−0.01	−0.01	0.02	0.01	−0.02	0.91	1.00

1.4.3 Comparison of Results

Now I compare the results of the Gini coefficient, distribution curves and lift curves of the standard approach using the logistic regression and the proposed repayment survival model using the Cox model on the corresponding data samples.

Random Sample

On the random sample we can see from table 1.24 that the two models have a very similar performance in the Gini coefficient on the development (training) and comparison (also testing or validation) sample, with the Cox model being a little more stable. As for the 10% lift, we can see that the Cox model performs better on the training sample and worse on the testing sample. From both the distribution and lift curves in figures 1.22 and 1.23 we see that the performance of the models is slightly different but in general I'd conclude the models as similar.

Table 1.24: Comparison of models on the random sample

Summary	Development	Comparison
Logistic regression Gini	0.51	0.39
Cox model Gini	0.50	0.40
Logistic regression lift 10%	2.73	2.78
Cox model lift 10%	2.96	2.32

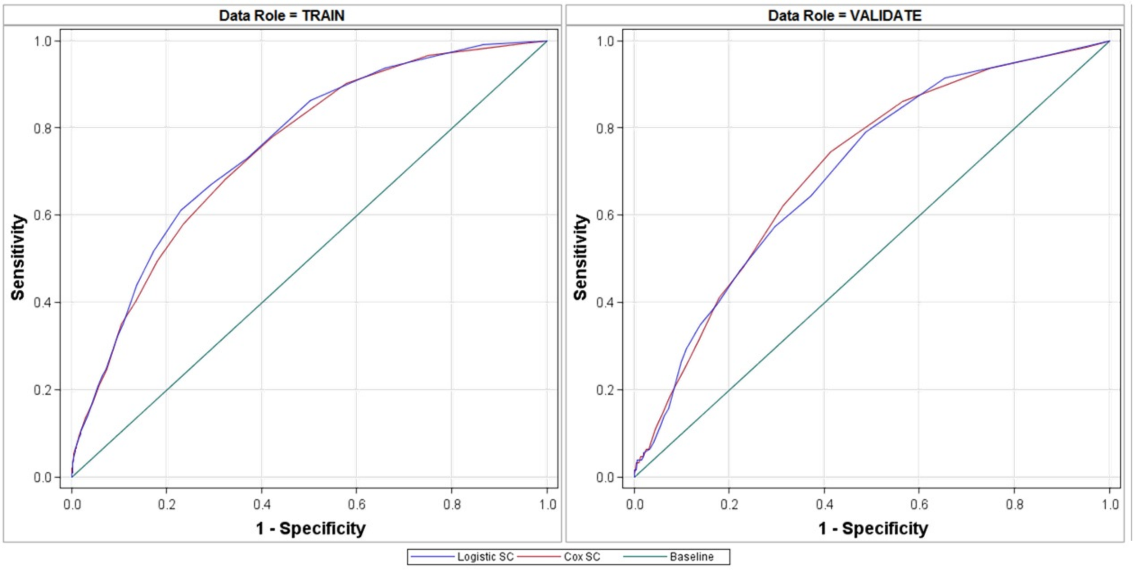

Figure 1.22: Distribution curve comparison on the random sample

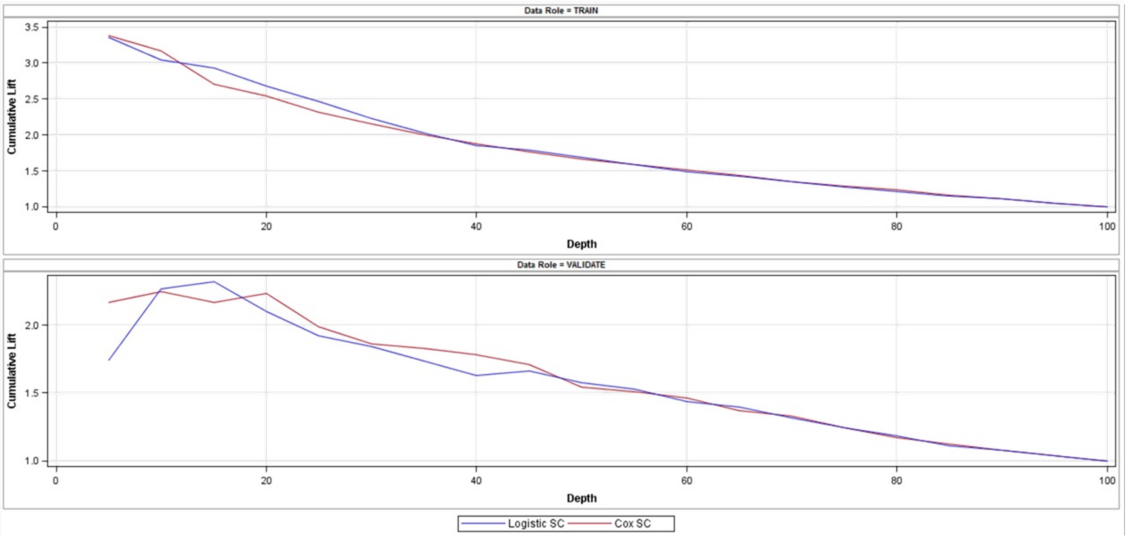

Figure 1.23: Lift curve comparison on the random sample

Progressive Time Sample

For the progressive time sample we can see from table 1.25 that whereas the Cox model is slightly more conservative on the 2002–2003 sample, it notably outperforms the logistic regression on the 2004–2005 sample. This is confirmed also on the distribution and lift curves in figures 1.24 and 1.25.

Table 1.25: Comparison of models on the progressive time sample

Summary	2002–2003	2004–2005
Logistic regression Gini	0.45	0.38
Cox model Gini	0.44	0.43
Logistic regression lift 10%	3.15	1.83
Cox model lift 10%	2.71	2.42

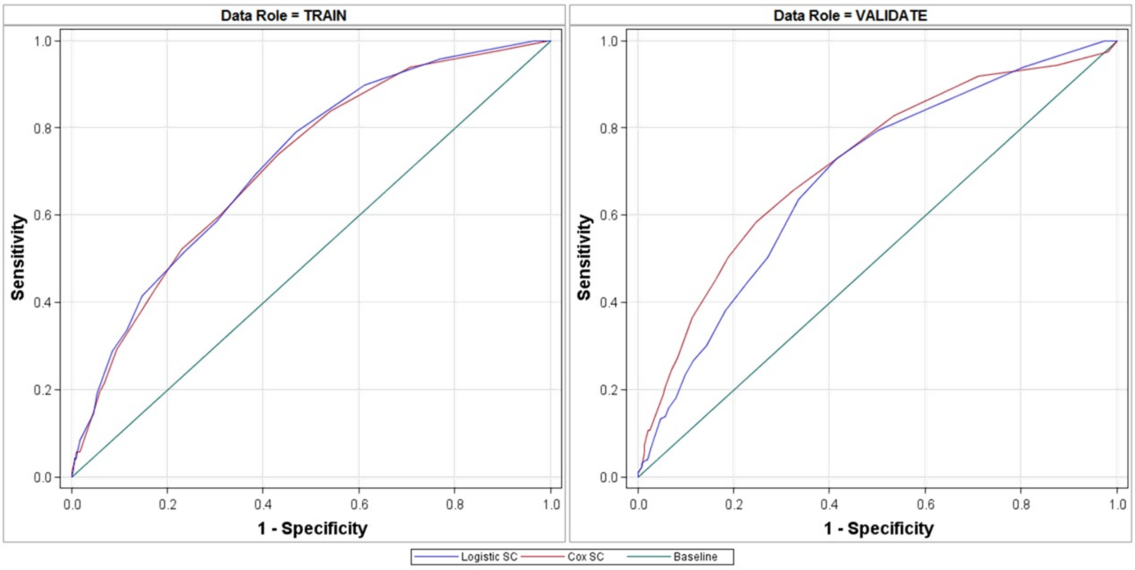

Figure 1.24: Distribution curve comparison on the progressive time sample

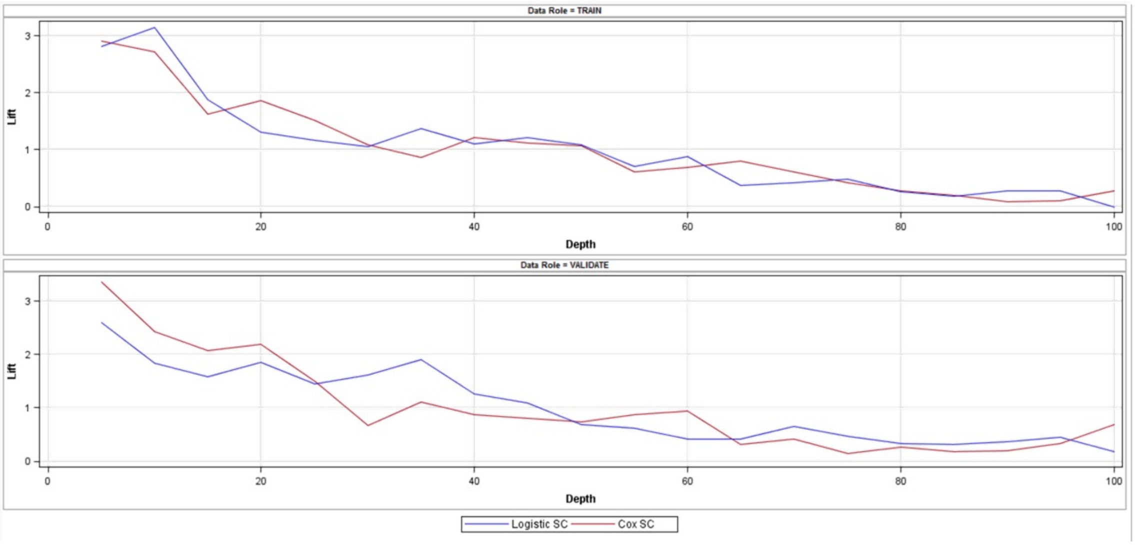

Figure 1.25: Lift curve comparison on the progressive time sample

1.5 Conclusions

In this chapter I dealt with the probability of default modeling and aimed to set the new performance criteria focusing on the predictive power of the models and compare the standard logistic regression model with the alternative of the survival-based Cox model on the real sample of Czech banking data.

I used the data and the Cox model from (Pazdera et al., 2009), I adjusted the data to contain a relevant subset of observations for modeling and divided the sample into the standard random development and validation part, as well as performed the newly proposed division into the development and ex-ante validation sample specially designed to measure the predictive power of the models. Then I implemented the logistic regression model alongside with the Cox model and compared their precision and predictive power using the Gini coefficient and lift characteristics.

As we can see from section 1.4.3, both models have similar performance on the random training and testing sample. This is in line with the existing research, e.g. (Stepanova and Thomas, 2002), (Cao et al., 2009) or (Bellotti and Crook, 2009), and I showed that the regional specific Czech fix-term unsecured loan banking data make no exception.

However, if compared by the new performance criteria measuring the predictive power of the model, the Cox model outperforms the logistic regression model in the progressive time sample comparison, and thus shows a better predictive power in extrapolating the last observable default vintages. This is a new result that can be beneficial for further research in this topic, as well as for banks and credit companies.

Here the survival analysis methodology was chosen on purpose, mainly because of the way it can cope with time-censored data. Therefore, it can incorporate the most recent observations into the model, and potentially improve its predictive power for the future.

Finally, the Cox model gives us the baseline and survival function for all times. This can be analyzed further (e.g. smoothed by a polynomial interpolation in figure 1.26) and used for a variety of additional analytical tasks including the calculation of expected profitability introduced in chapter 3.

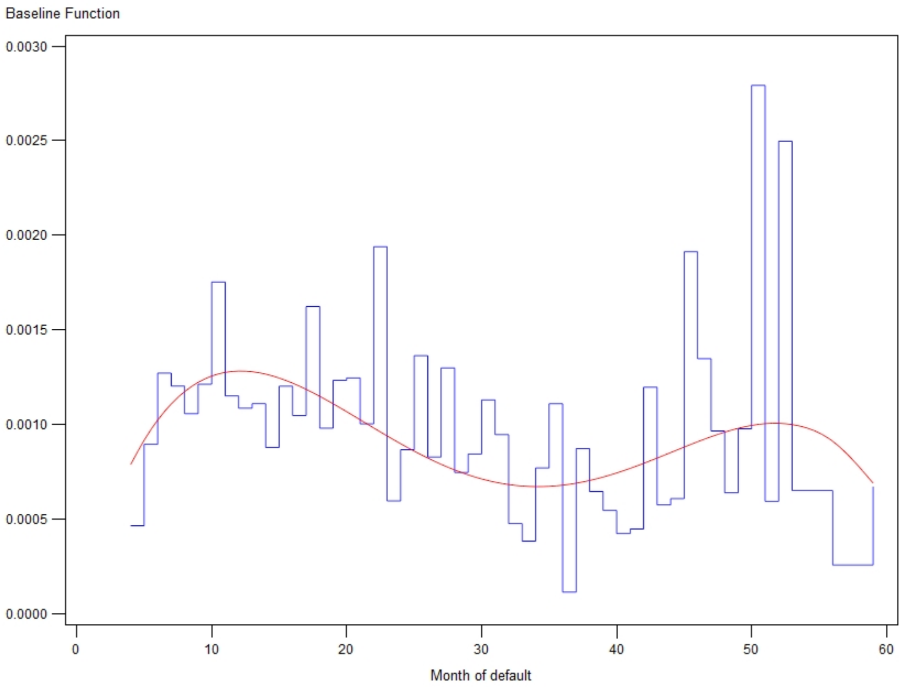

Figure 1.26: Baseline function with the weighted polynomial interpolation for the Cox model on the random sample

Chapter 2

Loss Given Default Modeling

In this chapter I concentrate on the loss given default (LGD) modeling and apply similar logic as in the first chapter, mainly by introducing the new performance criteria for the predictive power measuring and comparing the standard approaches using the linear and logistic regressions with two Cox-based models on the real Czech banking data.

This is a well discussed topic, especially after the (Basel II, 2001) regulation, and many comparisons of methods are presented in the literature, see e.g. (Gupton et al., 2002) or (Kim and Kim, 2006).

In my master thesis (Rychnovský, 2009) I focused on the problem of the recent censored observations and introduced several new methods including two models based on the survival analysis. I compared the methods with the standard approaches of the linear and logistic regression on the real banking data. Since the comparison was done on the development sample only, the standard linear and logistic regression methods performed better than the new proposed approaches.

Therefore, in my further doctoral research I concentrated on comparing of the predictive power of these methods on a time-censored data sample (similarly to the progressive time sample of chapter 1). Also I add the Full repayment survival model as a new model to the comparison. For this comparison I use the original data from (Rychnovský, 2009) and re-structure them for the predictive task (shorter time horizon, time-censored sample etc.). The comparison is then based on the modified coefficient of determination (originally assigned as R^2) introduced in (Rychnovský, 2009).

In this chapter I use the results published in (Witzany et al., 2012) and (Rychnovský, 2015).

2.1 Recovery Rate Models

In this section I would like to show two possible applications of the survival analysis theory for LGD modeling – the *full repayment survival model* and the *currency unit survival model.*[1]

First I introduce a little about the recovery process. In reality when a default occurs (e.g. 90 days past due), a collection process will start. During this period clients are contacted by the loan company, and according to their capacities, they either pay nothing, or repay one or more payments covering a part or the whole owed exposure. These payments are usually known as *recoveries.* If we now denote the recovered part of the whole exposure as *recovery rate* (RR), we can express the loss given default as

$$\mathrm{LGD} = 1 - \mathrm{RR}\,. \tag{2.1}$$

Then the problem transforms into modeling the expected clients' recoveries in time, denote it $\mathrm{RR}(t)$ conditional to clients' characteristics.

Before starting the explanation about the models, I briefly explain the structure of any available development sample. For the LGD modeling we can only use the accounts, that have actually defaulted, and we can observe the recovery process for some time. And here comes the biggest problem of the LGD or RR modeling – the longer we want to observe the recovery process, the older data we have to use. Therefore, if we want to model for example the recovery after 3 years, we can only use the 3 years old defaults, since the fresh data is not fully observable (we can call it censored). And it is mainly the dealing with the censored data which leads to the idea of using the theory of survival analysis.

2.1.1 Standard Approaches

Standard approaches usually consider an univariate target variable such as the recovery rate after a fixed time interval t, and model it conditional to the characteristics of the client. The most straightforward way one can think of is using the linear regression. Besides the already mentioned disadvantage of the old data sample, this approach also struggles with the limitations for the target variable (since the RR or LGD are numbers between 0 and 1 or the recovery amount between 0 and the full exposure at default, denoted as EAD) and usually very few positive recoveries (target more than 0).

Alternatively, one can transform the target variable into a binary target (e.g. 1 for full recovery and 0 otherwise) and use the logistic regression to estimate the

[1] The currency unit survival model has been already proposed in (Rychnovský, 2009).

probability of recovery. The weakness of this approach is the fact that it doesn't use the full information about the partial recoveries (as all or some of them are assigned to zero). This can be partially fixed by setting a certain positive threshold of recovery rate and considering the weighted average of the recoveries for the case of both target options. Then $\mathrm{RR}(t)$ can be computed as

$$\mathrm{RR}(t) = \pi_t \, \mathrm{RR}_1(t) + (1 - \pi_t) \, \mathrm{RR}_0(t),$$

where π_t is the probability that the case will be recovered over the threshold, $\mathrm{RR}_1(t)$ is the average recovery rate of the recovered accounts and $\mathrm{RR}_0(t)$ is the average recovery rate of other accounts. More information about using of these methods in LGD modeling can be found in (Rychnovský, 2009).

2.1.2 Full Repayment Survival Model

The first approach based on the survival analysis theory (shortly described in section 1.1.2) considers observing the clients and measuring their time until they fully repay or repay over some given threshold (e.g. 80%). Therefore, using the terminology of survival analysis, we can say that the subjects are the clients and exit is defined as a repayment of the exposure over the threshold. It is assumed that every client will repay eventually and every observation without a full recovery is considered as censored in time. Moreover, it is assumed that the baseline hazard function is the same for all accounts. This assumption is based on the practice, where there is usually a similar collection process for the accounts (e.g. calling, then letter, then court etc.), and therefore the shapes of the recovery curves follow similar trends. Finally, as the censoring is only caused by the shorter observation period since the default, we can expect the censoring to be non-informative, i.e. unrelated with the default event.

If we first assume the full repayment (i.e. the threshold set to 100%), using the Cox model we get for every t the survivor function $S(t)$, which stands for the probability that the client will "survive" time t, i.e. will not fully repay until time t. Therefore, for the expected recovery rate $\mathrm{RR}(t)$ we can use the probability of the full repayment in time t and get

$$\mathrm{RR}(t) = 1 - S(t)$$

and for $\mathrm{LGD}(t)$ directly

$$\mathrm{LGD}(t) = S(t).$$

Similarly as for the case of the logistic regression introduced above, this model does not take into account the fact, that for some observations the exposure was partially recovered. Again, this can be partially fixed by considering the weighted average of the recoveries for the case of both target options. Then $\mathrm{RR}(t)$ can be computed as

$$\mathrm{RR}(t) = (1 - S(t))\,\mathrm{RR}_1(t) + S(t)\,\mathrm{RR}_0(t),$$

where $\mathrm{RR}_1(t)$ is the average recovery rate of the accounts recovered over the threshold, and $\mathrm{RR}_0(t)$ is the average recovery rate of other accounts. These partial recoveries are also the motivation of the second approach.

2.1.3 Currency Unit Survival Model

The second survival-based model understands every currency unit (or alternatively percentage) of the owed exposure at default as a subject, and its repayment as the exit. Thus, every client is represented by a set of units and their repayment times, fully describing the client's repayment history. Then, for the repayed currency units (or percentage) we observe an exit, and the rest are censored to the maximal time of observation.

Again, it is assumed that every unit will be once repayed, the baseline hazard functions are the same for all the units and the censoring is non-informative. Then using the survival analysis theory we can understand the survivor function $S(t)$ as the probability that a currency unit with some client's characteristics will not be payed until time t. Then the expected recovered proportion of the client's exposure in time t can be expressed as

$$\mathrm{RR}(t) = 1 - S(t)$$

and thus

$$\mathrm{LGD}(t) = S(t).$$

For this model, we can use either the individual currency units (such as 1 EUR) or the proportions of the owed exposures (such as 1%) as subjects. Whereas the second approach is more balanced on the client level, the first currency unit approach can put more weight to high exposure cases (which can be preferred by financial companies in practice).

2.2 Real Data for Modeling

In this section I shortly describe the data used for modeling and and explain the structure of the development and comparison samples. The data transformation and modeling is performed in SAS.

2.2.1 Data Overview

The data sample for this research is taken from (Rychnovský, 2009) and was kindly provided by Česká spořitelna, a.s. It is a database of 4,000 defaulted accounts with the following properties:

- It is a homogenous product from a non-secured retail business.
- There is some unspecified definition of default identical for the whole portfolio.
- There was an unified collection process for all the accounts.
- For each account there is the history of net recoveries – all discounted by time and collection costs.
- For each account there is a set of characteristics $\boldsymbol{x} = (x^1, \ldots, x^8)'$, which are assumed to have predictive power. Here x^4 is categorical and the rest are numerical (presumably after some transformation).[2]

The modeling is done on the 26 month horizon. For the accounts where the full collection history up to 26 months can be observed, the cumulative recoveries are summarized in figure 2.1. As we can see from the graph, the recoveries of some accounts are greater than 1, which means that more than the owed amount was actually recovered in the collection process (due to paid collection fees etc.). On the other hand, some recoveries are negative, which could have been caused by some additional collection costs assigned to cases with small or no recoveries.

All the predictors are then examined to be used for the model. In figure 2.2 there is an example of the first numerical predictor that has been cut into 11 bins according to their values to check the relation with the recovery rate. From this figure we can assume that the suggested use as the numerical predictor is not contra-intuitive.

2.2.2 Data Structure for Modeling and Comparison

First I show the structure of data provided in the given sample. There is a time interval of months coded as 162–220 (without a real month reference) with a classical triangle structure – for the first month 162 we can observe the full history of 58 months, whereas for the last month we have no time to observe.

[2] It is not the aim of this work to further categorize or optimize the set of predictors or interpret the results (which even couldn't be responsibly done since no information about the meaning of the predictors is provided). Therefore, all of the predictors are used for the models without any transformation. The numerical predictors are designed in the way that enables their direct use in the model – this can be illustrated on the first predictor in figure 2.2.

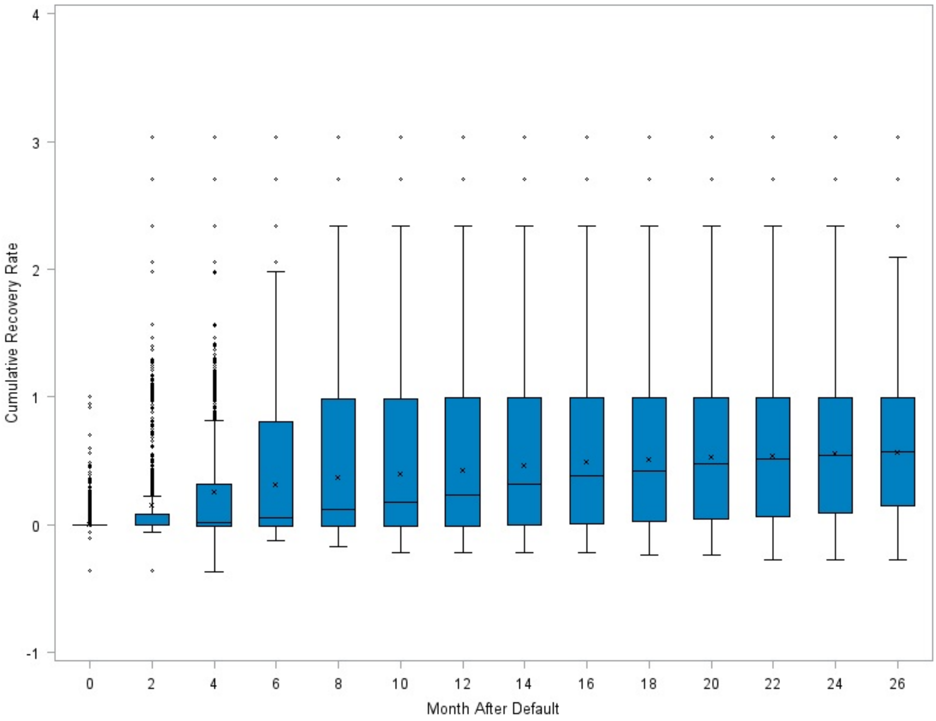

Figure 2.1: Dependence of recovery on month after default

Due to the sample size of our data, the 26 months recovery have been chosen for comparison. Therefore, we can only observe the full recovery for the vintages 162–194. See figure 2.3 for reference.

As I want to compare the predictive power of the models, I have to shorten the period for modeling and leave some data vintages for out-of-sample ex-ante prediction performance testing. Similar as in the progressive time sample construction in the first chapter I assume that only the events happening before month 194 are observable for the model development.

With that assumption I have the full history of vintages 162–168 (I denote this area as D1) and censored data of vintages 169–194 (I denote this area as D2). The rest of the vintages 169–194 with observations after month 194 (denoted as area D3) is left for comparison of the models. For more illustration about the data structure I again refer to figure 2.3.

2.3 Goodness of Fit Definition

Now I describe a measure that will be used for the comparison of the models. Since in this case all the binary variables are auxiliary and the original target variable recovery rate is real (not binary), I don't use the Gini coefficient or lift characteristics

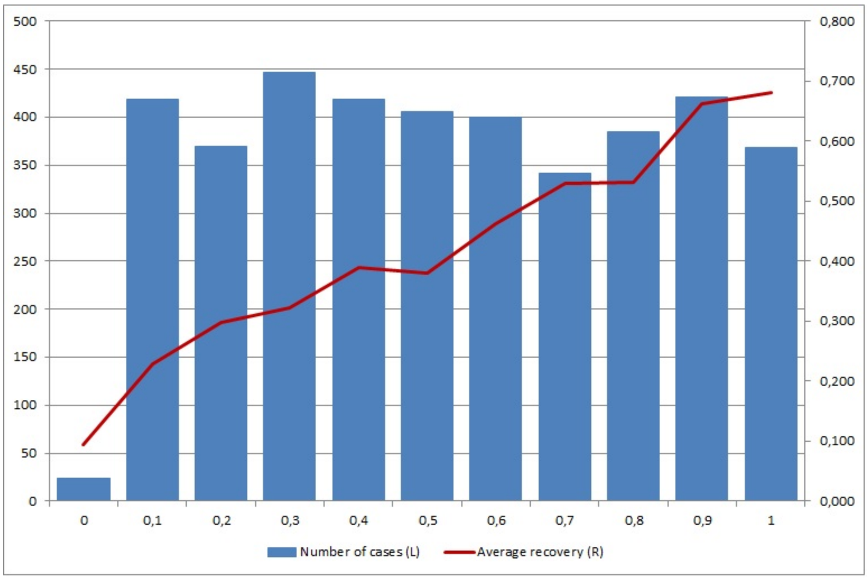

Figure 2.2: The number of cases and average recovery rate after 26 months for 11 bins of the first numerical characteristic

introduced in chapter 1. Instead to compare the results I use the modified coefficient of determination (MCD) introduced as R^2 in (Rychnovský, 2009).

2.3.1 Modified Coefficient of Determination

Each of the introduced models is developed on the data sample of D1 and D2 (D1 for the linear and logistic regression and D1+D2 for the survival models), and the recovery after 26 months is estimated for each account in vintages 162–194. Then I can compare the estimates with the real values using the weighted MCD defined as

$$\mathrm{MCD} = 1 - \frac{\sum_{i=1}^{n} w_i \big(\mathrm{RR}(26)_i - \widehat{\mathrm{RR}(26)}_i \big)^2}{\sum_{i=1}^{n} w_i \big(\mathrm{RR}(26)_i - \overline{\mathrm{RR}(26)}_{pool} \big)^2}, \tag{2.2}$$

where

$$\overline{\mathrm{RR}(26)}_{pool} = \sum_{i=1}^{n} w_i \, \mathrm{RR}(26)_i$$

is the weighted average of $\mathrm{RR}(26)_i$ and

$$w_i = \frac{\mathrm{EAD}_i}{\sum_{k=1}^{n} \mathrm{EAD}_k}$$

are the weights corresponding to the exposure at defaults (EAD) for individual cases; EAD_i is the unpaid principle of the i-th client at the moment of default.

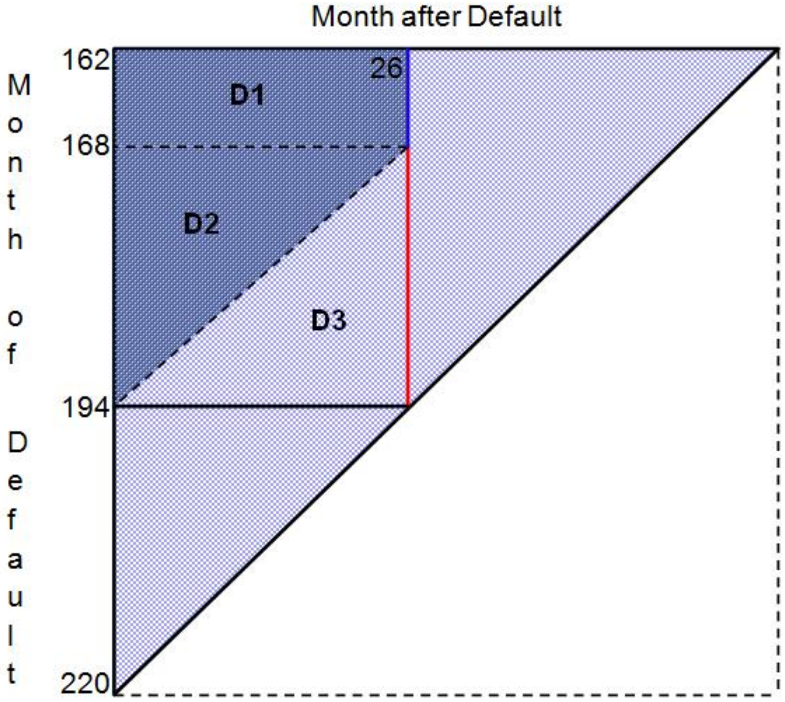

Figure 2.3: Triangle data and its structure for model development

The above defined MCD measure is usually in the interval $[0, 1]$ for the models that are performing at least as well as the weighted average constant $\overline{\mathrm{RR}(26)}_{pool}$, however it can also be negative if the model performs even worse than the $\overline{\mathrm{RR}(26)}_{pool}$ constant. The higher value of MCD, the more precise the model is on the given sample.

In this comparison I decided to use the weighted characteristic MCD to best simulate the needs of the credit companies that are more focused on the higher volume cases.[3] Then for each model I compute the weighted MCD separately on the area of known (development) recoveries D1, on the area of future (out-of-sample) recoveries D3 and the whole available data together D1+D3.

2.4 Results

In this section I compare the results of the respective models. Since the MCD is a weighed characteristic, I use the weighted linear and logistic regression models.

[3] The higher exposure at default the higher loss the company can suffer from. Also the overall loss for the company is calculated as the sum of all losses, i.e. the weighted average of LGDs multiplied by the total exposure at default.

2.4.1 Linear Regression Model

For the linear regression model (weighted by the exposures) only the data of D1 is used for model development. Then the model in the form

$$\mathrm{RR}(26) = \beta_0 + \boldsymbol{\beta}'\boldsymbol{x},$$

where β_0 is the intercept of the model and $\boldsymbol{\beta}$ is the vector of parameter estimates, is used to estimate the recoveries in the groups D1 and D3. The values of the MCD is summarized in table 2.1. We can see that the MCD is highest for the development sample and then rapidly decreases for the prediction.

Table 2.1: Characteristics of the linear regression model

D1	N	MCD	D3	N	MCD	D1+D3	N	MCD
	600	0.1621		1735	0.0064		2335	0.0558

2.4.2 Logistic Regression Model

The next tested approach for the recovery modeling is the weighted logistic regression model. The accounts from the development area D1 are divided into two groups according to their recovery rate (e.g. less than 0.1 and more then 0.1), and the probability $\pi(\boldsymbol{x})$ that an account with characteristics $\boldsymbol{x}$ will belong to the later category is modeled using the weighted logistic regression. Then for all accounts from D1 and D3 I compute the estimated recovery as

$$\mathrm{RR}(\boldsymbol{x}) = \pi(\boldsymbol{x})\,\mathrm{RR}_1 + (1 - \pi(\boldsymbol{x}))\,\mathrm{RR}_0,$$

where RR_0 is the weighted average recovery in the first group and RR_1 is the weighted average recovery in the second group (both computed on the D1 sample).

Since it is not clear what threshold would be optimal for using, I decided to try 11 options from 0 to 1 recovery rate values. Altogether, 11 models are computed with different values of the thresholds dividing the accounts to the recovered and non-recovered categories (see table 2.2). The results of all models can be found in table 2.3. We can see that the bound 0.1 has the best performance on D1+D3.

2.4.3 Full Repayment Survival Model

For the full repayment survival model introduced in section 2.1.2 I use all the accounts from D1 and D2 (i.e. censored by the observation time of 194) for development. Again I look for the optimal threshold to divide the accounts into two

Table 2.2: Number of cases for the logistic regression model

Threshold	Total (N)	Number of Non-Recovered (0)	Number of Recovered (1)
0	600	129	471
0.1	600	162	438
0.2	600	185	415
0.3	600	211	389
0.4	600	229	371
0.5	600	345	355
0.6	600	267	333
0.7	600	274	326
0.8	600	299	301
0.9	600	340	260
1	600	426	174

Table 2.3: Characteristics of the logistic regression models

Threshold	D1	N	MCD	D3	N	MCD	D1+D3	N	MCD
0		600	0.1249		1735	0.0205		2335	0.0562
0.1		600	0.1423		1735	0.0356		2335	0.0718
0.2		600	0.1510		1735	0.0116		2335	0.0566
0.3		600	0.1510		1735	-0.0253		2335	0.0297
0.4		600	0.1591		1735	-0.0055		2335	0.0463
0.5		600	0.1529		1735	0.0038		2335	0.0514
0.6		600	0.1512		1735	0.0117		2335	0.0567
0.7		600	0.1490		1735	0.0040		2335	0.0505
0.8		600	0.1431		1735	0.0216		2335	0.0618
0.9		600	0.1261		1735	-0.0008		2335	0.0410
1		600	0.0344		1735	0.0113		2335	0.0256

groups according to their recovery rate (e.g. less than 0.1 and more than 0.1), and the probability that an account will belong to the later category is understood as $1 - S(\boldsymbol{x}, 26)$, where $S(\boldsymbol{x}, 26)$ is the value of the survival function of an account with characteristics $\boldsymbol{x}$ in time 26. Then for all accounts from D1 and D3 I compute the estimated recovery as

$$\mathrm{RR}(\boldsymbol{x}) = (1 - S(\boldsymbol{x}, 26))\,\mathrm{RR}_1 + S(\boldsymbol{x}, 26)\,\mathrm{RR}_0,$$

where RR_0 is the weighted average recovery after 26 months in the first group and RR_1 is the weighted average recovery after 26 months in the second group (both computed on the D1 sample) – i.e. the same values as for the logistic regression model.

Again, 11 models are computed with different values of thresholds dividing the accounts to the recovered and non-recovered categories. The results of all models can be found in table 2.4. Same as for the logistic model, the bound 0.1 has the best performance on D1+D3. Moreover, in figure 2.4 we can see an example of the survival function for the first account in the selection, and in figure 2.5 there is the estimated baseline hazard function of the corresponding model.[4]

Table 2.4: Characteristics of the full repayment survival model

Threshold	D1	*N*	MCD	D3	*N*	MCD	D1+D3	*N*	MCD
0		600	0.0747		1735	0.0606		2335	0.0721
0.1		600	0.1012		1735	0.1355		2335	0.1336
0.2		600	0.0972		1735	0.1334		2335	0.1310
0.3		600	0.0925		1735	0.1091		2335	0.1121
0.4		600	0.0983		1735	0.0988		2335	0.1062
0.5		600	0.0922		1735	0.1024		2335	0.1071
0.6		600	0.0872		1735	0.1021		2335	0.1056
0.7		600	0.0874		1735	0.0994		2335	0.1037
0.8		600	0.0733		1735	0.0885		2335	0.0921
0.9		600	0.0668		1735	0.0792		2335	0.0836
1		600	0.0324		1735	0.0248		2335	0.0349

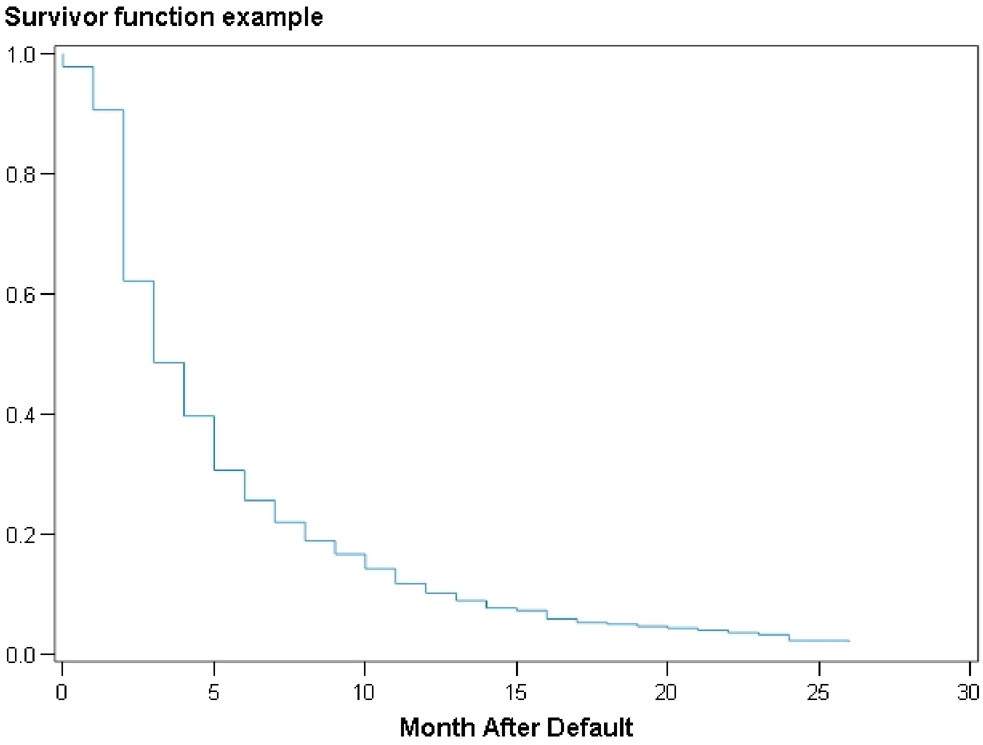

Figure 2.4: Estimated survivor function for the first account in the selection in the full repayment survival model

[4] Here we see a peak at 24 months that could be potentially connected to some special collection action.

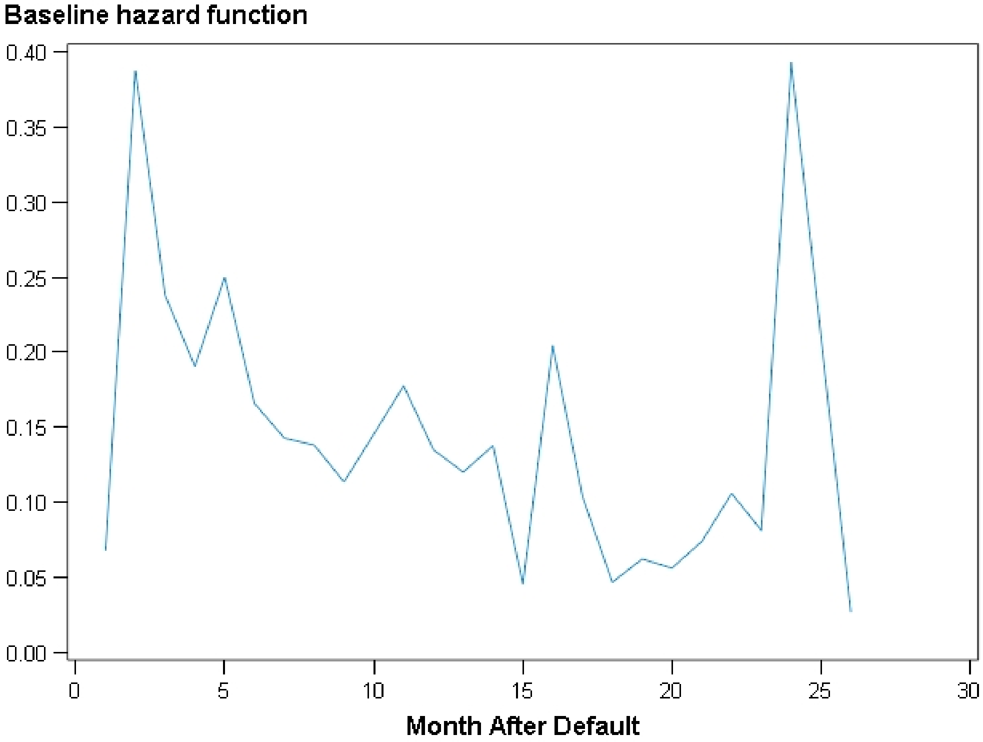

Figure 2.5: Estimated baseline hazard function in the full repayment survival model

2.4.4 Currency Unit Survival Model

The last tested model is the currency unit survival model introduced in section 2.1.3. Again, this model is developed on the full data of D1 and D2. In this model I observe the life cycle of every 100 CZK of the exposure as a subject. Then the survivor function $S(\boldsymbol{x}, 26)$ is interpreted as the probability that a particular amount of money (100 CZK) of an account with characteristics $\boldsymbol{x}$ will survive 26 months. Therefore, $1 - S(\boldsymbol{x}, 26)$ is the expected recovery rate after 26 months of such account with characteristics $\boldsymbol{x}$.

Again the MCD can be found in table 2.5. In figure 2.6 we can see an example of the survivor function for the first account in the selection and in figure 2.7 there is the estimated baseline hazard function. Even here we can see a very good performance on the D1+D3 area.

Table 2.5: Characteristics of the currency unit survival model

D1	N	MCD	D3	N	MCD	D1+D3	N	MCD
	600	0.0987		1735	0.1202		2335	0.1218

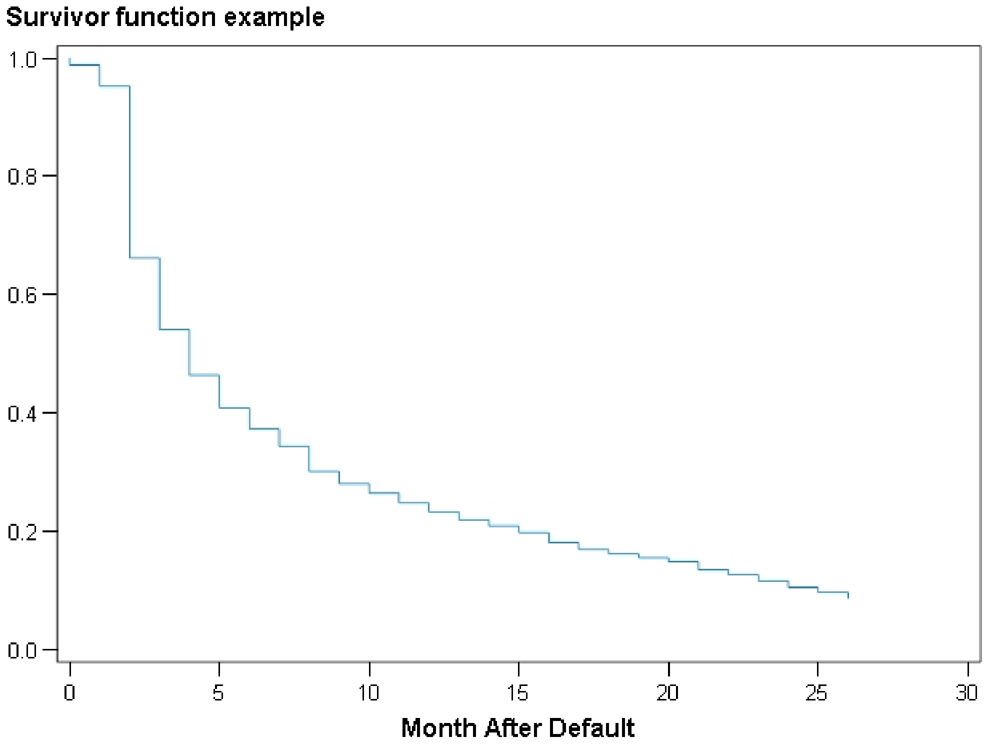

Figure 2.6: Estimated survivor function for the first account in the selection in the currency unit survival model

2.4.5 Comparison of Results

In table 2.6 we can find the comparison of the four tested models. We can see that for the estimation on the development data D1 the standard approaches of the linear and logistic regression models reach better results than the survival models. Here the linear regression performs a little better than the logistic regression (even when we take into account the best model on D1 with MCD = 0.1591).

However, when we compare the predictive power of the models in the ex-ante sample D3, we see that the additional information contained in the censored area D2 brought substantially better results to the survival models. Particularly the full repayment survival model with the bound of 0.1 seems very useful for this type of data.

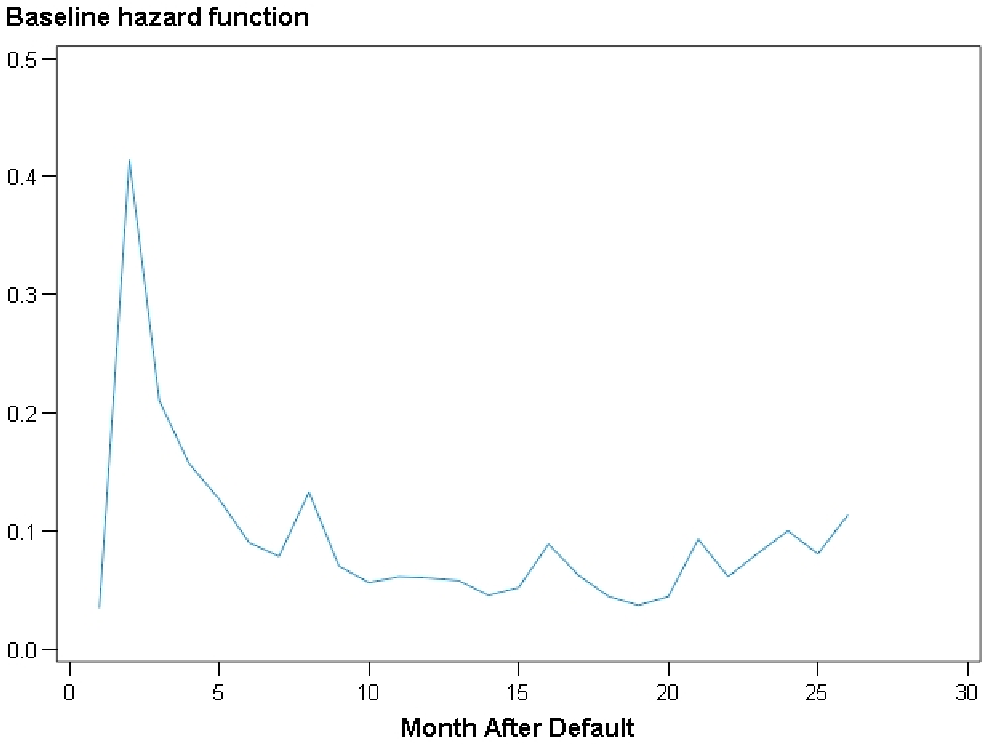

Figure 2.7: Estimated baseline hazard function in the currency unit survival model

Table 2.6: Comparison of the tested models

Model	MCD(D1)	MCD(D3)	MCD(D1+D3)
Linear regression	0.1621	0.0064	0.0558
Logistic regression (0.1)	0.1423	0.0356	0.0718
Full repayment survival model (0.1)	0.1012	0.1355	0.1336
Currency unit survival model	0.0987	0.1202	0.1218

2.5 Conclusions

In this chapter I aimed to introduce the new performance criteria for measuring the predictive power of loss given default models and compare the standard approaches using the linear and logistic regressions with two Cox-based models on the real Czech banking data.

For this chapter I used the standard linear and logistic models, as well as the Currency unit survival model introduced already in (Rychnovský, 2009) and added a newly proposed Full repayment survival model to the set. Then I re-structured the original real data sample from (Rychnovský, 2009), shortened the time horizon for prediction and divided the sample into the development sample and a time-censored ex-ante sample for prediction. Finally, I applied all the models to this data and compared their predictive power.

When compared on the development sample, the linear and logistic regression perform better than the survival analysis models – which is in line with the results from (Rychnovský, 2009). However when comparing the predictive power of the models on the time-censored sample, the new approaches clearly outperform the standard models in the terms the used goodness of fit measure.

Therefore, I believe that there is a good potential for further research and practical application in banks and financial institutions creating their own LGD models to decrease the capital requirement.

Moreover, both survival models give us formulas to compute the expected recovery for any time t within the observed period and thus a flexible information about the whole payment perspective. Seeing the whole recovery performance of the case in time can be used in practice to better understand the collection process for various accounts.

In this work I took the nonparametric Cox model as an example of widely used survival models; however, a parametric alternative (such as the Accelerated Failure Time (AFT) model) can be used instead. For more information about parametric models one can refer to (Kalbfleisch et al., 1980).

Chapter 3

Profitability Modeling

In this chapter I aim to use the models described in the previous chapters and construct a new comprehensive underwriting model that would be based on an estimation of loan profitability instead of the standard evaluation of the riskiness of the client. This idea is based mainly on (Allen et al., 2004) and (Stein, 2005) and my experience from the financial sector.

Therefore, the aim of this chapter is to generalize and describe the existing approaches to profitability modeling and derive the formulas needed for their application, as well as to propose the survival analysis models from chapters 1 and 2 to provide the most relevant inputs for the model.

Furthermore, I propose several more revenue streams and allocated costs to be incorporated in the model and increase the precision of the expected profitability estimation. Finally, I use the data set and results from chapter 1 to simulate the differences of using these profitability models compared to the standard probability of default model.

3.1 Introduction

In chapter 1 I discussed some models that aim to estimate the probability of default of applicants and compute the score of each client in order to evaluate his or her riskiness. Then the clients are often approved or rejected based on this score. However, is the probability of default the key criteria for approval?

Imagine a situation when there are two loan applicants with the following characteristics (see figure 3.1). Client A has the estimated probability of default at 5% and is applying for a 12 months loan with interest rate of 20%, whereas client B has the estimated probability of default also at 5% and is applying for a 24 months loan with interest rate of 30%. Then which client should the company prefer?

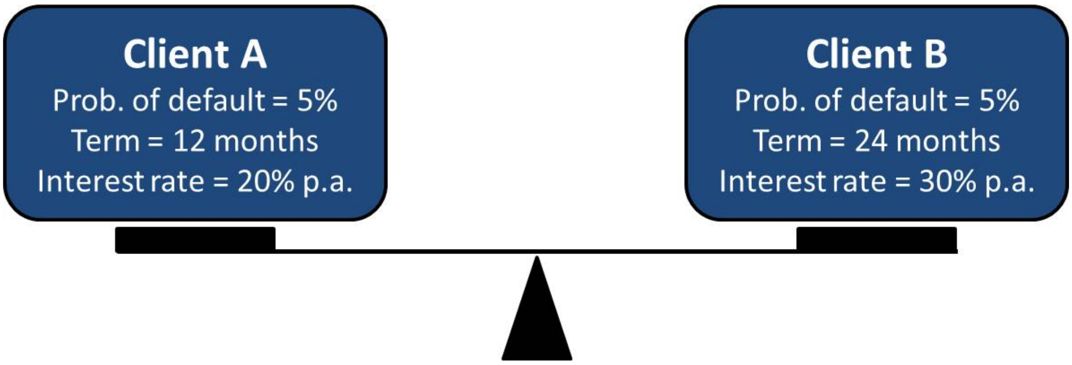

Figure 3.1: Example of two applicants with the same probability of default

Strictly from the probability of default point of view, the two clients are indifferent. However, from the information provided, it is likely that the second client would pay more interest and thus bring more profit to the company. Though, how would the situation change, if the first client is willing to take an insurance product or if the funding cost is different for the two clients (e.g. happening on two different markets)?

This is a motivation for building a complex profitability model that incorporates several more clients' and business' inputs to calculate the expected profitability from providing a loan to an applicant. Such a model can also serve as an engine for calculation of a personalized interest rate offer (satisfying company's profitability requirement) based on client's characteristics.[1]

In the following sections, I introduce the key components of the profitability model together with some characteristics for profitability evaluation.

3.2 Model Components

First I explain a little bit about the loan installment structure and the connected formulas, and then I already focus on the profitability model, its components and the topics of discounting and funding.

[1] This is called dynamic scoring and it is in practice often done by pre-defined combinations of probability of default and a corresponding interest rate range. However, such approach often does not incorporate other factors.

3.2.1 Loan Instalment Structure

For the use of this chapter I assume a loan with a constant interest rate, fix term and fix regular payments (e.g. monthly or quarterly) at the end of each period (also called ordinary annuities). Even though my assumptions about the loan installment structure are somehow standard on the market and their formulas are available in common financial mathematical handbooks, I prefer to derive the formulas myself in the way that can be mostly replicated even though the installment structure varies for specific cases and corresponding formulas are not available. In practice it indeed often happens that some advanced loan characteristics are not standard and it is not possible to find such formulas in the textbooks.

I use the following denotations:

- T is the term of the loan,
- A is the ordinary annuity of the loan covering the interest and principal part,
- i is the interest rate corresponding to the time interval between two consecutive payments (e.g. one twelfth of an annual rate for monthly installments),
- I_t is the interest part of the t-th installment,
- P_t is the principal part of the t-th installment,
- U_t is the unpaid principal after t-th installment, thus U_0 is the initial loan value and $U_T = 0$,
- F is the fix amount of monthly fees connected with the loan servicing, independent of all other components.

Then I have $I_t + P_t = A$ for all $t \in \{1, 2, \dots, T\}$ and $U_t = \sum_{k=t+1}^{T} P_k$ for all $t \in \{0, 1, \dots, T-1\}$. Now I derive some more formulas important for the following sections.

First I need to prove that

$$P_t = A\frac{1}{(1+i)^{T-t+1}}, \forall t \in \{1, 2, \dots, T\}. \tag{3.1}$$

This corresponds to the fact that using the given interest rate i the future values of the principal parts of all installments are equal. This can be easily shown by induction:

1. For $t = T$ I take the equation $P_T + I_T = A$ and get

$$\begin{aligned} P_T &= A - I_T \\ &= A - iU_{T-1} \\ &= A - iP_T \\ &= A\frac{1}{1+i}. \end{aligned}$$

2. Now assume that

$$P_k = A\frac{1}{(1+i)^{T-k+1}}, \forall k \in \{t, t+1, \dots, T\},$$

and for P_{t-1} I get

$$\begin{aligned} P_{t-1} &= A - I_{t-1} \\ &= A - iU_{t-2} \\ &= A - i\sum_{k=t-1}^{T} P_k \\ &= A - iP_{t-1} - i\sum_{k=t}^{T} P_k \\ &= A - iP_{t-1} - i\sum_{k=t}^{T} A\frac{1}{(1+i)^{T-k+1}} \\ &= A - iP_{t-1} - A\frac{i}{(1+i)}\sum_{k=0}^{T-t}\left(\frac{1}{1+i}\right)^k \\ &= A - iP_{t-1} - A\frac{i}{(1+i)}\frac{1-\left(\frac{1}{1+i}\right)^{T-t+1}}{1-\left(\frac{1}{1+i}\right)} \\ &= A - iP_{t-1} - A\frac{i}{(1+i)}\frac{1-\left(\frac{1}{1+i}\right)^{T-t+1}}{\frac{i}{1+i}} \\ &= A - iP_{t-1} - A\left(1 - \frac{1}{(1+i)^{T-t+1}}\right) \\ &= A\frac{1}{(1+i)}\frac{1}{(1+i)^{T-t+1}} \\ &= A\frac{1}{(1+i)^{T-(t-1)+1}}. \end{aligned}$$

Then using the same logic I get

$$U_t = A\frac{(1+i)^{T-t} - 1}{i(1+i)^{T-t}}, \forall t \in \{0, 1, \dots, T\}, \tag{3.2}$$

because

$$\begin{aligned}
U_t &= \sum_{k=t+1}^{T} P_k \\
&= \sum_{k=t+1}^{T} A\frac{1}{(1+i)^{T-k+1}} \\
&= A\frac{1}{(1+i)}\sum_{k=0}^{T-t-1}\left(\frac{1}{1+i}\right)^k \\
&= A\frac{1}{(1+i)}\frac{1-\left(\frac{1}{1+i}\right)^{T-t}}{1-\left(\frac{1}{1+i}\right)} \\
&= A\frac{1}{(1+i)}\frac{1-\left(\frac{1}{1+i}\right)^{T-t}}{\frac{i}{1+i}} \\
&= A\frac{1}{i}\left(1-\frac{1}{(1+i)^{T-t}}\right) \\
&= A\frac{(1+i)^{T-t}-1}{i(1+i)^{T-t}}.
\end{aligned}$$

Also, because $I_t + P_t = A$, I get

$$I_t = A\frac{(1+i)^{T-t+1}-1}{(1+i)^{T-t+1}}, \forall t \in \{1, 2, \ldots, T\}. \tag{3.3}$$

Finally using Formula (3.2) for $t = 0$, I get the annuity formula based on the loan amount U_0 and interest rate i as

$$A = U_0\frac{i(1+i)^T}{(1+i)^T-1}. \tag{3.4}$$

In figure 3.2 we can see an example of a 10.000 CZK loan with 24 monthly payments, interest rate 18% p.a. and an additional fix monthly fee of 50 CZK. The total monthly payment of such loan is approx. 550 CZK and consists of the decreasing interest part I_t, the increasing principal part P_t and the fix fee F. On the right axis we can see the decreasing unpaid principal U_t.

3.2.2 Profit Model Components

For the purpose of this chapter I use the term *expected absolute profit* of a loan as the expected value of profit that can a loan providing institution get from providing

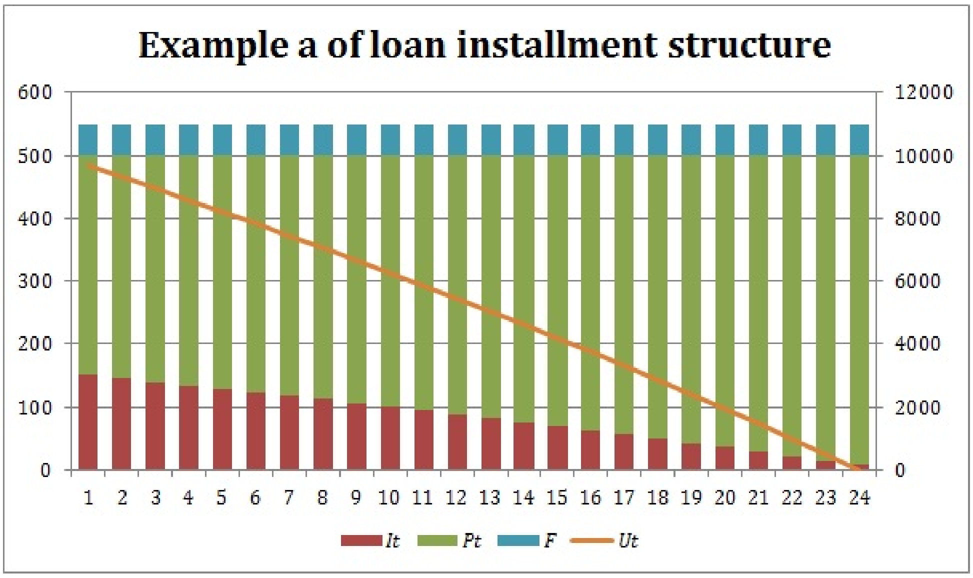

Figure 3.2: Example of a loan installment structure

a loan with specific characteristics (such as loan amount, term, interest rate, fee etc.) to a specific customer with given characteristics (such as application data, behavioral data, credit bureau data etc.). Only loans with a constant interest rate, fix term and fix regular payments at the end of each period are considered.

Then the expected absolute profit (EAP) can be simplified as the combination of the expected revenue (ER), expected loss (EL) and expected costs (EC) as

$$\mathrm{EAP} = \mathrm{ER} - \mathrm{EL} - \mathrm{EC}\,. \tag{3.5}$$

3.2.3 Discounting

Now I denote d the discount rate corresponding to the time interval between two consecutive payments. Then any future cash flow (including the expected cash flows) can be discounted to its present value using the discount factor as

$$\mathrm{PV} = \sum_{t=1}^{\infty} \frac{\mathrm{CF}_t}{(1+d)^t}, \tag{3.6}$$

where PV is the present value of the expected future cash flows CF_i in times i, $i \in \{1, 2, \dots\}$.

Even though this is an elementary financial mathematics theory, it is in reality quite an interesting topic to set a proper discount rate for this task. Especially when we consider the tight connection with the cost of funds and other characteristics introduced in later chapters. More information about discount rates and connected topics can be found in (Ho and Lee, 2004), (Homer and Sylla, 2011) and (Hull, 2009b).

3.2.4 Funding model

The last thing that should be set up before going into details about the profit components, is the funding model. In my model I assume that all the money issued as the loan amount are coming from two sources of investors:

1. Shareholders – put the capital into the company and require revenues in return. The minimum revenue corresponding to the time interval between two consecutive payments I denote i_S.
2. Funding partners – lend their money to the company and expect an interest i_F corresponding to the time interval between two consecutive payments.

For simplicity I assume that all the money is utilized (for the capital) or borrowed (for the funds) on a revolving basis.[2] This means that after each payment of the customer, the principal part of this payment is immediately repaid to the investors. Also if the event of default happens, the remaining principal is immediately repaid and accounted as a loss. Interest is paid regularly every period.

3.3 Expected Loss

For calculation of the expected loss I combine the methods discussed in chapters 1 and 2. If I denote π_t the estimated probability of default on the t-th payment and r_t the expected recovery rate after this default, then the present value of the expected loss can be expressed as

$$\mathrm{EL} = \sum_{t=1}^{T} \pi_t (1 - r_t) \frac{U_{t-1}}{(1+d)^t}, \tag{3.7}$$

i.e. as the sum of the unrecovered parts of the unpaid principals multiplied by the probabilities that such default happens, discounted to the time of loan providing. This is corresponding to the fact that in the case of default, the loan company has to repay all the unpaid principal of the loan, and this value is discounted to the time

[2] This is in reality substituted by the fact that loan companies often have a big and well planned loan portfolio, so as these expected revenues and losses are mostly compensated by a planned new business.

of loan issuing. Later the company may get some recoveries from the customer, that are again discounted with the same discount rate.

The definition of default for the probability of default estimation can be for example 90 to 180 days past due and has to be identical with the definition of default for the recovery estimation. In this thesis I work with 90 days definition of default.

3.3.1 Vector of Default Probabilities

The formula (3.7) contains a vector of probabilities of default for individual payments. Therefore, now we need for each client to estimate not one value of probability of default, but a vector of T values – one for each payment. Here the most straightforward method would be the application of the survival analysis model from chapter 1, however I also present some alternatives when using the standard logistic regression model.

When using the logistic regression model, of course, we could have T scoring functions estimating defaults on T different payments, but due to the fact that those scoring models would have distinct data samples for defaulted clients,[3] there might not be enough observations for model development. Moreover, it would be quite complex to maintain so many scoring functions. Therefore, several alternatives of extrapolation of several payment defaults to the whole installment structure are presented.

These concepts are not new and I have seen a specific combination of one constant interval with the exponentially distributed tail in practice. Therefore, the aim of this section is rather to generalize the existing models (e.g. by extrapolation from the constant intervals to the curve intervals or general estimation of the exponential tail in the following subsections), mainly to derive the specific formulas for these models and combine them with some practical hints based on my financial practice.

Finally, as new methods in this concept, I propose the log-normal model extrapolation and the method based on the survival analysis results.

Constant Intervals

The first proposed solution is to analyze the portfolio default rate on individual payments and combine those that seem to have similar probability of default. Then

[3] A client usually defaults 90 or more days past due on one payment only, because the contract is usually terminated in case of such default.

the PDs can be assumed to be constant in these intervals and modeled by the standard logistic regression models. In the case of figure 3.3 we could end up with for example three scoring functions – one for the first payment default, one for the second to fifth payment default and one for sixth to twenty fourth payment default. This example is not very far from the practical situation, because in practice it is sometimes the case that the first one or two payment defaults are higher due to the fraud or no intention to pay the loan back at all.

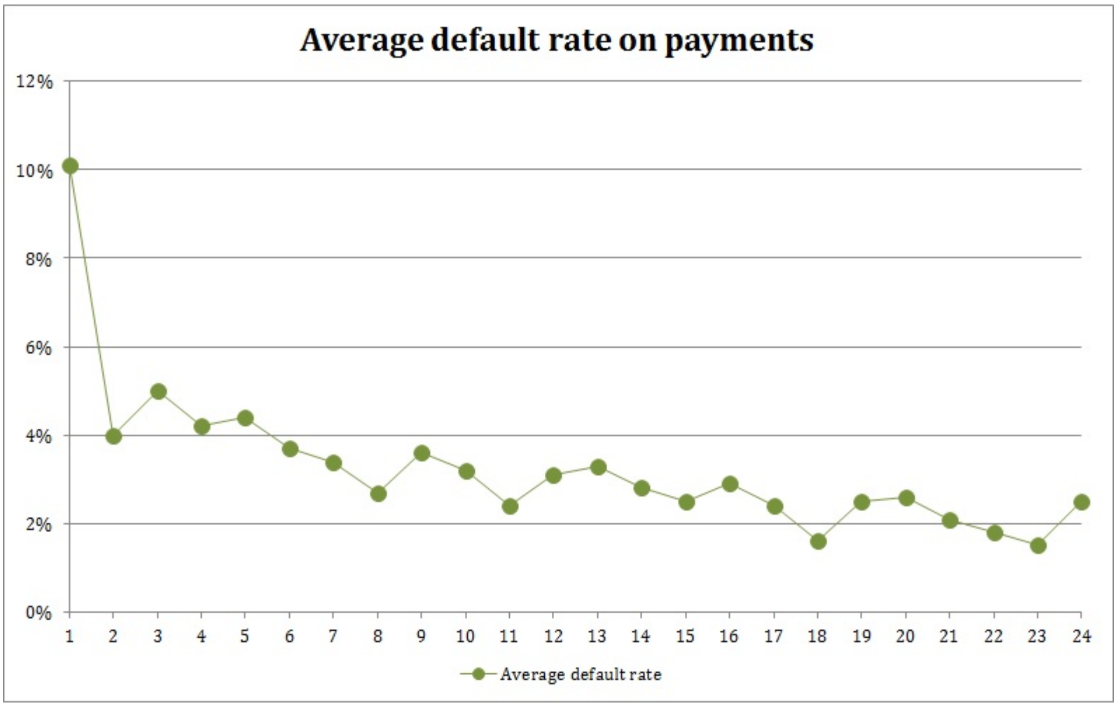

Figure 3.3: Example analysis of portfolio default rates on individual payments

Formally, I assume a positive random variable X representing the time of default. Then I say that the default occurred on t-th payment if $X \in (t-1, t\rangle$. Then for $s < t$ we can develop a scoring function for estimating $\pi_{s:t}$ the probability that a default occurred between the s-th and t-th payment, i.e. $\mathbb{P}(X \in (s-1, t\rangle)$ and for all $k \in \{s, s+1, \ldots, t\}$ set

$$\pi_k = \frac{\pi_{s:t}}{t-s+1}.$$

This approach is easy to understand and compute, but in reality the model can be quite weak. The reason is that usually there are similar factors to affect the probability of default on late payments as those affecting the default on early payments, thus the defaults occurring before the s-th payment considered as non-defaults could weaken the model.

Therefore, I recommend to rather model the probability that the default occurs between the s-th and t-th payment given the fact that it did not occur before the

s-th payment, i.e. estimating $\pi^*_{s:t}$ as

$$\pi^*_{s:t} = \mathbb{P}\left(X \in (s-1, t\rangle | X > s-1\right).$$

This can be achieved by an easy transformation, that all the observation with a default occurring before the s-th payment are removed from the sample. Thus we get a condition fulfilled for all the observations and we can model the probability that the default occurs between the s-th and t-th payment on this adjusted sample.

Now from the definition of conditional probability we get

$$\begin{aligned}\mathbb{P}\left(X \in (s-1, t\rangle | X > s-1\right) &= \frac{\mathbb{P}\left(X \in (s-1, t\rangle, X > s-1\right)}{\mathbb{P}(X > s-1)} \\ &= \frac{\mathbb{P}\left(X \in (s-1, t\rangle\right)}{1 - \mathbb{P}(X \leq s-1)},\end{aligned}$$

which gives us a recurrent formula

$$\pi_{s:t} = (1 - \pi_{1:s-1})\pi^*_{s:t}, \tag{3.8}$$

that can be applied consecutively for any sequence $t_1 < t_2 < \cdots < t_k < t_{k+1}$. Finally, for any $t \in \{t_k, t_k + 1, \ldots, t_{k+1}\}$ we get

$$\pi_t = \frac{\pi_{t_k:t_{k+1}}}{t_{k+1} - t_k + 1}. \tag{3.9}$$

Curve Intervals

In the previous section I assumed the probability of default of individual payments in a given interval to be constant. However, this assumption can be generalized to other curves as well. In general, various curves can be used for this concept to set the shape of the probability of default values; however, probably one of the most convenient approaches could be to estimate the probability of default on two payments (or some constant intervals) and connect them with a straight line. Formally assume that for any $s' \leq s < k \leq k'$ the probabilities $\pi_{s':s}$ and $\pi_{k:k'}$ are estimated by the constant interval method and thus the values of π_s and π_k are computed using formula (3.9). Then for any $t \in \langle s, k\rangle$ the probability of default on t-th payment can be estimated as

$$\pi_t = \pi_s + \frac{t-s}{k-s}(\pi_k - \pi_s). \tag{3.10}$$

Exponentially Distributed Tail

The first method assumed the probabilities to be constant in given intervals till the end of the loan. This method is often used for its simplicity; however, for high probabilities of default and a very long term it can happen that this extrapolation for the later payments can lead to the fact that the sum of default probabilities can be greater than one, which is not mathematically correct.

With the curve intervals the situation can be different, but it can also happen that the sum of probabilities will be greater than one or some probabilities will be out of the interval $\langle 0, 1\rangle$ (e.g. when using a non-constant line).

This problem can be solved by the operation that the tail of this probability vector (i.e. the probabilities π_t for all t greater than some given s) is approximated using a tail of some probability distribution. Generally, based on the shape of the observed probability distribution, many distributions can be used for this purpose. The process of computing the parameters of the distribution based on a set of estimated probability values followed by a derivation of the formulas for individual probabilities would be similar, so I use exponential distribution as an example.

Exponential distribution with parameter λ can be defined by its cumulative distribution function

$$F_E(x, \lambda) = \begin{cases} 1 - e^{-\lambda x}, & \text{if } x \geq 0, \\ 0, & \text{if } x < 0. \end{cases}$$

A well known property of the exponential distribution is its memorylessness, that in this case leads to the proportionality of the consecutive probabilities of defaults. This I demonstrate at the end of this section.

Now I assume that the probabilities of default $\pi_1, \pi_2, \ldots, \pi_s$ are estimated using the constant or curve interval method[4] and my aim is to find an infinite default probability sequence $\pi_{s+1}, \pi_{s+2}, \ldots$ so as

$$\sum_{k=1}^{\infty} \pi_k = 1.$$

To compute the parameter λ of the exponential distribution, we need one probability of default to be estimated from the data. For some $t > s$ we can estimate the conditional probability that the default occurs between time $s + 1$ and t, provided that it did not occur until time s, i.e. $\pi^*_{s+1:t}$.

First I compute $\pi_{1:s}$ as

$$\pi_{1:s} = \sum_{k=1}^{s} \pi_k \tag{3.11}$$

[4] In this case I consider $s = 1$ (i.e. only the probability of first payment default) or even $s = 0$ (i.e. no probabilities estimated and $\pi_0 = 0$) as a special case.

and $\pi_{1:t}$ as

$$\begin{aligned} \pi_{1:t} &= \mathbb{P}(X \leq t) \\ &= \mathbb{P}\big(X \leq t, X \leq s\big) + \mathbb{P}\big(X \leq t, X > s\big) \\ &= \mathbb{P}(X \leq s) + \mathbb{P}(X > s)\,\mathbb{P}\big(X \leq t | X > s\big) \\ &= \pi_{1:s} + (1 - \pi_{1:s})\pi^*_{s+1:t}. \end{aligned} \tag{3.12}$$

Then I assume that the random variable $(X - s)$ has exponential distribution with parameter λ, i.e. for every $x > s$ we get

$$\mathbb{P}\big(X \leq x | X > s\big) = 1 - e^{-\lambda(x-s)}. \tag{3.13}$$

To get the formula for λ I first evaluate the exponential part of the distribution function in the terms of $\pi_{1:s}$ and $\pi_{1:t}$ from formulas (3.11) and (3.12),

$$\begin{aligned} e^{-\lambda(t-s)} &= \mathbb{P}\big(X > t | X > s\big) \\ &= \frac{\mathbb{P}\big(X > t, X > s\big)}{\mathbb{P}(x > s)} \\ &= \frac{\mathbb{P}(x > t)}{\mathbb{P}(x > s)} \\ &= \frac{1 - \pi_{1:t}}{1 - \pi_{1:s}}. \end{aligned}$$

and then solving the equation above I get the formula for λ as

$$\lambda = -\frac{1}{t-s} \ln\left(\frac{1 - \pi_{1:t}}{1 - \pi_{1:s}}\right). \tag{3.14}$$

Similar formula as (3.14) can be also obtained by solving the equation for $\pi^*_{s+1:t}$ from expression (3.13). Then λ can be computed directly from the estimated value of $\pi^*_{s+1:t}$ as

$$\lambda = -\frac{1}{t-s} \ln\big(1 - \pi^*_{s+1:t}\big). \tag{3.15}$$

Finally, for every $x > s$ I evaluate the probability of default on x-th payment as

$$\begin{aligned} \pi_x &= \mathbb{P}\big(X \in (x-1, x\rangle\big) \\ &= \mathbb{P}\big(X \in (x-1, x\rangle, x > s\big) + \mathbb{P}\big(X \in (x-1, x\rangle, x \leq s\big) \\ &= \mathbb{P}(X > s)\,\mathbb{P}\big(X \in (x-1, x\rangle | x > s\big) \\ &= \big(1 - \mathbb{P}(X \leq s)\big)\Big[\mathbb{P}\big(X \leq x | x > s\big) - \mathbb{P}\big(X \leq x-1 | x > s\big)\Big] \\ &= (1 - \pi_{1:s})\left(e^{-\lambda(x-s-1)} - e^{-\lambda(x-s)}\right). \end{aligned} \tag{3.16}$$

This provides a practical guidance for approximating the tail of the probability distribution by exponential distribution. For a tail of $\pi_{s+1}, \pi_{s+2}, \ldots$, we take several of the first payments (e.g. 6 to 24), where we are still able to observe enough defaults in our sample (ideally more than 300), use a scoring model (incl. logistic regression models) to estimate $\pi^*_{s+1:t}$, and use formulas (3.11), (3.12) and (3.14) to estimate the parameter λ. Then all the probabilities $\pi_{s+1}, \pi_{s+2}, \ldots$ can be approximated by formula (3.16). These probabilities then fulfill the conditions given by the estimate of $\pi^*_{s+1:t}$ (i.e. the computed conditional probability of default will be equal to $\pi^*_{s+1:t}$). Moreover, the sum of all default probabilities will be equal to one.

If I now come back to the above mentioned proportionality of consequent probabilities of default. I can show that for every $k > s$ the proportion of probabilities π_k and π_{k+1} is constant. Using formula (3.16) I get

$$\begin{aligned}
\frac{\pi_k}{\pi_{k+1}} &= \frac{(1-\pi_{1:s})\left(e^{-\lambda(k-s-1)} - e^{-\lambda(k-s)}\right)}{(1-\pi_{1:s})\left(e^{-\lambda(k-s)} - e^{-\lambda(k-s+1)}\right)} \\
&= \frac{e^{-\lambda(k-s-1)} - e^{-\lambda(k-s)}}{e^{-\lambda(k-s)} - e^{-\lambda(k-s+1)}} \\
&= \frac{e^{\lambda} - 1}{1 - e^{-\lambda}} \\
&= e^{\lambda}.
\end{aligned}$$

This is a useful property of this distribution that is easy to understand and can serve for checking or computation purposes.

Finally, I take the above mentioned example of three intervals and demonstrate the difference between the constant interval tail and the exponentially distributed tail. Assume that there is a set of three logistic regression based scoring models that are used to create the vector of default probabilities. The first model estimates the probability of default on the first payment, π_1; the second model estimates the probability of default on the second to fifth payment, given there was no default on the first payment, $\pi^*_{2:5}$, and the third model estimates the probability of default on the sixth to twenty-fourth payment, given there was no default on the first five payments, $\pi^*_{6:24}$.

Then, as an example, for a specific client we estimate $\pi_1 = 10\%$, $\pi^*_{2:5} = 20\%$ and $\pi^*_{6:24} = 70\%$. Then using formulas (3.8), (3.11), (3.12) I get $\pi_{2:5} \doteq 18\%$, $\pi_{1:5} \doteq 28\%$ and $\pi_{1:24} \doteq 78\%$. Calculating λ from (3.14) I get $\lambda \doteq 0.063$. Finally, using formulas (3.9) and (3.16) I compute the probability vector for this specific client according to both approaches. In figure 3.4 we can see the estimated values of default probability vectors for a 60 months loan. On this extreme example we can see that the probability of default for the constant approximation sums to more than one for this loan.

In this example the probabilities of default are chosen very high to demonstrate the differences between these two approaches. In reality the probabilities of default

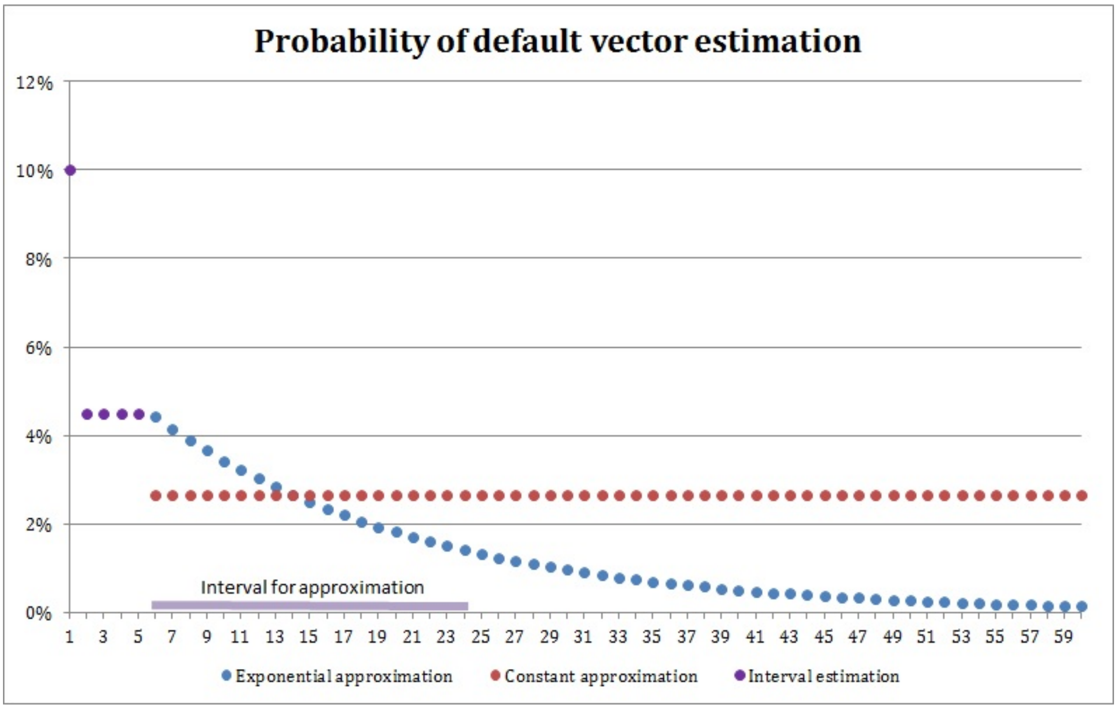

Figure 3.4: Comparison of constant and exponential probability vector approximations

are much lower and the two approaches give more similar results. Moreover, depending on the real shape of the portfolio default distribution the constant approach can be even better (especially for low values of probability of default and short terms), or one can use other probability distributions better fitting the observed shape of defaults.

Log-Normally Distributed Tail

As an alternative to the exponential distribution also the log-normal distribution can be used to extrapolate the tail probabilities.

We say that a positive random variable X is log-normally distributed with parameters μ and σ if its logarithm $\ln(X)$ follows the normal distribution with parameters μ and σ. Then the distribution function of X is

$$F_{LN}(x, \mu, \sigma) = \Phi\left(\frac{\ln(x) - \mu}{\sigma}\right), \tag{3.17}$$

where $\Phi(x)$ is the distribution function of the standard normal distribution.

Then similarly as for the exponential distribution, we assume that the random variable $(X - s)$ has the log-normal distribution with parameters μ and σ and for every $x > s$ we get

$$\mathbb{P}\big(X \leq x|X > s\big) = \Phi\left(\frac{\ln(x-s)-\mu}{\sigma}\right). \tag{3.18}$$

Since the log-normal distribution has two parameters, we need to use two estimates of probabilities $\pi^*_{s+1:t}$ and $\pi^*_{s+1:u}$ and solve the set of equations

$$\begin{aligned}\pi^*_{s+1:t} &= \Phi\left(\frac{\ln(t-s)-\mu}{\sigma}\right),\\ \pi^*_{s+1:u} &= \Phi\left(\frac{\ln(u-s)-\mu}{\sigma}\right).\end{aligned}$$

This way we get the parameter σ as

$$\sigma = \frac{\ln(u-s)-\ln(t-s)}{\Phi^{-1}\left(\pi^*_{s+1:u}\right)-\Phi^{-1}\left(\pi^*_{s+1:t}\right)} \tag{3.19}$$

and μ then as

$$\mu = \ln(t-s) - \sigma\Phi^{-1}\left(\pi^*_{s+1:t}\right). \tag{3.20}$$

Then again, for every $x > s$ the probability of default on x-th payment can be calculated using the expression

$$\pi_x = \big(1-\mathbb{P}(X \leq s)\big)\Big[\mathbb{P}\big(X \leq x|x > s\big) - \mathbb{P}\big(X \leq x-1|x > s\big)\Big],$$

and thus

$$\pi_x = (1-\pi_{1:s})\left[\Phi\left(\frac{\ln(x-s)-\mu}{\sigma}\right) - \Phi\left(\frac{\ln(x-s-1)-\mu}{\sigma}\right)\right] \tag{3.21}$$

for $x > s+1$, and

$$\pi_x = (1-\pi_{1:s})\Phi\left(\frac{\ln(x-s)-\mu}{\sigma}\right) \tag{3.22}$$

for $x = s+1$.

If I now come back to the example from the previous section, where for a specific client we had $\pi_1 = 10\%$, $\pi^*_{2:5} = 20\%$ and $\pi^*_{6:24} = 70\%$, I can calculate $\pi^*_{2:24} \doteq 76\%$ using the alternation of formula (3.12) in the form

$$\pi^*_{s+1:t} = \frac{\pi_{1:t}-\pi_{1:s}}{1-\pi_{1:s}}. \tag{3.23}$$

Then using formulas (3.19) and (3.20) I get the parameters $\mu \doteq 2.337$ and $\sigma \doteq 1.130$, and using (3.21) and (3.22) the probability vector for this client. In figure 3.5 we can see the estimated values compared with the constant intervals and exponential tail.

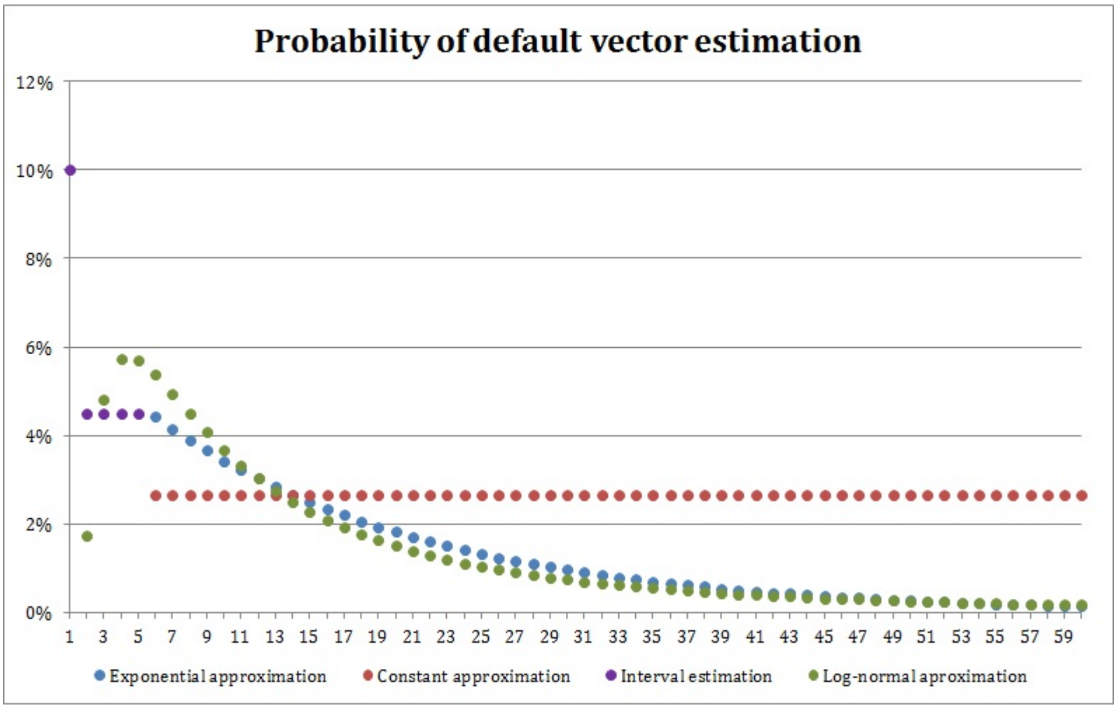

Figure 3.5: Comparison of log-normal probability vector approximation with the exponential and constant approach

Specific estimate

In the previous sections I dealt with several approaches to extrapolate the results of the logistic regression-based scoring functions to all the payments. However, as shown in chapter 1, there are methods that can estimate the probability of default of a specific client for every single payment from the observed sample.

If I take the repayment survival model introduced in section 1.1.2, I can get the vector of default probabilities directly from (1.7) as

$$\pi_t = S(t-1) - S(t). \tag{3.24}$$

This is one of the useful advantages of applying the Cox model for the default modeling. Not only it gives similar performance and apparently better predictions than the logistic regression model (see section 1.5),[5] it also gives the estimation for default on individual payments, that can be used in a profitability model. This can bring the additional dimension needed in the profitability model, and thus simplify the probability vector computations and enhance the model precision.

However, from the non-parametric Cox model we only get the estimation of probabilities of default for the times observed in the development sample, and then some

[5] This depends on the real default distribution.

parametric extrapolation needs to be done as well (e.g. using constant, exponential distribution or log-normal distribution from the previous sections).

To solve this problem, based on the default distribution, any parametric alternative can be used as well. Then for example the AFT (accelerated failure time) model assuming a log-normal distribution can be fitted from the data directly. A comparison of the AFT and Cox models can be found in (Pazdera et al., 2009).

3.3.2 Recoveries

In this section I come back to the equation (3.7) and explore a little more about the recovery rates r_t. For a specific client we need to estimate the after-default recovery in case the default happens. This is not only client specific but also depends on the payment when the default occurred. Similarly as for the probability of default estimation, in some portfolios there is a concentration of frauds and unwilling to pay the loan back, that usually leads to default on the first one or two payments and low recovery after that.

By the recovery rate r_t in the terms of the profitability model (3.5) I mean all the cash flows (CF) received from the customer after the default occurred, discounted to the time of default and divided by the unpaid principal at the time of default, i.e.

$$r_t = \frac{1}{U_{t-1}} \sum_{k=0}^{\infty} \frac{\mathrm{CF}_{t+k}}{(1+d)^k}. \tag{3.25}$$

For simplicity, I do not distinguish the principal, interest, fee and other payments that can occur during the collection process after default. This way I avoid using separate models for recovery of these components. This can be quite complicated due to the pairing algorithm that is based on the contract.[6] This way it can easily happen that a fully recovered loan would have recovery over one. This is in line with the way I defined the expected revenue in section 3.4, where I do not consider any interest and fees recoveies after default (since those are considered here).

I make one more remark about the recovery horizon. As per the definition in (3.25) I formally consider all the future recoveries without any time limitation. However, in reality our data sample is limited and in fact there are usually negligible recoveries after several years. Also the discount factor plays a role and the very late recoveries have a low impact on the total recovery rate. Therefore, in practice we can

[6] If the client pays some amount of money during the collection process, we would need to know exactly which debt transaction this money is paired to. Sometimes this algorithm is quite complicated (e.g. first it should be paired to the principal of the oldest unpaid installment, then to the interest and fees of this installment, then the same for the second oldest installment up to the full amount of the debt and penalties after the contract termination, late fees, late interest etc.).

take some relevant recovery horizon (usually 2–5 years) and consider the recovery process finished. Then we can either take a conservative approach and assume there are no more recoveries after this horizon, or we can make an expert estimate about the rate of future recoveries.

Portfolio Segments

The easiest way to estimate the recovery rate coefficients is to find the most relevant predictors (usually 1–3), cut the portfolio into several segments based on these predictors, and analyze the recovery rate separately for each segment. Then the estimated recovery rate for a given client on a given payment would be taken as the recovery rate of the corresponding segment.

Although this approach is not very scientific, it is from my experience sometimes used due to its simplicity and a lack of data.[7] As mentioned earlier, the payment number when the default occurred, should be considered as one of the predictors for segmentation.

When the segmentation is done, one can look at the recovery rate of individual segment vintages and make an expert estimation of the expected recovery in the infinite time. Another option is to suggest a curve that can be fitted to the observed data and can help with the recovery prediction. Even here the homogeneity of the recovery process needs to be fulfilled.

In my master thesis (Rychnovský, 2009) I applied some methods on the pool of recovery data, in order to estimate the recovery rate of the whole data sample.[8] One of the methods was to parameterize the recoveries using the following curve,

$$r_t = \breve{\mu}\frac{1-\breve{\nu}^t}{1-\breve{\nu}^T}, \tag{3.26}$$

where $\breve{\mu}$ a $\breve{\nu}$ are parameters.

Then I applied this curve on the pool of data for $T = 36$ months and used the weighted least squares method to estimate the parameters. In figure 3.6 we can see the cumulative recoveries of the whole pool together with the curve (3.26) with $T = 36$, $\breve{\mu} = 0.535$ and $\breve{\nu} = 0.914$.

[7] For a good recovery model the data sample needs to be very long (i.e. the defaults happening long time ago) and the collection process should be homogenous (i.e. not changing in time). These are conditions that are not so easy to satisfy, especially for institutions with short data history or dynamic processes.

[8] It was the same data sample as the one used in chapter 2.

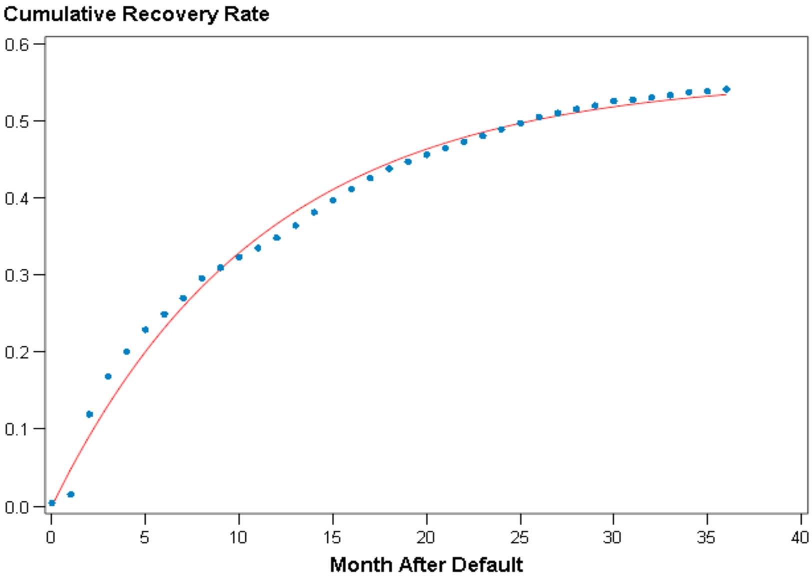

Figure 3.6: Recovery curve estimation for $T = 36$ using the model (3.26) from (Rychnovský, 2009)

Recovery Models

When we have a qualifying data sample with long history and homogenous collection process, we can use one of the recovery rate models described in chapter 2. For this purpose we can even use the number of payments before default as one of the predictors, or a stratification condition for the data sample.

Thus for a specific client we can estimate the recovery rate corresponding to every payment of the loan. Again, some reasonable time horizon has to be set and the ultimate recovery needs to be expertly adjusted.

This is one of the most straightforward applications of the recovery models described in chapter 2. Here the performance of the recovery model affects the total performance of the whole profitability model. Therefore, any improvement in the model accuracy, stability or prediction power (such as the application of the survival analysis model presented in this thesis) brings a direct benefit into a company's underwriting system and thus the profit of the company.

Insurance

There is a variety of insurance products on the market intending to cover the customer's expenses in case a pre-defined unfortunate event happens. Often these insurance products are sold together with the provided loan and intend to cover the loan payments in the case of death, injury or employment loss.

Not only the existence of the insurance needs to be considered in the expected revenue stream (since when sold as a byproduct with the loan it brings additional revenue), it can also affect the recovery for the insured customers.

Depending on the penetration of the insurance in our sample, the payout conditions and the length of the sample history, I propose several ways to work with insurance. If the history is sufficient, it is in my opinion best to use the insurance as one of the predictors for the probability model or segmentation and get the real payoff impact from the data directly. Then there is no need for further adjustments.

On the other hand, if the history of the product is insufficient, we can estimate the payoff probability of the insured customer – either by a probability model (e.g. using the logistic regression), or by a constant that we get from the historical data (either from our institution or from the insurance company), or by an expert estimation. In this case, the adjusted recovery rate for the insured customers can be calculated as

$$r_t = \kappa a + (1 - \kappa) r_t^*,$$

where κ is the probability of the insurance payoff, a is the recovery rate value in the case of insurance payoff (e.g. 120%, covering the principal and interest debt of the customer) and r_t^* is the original recovery rate without considering insurance.

3.4 Expected Revenue

Compared to the estimation of the expected loss, the expected revenue part is much more deterministic. I take the loan repayment structure from section 3.2.1 and express the expected part of the interest, fees and insurance using the estimated probabilities of default on individual payments. As mentioned already, the recovery part of the interest, fees and insurance is included in the recovery rate model and propagated to the expected loss calculation.

Therefore, for the purpose of the expected revenue, only revenues of non-defaulted payments should be considered. Furthermore, it follows that after a defaulted payment there are no revenues as well, since the case is already in the collection process and all additional interest, fees etc. are included in the recovery calculation as well. Therefore, the revenue is only considered if there was no default up to and including the corresponding payment.

I remind that the default can happen only once, the events are distinct and the probability that the event did not happen up to and including the t-th payment can be expressed as

$$\mathbb{P}(X > t) = 1 - \pi_{1:t} = 1 - \sum_{k=1}^{t} \pi_k.$$

Even though the expected revenue can consist of various components, for the purpose of this model, I simplify it as

$$\text{ER} = \text{ER}_I + \text{ER}_F + \text{ER}_C + \text{ER}_N, \tag{3.27}$$

where ER_I is the expected interest profit, ER_F is the expected profit from fees, ER_C is the expected profit from commission and ER_N is the expected profit from insurance. All these components are to be discounted appropriately.

3.4.1 Interest

Taking into account the interest part of each payment expressed in (3.3), I can compute the expected interest revenue ER_I as

$$\text{ER}_I = \sum_{t=1}^{T} \left(1 - \sum_{k=1}^{t} \pi_k\right) \frac{I_t}{(1+d)^t}.$$

Using the formula for interest of t-th payment, $I_t = iU_{t-1}$, this can also be written as

$$\text{ER}_I = \sum_{t=1}^{T} \left(1 - \sum_{k=1}^{t} \pi_k\right) \frac{iU_{t-1}}{(1+d)^t}.$$

3.4.2 Fees

Since fees are considered to be constant payments F for the services connected with the loan, the expected revenue from the fees can be expressed as

$$\text{ER}_F = \sum_{t=1}^{T} \left(1 - \sum_{k=1}^{t} \pi_k\right) \frac{F}{(1+d)^t}.$$

3.4.3 Commission

By a commission I mean the commission the loan company gets for providing a loan. For example when we are talking about consumer loans provided for buying certain goods, the retailer can give a commission to the loan company to offset a lower interest of the loan.[9]

The commission is often a one-off payment directly after the sales. Thus, there is no discounting and no default losses connected, and especially in high interest or risky markets this kind of a certain revenue can significantly improve the profitability of the loan.

The commission can be either fixed C or for example defined as a percentage c of product price or a loan amount. Then we get

$$\mathrm{ER}_C = C + cU_0.$$

3.4.4 Insurance

The last mentioned revenue component is insurance. Loan companies often sell insurance products alongside with loans and get a brokerage for the sales. Again, the product can be designed in various ways, where in my experience the most common are:

1. Charging a one-off brokerage at the beginning of the loan.
2. Charging a fix or unpaid principal based brokerage together with each payment.
3. Charging a one-off brokerage together with charging all the insurance fees to the customer at once at the beginning of the loan, usually by adding it to the credit amount (and therefore charging additional interest from it).

Either way, if I denote the N_0 the one-off brokerage at the beginning of the loan and N_t the insurance revenue from the t-th payment, I can express the expected revenue from the insurance as

$$\mathrm{ER}_N = N_0 + \sum_{t=1}^{T}\left(1 - \sum_{k=1}^{t}\pi_k\right)\frac{N_t}{(1+d)^t}.$$

3.5 Expected Costs

The last profit component in my model is the expected costs. Again, there are various costs that could be included in the model and their importance will always

[9] This is often the case when seeing zero interest loans or a cheaper product price when purchased with a loan.

depend on the business specifics. Generally, I divide the expected costs into the expected fix costs EC_F, expected variable costs EC_V, expected time dependent costs EC_T and expected collection costs EC_C as

$$\mathrm{EC} = \mathrm{EC}_F + \mathrm{EC}_V + \mathrm{EC}_T + \mathrm{EC}_C\,. \tag{3.28}$$

3.5.1 Fix Costs

By the fix costs I mean all the costs connected to the loan providing business that are not directly allocated to a specific loan or a collection activity (e.g. office rental costs, personal expenses except loan selling incentives, costs of technology etc.).

From the microeconomic point of view these costs should not be considered in the decision about an individual loan providing, since for profit optimization we should consider only marginal revenue and marginal cost of the tranche. Then the fix cost optimization is a separate task for the finance department. Even though some managers might disagree, for the purpose of this model I set

$$\mathrm{EC}_F = 0.$$

3.5.2 Variable Costs

By the variable costs I mean all the one-off costs at the beginning of the loan connected with issuing of the loan. It can include any absolute or relative commissions for the loan seller, retailer, costs for contracts postage etc. If I denote all these absolute costs as V and relative costs of the loan amount as v, I can put

$$\mathrm{EC}_V = V + vU_0.$$

3.5.3 Time Dependent Costs

Apart of the fix costs and the one-off variable costs, there are further costs connected with the duration of the loan or amount of the borrowed money. As an example I can mention a servicing cost s that I count for every period that the loan is not defaulted (after default I assume the collection cost mentioned in the next section), and the cost of funds that the company needs to pay to the investors providing the funding of the loan. This is assumed to be the minimal required revenue of the shareholders i_S and the interest rate i_F that the company pays to the funding partners. If a default occurs, I assume that the money is paid back immediately as the default losses.

Furthermore, I assume that ρ is a constant ratio of capital in the loan principal.[10] Then using the servicing cost s and the total funding interest rate $\rho i_S + (1-\rho)i_F$, both only in the case of no default, I can evaluate the expected time dependent costs as

$$\mathrm{EC}_T = \sum_{t=1}^{T}\left(1 - \sum_{k=1}^{t}\pi_k\right)\frac{s + \big(\rho i_S + (1-\rho)i_F\big)U_{t-1}}{(1+d)^t}.$$

3.5.4 Collection Costs

Finally, by the collection costs I mean the costs that need to be paid in the case that the client defaults. These cost should cover all the unit costs Γ, like costs of phone calls, postage, time of the operators etc, as well as the costs connected with the unpaid principal of the loan, like court fees, executory fees etc. For simplicity I assume it to be a percentage γ from the unpaid principal at the time of default. Then the collection costs can be expressed as

$$\mathrm{EC}_C = \sum_{t=1}^{T}\pi_t\frac{\Gamma + \gamma U_{t-1}}{(1+d)^t}.$$

3.6 Expected Profit

Coming back to the expected absolute profit formula (3.5) I combine the formulas in the previous sections to the final expression

$$\begin{aligned}\mathrm{EAP} &= C + N_0 - V + (c-v)U_0 + \\ &+\sum_{t=1}^{T}\left(1 - \sum_{k=1}^{t}\pi_k\right)\frac{F + N_t - s + \big(i - \rho i_S - (1-\rho)i_F\big)U_{t-1}}{(1+d)^t} + \\ &+\sum_{t=1}^{T}\pi_t\frac{-\Gamma - (1 - r_t + \gamma)U_{t-1}}{(1+d)^t},\end{aligned} \tag{3.29}$$

where π_t and r_t are properly modeled values of the probability of default and recovery rate of the customer on each payment, considering possible insurance payoff.

Now putting aside any other manual activities (like documents checking, employment verification, manual underwriting etc.), with the expected absolute profit formula (3.29) one can setup the automated underwriting system as

[10] The minimal proportion of the capital can by required by the funding partner and is usually required by the regulation.

- If EAP > 0 then APPROVE.
- If EAP ≤ 0 then REJECT.

This means that the company approves all loans that are expected to bring profit, and reject all loans that are expected to make losses. Strictly speaking, this is a mathematically correct approach to maximize the expected profit of the company from the loan providing business.

However, in the real business many more aspects are usually considered (e.g. imperfection of the model, overall company strategy and risk appetite or an increased requirement on the profitability of the business) that lead to the fact that one needs to set a threshold for the approval or rejection other than zero profit.[11]

In this case I start questioning the above defined expected absolute profit to be the only and correct measure for approval and rejection. Therefore, I provide also several alternatives to be considered.

3.6.1 Absolute Profit

The EAP characteristics from formula (3.29) answers the question what is the absolute amount of the risk adjusted profit the company gets from providing this loan. In my opinion, this characteristics should be considered for the companies, where funding is no limitation and a big issue is attracting new customers. Then using EAP maximizes the absolute profit from each customer.

3.6.2 Relative Profit

By the expected relative profit I mean the expected absolute profit divided by the original loan amount, i.e.

$$\mathrm{ERP} = \frac{\mathrm{EAP}}{U_0}. \tag{3.30}$$

I would suggest using this approach for the cases when there is a limitation of funding and the company wants to maximize the profit coming from providing one currency unit in a loan for customers in the present time, i.e. not considering repeated loans.

[11] For example negative, if a company strategy is to achieve a high market share in a short time and is willing to do some unprofitable business to serve more customers; or positive, if an economic crisis is expected.

3.6.3 Internal Rate of Return

Whereas the absolute and relative profit characteristics are taken purely from the perspective of the loan providing company, the internal rate of return is a characteristic that should be considered in the case of investing the company's own money. Therefore, I consider the net present value (NPV) and the internal rate of return (IRR) of such investment. More information about investment metrics can be found e.g. in (Promislow, 2014).

If I look at the provided loan as an investment of U_0 at the time $t = 0$, I can rearrange the formula (3.29) in the form of the risk-adjusted net present value (NPV) as a function of discount rate δ.

The NPV formula should consist of the negative value of the initial investment U_0 and the expected value of all the future cash flows between the company and the customer. In this case I do not need to distinguish the interest part and the principal part, but simply replace it by the annuity A, that is paid in the case of no default occurring before the payment, and the recovery amount $r_t U_{t-1}$, that is recovered in the case of default. Also the cost of funding part is excluded.

Then the risk-adjusted expected net present value of the loan with discount rate δ can be computed as

$$\begin{aligned} \mathrm{NPV}(\delta) &= -U_0 + C + N_0 - V + (c - v)U_0 + \\ &+ \sum_{t=1}^{T} \left(1 - \sum_{k=1}^{t} \pi_k\right) \frac{F + N_t - s + A}{(1+\delta)^t} + \\ &+ \sum_{t=1}^{T} \pi_t \frac{-\Gamma + (r_t - \gamma)U_{t-1}}{(1+\delta)^t}. \end{aligned} \tag{3.31}$$

Finally, the risk-adjusted expected return on investment corresponding to the time interval between two payments, is the value of the discount rate δ that solves the equation

$$\mathrm{NPV}(\delta) = 0. \tag{3.32}$$

This equation is usually solved numerically by the bisection method or the Newton-Raphson method (see e.g. in (Brandimarte, 2003), (Ypma, 1995) or (Verbeke and Cools, 1995)). In some cases the equation might have more than one solution and the convergence algorithm must be slightly adjusted (see e.g. (Cannaday et al., 1986), (Flemming and Wright, 1971) or (Colwell, 1995)).

The risk-adjusted expected internal rate of return is a characteristic I would suggest using in the situation when the company is in the position of the only investor, there is no strict limitation of the clients or capital and the company wants to achieve best return of their investment.

3.6.4 Return on Equity

The last characteristic I propose, is the risk-adjusted expected return on equity (ROE) from this particular loan, corresponding to one period. This characteristics is very similar to the above specified internal rate of return, just assumes that the equity is only used to cover a specific part of the loan amount.

Similarly as in section 3.5.3 I assume that a proportion ρ of the loan amount is covered by the equity from the shareholders and the rest of the loan amount, i.e. $(1-\rho)U_0$, is funded from some external sources with interest rate i_F. In this case I change in formula (3.31) the initial investment to ρU_0.

Now I need to calculate the interest and principal repayments of the funding. Here I consider two alternatives. First alternative is a standard loan from the funding partner with the loan value $(1-\rho)U_0$, term T and interest i_F. Then independently of the client's behavior, the company needs to repay every time interval a fixed annuity $A_{(1-\rho)U_0,T,i_F}$, that is calculated as

$$A_{(1-\rho)U_0,T,i_F} = (1-\rho)U_0 \frac{i_F(1+i_F)^T}{(1+i_F)^T-1}$$

and the final NPV formula can be expressed as

$$\begin{aligned} \text{NPV}(\delta) &= -\rho U_0 + C + N_0 - V + (c-v)U_0 + \\ &+ \sum_{t=1}^{T}\left(1-\sum_{k=1}^{t}\pi_k\right)\frac{F+N_t-s+A}{(1+\delta)^t} + \\ &+ \sum_{t=1}^{T}\pi_t\frac{-\Gamma+(r_t-\gamma)U_{t-1}}{(1+\delta)^t} - \\ &- \sum_{t=1}^{T}\frac{(1-\rho)U_0}{(1+\delta)^t}\frac{i_F(1+i_F)^T}{(1+i_F)^T-1}. \end{aligned} \tag{3.33}$$

The second alternative is the funding scheme described in section 3.2.4, i.e. a revolving loan with the initial loan amount of $(1-\rho)U_0$ and interest rate i_F. This loan is paid according to client's repayment. This means that every time period t when the client repays the principal part of the loan P_t, the company will repay the appropriate part of the principal $(1-\rho)P_t$ to the funding partner, together with the interest payment $i_F(1-\rho)U_{t-1}$ corresponding to the period. This happens only if the client has not defaulted up to this payment. On the other hand, if the client defaults on t-th payment, all the remaining principle of $(1-\rho)U_{t-1}$ is repaid at once, together with the corresponding interest $i_F(1-\rho)U_{t-1}$.

Then the NPV formula changes to the form

$$\begin{aligned}
\mathrm{NPV}(\delta) &= -\rho U_0 + C + N_0 - V + (c-v)U_0 + \\
&+ \sum_{t=1}^{T} \left(1 - \sum_{k=1}^{t} \pi_k\right) \frac{F + N_t - s + A}{(1+\delta)^t} + \\
&+ \sum_{t=1}^{T} \pi_t \frac{-\Gamma + (r_t - \gamma)U_{t-1}}{(1+\delta)^t} - \\
&- \sum_{t=1}^{T} \left(1 - \sum_{k=1}^{t} \pi_k\right) \frac{i_F(1-\rho)U_{t-1} + (1-\rho)P_t}{(1+\delta)^t} - \\
&- \sum_{t=1}^{T} \pi_t \frac{(1+i_F)(1-\rho)U_{t-1}}{(1+\delta)^t},
\end{aligned}$$

that can be re-arranged as

$$\begin{aligned}
\mathrm{NPV}(\delta) &= -\rho U_0 + C + N_0 - V + (c-v)U_0 + \qquad (3.34) \\
&+ \sum_{t=1}^{T} \left(1 - \sum_{k=1}^{t} \pi_k\right) \frac{F + N_t - s + A - i_F(1-\rho)U_{t-1} - (1-\rho)P_t}{(1+\delta)^t} + \\
&+ \sum_{t=1}^{T} \pi_t \frac{-\Gamma + (r_t - \gamma - (1-\rho) - i_F(1-\rho))U_{t-1}}{(1+\delta)^t}.
\end{aligned}$$

Finally, based on the funding model, the NPV of (3.33) or (3.34) can be set to zero and the root of this equation I take as the final risk-adjusted expected return on equity from this particular loan, corresponding to one period. Same as for the IRR, this equation is usually solved numerically and the proper solution needs to be selected.

I would suggest using this characteristic in the situation, when there are no strict limitations on the number of clients or funding, and our main priority is to maximize the shareholders' return on equity.

3.7 Data Simulation

In this section I aim to implement the proposed profitability models on the sample of banking data to simulate the impact of the profitability model. There are no complete banking data available for this research, so I decided to use the data sample from chapter 1 and simulate the values that are not known. Even though this approach will not give the fully authentic results for this sample, it enables me to illustrate the impact and simulate the sensitivity on two different settings of the values.

Therefore, since some of the key characteristics are not provided in the sample, I simulate the potential values and calculate the profitability of every single loan. Then I simulate the automated approval process based on the probability of default (PD) as well as on the four profitability measures (EAP, ERP, IRR, ROE) introduced in the previous section. Finally, I compare the methods and show the sensitivity on some of these parameters. The whole model is implemented in MS Excel.

3.7.1 Data Overview and Sample Preparation

For the purpose of this analysis I take the comparison random sample of 2,835 clients from chapter 1 together with the variable ID as the identifier, credit limit indicating the credit amount of the particular loan, variable effect indicating the month when the loan was issued and variable maturity indicating the month of loan maturity (all the selected loans in this sample have a fixed end). For all of these clients I also use the Cox estimations of the survival probabilities for all times.

For all of the loans I calculate the term in months and limit it to maximum of 72 months (there were a few outliers). Thus I get a sample of 2,835 loans with the loan value U_0, term T and the probability of default estimations by the Cox model.

3.7.2 Estimation of the Vector of Default Probabilities

Now I need to set the default probability vectors for individual payments of each client. Since in this sample the Cox model gives relevant individual survival probability estimations only up to about 48 months (due to a low number of observations defaulting in higher terms), I need to extrapolate the probability vector to the maximum term (i.e. up to 72 months for some loans).

For this purpose, I use the value of the survival function for 48 and 60 months to estimate the parameter λ of the exponential distribution for each loan. Then the probabilities of default for individual payments are calculated as

$$\pi_t = S(t-1) - S(t), \qquad \text{for } t \in \{1, 2, \dots, 48\},$$

and

$$\pi_t = S(48)\left(e^{-\lambda(t-49)} - e^{-\lambda(t-48)}\right), \qquad \text{for } t \in \{49, 50, \dots, 72\}.$$

In figure 3.7 we can see the vectors of default probabilities for the first three loans in the sample.

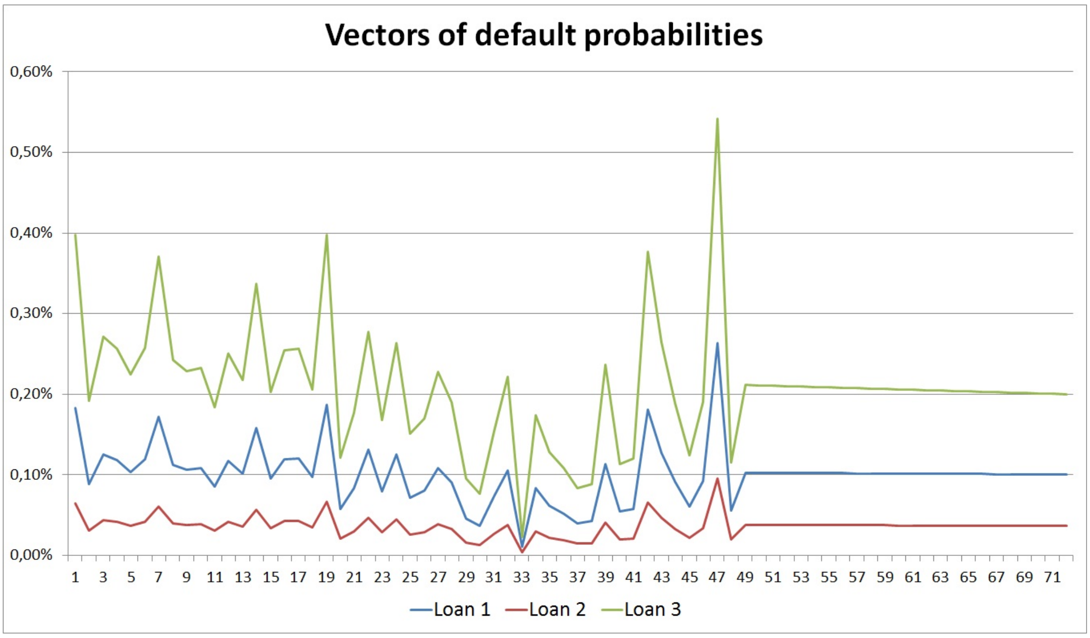

Figure 3.7: Vectors of default probabilities for the first three loans in the sample

3.7.3 Simulation of Characteristics

Now I need to simulate the rest of inputs for the profitability model. First I simulate all the values to be the same for all clients and then I simulate some progression in the interest rate i and the capital share ρ. Also I assume that some of the clients took the insurance with monthly payment N.

Similar Loan Characteristics

All the loans in the sample have the real loan value and term, together with the personalized estimated probability of default. However, the following parameters I set the same for the whole portfolio,

- $i = 0.8\%$ – interest rate 0.8% per month (9.6% p.a.),
- $F = 10$ – fee charge 10 CZK per month,
- $N = 0$ – zero insurance income,
- $C = 0$ – zero commission from retailer,
- $c = 0\%$ – zero commission from retailer,
- $V = 200$ – variable cost 200 CZK per approved contract,

- $v = 0.5\%$ – variable cost 0.5% of approved credit amount,
- $s = 50$ – servicing cost 50 CZK per month,
- $i_F = 0.1\%$ – funding interest 0.1% per month (1.2% p.a.),
- $i_S = 1\%$ – shareholders' interest 1% per month (12% p.a.),
- $\rho = 10\%$ – share of capital 10% of the principal amount,
- $\Gamma = 500$ – collection cost 500 CZK in case of default,
- $\gamma = 1\%$ – collection cost 1% of unpaid principle in case of default,
- $d = 0.1\%$ – discount rate 0.1% per month (1.2% p.a.).

Now I am able to calculate the loan annuity A, outstanding principle structure U_t, principle payment structure P_t, as well as all the profitability measures EAP, ERP, IRR and ROE for each loan.[12] In figure 3.8 we can see the relation between the probability of default and the expected relative profit for this simulation. As we can see from this chart, there is a visible negative correlation of these two measures (i.e. the higher probability of default, the lower expected relative profit). The Pearson correlation coefficient of this pair is -0.60. Given the equal simulated values, this is in line with my expectation and it suggests that the simulation works as expected.

In table 3.1 we can see the correlations of all the characteristics that can be used for approval or rejection of the loans. As we can see from the table, the characteristics IRR and ROE are fairly similar in this comparison.

Table 3.1: Pearson correlation structure in the similar loan characteristics simulation

Pearson	PD	EAP	ERP	IRR	ROE
PD	1.00	−0.43	−0.60	−0.65	−0.65
EAP	−0.43	1.00	0.71	0.71	0.72
ERP	−0.60	0.71	1.00	0.90	0.90
IRR	−0.65	0.71	0.90	1.00	0.99
ROE	−0.65	0.72	0.90	0.99	1.00

Progressive Loan Characteristics

As an alternative to the similar loan characteristics approach I simulate the progressive risk-based pricing, which is closer to the market standard in the Czech Republic. In this approach I assume that the loan interest rate and the capital ratio are functions of the risk estimation of the client (particularly the probability of

[12] For calculation of IRR and ROE I use a simple macro in Visual Basic.

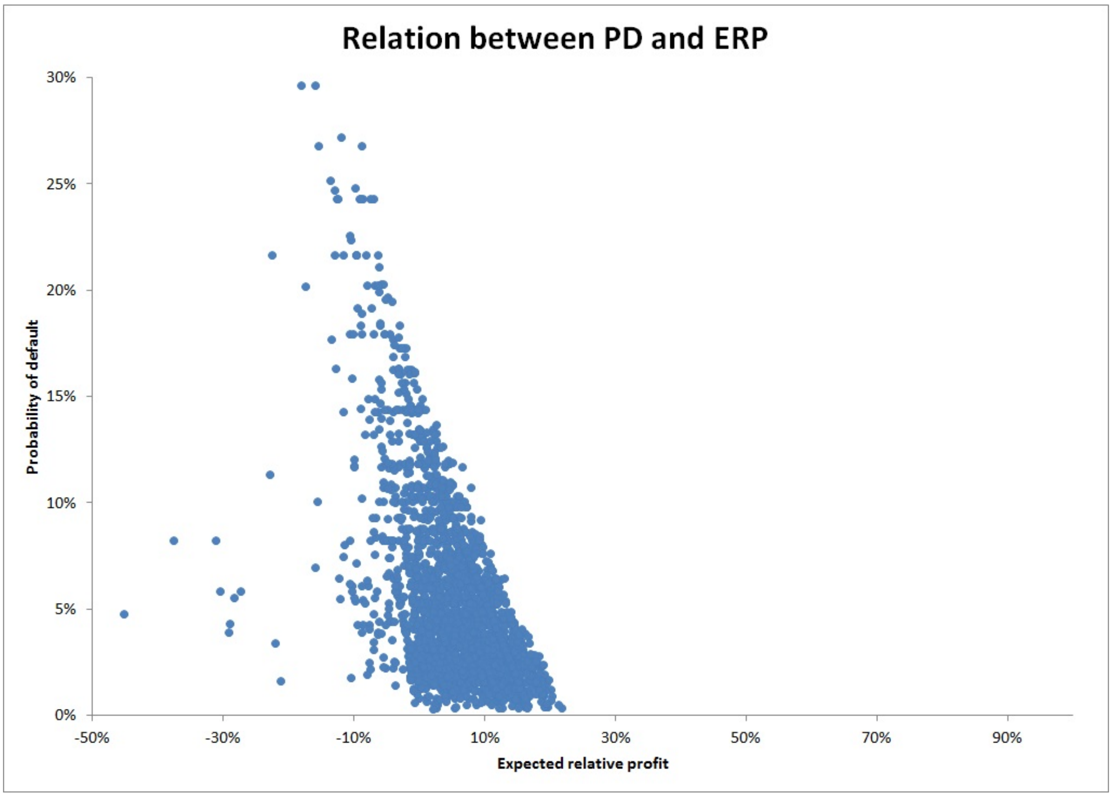

Figure 3.8: Relation between the probability of default and the expected relative profit for the similar loan characteristics simulation

default on the first 24 months) and approximately one third of the clients would take insurance.

Therefore, I keep the values as for the similar loan characteristics simulation, just change i, ρ and N as follows,

- $i = \frac{4\%+\pi_{1:24}}{12}$ – interest rate as a function of the estimated probability of default of the specific client,
- $N = \begin{cases} 200, & \text{if ID} \mod 3 = 0, \\ 0, & \text{if ID} \mod 3 \neq 0. \end{cases}$ – insurance payment of 200 CZK monthly for one third of clients and no payment for the rest.
- $\rho = 2.4\% + \pi_{1:24}$ – share of capital as a function of the estimated probability of default of the specific client.

Now I again calculate the loan annuity A, outstanding principle structure U_t, principle payment structure P_t, as well as all the profitability measures EAP, ERP, IRR and ROE for each loan.

For comparison, in picture 3.9 we can again see the relation between the probability of default and the expected relative profit for this simulation. Here the characteristics show a slightly positive correlation (i.e. the higher probability of default, the higher expected relative profit) and the Pearson correlation coefficient of this pair is 0.28. This is mainly connected to the fact that the higher default probabilities are associated with higher interest rates in this simulation.

In table 3.2 we can again see the correlations of all the characteristics. As we can see from the comparisons of tables 3.1 and 3.2, with the progressive simulation the characteristics are generally less correlated.

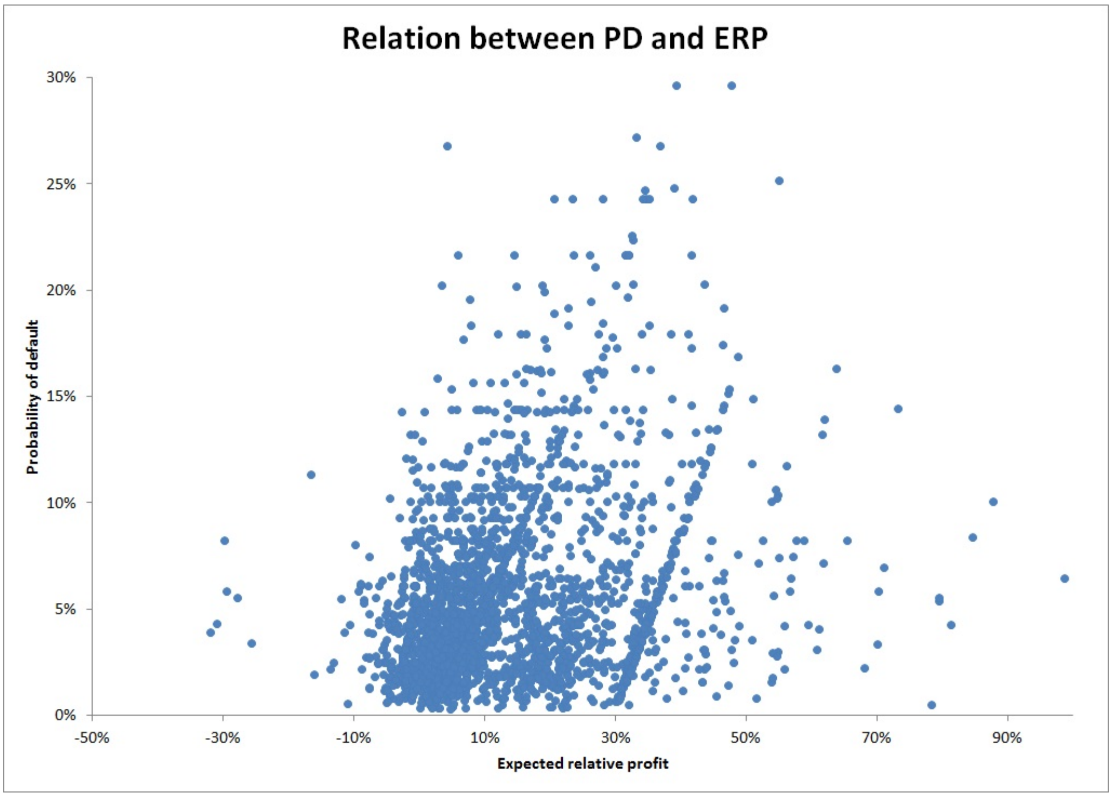

Figure 3.9: Relation between the probability of default and the expected relative profit for the progressive loan characteristics simulation

Table 3.2: Pearson correlation structure in the progressive loan characteristics simulation

Pearson	PD	EAP	ERP	IRR	ROE
PD	1.00	0.30	0.28	0.32	−0.12
EAP	0.30	1.00	0.59	0.44	0.35
ERP	0.28	0.59	1.00	0.87	0.75
IRR	0.32	0.44	0.87	1.00	0.85
ROE	−0.12	0.35	0.75	0.85	1.00

3.7.4 Approval Process Simulation

In this section I simulate the automated underwriting process and compare its results when using different characteristics as the driver for approval decision. With all the characteristics calculated, the simulation is done in the way that it is assumed that the current approval process is approving 75% of applicants based on their probability of default, and the alternatives should be approving either the same number of applicants, or similar volume of loans based on individual profitability measures.[13]

Then for each of the measures I compute one threshold, such as 75% of applicants have the expected value of this measure over the threshold and are assumed to be approved. In this case the threshold is actually given by the upper quartile of all the values of this measure in the sample. Then I compute a second threshold, such as the total loan volume of the approved customers is similar to the volume of the 75% of clients approved by the probability of default. In this case such value is found numerically by a simple macro.

Then for all of the characteristics and both schemes I know which specific clients are assumed to be approved and which are rejected, so I am able to easily compute the key expected business, profitability and risk characteristics of the approved portfolio, which are

- approved volume (apr. volume);
- number of approved applications (num. apr.);
- average probability of default of the approved clients (avg. PD);
- average probability of default of the approved clients weighted by the loan amounts (avg. vol. PD);
- average expected absolute profit of the approved clients (avg. EAP);
- total expected absolute profit of the approved clients (sum. EAP);
- average expected relative profit of the approved clients (avg. ERP);
- average expected relative profit of the approved clients weighted by the loan amounts (avg. vol. ERP);
- average internal rate of return of the approved clients (avg. IRR);

[13] Similar approval process simulations are often done in financial practice, whenever there is a change in the underwriting model or process. In this case I simulate the impact of the change of the whole underwriting model under the assumption of the same approved number of clients or the same volume. Of course, in reality the real approved volume could be adjusted before the real implementation based on the results.

- average internal rate of return of the approved clients weighted by the loan amounts (avg. vol. IRR);
- average return of equity of the approved clients (avg. ROE);
- average return of equity of the approved clients weighted by the loan amounts (avg. vol. ROE).

Similar Loan Characteristics

First for the case of the similar loan characteristics simulation the results are summarized in table 3.3 for the case of approving similar number of cases and in table 3.4 for the case of similar approved volume.

Table 3.3: Results of the similar loan characteristics simulation and similar number of approved clients

Method	PD	EAP	ERP	IRR	ROE
Q75	6.55%	751	2.14%	0.29%	1.95%
Apr. volume	130,722,330	148,914,232	147,462,032	147,185,232	146,668,232
Num. apr.	2,124	2,126	2,126	2,124	2,126
Avg. PD	3.29%	4.13%	4.02%	3.93%	3.92%
Avg. vol. PD	3.22%	4.17%	4.09%	4.01%	3.98%
Avg. EAP	6,559	7,198	7,190	7,186	7,175
Sum. EAP	13,930,454	15,303,190	15,286,101	15,262,184	15,253,516
Avg. ERP	8.47%	9.60%	9.61%	9.60%	9.60%
Avg. vol. ERP	10.66%	10.28%	10.37%	10.37%	10.40%
Avg. IRR	0.48%	0.51%	0.51%	0.51%	0.51%
Avg. vol. IRR	0.55%	0.53%	0.54%	0.54%	0.54%
Avg. ROE	3.81%	4.10%	4.11%	4.12%	4.12%
Avg. vol. ROE	4.45%	4.31%	4.34%	4.34%	4.35%

As we can see from table 3.3, when using the profitability measures for approval, the risk of the portfolio generally rises (measured by the avg. PD or avg. vol. PD), but the volume and profitability of the business rise as well (except for the Avg. vol. IRR). Moreover, we can see that the use of each characteristic maximizes the its average value on the approved portfolio (and Sum. EAP respectively).

On the other hand, when we simulate the approved volume to be on the same level as for the PD approval (results in table 3.4), the number of approved cases decreases and the measures optimize their weighted average (and Avg. EAP respectively). For example when using the ERP measure, we can conclude that this approach can bring a relative increase of 5.4% in the expected relative profitability of the business (as 11.24%/10.66% = 1.0544).

Table 3.4: Results of the similar loan characteristics simulation and similar approved volume

Method	PD	EAP	ERP	IRR	ROE
Q75	6.55%	2 745	4.79%	0.39%	2.93%
Apr. volume	130,722,330	130,722,456	130,801,178	130,787,018	130,750,018
Num. apr.	2,124	1,703	1,793	1,791	1,808
Avg. PD	3.29%	3.80%	3.65%	3.40%	3.40%
Avg. vol. PD	3.22%	3.90%	3.70%	3.50%	3.49%
Avg. EAP	6,559	8,570	8,196	8,134	8,068
Sum. EAP	13,930,454	14,594,614	14,695,586	14,567,706	14,587,449
Avg. ERP	8.47%	10.93%	10.75%	10.63%	10.60%
Avg. vol. ERP	10.66%	11.16%	11.24%	11.14%	11.16%
Avg. IRR	0.48%	0.54%	0.54%	0.54%	0.54%
Avg. vol. IRR	0.55%	0.56%	0.56%	0.56%	0.56%
Avg. ROE	3.81%	4.42%	4.38%	4.42%	4.40%
Avg. vol. ROE	4.45%	4.53%	4.55%	4.58%	4.58%

Progressive Loan Characteristics

Now for the case of the progressive loan characteristics simulation the results are summarized in table 3.5 for the case of approving similar number of cases and in table 3.6 for the case of similar approved volume.

Table 3.5: Results of the progressive loan characteristics simulation and similar number of approved clients

Method	PD	EAP	ERP	IRR	ROE
Q75	6.55%	1,945	4.02%	0.30%	3.81%
Apr. volume	130,722,330	147,141,936	143,330,853	143,144,242	148,097,682
Num. apr.	2,124	2,126	2,126	2,124	2,126
Avg. PD	3.29%	5.53%	5.73%	5.91%	5.34%
Avg. vol. PD	3.22%	5.24%	5.41%	5.51%	5.13%
Avg. EAP	6,952	10,475	10,429	10,384	10,459
Sum. EAP	14,766,331	22,270,030	22,171,502	22,055,569	22,235,567
Avg. ERP	11.46%	17.90%	17.96%	17.88%	17.80%
Avg. vol. ERP	11.30%	15.14%	15.47%	15.41%	15.01%
Avg. IRR	0.57%	0.81%	0.82%	0.82%	0.81%
Avg. vol. IRR	0.54%	0.68%	0.69%	0.70%	0.68%
Avg. ROE	6.94%	8.23%	8.19%	8.15%	8.24%
Avg. vol. ROE	6.74%	7.11%	7.16%	7.15%	7.10%

From this comparison we can see a much higher differentiation of the outcome portfolio characteristics. Here both the risk and profitabilities increase substantially when using the profitability measures.

To compare with the first example, when again using the ERP measure to ap-

Table 3.6: Results of the progressive loan characteristics simulation and similar approved volume

Method	PD	EAP	ERP	IRR	ROE
Q75	6.55%	3,437	5.09%	0.33%	4.34%
Apr. volume	130,722,330	130,742,819	130,748,422	130,961,742	130,821,602
Num. apr.	2,124	1,795	1,932	1,947	1,783
Avg. PD	3.29%	5.78%	5.96%	6.18%	5.55%
Avg. vol. PD	3.22%	5.44%	5.66%	5.81%	5.29%
Avg. EAP	6,952	11,919	11,177	11,016	11,911
Sum. EAP	14,766,331	21,394,748	21,593,532	21,448,993	21,238,048
Avg. ERP	11.46%	20.07%	19.30%	19.07%	20.04%
Avg. vol. ERP	11.30%	16.36%	16.52%	16.38%	16.23%
Avg. IRR	0.57%	0.87%	0.86%	0.87%	0.89%
Avg. vol. IRR	0.54%	0.72%	0.73%	0.73%	0.72%
Avg. ROE	6.94%	8.70%	8.58%	8.51%	9.04%
Avg. vol. ROE	6.74%	7.41%	7.40%	7.37%	7.50%

prove a similar volume of loans, we can conclude that this approach can bring a relative increase of 46.2% in the expected relative profitability of the business. On the other hand the portfolio risk would increase 75.8%. This is mainly connected to fact that the risk based pricing was set in the way that the most risky clients got such interest rate that they become more profitable than the low risk clients (this follows from the negative correlation).

However, from this comparison we can see that for specific market conditions the difference between risk and profitability of loans can be big and a profitability based underwriting model can bring a substantial increase of loan profitability. Here it is important to note that the situation is more complicated in the real business and any potential implementation of such model needs to be well evaluated based on various business aspects.

3.8 Conclusions

The aim of this chapter was to create a complex model that estimates the profitability of every single consumer loan in the approval process, based on the clients' and loans' characteristics. For this model I suggest four profitability measures that can be used for the approval decision and simulate the outcome of the approval process based on these measures.

I believe that these simulations in the consumer loan area contribute to the recent research about profitability models. Moreover, the specifically derived formulas simplify the implementation of such model and enable the loan providing companies to optimize the expected profit from the customer base instead of just minimizing the risk. Such an approach can bring additional profitability for the loan providing company as well as for its shareholders.

For the model building I used a lot of inspiration from my experience in the consumer loan business and managed to combine it with my research in the probability of default estimation and recovery rate modeling. Especially the Cox model used in the probability of default modeling enables to estimate the probability of default of a specific customer on every payment, which is one of the key inputs of the profitability model.

I took the basic idea of the profitability models described in (Allen et al., 2004) and (Stein, 2005), explored the theory and applied its logic into the consumer loan scheme that I know from my professional experience. I derived all the formulas in this chapter and incorporated several other potential costs and revenue streams. Then I defined four profitability measures and assessed them for practical use. Finally, I constructed a data analysis where I compared the performance of the profitability model with the standard risk based approval approach. Even though the data was partially simulated, the analysis shows that the profitability model brings an additional profit to the company, especially under some specific market conditions. On the other hand the risk of the portfolio rises as well, which needs to be well evaluated before such approach is used.

In the real business, these profitability measures can be used when the system is automatically deciding about the approval or rejection of an existing loan application; however, often the system is just pre-approving some applications and deciding about their further approval process. One of modern trends is then a dynamic calculation of the loan interest rate based on clients' characteristics, where this interest rate is then offered to the client. This is called dynamic scoring or dynamic pricing, and it can be done either after a specific loan application, or continuously calculated for the whole existing customer portfolio.

With this profitability model, the use of dynamic pricing is very straightforward. The loan providing company can set a minimal or optimal value of any chosen profitability measure, and the corresponding interest rate can be computed automatically. Then clients can get their tailor-made interest rate immediately after the loan application.

This approach can be further enhanced by considering other potential inputs for the model. One potential input to be considered can be estimation of the probability of early repayment, that can decrease the expected revenue quite significantly. Another improvement can be a calculation of the expected revenues from future loans of the customers. This is called cross selling and it is quite common that the company is willing to sacrifice some profit in the first loan (e.g. by a decreased pricing) to get more profit from the cross sell loans. Finally, this presented approach only works with the expected characteristics and not the full distribution of the potential outcomes. By considering the full distribution (often assumed to be normal) one could get to the value at risk approach, i.e. estimation of profit with a confidence level. Such approach is used for example in (Crouhy et al., 2000).

Chapter 4

Default Concentrations Discovery

In this chapter I aim to discuss some of the weaknesses of the scoring and profitability models described in the previous chapters and propose alternative methods that can partially treat the consequences of these weaknesses. Consequently, I also aim to solve the fraud discovery problem from the credit risk practice. My solution consists of two main parts – first I define a measure that in my opinion best distinguishes the fraudulent segments, and second I discuss and compare several methods for finding such segments in a big portfolio of data.

When using the models of the previous chapters, one could get under impression that such comprehensive models will maximize the profit and prevent the company from all potential losses. This is in reality not true and many more further actions need to be taken to discover default concentrations in the portfolio and prevent various forms of individual or organized fraud.

In the financial practice this fraud monitoring and prevention scheme is standardly based on a set of one-dimensional or two-dimensional analyses or reports. However, these simple conditions are often unable to discover a more complicated fraud pattern and more advanced data mining methods are needed. In the scholarly literature, this topic is usually neglected and up to my knowledge, there are no publications solving this particular problem.

In this chapter I aim to define three default severity measures, that can be used to identify severe default concentrations, and discuss several approaches to find these concentrations in a big data set. Finally, two methods are compared on a sample of real financial data.

4.1 Introduction

If we shortly have a look at how the risk or profit of the client is evaluated by the models, we can see that most of the model setup is based on the statistics and expected values. Specifically, a scoring function combines the risk characteristics of the clients to compute the score – then for example according to figures 1.4, 1.5 and 1.6 we can conclude that a highly educated middle-aged married person will have statistically a very low risk, but does is mean there cannot be any default concentrations in some region or product?

The statistics of the scoring models works well on the whole portfolio; however, it can easily happen that a relatively small segment of the portfolio will not be reflected by the scoring model, even if it is already in the scoring development sample. This could be illustrated by the fact that even if this segment with very high delinquency is discovered and a binary segment indicator is created, this predictor would have a very low Gini value due to the low number of observations in the risky category. This also means that such a predictor will not be selected to the final model, because the Wald tests would not reject the null hypothesis.

Moreover, in reality it can happen that there is an organized group of applicants that uses the weaknesses of the underwriting scheme and intentionally provide the application data in the way that it increases the chance of approval or increases the credit amount the company is willing to lend (e.g. by providing fake information that is difficult to verify, or by increasing the expected profit by a higher interest rate or an insurance). This is considered as a *fraud behavior* and it is often connected with high delinquencies of this group (either because of repayment problems or because of no intention to pay at all). And vice versa, a very high concentration of default often points to a fraud activity.

This is the main motivation for this chapter. Whereas the underwriting process, based on the models from the previous chapters and other manual activities, focuses on revealing fraudulent and insolvent clients prior approval to reject their loan application in time, there are other processes more oriented to discover fraud patterns and default concentrations after the loans are issued. For this purpose, clients' delinquencies are measured on daily basis, and concentrations on the most important dimensions (such as products, regions, branch offices, credit agents etc.) are reported by automated alarm systems.

However, in reality fraud schemes can be very sophisticated to avoid these basic concentration triggers, and the fraud is thus not detected by the alert system. Moreover, in a big portfolio the delinquency level can be "diluted" by a large number of good clients in the same category.[1] Therefore, one would like to find a method

[1] As an example, imagine there is a fraud attack on one specific branch office for a particular product. However, since this branch office has a big portfolio of clients and products, the overall delinquency level of the office is not that high, and thus does not activate the concentration trigger.

to identify specific segments of the portfolio (represented by various combinations of clients' and products' characteristics) with very high level of delinquency (i.e. a probable concentration of fraud).

This is an uneasy task that consists of two main questions. First, how to distinguish the high default rate segments. And second, how to systematically search and find them.

Before answering the questions, I make the following denotations in the four-fold table 4.1 defined by the investigated segment and the number of defaults. In this table a is the number of defaults in the segment, b is the number of non-defaults in the segment, c is the number of defaults outside of the segment, and d is the number of non-defaults outside of the segment. Then $k = a + c$, $l = b + d$, $r = a + b$ and $s = c + d$ are the marginal sums and $n = a + b + c + d$ is the total number of cases. Then obviously $\frac{a}{r}$ is the default rate of the segment and $\frac{k}{n}$ is the default rate of the whole portfolio.

Table 4.1: Four-fold table for segments

Segmentation	Number of defaults	Number of non-defaults	Total
Segment	a	b	r
The rest of portfolio	c	d	s
Total	k	l	n

4.2 Default Severity Measures

In this section I aim to solve the first question and provide some techniques to recognize the segments with high default rate and possible fraud. The problem is that the segments are generally small and a simple default rate is not sufficient (obviously, if we observe 100% default rate on one observation, we cannot conclude that there is a significant fraud). Therefore, I propose several alternatives combining the default rate with the number of cases.

4.2.1 Basic Trigger

The simplest and most understandable way of selecting the severe segments is a trigger based on the default rate and the number of cases, i.e. fulfill the trigger if

$$\frac{a}{r} > q_1 \qquad \text{and} \qquad r > q_2,$$

or alternatively

$$\frac{a}{r} > q_1 \qquad \text{and} \qquad a > q_3,$$

where q_1, q_2 and q_3 are some pre-set constants.

The logic of this trigger is based on the fact that the higher default rate in a segment the more serious default or fraud we observe; combined with the requirement for a minimal sample size, where the default rate is giving us reliable information.[2] Moreover, the more default cases in the segment, the higher loss can be potentially cured.

This is a very easy and well understandable trigger. On the other hand, I discuss some disadvantages as well. Apart from the fact that this trigger cannot order the data by fraud severity (i.e. there is no difference between 30% default rate on a 1,000 and a 10,000 segment), it is not even guaranteed that the triggered segments are really the most relevant in the portfolio. I illustrate this by an example. Assume that I define a trigger as $q_1 = 10\%$ default rate with at least $q_3 = 10$ defaults. Then a segment of 10 defaults out of 100 observations is triggered, whereas a segment with 9 observations and all of them defaults is not. From the rational point of view I would expect that a segment with 9 observations where all of them defaulted is more suspicious than the first segment.

Even though more advanced measures are introduced in the following sections, the basic trigger is very often used in practice for its simplicity and understandability for all the people involved in the system.

4.2.2 χ^2 and Fisher's Factorial Tests

In most of the common statistics textbooks, e.g. (Anděl, 2007), we can find the χ^2 independence test's variant for a four-fold table, that can be applied to table 4.1. It uses the fact that for i.i.d. random variables the following formulas have asymptotically the χ^2_1 distribution,

$$\chi^2 = n\frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}} \sim \chi^2_1,$$

which can be rewritten in my denotation as

$$\chi^2 = n\frac{(ad - bc)^2}{rskl} \sim \chi^2_1.$$

Then it is required that all of the theoretical frequencies $\frac{n_{i\cdot}n_{\cdot j}}{n}$ are greater than 5. For the cases when this condition is not fulfilled, one can use the Fisher's exact factorial test as an alternative, see (Anděl, 2007).

[2] As the matter of fact many business analysts in practice actually use the default rate only for the severity identification, but if the observation is done on an insufficient number of cases the outcomes might be misleading.

For both of these tests we can understand the p-value of the test as some kind of significance score and then order the data by this score to select the most significant segments for investigation. Alternatively, we can construct a trigger as

$$\text{p-val}_{\chi^2} < q_4,$$

where q_4 is a pre-set constant.

These tests can now solve some of the problems of the basic trigger; however, the property of these tests is that they test the significance of the difference of the segment's default rate comparing to the overall data, which does not necessarily mean a higher proportion of fraud. As an example we can take the credit amount predictor with two categories (lower amount and higher amount) of similar category size. Then in the data we can observe that the default rate of the higher amount category is slightly higher compared to the overall portfolio. Then due to the fact that both categories have a high number of observations, the difference is very significant and the p-value of the test is small. However, this slight difference of default rate is natural (can be caused by people having problems to repay higher payments connected to higher credit amounts) and does not suggest a fraud presence at all. This is the motivation for constructing the following test.

4.2.3 Expert Binomial Test

For this test I incorporate an expert evaluation of the fraud severity level and combine it with a statistical test. The merit of this test is as follows. An expert sets a value f to be a *fraud rate* or *fraud threshold*, i.e. such a value that any segment with default rate significantly exceeding this threshold will be considered as *affected by fraud*. This fraud rate has to be reflecting the portfolio characteristics (e.g. three times of the portfolio average) and the purpose of the particular analysis.

Additionally, I assume that if there is no fraud in the segment (i.e. the segment default rate is lower or equal to f), then the defaults happen independently with probability p. With this assumption, for a segment with an observed default rate $p_s > f$ I try to evaluate the probability that this segment is affected by fraud.

More rigorously, I construct a statistical test with $H_0 : p = f$ and alternative $H_A : p > f$. Then if a is the number of defaults and r is the number of observations in the segment, I use the assumptions introduced in the previous paragraph to compute the p-value of the test. Under the null hypothesis I know that the defaults in the segment follow the binomial distribution with parameters f and r, $\text{Bi}(f, r)$. Then the p-value of his test is the probability that such or worse situation happens under the null hypothesis (by accident), i.e. that under the null hypothesis there will happen at least a defaults out of r events. Then using the binomial distribution the p-value can be computed as

$$\text{p-val} = \sum_{i=a}^{r} \binom{r}{i} f^i (1-f)^{r-i}.$$

Here the p-value could be interpreted as the probability that such high default rate in the segment happened by accident. Therefore, the lower the p-value the more significant fraud suspicion we have. Then with the above definition we can interpret $1 - \text{p-val}$ as the probability that the segment is affected by fraud. Based on the p-value we can again construct a trigger

$$\text{p-val}_f < q_5,$$

where q_5 is a pre-set constant.

With this measure we can again construct the significance score and order the segments by their relevance (unlike with the basic trigger), we do not have any requirements on the sample size (unlike with the four-fold table χ^2 test), and we define our own fraud rate to prevent triggering small differences in default rate on very big segments (unlike with the χ^2 test and Fisher's factorial test). Moreover, the p-value is simple to be computed in MS Excel on any sample of data.[3] On the other hand, in practice, the statistical tests and their p-values are more complicated and their usage can cause some interpretation difficulties (unlike with the basic trigger). Anyway, I personally consider this method to be the most relevant for the fraud or default concentrations discovery purpose.

4.3 Severe Segments Finding Method

In this section I describe a practical method that can be used for finding the segments with fraud or default concentrations and discuss its differences to other methods commonly used methods. Usually, the input for this task can be a table (e.g. in Excel or database) with a list of clients or loans in rows and a lot of categorized characteristics in columns. Moreover, for every client or loan there is a default indicator (e.g. 1 if the client defaulted on one of the first payments for more than 30 days past due, or 0 otherwise).

Then all segments are created as various combinations of the characteristics' categories. For each segment se can see the number of cases, the number of defaults and the segment's default rate. Then we can use the default severity measures from the previous section to decide, whether an evaluated segment satisfies the trigger and should be investigated or not.

[3] The p-value can be expressed by the formula `1 - BINOMDIST(a,r,f,TRUE) + BINOMDIST(a,r,f,FALSE)`.

Now one could try to look for the affected segments manually, use some heuristic method like decision trees, or take advantage of some data mining tools. As this is a special case of supervised learning or anomaly detection, some kind of big data methods can be used. From the great variety of data mining tools (see e.g. (Fayyad et al., 1996), (Witten and Frank, 2005) or (Han et al., 2011)) I use the GUHA method in association analysis, that I shortly describe bellow. For more information about the big data topic I refer to (Baesens, 2014), (Mayer-Schonberger and Cukier, 2013), (Zikopoulos et al., 2011).

4.3.1 GUHA Method Introduction

In this section I briefly describe the logic of the association analysis based on the *General Unary Hypotheses Automaton Method (GUHA)*. Most information about this method and the used terminology is taken from (Rauch and Šimůnek, 2005a), (Rauch and Šimůnek, 2005b) and the manuals to the LISP Miner software.

GUHA is originally a Czech data mining method introduced in (Hájek et al., 1966). Its aim is to systematically formulate all hypotheses of a suggested structure and evaluate them using a given data sample and a pre-defined trigger condition. The hypotheses or *association rules* are expressions of the form $X \longrightarrow Y$, where the fulfillment of the condition X (also called *antecedent*) tends to the result Y (also called *succedent*). From all the hypotheses only the rules with required *support* (i.e. number of cases of such property) and *confidence* (the percentage of cases of property X leading to result Y) are chosen to the final output.

More specifically, the four-fold table 4.1 can be now written in the form of table 4.2. Then a is the support and $\frac{a}{r}$ is the confidence of the association rule $X \longrightarrow Y$.

Table 4.2: Four-fold table for GUHA method

Attribute	Y	$\neg Y$	Total
X	a	b	r
$\neg X$	c	d	s
Total	k	l	n

Then if we define the minimum required support and minimum required confidence, the method filters only those segments fulfilling this condition. Specific implementations of the GUHA method allow other conditions as well.

The GUHA method is based on the principle that *all* the *relevant* segments of data are systematically evaluated – “all” in the meaning that no segment fulfilling the trigger condition is omitted, and “relevant” in the meaning that the algorithm is optimized to skip the creation of segments that cannot possibly fulfill the trigger condition (e.g. the segment is already too small or has too few defaults that it makes no sense to split it according to any further categories). Moreover the method can usually work with big data samples exploring millions of hypotheses.

4.3.2 Use in Fraud Discovery

The GUHA method thus gives us a good framework to discover default concentrations (possible fraud) hidden in some smaller segments of the data. For the succedent attribute we take default indicated from the data sample and for the antecedent attributes X we take all possible combinations of predictors up to some limitation (e.g. no more than 4 predictors to be combined in the same hypothesis).

The GUHA method is implemented for example in the 4ft-Miner procedure of the LISP-Miner software[4] developed at the University of Economics in Prague.

The LISP-Miner software allows the user to define additional settings of the task including various modifications of the trigger conditions, ways to combine and merge the predictors' categories (e.g. combine only the neighboring categories of ordinal predictors etc.). Moreover, any follow-up tasks can be done in MS Excel, including ordering by the significance score based on the p-value of the expert binomial test.

4.3.3 Theoretical Difference to Other Methods

Now what makes this method different from the more commonly used methods such as discrimination analysis, logistic regression or decision trees? Mainly it is the fact that these methods are designed to work well on the whole sample and thus use the strongest predictors in the terms of the overall discrimination power (often represented by the Gini or lift characteristics). As an example, we can say that the predictor gender would be selected to the scoring model, whereas the predictor of business branch name wouldn't (since it can have thousands of small categories quite impossible to categorize into bigger reasonable segments). Moreover, for many models the predictors are not easily combined and selected combinations have to be driven manually.

From the perspective of finding the interesting "outlying" segments, a good help can be expected from the decision (also regression or classification) tree model, where there are small segments found in the leaves. However, the regression three algorithm is a heuristic method growing the tree according some predefined measure, and thus cannot check all possible combinations (as the GUHA method does). Therefore, it can happen that the tree is right in the root divided into several branches according to some strong predictor (e.g. the gender, mentioned above) and a significant fraud

[4] More information can be found at http://lispminer.vse.cz.

segment (which can be independent on that predictor) is then cut into smaller pieces that are not significant anymore (due to their size).

On the other hand, for the practical use, the decision three method gives us a distinct set of segments that are easily evaluated, whereas the GUHA methods gives us the full set of all possible overlapping segments that we need to analyze further. In the next section I present a practical comparison of the GUHA method with the classification three on the real financial data.

4.4 Data Analysis

In this section I aim to compare the GUHA method with the classification tree model on a real financial data sample. The data for this purpose is provided by a financial company operating on a foreign consumer finance market. Some of the original characteristics' labels are not provided to prevent the company's know-how. Therefore, I will treated them as undisclosed.

4.4.1 Data Structure

The provided sample contains the data about 161,786 approved loans, where for each loan we have the target variable defining the default for this task as 30 days past due on the first payment and 18 categorized predictors. This sample contains 5,328 defaults representing 3.3% default rate. Any missing data is categorized as a special category. See table 4.3 for details.

4.4.2 Comparison of Methods

On this data sample I run the GUHA method and the classification tree to find the default concentrations.

GUHA Method

For this comparison I used the GUHA method implemented in the 4ft-Miner procedure of the LISP-Miner software. For the purpose of this analysis, the LISP-Miner was set to try all the possible combinations of 1–4 out of the 18 predictors.[5]

[5] I need to limit the number of predictors to set the range of the task. The used combination of up to 4 predictors seems most relevant given the sample size and interpretability of the results.

Table 4.3: Data structure

Predictor	Number of categories
Age	9
Credit amount	9
Distribution channel	2
Down payment	6
Family state	3
Goods type	6
House type	5
Income	7
Insurance	2
Price	9
Region	29
Term	9
Undisclosed predictor 1	2
Undisclosed predictor 2	2
Undisclosed predictor 3	5
Undisclosed predictor 4	2
Undisclosed predictor 5	3
Undisclosed predictor 6	2
All 18 predictors	112

Then the model evaluated 4,047 combinations of predictors with over 900,000 relevant hypotheses and returned 5,467 segments with the required confidence of 10 defaults and support at the level of 10% (i.e. about three times the portfolio average). The whole task took 16 minutes on a common laptop.

Classification Tree

For this comparison I used the classification tree method implemented in the HPSPLIT procedure of SAS 9.4., where I selected the maximum tree depth 4 (to be in line with 1–4 predictors in the GUHA method), the maximum number of children per node 30 (to enable the split to individual regions) and the decrease in entropy as the splitting criterion.

This task took about 30 seconds in SAS on a common laptop. The result was a tree with 4 levels, 770 nodes and 3,178 leaves. Among these leaves and nodes there were 30 segments with the required confidence of 10 defaults and 10% support.

Comparison

I exported the outputs of both methods for further evaluation in MS Excel, were I also computed the p-value of the binomial trigger with the fraud rate $f = 10\%$ to identify the most relevant segments.

The two most relevant segments from the classification tree analysis contained

- 49 defaults from 91 cases, i.e. 54% default rate with the binomial test p-value about $1.9 \cdot 10^{-25}$ and
- 79 defaults from 235 cases, i.e. 34% default rate with the binomial test p-value about $5.8 \cdot 10^{-23}$.

When I searched for these segments in the GUHA analysis output, these were found as the fifth and the tenth most relevant segments.

On the other hand, when analyzing the most relevant segment of the GUHA method containing

- 99 defaults from 229 cases, i.e. 43% default rate with the binomial test p-value about $6.3 \cdot 10^{-39}$,

I found out that such segment was not directly found by the classification tree method; however, 98 of these 99 defaults were actually included in the first two classification tree segments shown above. Therefore, I conclude that the most severe fraud pattern was identified by both methods.

Whereas the GUHA method identified the fraud pattern more precisely and with a proper setting of the selection and category combining criteria it can give us a comprehensive and “assuring” information about all the possible fraud patterns in the portfolio, the classification tree method runs much faster and provides cleaner output organized in distinct leaves.

From the results several important patterns were identified – such as a specific combination of the region, credit amount category, down payment and goods type. This was an example where a hidden factor (or fraud pattern) was found by combining the predictors’ categories, even though it was not directly included in the set of predictors.

4.5 Conclusions

The aim of this chapter was to react on the probability-of-default-based scoring models and profitability models described in this thesis, discuss their weaknesses, and propose some tools to identify default concentrations and potential fraud patterns in big portfolios of data. Once those concentrations are found, the underwriting process can be adjusted to prevent further losses.

For this purpose I define several measures that can be used for evaluation of the significance of the default rate in each segment. Here the definition and use of the expert binomial test for the purpose of default concentration identification I consider as novel according to the studied literature.

Then for the task of finding the segments with default concentrations, robust data mining tools can bring more reliable results than the classical approaches of the discrimination analysis, logistic regression or decision trees. Here the GUHA method has been shortly described as a representant of the supervised machine learning techniques, and the LISP-Miner as a convenient free-ware software where this method is implemented. Moreover, some conceptual differences to the classical approaches have been discussed.

Finally, the GUHA method and the classification tree method were applied on a sample set of financial data, where both methods managed to discover suspicious default concentrations. Comparing these two methods the modeler needs to choose between the GUHA method's assurance that really all the potential segments are discovered, and speed and simplicity of the classification tree.

From my professional experience the problem of fraud discovery is a very important and closely watched topic for all credit providing institutions, and all innovations, tools and results that positively affect the fraud prevention mechanisms are well accepted.

Conclusions

In this thesis I aimed to propose new mathematical and statistical methods to enhance the standard credit underwriting automated scoring. Particularly, I sought to challenge the performance criteria based on the ex-post random testing samples and proposed comparing the predictive power of the models on an ex-ante sample of the most recent data instead. Then I wanted to use this new criteria and a real Czech banking data sample to compare the standard models performance with some suggested alternatives. Finally, I aimed to construct a new comprehensive underwriting model that would be based on an estimation of loan profitability instead of the standard evaluation of the riskiness of the client. Such model should have been described in detail, the results simulated and compared with the standard approach and its weaknesses treated by proposed alternative methods.

In the first chapter I dealt with the probability of default modeling and set the new performance criteria focusing on the predictive power of the models, and compared the standard logistic regression model with the alternative of the survival-based Cox model on the real sample of Czech banking data. Based on this comparison I concluded that in this sample both models have similar performance on the random training and testing sample, which is in line with the existing research of (Stepanova and Thomas, 2002), (Cao et al., 2009) or (Bellotti and Crook, 2009); however, if compared by the new performance criteria measuring the predictive power of the model, the Cox model notably outperforms the logistic regression model. This is a new result contributing to the academic debate and showing the Cox model in the new light.

In the second chapter I introduced similar logic as in the first chapter and applied the new performance criteria for measuring the predictive power of loss given default models. Again, I compared the standard approaches using the linear and logistic regressions with one existing and one new Cox-based model on the real Czech banking data. Even this time, when comparing the predictive power of the models on the time-censored sample, the new approaches clearly outperform the standard models in the terms the used goodness of fit measure. Therefore, I believe that there is a good potential for further research as well as for practical application in banks and financial institutions creating their own LGD models to decrease the capital requirement.

The red thread of this thesis was then the creation of the comprehensive underwriting model that estimates the profitability of every single consumer loan in the approval process, based on the clients' and loans' characteristics. For this model, inspired by (Stein, 2005) and (Allen et al., 2004), I derived all the formulas, suggested several alternatives for the probability of default on individual payments estimation (including the straightforward application of the Cox-based models from the first two chapters) and enriched the concept for multiple potential costs and revenue streams that are important in the business. Moreover, I calculated four alternative profitability measures coming from this model, that can each support the priorities of the company. Especially, the risk-adjusted expected return on equity from a specific loan could be a really beneficial measure on some markets, and I have not seen it implemented in practice nor suggested in literature. Finally, to simulate the impact and compare the new model with the standard risk-based approach, I conducted a data analysis, where I compared the performance of the profitability model with the standard risk-based approval model calculating and comparing various risk and profitability characteristics. In this simulation it was shown that under specific market situations the difference between various performance criteria can be crucial and the choice of the right model for the right criteria is substantial. Particularly, when optimizing the expected profitability, the new model brings a significant added value.

Finally, in the last chapter I reacted to the probability-of-default-based scoring models and profitability models described in this thesis, discussed their weaknesses, and proposed some tools designed to treat them by identifying default concentrations and potential fraud patterns in big portfolios of data. For this purpose I defined several measures that can be used for the evaluation of significance of the default rate in each segment, where the definition and use of the expert binomial test and trigger I consider as best fitting the needs of this task and novel according to the studied literature. Also, I discussed several methods for the fraud or default concentration search. Finally, I compared the GUHA method with the classification tree model on the real financial data sample.

By the definition of the new performance criteria, the thesis sought to contribute to the scholarly debates on mathematical modeling in finance and showed how this new criteria can change the outcomes of the probability of default and the loss given default models comparisons. This paves the way for further research to study the strengths and weaknesses of the survival analysis models and apply it to a broader variety of financial data to test the robustness of the claim. Furthermore, by constructing a comprehensive analytical framework for modeling the loan-cycle and default structure, the thesis offered an innovative way of operationalizing the concept of profitability modeling. Last but not least, the thesis identifies a common problem of regression-based models which is the inability to address concentrated outliers and provides a solution in the area of default concentration discovery.

Apart from the academic contributions, the thesis points out to the advanced predictive power of the survival-based models used for probability of default modeling and provides a new model for loss given default modeling with better predictive power. Together with the specific description of the profitability model and innovations in the field of fraud and default concentration discovery, the application of this research can prevent substantial losses and bring added profit to the financial companies. In effect, such improvements might enable more personalized approach and lower interest rates in the consumer finance industry as a whole.

As the main directions for further research I see the testing and comparison of all the models on multiple data samples (both real and simulated) to be able to draw more general conclusions about the out-performance of the models. Moreover, the prediction horizons should be extended on the real data to understand the stability of such models in time. Finally, a big potential for any further research I personally see in the direction of the proper evaluation of the whole loan profit probability distribution, not only the expected value as is done in this model. This is a very complex task that involves mainly a proper model to capture the correlation structure in the client portfolio. Even though this approach is more important for evaluation of financial derivatives, see e.g. (Hull, 2009a), (Embrechts et al., 2002) or (Rychnovský, 2012), I believe that similar ideas can be applied to retail risk modeling as well.

Bibliography

Agresti, A. (1990). *Categorical data analysis*. John Wiley & Sons, Inc. ISBN 978-04-718-5301-5.

Allen, L., DeLong, G., and Saunders, A. (2004). Issues in the credit risk modeling of retail markets. *Journal of Banking & Finance*, volume 28, no. 4, pages 727–752.

Anděl, J. (2007). *Základy matematické statistiky*, volume 1. Matfyzpress.

Asarnow, E. and Edwards, D. (1995). Measuring loss on defaulted bank loans: A 24-year study. *The Journal of Commercial Lending*, volume 77, no. 7, pages 11–23.

Baesens, B. (2014). *Analytics in a big data world: The essential guide to data science and its applications*. John Wiley & Sons.

Banasik, J., Crook, J.N., and Thomas, L.C. (1999). Not if but when will borrowers default. *Journal of the Operational Research Society*, volume 50, no. 12, pages 1185–1190.

Basel II (2001). The new Basel capital accord [online]. Basel Committee on Banking Supervision, Bank for International Settlements, http://www.bis.org/publ/bcbsca03.pdf (2011-08-05).

Bellotti, T. and Crook, J. (2009). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society*, volume 60, no. 12, pages 1699–1707.

Belyaev, K., Belyaeva, A., Konecný, T., Seidler, J., and Vojtek, M. (2012). Macroeconomic Factors as Drivers of LGD Prediction: Empirical Evidence from the Czech Republic. *Working Paper Series of the Czech National Bank*, volume 12.

Bonini, S. and Caivano, G. (2012). Beyond basel2: Modeling loss given default through survival analysis. In *Mathematical and Statistical Methods for Actuarial Sciences and Finance*, pages 43–52. Springer.

Bonini, S. and Caivano, G. (2013). The survival analysis approach in Basel II credit risk management: modeling danger rates in the loss given default parameter. *The Journal of Credit Risk*, volume 9, no. 1, page 101.

Brandimarte, P. (2003). *Numerical methods in finance: a MATLAB-based introduction*, volume 489. John Wiley & Sons.

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

Cannaday, R.E., Colwell, P.F., and Paley, H. (1986). Relevant and irrelevant internal rates of return. *The Engineering Economist*, volume 32, no. 1, pages 17–38.

Cao, R., Vilar, J.M., and Devia, A. (2009). Modelling consumer credit risk via survival analysis. *SORT*, volume 33, no. 1, pages 3–30.

Colwell, P.F. (1995). Solving the dual irr puzzle. *Journal of Property Management*, volume 60, no. 2, pages 60–62.

Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, volume 34, no. 2, pages 187–220. ISSN 1467-9868.

Crouhy, M., Galai, D., and Mark, R. (2000). A comparative analysis of current credit risk models. *Journal of Banking & Finance*, volume 24, no. 1, pages 59–117.

Derksen, S. and Keselman, H. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, volume 45, no. 2, pages 265–282.

Embrechts, P., McNeil, A., and Straumann, D. (2002). Correlation and dependence in risk management: properties and pitfalls. *Risk management: value at risk and beyond*, pages 176–223.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (1996). *Advances in knowledge discovery and data mining*, volume 21. AAAI press Menlo Park.

Flemming, J. and Wright, J. (1971). Uniqueness of the internal rate of return: A generalisation. *The Economic Journal*, volume 81, no. 322, pages 256–263.

Glennon, D.C. and Nigro, P. (2005). Measuring the default risk of small business loans: A survival analysis approach. *Journal of Money, Credit, and Banking*, volume 37, no. 5, pages 923–947.

Gupton, G.M. (2005). Advancing loss given default prediction models: how the quiet have quickened. *Economic Notes*, volume 34, no. 2, pages 185–230.

Gupton, G.M., Gates, D., and Carty, L.V. (2000). Bank loan loss given default. *Moody's Investors Service, Global Credit Research, November.*

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

Hájek, P., Havel, I., and Chytil, M. (1966). The guha method of automatic hypotheses determination. *Computing*, volume 1, no. 4, pages 293–308.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hanley, J., McNeil, B., et al. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, volume 148, no. 3, pages 839–843. ISSN 1527-1315.

Ho, T. and Lee, S. (2004). *The Oxford Guide to Financial Modeling: Applications for Capital Markets, Corporate Finance, Risk Management and Financial Institutions*. Oxford University Press. ISBN 9780199727704.

Homer, S. and Sylla, R. (2011). *A History of Interest Rates*. Wiley Finance. Wiley. ISBN 9781118046227.

Hosmer, D. and Lemeshow, S. (2000). *Applied logistic regression*. Wiley-Interscience. ISBN 978-04-713-5632-5.

Huang, X. and Oosterlee, C.W. (2011). Generalized beta regression models for random loss given default. *The Journal of Credit Risk*, volume 7, no. 4, page 45.

Hull, J. (2009a). *Options, Futures and Other Derivatives*. Options, Futures and Other Derivatives. Pearson/Prentice Hall. ISBN 9780136015864.

Hull, J. (2009b). *Risk management and financial institutions*. Pearson Prentice Hall, 2nd edition. ISBN 978-01-361-0295-3.

Kalbfleisch, J., Prentice, R., and Kalbfleisch, J. (1980). *The statistical analysis of failure time data*. Wiley New York. ISBN 978-04-710-5519-8.

Kim, J. and Kim, H. (2006). Loss given default modelling under the asymptotic single risk factor assumption. *MPRA Paper*.

Le Cessie, S. and Van Houwelingen, J. (1994). Logistic regression for correlated binary data. *Applied Statistics*, pages 95–108.

Lehmann, E. and Casella, G. (1998). *Theory of point estimation*. Springer Verlag. ISBN 978-03-879-8502-2.

Loterman, G., Brown, I., Martens, D., Mues, C., and Baesens, B. (2012). Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting*, volume 28, no. 1, pages 161–170.

Louzada, F., Cancho, V.G., De Oliveira Jr, M.R., and Bao, Y. (2014). Modeling time to default on a personal loan portfolio in presence of disproportionate hazard rates. *Journal of Statistics Applications & Probability, An International Journal. J. Stat. Appl. Pro*, volume 3, pages 1–11.

Mayer-Schonberger, V. and Cukier, K. (2013). *Big data: a revolution that will transform how we live, work and think.* John Murray, London.

Narain, B. (1992). Survival analysis and the credit granting decision, In L.C. Thomas, J. Crook N., D.B. Edelman (eds.), Credit scoring and credit control. OUP, Oxford, UK.

Nehrebecka, N. et al. (2016). Approach to the assessment of credit risk for non-financial corporations. Evidence from Poland. *IFC Bulletins chapters*, volume 41.

Pazdera, J., Rychnovský, M., and Zahradník, P. (2009). *Survival analysis in credit scoring.* Unpublished seminar paper, Faculty of Mathematics and Physics, Charles University, Prague.

Persson, I. (2002). *Essays on the assumption of proportional hazards in Cox regression.* Acta Universitatis Upsaliensis.

Peto, R. (1972). Contribution to the discussion of a paper by D.R. Cox. *Journal of the Royal Statistical Society*, volume 34, pages 205–207.

Prívara, S., Kolman, M., Witzany, J., et al. (2014). Recovery rates in consumer lending: Empirical evidence and the model comparison. *Bulletin of the Czech Econometric Society*, volume 21.

Promislow, S. (2014). *Fundamentals of Actuarial Mathematics.* Wiley Desktop Editions. Wiley. ISBN 9781118782521.

Rauch, J. and Šimůnek, M. (2005a). An alternative approach to mining association rules. In *Foundations of Data Mining and Knowledge Discovery*, pages 211–231. Springer.

Rauch, J. and Šimůnek, M. (2005b). Guha method and granular computing. In *2005 IEEE International Conference on Granular Computing*, volume 2, pages 630–635. IEEE.

Reisnerová, S. (2004). Analýza přežití a coxův model pro diskretní čas. In *Robust*, volume 13, pages 339–346.

Řezáč, M., Řezáč, F., et al. (2011). How to measure the quality of credit scoring models. *Finance a úvěr: Czech Journal of Economics and Finance*, volume 61, no. 5, pages 486–507.

Rychnovský, M. (2009). *Matematické modely LGD (Mathematical models for LGD).* Master Thesis, Faculty of Mathematics and Physics, Charles University, Prague.

Rychnovský, M. (2011). *Scoring Models in Finance.* Master Thesis, Faculty of Informatics and Statistics, University of Economics, Prague.

Rychnovský, M. (2012). *Portfolio credit risk models.* LAP LAMBERT Academic Publishing. ISBN 978-3-8454-4137-5.

Rychnovský, M. (2015). Mathematical models for loss given default estimation. *Sborník prací vedeckého semináre doktorského studia FIS VŠE*, pages 103–111.

Schuermann, T. (2004). What do we know about loss given default?

Somers, R. (1962). A new asymmetric measure of association for ordinal variables. *American Sociological Review*, volume 27, no. 6, pages 799–811. ISSN 0003-1224.

Stein, R.M. (2005). The relationship between default prediction and lending profits: Integrating roc analysis and loan pricing. *Journal of Banking & Finance*, volume 29, no. 5, pages 1213–1236.

Stepanova, M. and Thomas, L. (2002). Survival analysis methods for personal loan data. *Operations Research*, volume 50, no. 2, pages 277–289.

Therneau, T.M. and Grambsch, P.M. (2000). *Modeling survival data: extending the Cox model.* Springer Science & Business Media.

Thomas, L.C., Matuszyk, A., So, M.C., Mues, C., and Moore, A. (2016). Modelling repayment patterns in the collections process for unsecured consumer debt: A case study. *European Journal of Operational Research*, volume 249, no. 2, pages 476–486.

Van der Vaart, A. (2000). *Asymptotic statistics.* Cambridge University Press. ISBN 978-05-217-8450-4.

Verbeke, J. and Cools, R. (1995). The Newton-Raphson method. *International Journal of Mathematical Education in Science and Technology*, volume 26, no. 2, pages 177–193.

Witten, I.H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann.

Witzany, J. (2009). Definition of default and quality of scoring functions. *Available at SSRN 1467718.*

Witzany, J. (2010). *Credit risk management and modeling.* Oeconomica, Prague. ISBN 978-80-245-1682-0.

Witzany, J. (2017). *Credit Risk Management: Pricing, Measurement, and Modeling.* Springer International Publishing. ISBN 978-33-194-9800-3.

Witzany, J., Rychnovský, M., and Charamza, P. (2010). Survival Analysis in LGD Modeling. *IES Working Papers 2/2010.*

Witzany, J., Rychnovský, M., and Charamza, P. (2012). Survival Analysis in LGD Modeling. *European Financial and Accounting Journal*, volume 7, no. 1, pages 6–27.

Ypma, T.J. (1995). Historical development of the Newton-Raphson method. *SIAM review*, volume 37, no. 4, pages 531–551.

Zhang, J. and Thomas, L.C. (2012). Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling lgd. *International Journal of Forecasting*, volume 28, no. 1, pages 204–215.

Zikopoulos, P., Eaton, C., et al. (2011). *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media.