Vysoká škola ekonomická v Praze
Fakulta informatiky a statistiky

DIPLOMOVÁ PRÁCE

Michal Rychnovský

Scoring Models in Finance
(Skóringové modely ve financích)

Katedra ekonometrie

Vedoucí práce: Ing. Jan Zouhar, Ph.D.
Studijní obor: Matematické metody v ekonomii

2011

Prohlášení

Prohlašuji, že jsem diplomovou práci na téma Scoring Models in Finance (Skóringové modely ve financích) zpracoval samostatně. Veškerou použitou literaturu a další podkladové materiály uvádím v seznamu použité literatury.

V Šen-čenu dne 7. srpna 2011 Michal Rychnovský

Poděkování

Rád bych na tomto místě poděkoval svému vedoucímu Ing. Janu Zouharovi, Ph.D. za celkovou pomoc a vstřícný přístup při vedení práce.

Michal Rychnovský

Abstrakt

Název práce: Skóringové modely ve financích
Autor: Michal Rychnovský
Katedra: Katedra ekonometrie
Vedoucí práce: Ing. Jan Zouhar, Ph.D.

Abstrakt: Cílem této práce je popsat aplikace modelu logistické regrese pro odhad pravděpodobnosti defaultu klienta a stručně nastínit proces vývoje skóringových funkcí ve finanční praxi. Nejdříve uvádíme teoretický popis logistické regrese, následovaný postupným odvozením tří nejpoužívanějších skóringových modelů. Poté přicházíme s formální definicí Giniho koeficientu jako míry diverzifikační schopnosti modelu a odvozujeme výpočetní formule (Somersova typu) pro jeho odhad. Hlavní částí práce je potom popis úplného procesu vývoje skóringových funkcí, ilustrovaný na reálných příkladech z praxe.

Klíčová slova: Skóringové modely, kreditní riziko, logistická regrese.

Abstract

Title: Scoring Models in Finance
Author: Michal Rychnovský
Department: Department of Econometrics
Supervisor: Ing. Jan Zouhar, Ph.D.

Abstract: The aim of the present work is to describe the application of the logistic regression model to the field of probability of default modeling, and provide a brief introduction to the scoring development process used in financial practice. We start by introducing the theoretical background of the logistic regression model; followed by a consequent derivation of three most common scoring models. Then we present a formal definition of the Gini coefficient as a diversification power measure and derive the Somers-type formulas for its estimation. Finally, the key part of this work gives an overview of the whole scoring development process illustrated on the examples of real business data.

Keywords: Scoring models, credit risk, logistic regression.

Contents

Introduction

Along with the extensive growth of the financial industry all around the world, it has become especially important to incorporate various advanced mathematical and statistical methods to evaluate the possible risks resulting from different investment activities. These technical solutions provide quick and fully automatic tools that help the market actors make effective decisions. In this thesis we concentrate on the field of banking and consumer finance companies providing personal loans to their customers, and introduce several analytical tools to evaluate the potential risk of the applicants.[1]

In practice, all the loan providing institutions have a complete underwriting process to evaluate the credit risk of an applicant before issuing the loan. This process usually consists of two main parts – verification of customers' personal data (i.e. ID check etc.) and evaluating the customers' risk – this part is called scoring. In the terms of correct definition, we can define scoring as an estimation of the conditional probability of default, given the client's characteristics.

This paper works with the standard theory of logistic regression introduced for example in Agresti (1990) or Hosmer and Lemeshow (2000) and its connection to the probability of default modeling (see e.g. Aspey et al. (2003)). Whereas the theoretical part about scoring models has been already described in Rychnovský (2008), this work outlines the practical aspects of the application in the real financial market, and thus provide a comprehensive perspective on this issue. The thesis aims to serve as a brief introductory guide for beginning underwriting analysts as well as for anybody interested in this field.

In the first chapter we introduce the basic statistical theory of the logistic regression model and its application to the probability of default modeling. In the second chapter we use the assumption of independence of predictors to derive three odds-based scoring models. Third chapter is then dedicated to the quality of scoring evaluation in the terms of diversification power, and the most emphasis is put on

[1] This risk is called credit risk – i.e. the risk carried by the creditor – and is connected to the event called default. Default is usually defined as a violation of debt contract conditions, such as a lack of will or a disability to pay the loan back. Then, in the case of client's default, the creditor suffers a loss.

the definition and use of the Gini coefficient. Finally, in the last chapter we describe more details about the development and implementation of scoring models in practice. Here we offer a complete guideline from the definition of default and data sample preparation, over the predictors' selection and categorization, up to the final model building and testing. Furthermore, to demonstrate the whole process on the examples from the real world financial market we analyze the existing business data and provide possible results of the model.

Chapter 1

Logistic Regression

In this chapter, mainly based on Agresti (1990) and Hosmer and Lemeshow (2000), we introduce the logistic regression model as a widely used tool for probability of default estimation. First we describe the logic of the model and then the maximum likelihood approach to estimate its parameters and to test their significance.

1.1 Logistic Regression Model

For a given vector $\boldsymbol{x} = (x^0, \dots, x^p)'$ of client's characteristics we consider a random variable $Y_{\boldsymbol{x}}$ with an alternative distribution (where $Y_{\boldsymbol{x}} = 1$ for default and $Y_{\boldsymbol{x}} = 0$ otherwise). Then the expected value of $Y_{\boldsymbol{x}}$ can be written as

$$\mathbb{E}(Y_{\boldsymbol{x}}) = 1 \cdot \mathbb{P}(Y_{\boldsymbol{x}} = 1) + 0 \cdot \mathbb{P}(Y_{\boldsymbol{x}} = 0) = \mathbb{P}(Y_{\boldsymbol{x}} = 1) = \pi(\boldsymbol{x}),$$

where $\pi(\boldsymbol{x}) = \mathbb{P}(Y_{\boldsymbol{x}} = 1)$ is the conditional probability of default given the vector of predictors $\boldsymbol{x}$.

The aim of this section is to find a convenient model to describe the dependence of the probability of default $\pi(\boldsymbol{x})$ on the vector of clients characteristics $\boldsymbol{x}$. The first model we might think of is the linear regression model

$$\pi(\boldsymbol{x}) = \boldsymbol{\beta}'\boldsymbol{x}.$$

with a vector of parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$.

Even though this model is in practice for its simplicity sometimes used as well, it is generally not suitable for binary target variables. This is mainly because of the fact that $\pi(\boldsymbol{x})$ is a value of probability in the interval $[0, 1]$, whereas the linear regression estimated values (i.e. $\boldsymbol{\beta}'\boldsymbol{x}$) can be any real numbers.

Therefore, we define a function called *odds* as the ratio of the probability of default and its complement (i.e. the probability of recovery),

$$\operatorname{odds}(\boldsymbol{x}) = \frac{\mathbb{P}(Y_{\boldsymbol{x}} = 1)}{\mathbb{P}(Y_{\boldsymbol{x}} = 0)} = \frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}. \tag{1.1}$$

The values of this function are then in the interval $[0, \infty)$. Now, in order to get the values in all $\mathbb{R}$, we use logarithmic transformation. This function is then called *logit* and is defined as

$$\operatorname{logit}(\boldsymbol{x}) = \ln\big(\operatorname{odds}(\boldsymbol{x})\big) = \ln\left(\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}\right). \tag{1.2}$$

If we finally set $\operatorname{logit}(\boldsymbol{x}) = \boldsymbol{\beta}'\boldsymbol{x}$, we get the specific formula for the logistic regression model in the form

$$\pi(\boldsymbol{x}) = \frac{e^{\boldsymbol{\beta}'\boldsymbol{x}}}{1 + e^{\boldsymbol{\beta}'\boldsymbol{x}}}. \tag{1.3}$$

Generally, instead of the logit function (1.2) (and its inverse (1.3) respectively) any other transformation from $[0, 1]$ to $\mathbb{R}$ can be used. For example for the distribution function Φ of the standard normal distribution, we can define the *probit* function as

$$\operatorname{probit}(\boldsymbol{x}) = \Phi^{-1}\big(\pi(\boldsymbol{x})\big).$$

However, the logit function is often preferred for its simplicity – formulas (1.2) and (1.3) are closed and easy to compute. Moreover, the parameters have a straightforward interpretation, since for a positive unit change of a characteristic x_i, the odds of the client is multiplied by e^{β_i}.

1.2 Parameters Estimation

Generally, we assume a sample of n independent clients where for each client k we have a vector of characteristics (also called predictors or regressors) $\boldsymbol{x}_k = (x_k^0, \ldots, x_k^p)'$ and the target value y_k, where $y_k = 1$ in the case of default and $y_k = 0$ otherwise. Our aim is now to estimate the parameters of the model, i.e. vector $\boldsymbol{\beta}$ from formula (1.3).

For estimating of parameters of the linear regression model we usually use the *least squares method* based on the minimization of the sum of squared differences between the real and the estimated values. On the other hand, for the logistic model, we often refer to the *maximum likelihood method* as an alternative. For more details

about the maximum likelihood method see for example Lehmann and Casella (1998) or Van der Vaart (2000).

The maximum likelihood method is based on the construction of the *likelihood function*. For every vector of parameters $\boldsymbol{\beta}$ this function expresses the probability that exactly all the observations happen. The maximum likelihood estimate of $\boldsymbol{\beta}$ is then such a vector $\widehat{\boldsymbol{\beta}}$ for which this probability is maximal.

Let's construct the likelihood function for our regression. Using the expression $\pi(\boldsymbol{x})$ from (1.3), we can write the conditional probability that the client k will have the target value y_k as

$$\mathbb{P}(Y_{\boldsymbol{x}_k} = y_k) = \pi(\boldsymbol{x}_k)^{y_k}\big(1 - \pi(\boldsymbol{x}_k)\big)^{1-y_k}.$$

This means that for $y_k = 1$ it is the probability $\pi(\boldsymbol{x}_k)$, and for $y_k = 0$ the probability $1 - \pi(\boldsymbol{x}_k)$.

As the observations are assumed to be independent, we can define the likelihood function $\mathrm{l}(\boldsymbol{\beta})$ as the product of the conditional probabilities for independent clients,

$$\mathrm{l}(\boldsymbol{\beta}) = \prod_{k=1}^{n} \pi(\boldsymbol{x}_k)^{y_k}\big(1 - \pi(\boldsymbol{x}_k)\big)^{1-y_k}. \tag{1.4}$$

In order to find the maximum of this function, we first use logarithmic transformation. This transformation does not affect the point of the extreme and makes the function more convenient for differentiation. Thus we get

$$\mathrm{L}(\boldsymbol{\beta}) = \ln\big(\mathrm{l}(\boldsymbol{\beta})\big) = \sum_{k=1}^{n}\Big(y_k \ln\big(\pi(\boldsymbol{x}_k)\big) + (1 - y_k)\ln\big(1 - \pi(\boldsymbol{x}_k)\big)\Big). \tag{1.5}$$

Now we can compute the partial derivatives of (1.5) with respect to $\beta_0, \beta_1, \ldots, \beta_p$ and set them equal to zero. For this we consider the function π from (1.3) as a function of $\boldsymbol{\beta}$ and $\boldsymbol{x}$. This way we get a set of so called *likelihood equations* in the form

$$\sum_{k=1}^{n}\big(y_k - \pi(\boldsymbol{x}_k)\big) = 0 \tag{1.6}$$

$$\sum_{k=1}^{n} x_k^i\big(y_k - \pi(\boldsymbol{x}_k)\big) = 0, \tag{1.7}$$

for $i = 1, 2, \ldots, p$ where x_k^i is the i-th component of the vector $\boldsymbol{x}_k$.

This nonlinear set of equations is usually solved numerically using a special statistical software (e.g. SAS, SPSS, EViews etc.). By solving these equations we get the *maximum likelihood estimate* $\widehat{\boldsymbol{\beta}}$ of the vector of parameters $\boldsymbol{\beta}$.

From the asymptotic properties of the maximum likelihood estimates (see for example Rao (1973)) we can also get the asymptotic estimates $\widehat{\mathrm{SE}}(\widehat{\beta}_i)$ of the standard errors of the estimated parameters $\widehat{\beta}_i$. These are based on the information matrix $\mathrm{I}(\boldsymbol{\beta}) = \big(\mathrm{i}(\boldsymbol{\beta})_{ij}\big)_{i,j=0}^{p}$ defined as

$$\mathrm{i}(\boldsymbol{\beta})_{ij} = -\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_i \partial\beta_j} = \sum_{k=1}^{n} x_k^i x_k^j \pi(\boldsymbol{x}_k)\big(1 - \pi(\boldsymbol{x}_k)\big).$$

The variation matrix $\mathrm{Var}(\boldsymbol{\beta})$ we then get as the inverse of the information matrix, $\mathrm{Var}(\boldsymbol{\beta}) = \mathrm{I}^{-1}(\boldsymbol{\beta})$. Then the variance $\mathrm{Var}(\beta_i)$ of the i-th component of $\boldsymbol{\beta}$ is the i-th diagonal component of the variation matrix $\mathrm{Var}(\boldsymbol{\beta})$. Finally, by substitution of $\widehat{\boldsymbol{\beta}}$ we get the asymptotic estimate of the variance $\widehat{\mathrm{Var}}(\widehat{\beta}_i)$ and also the asymptotic estimate of the standard error of i-th parameter,

$$\widehat{\mathrm{SE}}(\widehat{\beta}_i) = \sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_i)}. \tag{1.8}$$

1.3 Significance of Parameters

After we have the maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$ of parameters, we focus on the statistical significance of the model, as well as the statistical significance of its individual parameters. Our aim is not to deal with goodness of fit characteristics (in an absolute sense), but to evaluate how much the individual coefficients contribute to the fit of the model (in a relative sense).

We start with the tests about individual parameters. From the asymptotic normality of the maximum likelihood estimates (see for example Rao (1973)) we know that

$$\frac{\widehat{\beta}_i - \beta_i}{\widehat{\mathrm{SE}}(\widehat{\beta}_i)} \overset{a}{\sim} N(0,1).$$

Thus we can construct the standard *Wald test* as the ratio of the maximum likelihood estimate $\widehat{\beta}_i$ and its standard error $\widehat{\mathrm{SE}}(\widehat{\beta}_i)$ as

$$W = \frac{\widehat{\beta}_i}{\widehat{\mathrm{SE}}(\widehat{\beta}_i)}. \tag{1.9}$$

Under the null hypothesis that $\beta_i = 0$, W has the standard normal distribution. Therefore, if for a given significance level α, $|W|$ is greater than the correspondent quantile $z_{1-\frac{\alpha}{2}}$ of the standard normal distribution, we reject the null hypothesis, and this parameter is significant (with a confidence level of $1 - \alpha$).

Based on the Wald test we can also construct the *confidence intervals* for individual parameters for a given α as

$$\beta_i \in \big(\widehat{\beta}_i - z_{1-\frac{\alpha}{2}}\widehat{\mathrm{SE}}(\widehat{\beta}_i), \widehat{\beta}_i + z_{1-\frac{\alpha}{2}}\widehat{\mathrm{SE}}(\widehat{\beta}_i)\big). \tag{1.10}$$

Now let's have a look at the statistical significance of a set of parameters or the model as whole. For the linear regression model we use the *residual sum of squares* given by

$$RSS = \sum_{k=1}^{n}(y_k - \widehat{y}_k)^2$$

and construct the F test as the ratio of the original (also *unrestricted*) RSS and the *restricted* RSS.

For the logistic regression model we define a similar test based on the likelihood function. Assume that $\widehat{\boldsymbol{\beta}}_u$ is an estimated (unrestricted) vector of parameters,

$$\pi_u(\boldsymbol{x}) = \frac{e^{(\widehat{\boldsymbol{\beta}}_u)'\boldsymbol{x}}}{1 + e^{(\widehat{\boldsymbol{\beta}}_u)'\boldsymbol{x}}}$$

is its model function and $\mathrm{l}(\widehat{\boldsymbol{\beta}}_u) = \prod_{k=1}^{n}\pi_u(\boldsymbol{x}_k)^{y_k}\big(1-\pi_u(\boldsymbol{x}_k)\big)^{1-y_k}$ the corresponding likelihood function.

If we want to test the statistical significance of a set of q parameters (where $q \leq p$), we denote $\widehat{\boldsymbol{\beta}}_r$ the restricted vector of parameters, and again

$$\pi_r(\boldsymbol{x}) = \frac{e^{(\widehat{\boldsymbol{\beta}}_r)'\boldsymbol{x}}}{1 + e^{(\widehat{\boldsymbol{\beta}}_r)'\boldsymbol{x}}}$$

is its model function and $\mathrm{l}(\widehat{\boldsymbol{\beta}}_r) = \prod_{k=1}^{n}\pi_r(\boldsymbol{x}_k)^{y_k}\big(1-\pi_r(\boldsymbol{x}_k)\big)^{1-y_k}$ the corresponding likelihood function.

Then, under the null hypothesis that all the q parameters are equal to zero, the statistic

$$G = -2\ln\left(\frac{\mathrm{l}(\widehat{\boldsymbol{\beta}}_r)}{\mathrm{l}(\widehat{\boldsymbol{\beta}}_u)}\right) \tag{1.11}$$

has the χ^2 distribution with q degrees of freedom. Therefore, if G is greater than the quantile $\chi^2_{1-\alpha}(q)$, we reject the null hypothesis with the confidence level $1-\alpha$, and conclude that with probability at least $1-\alpha$ one of those parameters is not equal to zero.

Chapter 2

Scoring Models

In this chapter we would like to describe the basic principle of the most common scoring models used in finance. First, we start with some important definitions.

2.1 Odds Ratio Definition

Suppose that our database contains s explanatory categorical[1] variables (predictors) for each client, where the i-th variable consists of p_i categories. Then put

$$Z = \big\{(i,j) : i \in \{1,\dots,p\}, j \in \{1,\dots p_i,\}\big\} \tag{2.1}$$

the set of all ordered pairs (i,j) of variables i and their categories j.

Then for each client k we have the vector

$$\boldsymbol{x}_k = \big((x_j^i)_k : (i,j) \in Z\big) \tag{2.2}$$

of dummy variables (i.e. $(x_j^i)_k = 1$ if the client k lies in the category j of the variable i, and $(x_j^i)_k = 0$ otherwise). Then we denote by B the index set of all defaulted clients (we also call them *bad* clients) and G the index set of all non-defaulted clients (*good* clients), and in the same spirit we define

$$B_j^i = \big\{k : k \in B, (x_j^i)_k = 1\big\}$$

as the index set of all defaulted clients k lying in the category j of the variable i, and

$$G_j^i = \big\{k : k \in G, (x_j^i)_k = 1\big\}$$

[1] As we will see in the practical examples in Chapter 4 the scoring models very often consist of categorical predictors only.

as the index set of all non-defaulted clients k lying in the category j of the variable i.

Now, based on the database of the observed clients, we define the total *odds* as the ratio of the number of defaulted vs. non-defaulted clients in the sample,

$$odds = \frac{|B|}{|G|}, \tag{2.3}$$

and also for individual categories j of variable i the $odds_j^i$ for the certain category,

$$odds_j^i = \frac{|B_j^i|}{|G_j^i|}. \tag{2.4}$$

Finally we define the *odds ratio* (OR_j^i) as the ratio of categorical $odds_j^i$ and total *odds*,

$$OR_j^i = \frac{odds_j^i}{odds}. \tag{2.5}$$

2.2 Fundamental of Scoring Models

In the beginning we remark that the notation of the variable *odds* from (2.3) is consistent with the function odds($\boldsymbol{x}$) from (1.1) because with the usual estimates of algebraic probabilities we can write

$$odds = \frac{|B|}{|G|} = \frac{\frac{|B|}{|B\cup G|}}{\frac{|G|}{|B\cup G|}} \approx \frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)}.$$

Now estimate the value of the theoretical function odds($\boldsymbol{x}$) from definition (1.1) based on the empirical values for the introduced variables. We write

$$\text{odds}(\boldsymbol{x}) = \frac{\mathbb{P}(Y_{\boldsymbol{x}}=1)}{\mathbb{P}(Y_{\boldsymbol{x}}=0)} \approx \frac{\frac{|B_{\boldsymbol{x}}|}{|B_{\boldsymbol{x}}\cup G_{\boldsymbol{x}}|}}{\frac{|G_{\boldsymbol{x}}|}{|B_{\boldsymbol{x}}\cup G_{\boldsymbol{x}}|}} = \frac{|B_{\boldsymbol{x}}|}{|G_{\boldsymbol{x}}|}, \tag{2.6}$$

where we denote

$$B_{\boldsymbol{x}} = \big\{k : k \in B, \boldsymbol{x}_k = \boldsymbol{x}\big\}$$

as the index set of all defaulted clients with the vector of characteristics $\boldsymbol{x}$, and

$$G_{\boldsymbol{x}} = \big\{k : k \in G, \boldsymbol{x}_k = \boldsymbol{x}\big\}$$

as the index set of all non-defaulted clients with the vector of characteristics $\boldsymbol{x}$.

As the values $|B_{\boldsymbol{x}}|$ and $|G_{\boldsymbol{x}}|$ are dependent on concrete combinations of values of the vector $\boldsymbol{x}$, it is usually not convenient to estimate $\text{odds}(\boldsymbol{x})$ using (2.6).[2] Therefore, we alter this expression

$$\text{odds}(\boldsymbol{x}) = \frac{|B_{\boldsymbol{x}}|}{|G_{\boldsymbol{x}}|} = \frac{|B|}{|G|} \frac{\frac{|B_{\boldsymbol{x}}|}{|B|}}{\frac{|G_{\boldsymbol{x}}|}{|G|}}. \tag{2.7}$$

As the expression $\frac{|B_{\boldsymbol{x}}|}{|B|}$ can be interpreted as an empirical estimate of the probability that a defaulted client will have the vector of characteristics $\boldsymbol{x}$, we can under the assumption of independence of predictors[3] rewrite this probability as the product of individual probabilities for individual predictors as

$$\frac{|B_{\boldsymbol{x}}|}{|B|} = \prod_{(i,j)\in Z} \left(\frac{|B_j^i|}{|B|}\right)^{x_j^i}.$$

This means that we compute the product for all categories where $x_j^i = 1$ (i.e. all relevant categories). In the same spirit we can express $\frac{|G_{\boldsymbol{x}}|}{|G|}$ and substitute into (2.7). Thus we get

$$\text{odds}(\boldsymbol{x}) = \frac{|B|}{|G|} \frac{\prod_{(i,j)\in Z} \left(\frac{|B_j^i|}{|B|}\right)^{x_j^i}}{\prod_{(i,j)\in Z} \left(\frac{|G_j^i|}{|G|}\right)^{x_j^i}} = \frac{|B|}{|G|} \prod_{(i,j)\in Z} \left(\frac{\frac{|B_j^i|}{|G_j^i|}}{\frac{|B|}{|G|}}\right)^{x_j^i}. \tag{2.8}$$

Finally, using the introduced notation we get the estimate of $\text{odds}(\boldsymbol{x})$ in the form

$$\text{odds}(\boldsymbol{x}) = odds \prod_{(i,j)\in Z} \left(\frac{odds_j^i}{odds}\right)^{x_j^i} = odds \prod_{(i,j)\in Z} (OR_j^i)^{x_j^i}. \tag{2.9}$$

This expression, together with the assumption of independence of predictors, forms the basics of the *Independence model*.

2.3 Independence Model

The *Independence model* is the simplest from the three introduced scoring models. The scoring function is based only on the computed values of *odds* and OR_j^i and

[2] There are $\prod_{i=1}^{p} p_i$ of those combinations which leads generally to very few observations in the single segments; and thus, the estimation of the probabilities would be very inaccurate.

[3] This assumption can be in practice sometimes difficult to fulfill (especially with a higher number of predictors); and therefore, we often choose more robust models instead.

can be according to the previous theory represented in the following way

$$\mathrm{S}^{IM}(\boldsymbol{x}) = odds \prod_{(i,j)\in Z} (OR_j^i)^{x_j^i}, \tag{2.10}$$

where $\boldsymbol{x} = \left(x_j^i : (i,j) \in Z\right)$ is the set of dummy variables representing the client. Sometimes also a logarithm of this function is used as a scoring function – in this case the score value corresponds to $\mathrm{logit}(\boldsymbol{x})$,

$$\ln\left(\mathrm{S}^{IM}(x)\right) = \ln(odds) + \sum_{(i,j)\in Z} x_j^i \ln(OR_j^i). \tag{2.11}$$

The main disadvantage of this model is the assumption of independence and the fact that all categories have the same weights. Even though the assumption of independence is in practice seldom completely fulfilled, this model is for its simplicity often used (especially with a low number of predictors which are not much dependent). There are two more models generalizing this approach by adding a nonnegative weight to each variable or even to each category.

2.4 WOE Model

Another approach to model the probability of default is the *WOE model* (for *Weight of Evidence*) as a generalization of the function (2.10), where to all variables we assign a weight according to their statistical importance in contribution to the final fit. This way we get the scoring function in the form

$$\mathrm{S}^{WOE}(\boldsymbol{x}, \boldsymbol{\lambda}) = odds \prod_{(i,j)\in Z} (OR_j^i)^{\lambda^i x_j^i}, \tag{2.12}$$

where $\boldsymbol{x} = \left(x_j^i : (i,j) \in Z\right)$ is again the set of predictors and $\boldsymbol{\lambda} = \left(\lambda^i : i \in \{1,\ldots,p\}\right)$ is a vector of parameters (weights) for individual predictors.

Again, the scoring function (2.12) is the estimation of the function $\mathrm{odds}(\boldsymbol{x})$. Thus, for $\mathrm{logit}(\boldsymbol{x})$ we get a logarithm of the scoring function in the form

$$\ln\left(\mathrm{S}^{WOE}(\boldsymbol{x}, \boldsymbol{\lambda})\right) = \ln(odds) + \sum_{(i,j)\in Z} \lambda^i x_j^i \ln(OR_j^i). \tag{2.13}$$

From this form of $\mathrm{logit}(\boldsymbol{x})$ we can then estimate the vector of parameters $\boldsymbol{\lambda} = \left(\lambda^i : i \in \{1,\ldots,p\}\right)$ using the logistic regression introduced in Chapter 1.

This model is computationally more difficult than the Independence model but especially for the databases with higher number of defaults it can provide more precise estimates of probabilities of defaults. According to Aspey et al. (2003), the WOE model is suitable for databases with at least 150 defaults.

2.5 Full Logistic Model

Finally, in the *Full logistic model* we put a certain weight to each category of the categorical variables (i.e. to each dummy variable). The scoring function is then in the form

$$\mathrm{S}^{FLM}(\boldsymbol{x}, \boldsymbol{\lambda}) = odds \prod_{(i,j)\in Z} (OR_j^i)^{\lambda_j^i x_j^i}, \tag{2.14}$$

where $\boldsymbol{x} = \left(x_j^i : (i,j) \in Z\right)$ is again the set of predictors and $\boldsymbol{\lambda} = \left(\lambda_j^i : (i,j) \in Z\right)$ is a vector of parameters for all the individual categories of predictors.

These parameters we again estimate using the logistic regression from the form for $\mathrm{logit}(\boldsymbol{x})$, i.e. from the logarithm of the scoring function,

$$\ln\left(\mathrm{S}^{FLM}(\boldsymbol{x}, \boldsymbol{\lambda})\right) = \ln(odds) + \sum_{(i,j)\in Z} \lambda_j^i x_j^i \ln(OR_j^i). \tag{2.15}$$

This model is the most flexible but also the most complicated from the three introduced models. According to Aspey et al. (2003), the full logistic model is suitable for databases with at least 1200 defaults. For a comparison of these three models we refer to Rychnovský (2008).

From the form (2.15) we can see that due to the used logit function this approach is equivalent to the logistic regression model introduced in Chapter 1. In the following text we refer to this model and in Chapter 4 we show a real example of this model and introduce its development.

Chapter 3

Quality of Scoring

In this chapter we describe one of the most important characteristics of a scoring model – its diversification power – and introduce two different measures to evaluate and compare this quality of different models.

3.1 Diversification Power

By *diversification power* we mean the ability of a scoring model to distinguish bad clients (i.e. the clients who will default) from good clients. We know that 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.

For an ideal scoring model we would get a line where all the clients with low score would be good and all the clients with high score would be bad (see Figure 3.1). On the other hand for a random model (i.e. very bad scoring) we would get a random distribution of good and bad clients in the whole scoring scale (see Figure 3.2).

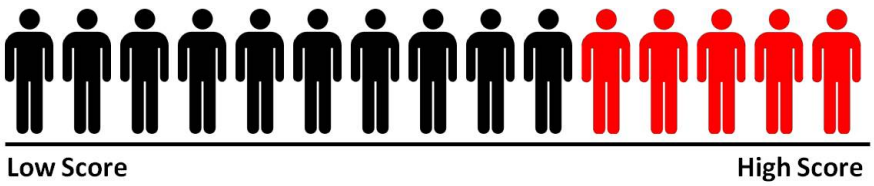

Figure 3.1: Example of an ideal scoring model.

However, even for a real scoring model it can happen that a client defaults, even though his score is low; and on the other hand, a client with high score might pay

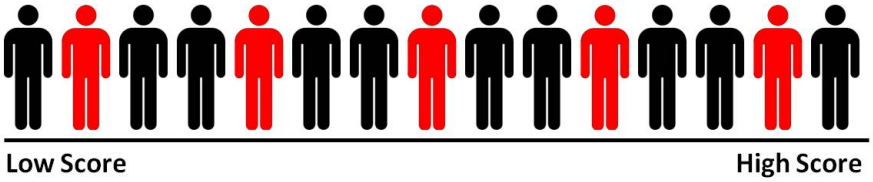

Figure 3.2: Example of a random scoring model.

everything in time. Thus, in reality we get a line where for low score values there are majority of good clients with a few bad clients among them, and for increasing score values we get higher proportion of bad clients with some good clients among them (see Figure 3.3).

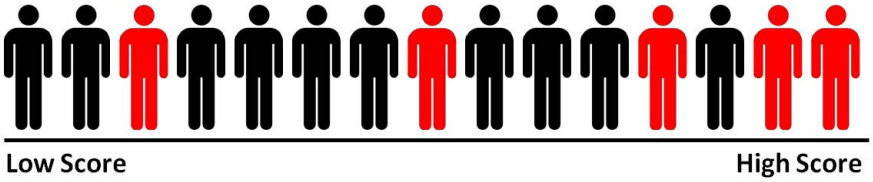

Figure 3.3: Example of a real scoring model.

From these examples we can see that the ideal model has a very good diversification power, whereas the random model is not useful at all. Our aim is then to find a model which is as close as possible to the ideal model and as far as possible from the random model. Therefore, for measuring the diversification power of scoring models, we introduce the Gini coefficient.

3.2 Gini Coefficient

The term of the Gini coefficient is well known from economics where – together with the Lorenz curve – it is used for measuring the inequality of income or wealth in some populations, usually states or regions. However, in the following text we concentrate on its use in the scoring diversification power measurement. First we present a formal definition and then we introduce several ways to compute its value for the whole scoring model performance, as well as for a single predictor.

3.2.1 Definition of the Gini Coefficient

First denote $S = \{S(\boldsymbol{x}), \boldsymbol{x} \in \boldsymbol{X}\}$ the set of all values of a scoring function $S(\boldsymbol{x})$. Then for every value of score $s \in S$ we 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*[1] as the connection of the set

$$L = \Big\{ \big[\mathrm{F}^B(s), \mathrm{F}^G(s) \big] \in \mathbb{R}^2 : s \in S \Big\}, \tag{3.1}$$

for all values $s \in S$ of the scoring function. Then this curve lies in the unit square connecting the opposite corners (see Figure 3.4).

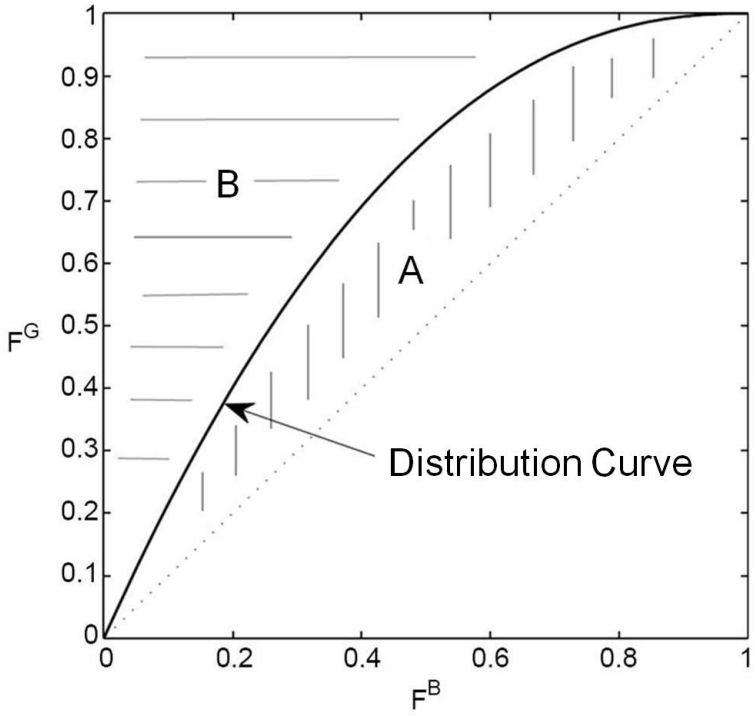

Figure 3.4: Distribution curve.

We can see that the better diversification power of the model, the closer the distribution curve is to the edges of the unit square. Therefore, we can describe the *Gini coefficient* 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 Figure 3.4).

[1] This curve is often called ROC curve (from Receiver Operating Characteristic, for more information see for example Hanley et al. (1983) or Witzany (2010)) or Lorenz curve.

Using the fact that the total area above the diagonal is one half of the unit square ($A + B = \frac{1}{2}$), we can reformulate the expression as $GC = 2A$. Thus we get the formal definition in the form

$$GC = 2 \int_S \big(\mathrm{F}^G(s) - \mathrm{F}^B(s) \big) \, \mathrm{dF}^B(s) \tag{3.2}$$

or equivalently

$$GC = 2 \int_S \mathrm{F}^G(s) \, \mathrm{dF}^B(s) - 1. \tag{3.3}$$

The value of the Gini coefficient is then in the interval $[-1, 1]$, where $GC = 1$ for an ideal diversification power (corresponding to the ideal model from Figure 3.1); $GC = 0$ for a zero diversification power (for example the random model from Figure 3.2), and negative values (i.e. the distribution curve below the diagonal) for a reversal model (i.e. with a contradictory classification).

3.2.2 Computing Gini Coefficient for Model

Assume now that in the model there are no two clients having the same score value. We take the formal definition of the Gini coefficient introduced in the previous paragraph and explore a bit more about its real meaning.

Again, as on page 8, we denote by B the index set of all bad clients and G the index set of all good clients. Then denoting s_k the score of the k-th client we can estimate the distribution functions of bad and good clients as

$$\mathrm{F}^B(s) = \mathbb{P}(s_k < s | k \in B) = \frac{|\{k : k \in B, s_k < s\}|}{|B|}$$

and

$$\mathrm{F}^G(s) = \mathbb{P}(s_k < s | k \in G) = \frac{|\{k : k \in G, s_k < s\}|}{|G|}.$$

Then the integral from the definition of the Gini coefficient (3.3) can be expressed as the sum

$$\int_S \mathrm{F}^G(s) \, \mathrm{dF}^B(s) = \sum_{l=1}^{n} \mathrm{F}^G(s_l) \, \mathbb{P}(s_k = s_l | k \in B).$$

Because we assume that no two clients have the same score, we have $\mathbb{P}(s_k =$

$s_l|k \in B) = 0$ for all $l \in G$ and $\mathbb{P}(s_k = s_l|k \in B) = \frac{1}{|B|}$ for all $l \in B$. Thus we get

$$\begin{aligned}\int_S \mathrm{F}^G(s)\,\mathrm{dF}^B(s) &= \sum_{l=1}^{n} \mathrm{F}^G(s)\,\mathbb{P}(s_k = s_l|k \in B) = \frac{1}{|B|}\sum_{l\in B}\mathrm{F}^G(s_l) = \\ &= \frac{1}{|B|}\sum_{l\in B}\frac{|\{k : k \in G, s_k < s_l\}|}{|G|} = \\ &= \frac{1}{|B|\cdot|G|}\sum_{l\in B}|\{k : k \in G, s_k < s_l\}|.\end{aligned}$$

In the last expression we can denote $a = \sum_{l\in B}|\{k : k \in G, s_k < s_l\}|$ as the 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). If we moreover define $b = \sum_{l\in B}|\{k : k \in G, s_k > s_l\}|$ as 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), we get that $|B|\cdot|G| = a + b$ is the number of all pairs of good and bad clients and the integral can be expressed in the form

$$\int_S \mathrm{F}^G(s)\,\mathrm{dF}^B(s) = \frac{a}{a+b}. \tag{3.4}$$

After substitution to (3.3) we get

$$GC = 2\frac{a}{a+b} - 1 = \frac{a-b}{a+b}. \tag{3.5}$$

Example 3.1. *Using the formula* (3.5) *we estimate the Gini coefficient of the example model from Figure 3.3, as*

$$\begin{aligned}a &= 2\cdot 5 + 4\cdot 4 + 3\cdot 3 + 1\cdot 2 = 37,\\ b &= 4\cdot 1 + 3\cdot 2 + 1\cdot 3 = 13;\end{aligned}$$

and thus, $GC = 0.48$.

If we now abandon the assumption that no two clients have the same score, we can estimate the Gini coefficient by the *Somers' d statistic* as

$$d = \frac{a-b}{a+b+c}, \tag{3.6}$$

where again $a = \sum_{l\in B}|\{k : k \in G, s_k < s_l\}|$ is the 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*); $b = \sum_{l\in B}|\{k : k \in G, s_k > s_l\}|$ 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 $c = \sum_{l \in B} |\{k : k \in G, s_k = s_l\}|$ 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 please refer to Somers (1962).

Sometimes, the Gini coefficient is described using the number of interchanges of neighboring clients necessary to achieve the ideal model. Then we can compute the Gini coefficient as

$$GC = \frac{m_r - m}{m_r},$$

where m is the number of interchanges necessary for the measured model, and $m_r = \frac{|G| \cdot |B|}{2}$ is the number of interchanges necessary for a random model. Considering the fact that $m = b$ and $m_r = \frac{a+b}{2}$, this approach is equivalent to the formula (3.5).

Example 3.2. *For the example model from Figure 3.3 we see 13 interchanges in Figure 3.5). For a random model we need* $\frac{1}{2} \cdot |G| \cdot |B| = 25$ *interchanges. Then* $GC = \frac{25-13}{25} = 0.48$.

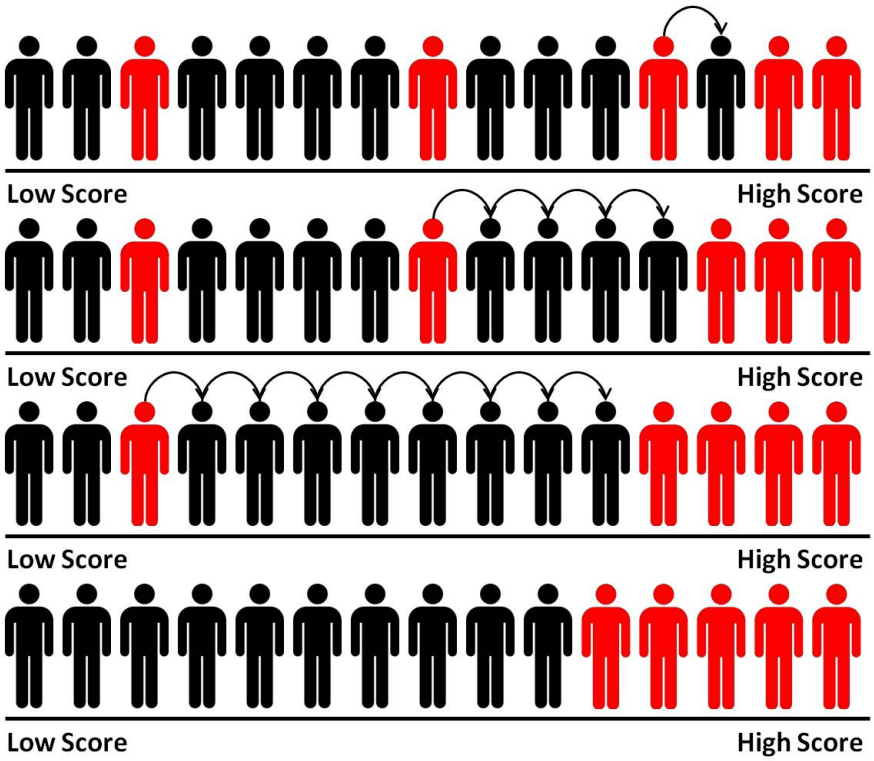

Figure 3.5: Gini calculation example.

3.2.3 Computing Gini Coefficient for Predictors

So far we introduced the Gini coefficient as a measure of the diversification power of scoring models. However, the Gini coefficient is not only useful for comparing the quality of different models, but also to compare the diversification power of single predictors. Based on this we can do the preliminary data exploration to understand which predictors are going to be the most powerful for modeling. Moreover, the Gini information is also very helpful for categorization and combination of predictors as we see in Section 4.2.

Even in this case the Gini coefficient is computed using the Somers' d statistics (3.6). Imagine for example that we have a predictor with three categories with different risk performance (let's call them *high*, *middle* and *low*). Then again denote G the number of good clients, B the number of bad clients and G_1, G_2, G_3, B_1, B_2, B_3 the number of good and bad clients in the respective categories (see Table 3.1).

Category	Good	Bad
high	G_1	B_1
middle	G_2	B_2
low	G_3	B_3
Total	G	B

Table 3.1: Categorical predictor distribution.

Then all the B_1 bad clients from the *high* category form concordant pairs with all the G_2 and G_3 good clients from the better categories. Also the B_2 bad clients from the *middle* category form concordant pairs with the G_3 good clients from the *low* category. Therefore, there are $a = B_1(G_2 + G_3) + B_2G_3$ concordant pairs in the model. Analogically, we can compute the $b = B_3(G_1 + G_2) + B_2G_1$ discordant pairs. Finally, we know that for all the pairs in the model we have $a + b + c = GB$; and thus,

$$d = \frac{a - b}{a + b + c} = \frac{B_1(G_2 + G_3) + B_2G_3 - B_3(G_1 + G_2) - B_2G_1}{GB}. \tag{3.7}$$

If we now define $g_1 = \frac{G_1}{G}$, $g_2 = \frac{G_2}{G}$, $g_3 = \frac{G_3}{G}$ and $b_1 = \frac{B_1}{B}$, $b_2 = \frac{B_2}{B}$, $b_3 = \frac{B_3}{B}$ as the proportions of good and bad clients in the respective categories, we get the Somers' d from (3.7) in the form

$$d = b_1(g_2 + g_3) + b_2g_3 - b_3(g_1 + g_2) - b_2g_1. \tag{3.8}$$

If we finally denote $\boldsymbol{g} = (g_1, g_2, g_3)'$, $\boldsymbol{b} = (b_1, b_2, b_3)'$ and U be the upper triangular unit matrix, we get

$$d = \boldsymbol{g}'(U' - U)\boldsymbol{b}, \tag{3.9}$$

because we have

$$\boldsymbol{g}'(U'-U)\boldsymbol{b} = (g_1, g_2, g_3)\begin{pmatrix} 0 & -1 & -1 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \end{pmatrix}\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = b_1(g_2+g_3)+b_2g_3-b_3(g_1+g_2)-b_2g_1.$$

In the general case of a predictor with p_i categories we again define $\boldsymbol{g} = (g_1, \ldots, g_{p_i})'$, $\boldsymbol{b} = (b_1, \ldots, b_{p_i})'$ as the proportions of good and bad clients in the respective categories, and U as the upper triangular unit matrix. Then the Somers' d can be computed using formula (3.9) – see e.g. Lucas (2004).

In the case of a binary predictor, we can simplify the formula (3.9) using the fact that $b_1 + b_2 = 1$ and $g_1 + g_2 = 1$,

$$d = \boldsymbol{g}'(U' - U)\boldsymbol{b} = (g_1, 1 - g_1)\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} b_1 \\ 1 - b_1 \end{pmatrix} = b_1 - g_1. \tag{3.10}$$

Example 3.3. *Let's take the family status as a predictor with three categories. See the distribution in Table 3.2.*

Category	Good	Bad	Good (%)	Bad (%)	Default rate
Others	2,944	84	2.0%	3.7%	2.77%
Single	119,009	1802	80.5%	80.1%	1.49%
Married	25,797	364	17.5%	16.2%	1.39%
Total	147,750	2,250	100%	100%	1.50%

Table 3.2: Family status distribution.

Then the Somers' d can be computed using the general formula (3.9) *as*

$$d = (0.020, 0.805, 0.175)\begin{pmatrix} 0 & -1 & -1 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \end{pmatrix}\begin{pmatrix} 0.037 \\ 0.801 \\ 0.162 \end{pmatrix} = 0.026765.$$

Example 3.4. *Let's take the gender of the client as a binary predictor (see Table 3.3). Then we can compute the Somers' d using* (3.10) *as* $d = 0.874 - 0.741 = 0.133$.

Category	Good	Bad	Good (%)	Bad (%)	Default rate
Male	109,475	1,966	74.1%	87.4%	1.76%
Female	38,275	284	25.9%	12.6%	0.74%
Total	147,750	2,250	100%	100%	1.50%

Table 3.3: Sex distribution.

If we want to use a predictor as numerical (in the case that the default rate is proportional to the values), we can compute the Gini coefficient using the Somers'

d statistics directly from (3.6) where we consider the ordering according to this predictor. Alternatively, we can categorize this predictor by deciles (i.e. ten groups of equal size from the ordered data) and compute the Somers' d for these categories. This approach is also useful to see the monotonicity and linearity of the default rate performance. For some examples see Section 4.2.

3.3 Lift

The last introduced characteristics in this chapter is *lift*. For the purpose of scoring modeling, we define the $P\%$ value of lift[2] as the ratio of the default rate for the $P\%$ worst cases divided by the default rate for the whole population.

Example 3.5. *Let's compute the* 20% *lift for the example model from Figure 3.3. There are total 15 clients in the sample with the overall default rate* $\frac{1}{3}$*. If we take the 20% worst clients, we get the default rate* $\frac{2}{3}$ *(see Figure 3.6). Therefore, the 20% lift of this model is 2.*

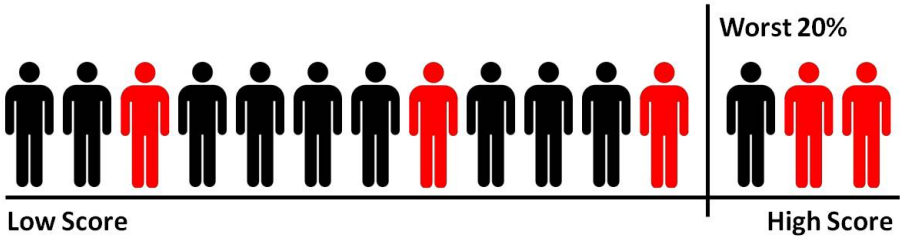

Figure 3.6: Example of 20% lift.

This characteristic is then very helpful for understanding the impact of using this scoring function. If we know the original default rate d_o in population and we would like to reject $P\%$ worst clients according to some scoring function, we can use the $P\%$ lift $l(P)$ of this function to compute the resulting default rate d_a of the approved population.

If we denote d_r the default rate of the rejected $P\%$ population, we get from the definition of lift that

$$l(P) = \frac{d_r}{d_o}. \tag{3.11}$$

[2]In practice, the most used are 10% and 20% lifts; however, for concrete purposes other values are chosen. Sometimes even the complete lift curve (e.g. the values for all $P \in [10, 90]$) is used. In this case we get some information about the overall model performance – similarly to the distribution curve.

As the original default rate d_o is a weighted average of the default rate d_a of the approved part and the default rate d_r of the rejected part, we can use the equation

$$d_o = \frac{Pd_r + (100 - P)d_a}{100}$$

to derive the formula

$$d_a = \frac{100 - Pl(P)}{100 - P} d_o. \tag{3.12}$$

Example 3.6. *In the population with default rate $\frac{1}{3}$ we would like to use a scoring function with 20% lift 2 and reject 20% worst clients. Then the default rate of the approved population will be*

$$d_a = \frac{100 - 20 \cdot 2}{100 - 20} \cdot \frac{1}{3} = \frac{1}{4}.$$

This is possible to compare with Example 3.5; even though, the setting is much more general.

Chapter 4

Scoring Development Process

In this chapter we describe a basic approach to develop a scoring function in practise. Of course, many developers can have different experience and thus different methods to achieve good results; however, this methodology should help to understand the basic theory of scoring development, and perhaps serve as an elementary manual for beginning underwriting analysts. The development process is also illustrated on real data examples from financial practise.[1] The data is processed in SQL, SAS and MS Excel.

4.1 Sample Preparation

The first step in the scoring development process is the preparation of the development sample, together with the definition of the target and explanatory variables.

4.1.1 Sample and Target Variable Selection

For development we use a sample of historical data where all defaults are known, i.e. where we know whether the default occurred in the defined time period or not. This is then connected with the definition of default and available sample size.

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 closed-end consumer

[1]The data is provided by an unspecified financial company operating on a foreign consumer finance market. Some characteristics are therefore excluded or undisclosed to protect the anonymity and know-how of the company. Consequently, neither the results nor the performance is fully correspondent with the real market situation.

loans[2] we define defaults for single payments. For example by FPD 30 we denote the *first payment default* of 30 days, i.e. the case when the customer didn't pay their first instalment within 30 days after the due date. Similarly, we can define FPD 90 as a 90 days default on first payment or SPD 30 and SPD 90 as defaults on the second payment.

According to the default definition, a corresponding development sample has to be chosen. For example, if we want to estimate the probability of 30 days default on the first payment (i.e. FPD 30), we can use all the data which is more than 30 days after the first due date (i.e. a fresh sample).[3] On the other hand, if we want to predict a 90 days default on the first four payments, we need the data which is at least 90 days after the fourth payment (i.e. much older sample).[4]

Of course, the population usually change in time and thus the fresh data is always more valuable. This is the reason that we seldom consider a life-time default (i.e. a default on any payment) for estimation, because of the old sample. Instead, we usually define a so called *fraud scoring function* for defaults on the first payment and a *default scoring function* for defaults on some other payments (e.g. the second to the fourth payment). This solution then offers a suitable compromise to analytically estimate the life-time default behavior. In our examples we consider a fraud scoring function with the FPD 90 definition of default.

Another aspect is the sample size. Again we have to face the problem of balancing the sample size (the bigger sample the better for the statistical results) and the age of the sample (bigger sample usually means older data). The chosen solution then depends on the homogeneity of the market – in a fast growing market the freshness of the sample is very important whereas for a stable market we can allow older sample with more data. For our example we use a four months sample of 150,000 loans with 2,500 defaults.[5]

4.1.2 Predictors – Collection, Definition and Coding

After we have the sample defined, we have to find a set of potential predictors for the default modeling. The first step is to collect all ideas for potential predictors and focus on finding as many new predictors as possible. Here we can say that predictors can be usually found in the following categories.

[2] A *closed-end consumer loan* or also *consumer installment loan* is a loan which the customer has to repay in a predefined number of payments. Opposite to closed-end loans, there can be *open-end* (or *revolving*) loans where the loan conditions change in time (e.g. credit cards).

[3] The first due date is usually one month after issuing the loan; and therefore, we can use the data older than approximately 2 months.

[4] At least 7 months after the loan issuing.

[5] Neither the sample size nor the default rate is correspondent to the original sales volume and default rate of the data providing company.

- *Application data* – this is the basic data about the client asked in the loan application form. In this category we find the demographic predictors (e.g. gender, education, marital status, number of children, region or city etc.), employment information (e.g. occupation industry, occupation position, length of employment, monthly income etc.), credit information (e.g. credit amount, number of payments, monthly annuity etc.) and a lot of additional information about the client (e.g. whether the client has a home fix line, how many contact numbers the client provided and more).

- *Historical and behavioral data* – if it is a known client (for example a current client of a bank) or a repeated client (for a consumer finance company), we can even find some information about the account and credit history of the client. For a current client of a bank it can be for example information about turnover in the current account or the volume of deposits; for a repeated loan client it can be last credit amount, current debt amount, number of payments successfully paid in time, maximal number of days past due date etc.

- *Data from databases* – in this category we can have all the available data from credit bureau (e.g. number of active loans in other campanies, current debt amount, defaults on other loans etc.) as well as the data from other databases potentially used in underwriting (e.g. health and social security information, tax information, availability of the provided contacts in yellow pages etc.).

- *Additionally collected data* – in this category we can consider all the information provided by the employee personally dealing with the customer (e.g. internal comments) as well as the results of verification of the customer data.

- *Other and combined predictors* – in this category we can classify all the predictors not fitting into the previous categories (like selling place information, application time etc.) as well as the predictors combined using more different information (like total debt-income ratio, affordability limit etc.).

The next step is then a proper and unique definition of all the variables and database scripts (usually in an SQL, Oracle, SAS or other syntax) to compute these predictors from the available database. Here we have to make sure that all the predictors' values were really known at the moment of approval of the loan. Even though this is an elementary message, it is very important to keep in mind – especially when computing historical and behaviorial characteristics – that all the information has to be censored to the moment of application of the loan.[6] Otherwise we would face a data inconsistence after implementation of such scoring function.

[6]This means that even though in our database we can see all the successful payments of the client up to now, we can use only the number of successful payments paid before the application date of the loan (the same for other predictors like the maximal number of days past due etc.).

4.1.3 Training, Testing and Validation Sample

After we have all the predictors for the whole sample, we divide the data into three parts – training, testing and validation.

- *Training sample* (usually 50%–70% of the data) is the sample for model development.
- *Testing sample* (usually 30%–50% of the data) is the sample for testing the performance of the model. This is important mainly for testing the stability of the model in the sense of overfitting (see Section 4.3).
- *Validation sample* (usually last period of the data – 14–28 days[7]) is used for testing the time stability of the model, i.e. testing the model performance on a future sample (see Section 4.4).

4.1.4 Inconsistence with Original Data

At the end of this section we would like to mention a short comment about inconsistence of the sample data with the original population. Imagine first that we currently have a process where every client is approved. Then we have a sample of full population where we can observe defaults – i.e. a consistent sample. However, in reality we usually have some underwriting process already in place (either an old scoring function or another set of rules to decide which client should be approved); and thus, some clients are rejected. And as the development sample consists of approved clients only (because otherwise we cannot observe a default), it is not fully consistent with the original population. Therefore, monitoring of the newly implemented scoring function is very important.

There exists one way to eliminate this inconsistence. In some developed systems there is a small proportion of loans (e.g. 1%–5%) which are approved automatically regardless the approval conditions. If the proportion is small enough, it does not harm the overall risk performance, and it can provide some information about defaults on the original population.

4.2 Data Exploration and Variable Categorization

The next step of the scoring development process is the data exploration and categorization. It means that for every potential predictor we look at its diversification

[7]Testing on the validation sample is one of the last parts of the scoring development process. Therefore, if the development is expected to last more than two weeks, it is possible to compute the validation data later on the new sample of the last two weeks.

power and try to propose some efficient categorization.

4.2.1 First-Step Data Exploration

At the beginning we look at every single predictor to see its performance – i.e. a distribution of individual values or a share of categories together with the default rate on individual segments. At the same time we can compute the predictor Gini introduced in Section 3.2.3. For this we can use for example SQL, SAS or MS Excel/VBA.

In the case of categorical predictors we get the share and default rate directly (see Example 4.1) together with the Gini coefficient. For numerical predictors we can start the data exploration with histograms (see Example 4.2) or deciles of the predictor values (see an analogy in Example 4.5).

Example 4.1. *Education is a categorical predictor with 7 categories (see the distribution and default rates in Table 4.1 and Figure 4.1) and Gini value of 14.5%. We can see that the higher education the better risk performance of the clients. For the category Master and Above the default rate is higher – this can be caused by two reasons. Firstly, the sample of the clients with a Master degree is very small (148 clients) and the default rate is caused by only 2 defaults. Secondly, as there is no verification of the achieved education, it is possible that those (defaulted) clients are cheating with their education.*

Education	Population	Share	Default rate
Elementary School	774	0.5%	3.49%
Junior High School	43,526	29.0%	1.92%
Senior High School	38,268	25.5%	1.61%
Technical Secondary School	41,845	27.9%	1.35%
Junior College	17,059	11.4%	0.98%
Bachelor	8,380	5.6%	0.49%
Master and Above	148	0.1%	1.35%
Total	150,000	100.0%	1.50%

Table 4.1: Data exploration example – Education (Gini 14.5%).

Example 4.2. *Employment length (i.e. how many months is the client employed with the current employer) is a numerical predictor with values from 0 to 980. In this case the observation of single values does not give us much information and it is better to see a histogram of values for defined time intervals (see Table 4.2 and Figure 4.2). With this categorization the Gini value of this predictor is 7.2% (i.e. lower than for Education). Generally, we can see that with higher employment length the risk is lower. This is again not that true for very high values where the clients can be cheating. Moreover, we can see that there are significant peaks for the values with complete years which suggest that the clients are often rounding this information.*

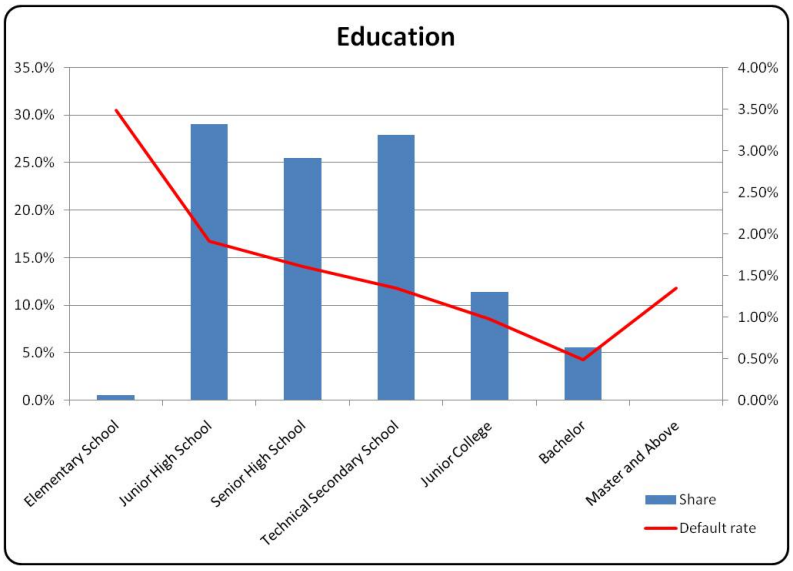

Figure 4.1: Data exploration example – Education (Gini 14.5%).

Employment length	Population	Share	Default rate
less than 3 months	28,181	18.8%	1.80%
less than 6 months	33,112	22.1%	1.49%
less than 9 months	10,841	7.2%	1.47%
less than 12 months	20,798	13.9%	1.52%
less than 15 months	6,201	4.1%	1.60%
less than 18 months	5,498	3.7%	1.64%
less than 21 months	2,136	1.4%	1.64%
less than 24 months	12,508	8.3%	1.35%
less than 27 months	1,967	1.3%	1.37%
less than 30 months	2,426	1.6%	0.95%
less than 33 months	1,122	0.7%	0.71%
less than 36 months	8,223	5.5%	1.34%
less than 39 months	726	0.5%	1.24%
less than 42 months	1,063	0.7%	1.03%
less than 45 months	594	0.4%	0.34%
less than 48 months	3,437	2.3%	1.28%
more than 48 months	11,167	7.4%	1.31%
Total	150,000	100.0%	1.50%

Table 4.2: Data exploration example – Employment length (Gini 7.2%).

4.2.2 Categorization of Predictors

Usually the variables have many initial values or categories. In that case we cannot use them as predictors directly (because of the overfitting problem) and we

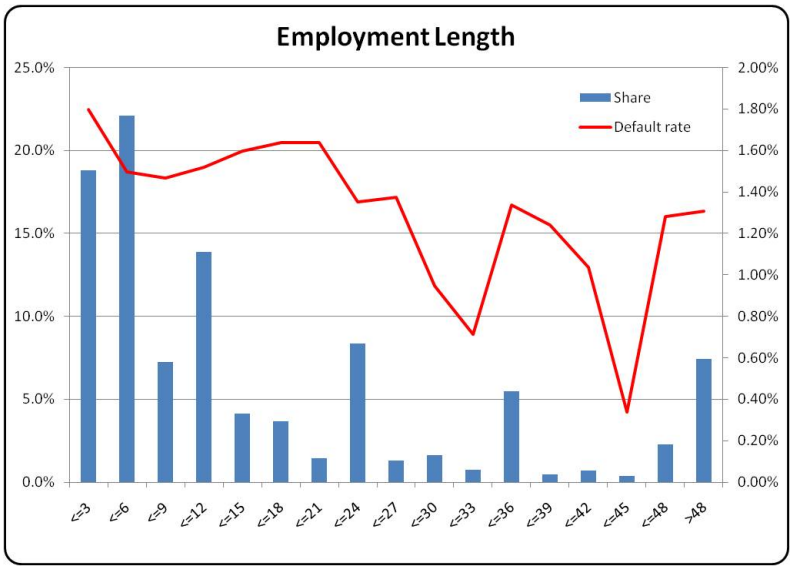

Figure 4.2: Data exploration example – Employment length (Gini 7.2%).

have to categorize the output into a smaller number of relevant categories. For this process we can use the tables of the data exploration step. When grouping the results we should keep the following principals.

- *Default rate* – according to this principal we try to group the results with similar default rates together in the same category, and at the same time achieve a big difference of average default rate among different categories. This then helps the diversification power of the categorized predictor. When doing this we have to always mind the sample size for every value or category – if the sample size is small then we had better follow the business logic instead.
- *Business logic* – the second requirement for the categories is to keep the business logic and understanding. At every moment we should be able to understand and explain why it makes sense to create this particular categorization. Especially for small sample categories it is more important to keep the business logic than the default rate principal.
- *Reasonable share* – last we try to create categories with reasonable shares. If the share of a category is very small than the diversification power might be small as well, even if the difference in default rate is significant.[8]

Example 4.3. *As an example we take client's age as a predictor to be categorized. As of the local requirements of the company, a client has to be older than 18 years*

[8]E.g. we have two categories of equal share 50% and default rates 1.3% and 1.7%. Then the Gini value of this predictor is 6.7% (using formula 3.10). Whereas using a categorization of 5% share with 0.1% default rate and 95% share with 1.57% default rate, we would get a lower Gini of 4.7%.

and younger than 56 years to be eligible for a loan. Therefore, we can observe all values of age between 18 and 55 (see Table 4.3 and Figure 4.3). If we compute the Gini value for the original predictor (with 38 categories) we get 15.33%. Our aim is now to propose a categorization into some smaller number of categories (let's say 2–5) with an understandable business logic and a high value of Gini.[9]

The business logic in this case suggests that the clients with similar age should have a similar default rate. Therefore, we start our categorization by ordering the data by age. Now we can look for age segments with relevant shares and similar default performance. From Table 4.3 we can see that the segment from 18 to 23 years has similar risk performance and is less risky comparing to the risk performance of the segment from 24 to 27 years and the rest of the data. As for the segment 28+ there are no significant differences (i.e. differences on significant samples and explainable by a business logic), we can conclude that we have found a reasonable categorization into three categories – we clearly see the default rate similarity, business logic and sample relevance. See the results of this categorization in Table 4.4 and Figure 4.4. The Gini value for this categorization is 12.55%.

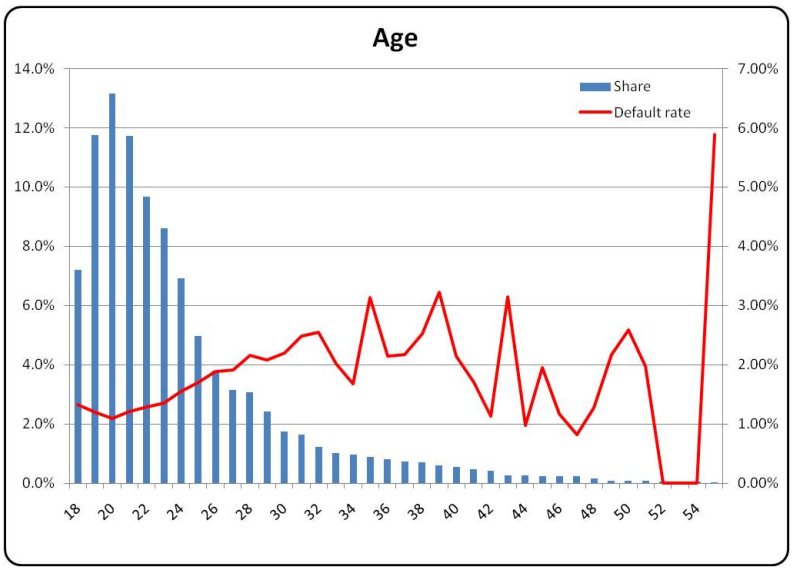

Figure 4.3: Variable categorization example – Age (Gini 15.33%).

Especially from Figure 4.4 notice that the first category 18–23 is still over 60% of the total population; and therefore, it might make sense to divide it into two categories. If we again look at Table 4.3 we can see that the segment 19–21 is less risky than the rest of this category. This has a business logic as well as the 18 years old clients are too young (usually fresh high school graduates) and their risk is therefore a little higher (comparable with the group 22-23). This way we created

[9]It is clear from the definition of Gini that by merging different categories the original Gini will decrease.

Age	Population	Share	Default rate	Category 1	Category 2
18	10,794	7.2%	1.32%	18–23	18&22–23
19	17,627	11.8%	1.20%	18–23	19–21
20	19,746	13.2%	1.10%	18–23	19–21
21	17,591	11.7%	1.21%	18–23	19–21
22	14,513	9.7%	1.28%	18–23	18&22–23
23	12,912	8.6%	1.36%	18–23	18&22–23
24	10,392	6.9%	1.55%	24–27	24–27
25	7,458	5.0%	1.70%	24–27	24–27
26	5,682	3.8%	1.88%	24–27	24–27
27	4,706	3.1%	1.91%	24–27	24–27
28	4,592	3.1%	2.16%	28+	28+
29	3,612	2.4%	2.08%	28+	28+
30	2,598	1.7%	2.19%	28+	28+
31	2,462	1.6%	2.48%	28+	28+
32	1,845	1.2%	2.55%	28+	28+
33	1,540	1.0%	2.01%	28+	28+
34	1,435	1.0%	1.67%	28+	28+
35	1,343	0.9%	3.13%	28+	28+
36	1,209	0.8%	2.15%	28+	28+
37	1,104	0.7%	2.17%	28+	28+
38	1,070	0.7%	2.52%	28+	28+
39	899	0.6%	3.23%	28+	28+
40	841	0.6%	2.14%	28+	28+
41	707	0.5%	1.70%	28+	28+
42	621	0.4%	1.13%	28+	28+
43	381	0.3%	3.15%	28+	28+
44	410	0.3%	0.98%	28+	28+
45	358	0.2%	1.96%	28+	28+
46	341	0.2%	1.17%	28+	28+
47	366	0.2%	0.82%	28+	28+
48	236	0.2%	1.27%	28+	28+
49	138	0.1%	2.17%	28+	28+
50	116	0.1%	2.59%	28+	28+
51	101	0.1%	1.98%	28+	28+
52	98	0.1%	0.00%	28+	28+
53	77	0.1%	0.00%	28+	28+
54	62	0.0%	0.00%	28+	28+
55	17	0.0%	5.88%	28+	28+
Total	150,000	100.0%	1.50%	3 Categories	4 Categories

Table 4.3: Variable categorization example – Age (Gini 15.33%).

Age category	Population	Share	Default rate
18–23	93,183	62.1%	1.23%
24–27	28,238	18.8%	1.72%
28+	28,579	19.1%	2.17%
Total	150,000	100.0%	1.50%

Table 4.4: Variable categorization example – Age in 3 categories (Gini 12.55%).

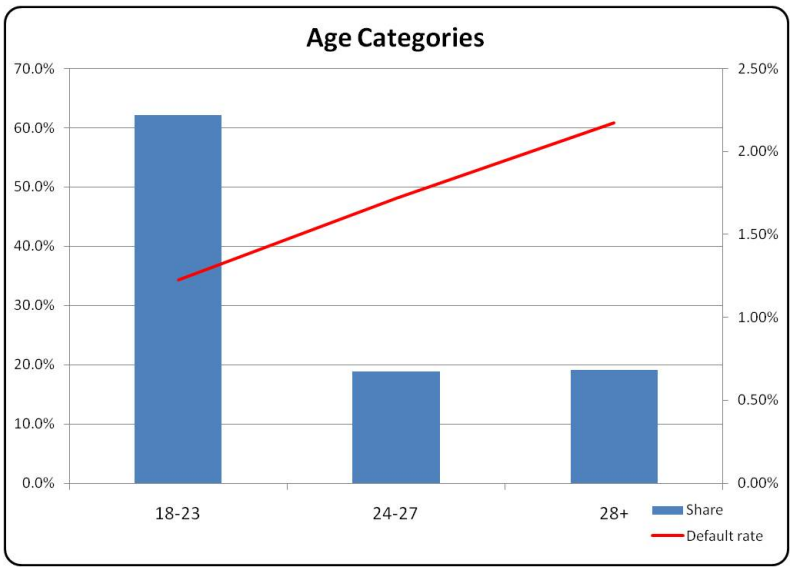

Figure 4.4: Variable categorization example – Age in 3 categories (Gini 12.55%).

a categorization into 4 categories according to the above stated principals. See Table 4.5 and Figure 4.5 for the results. The Gini value for this categorization is 13.53%.

Age category	Population	Share	Default rate
18&22–23	38,219	25.5%	1.32%
19–21	54,964	36.6%	1.16%
24–27	28,238	18.8%	1.72%
28+	28,579	19.1%	2.17%
Total	150,000	100.0%	1.50%

Table 4.5: Variable categorization example – Age in 4 categories (Gini 13.53%).

For the variable age we now have two proposed categorizations – 4 categories with Gini 13.53% and 3 categories with Gini 12.55%. Similarly we can propose several categorizations for every variable and include them all in the modeling. Then in the model development stage (see Section 4.3) we can decide which categorization performs best for the concrete model.

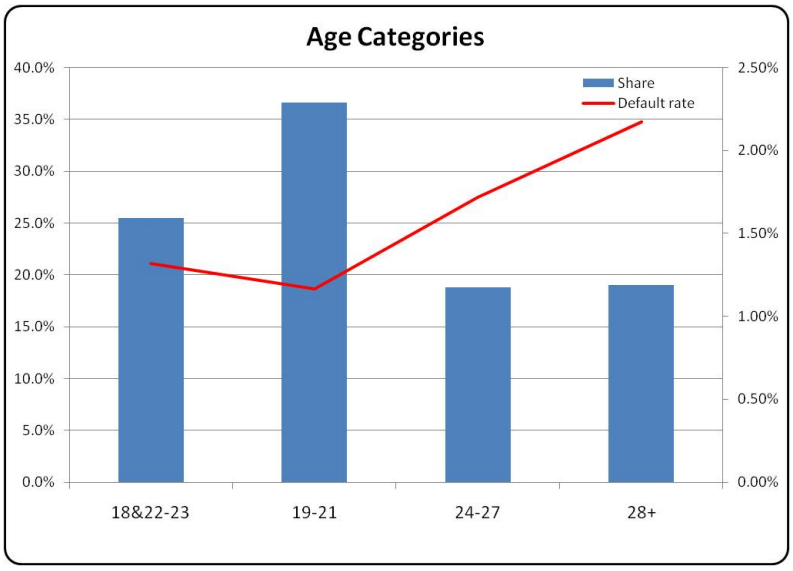

Figure 4.5: Variable categorization example – Age in 4 categories (Gini 13.53%).

4.2.3 Combined Predictors

Often it is useful to combine two or more variables together. This is mainly the case when the variables (or some of their categories) are correlated, and the information of their combination can help to achieve better results. Sometimes this approach is also used to reduce the number of parameters of the model. When constructing the combined variables, we consider all the combinations of the original variables and then categorize them into a new variable following the same principles as in the previous paragraph.

Example 4.4. *In Table 4.6 and Figure 4.6 we can see all the combinations of client's sex and education and their proposed categorization into 4 categories. Thus we can get a combined predictor with four categories and Gini 21.12% (see the performance in Table 4.7 and Figure 4.7.*

4.2.4 Numerical Predictors

So far we have been talking only about categorical predictors. Although categorical predictors are much more common in the scoring modeling, there are some situations where we can use a numerical predictor instead. An important condition for using a numerical predictor is the monotonicity of the default rate along the predictor values.[10]

[10]This means that with increasing values of the predictor the default rate has to increase (resp. decrease) for all the predictor's range. If the default rate is for example first decreasing and then increasing (i.e. a "V" shape), it is recommended to categorize it into several categories.

Sex & Education	Population	Share	Default rate	Category
Male: Elementary	638	0.4%	4.23%	M_1
Male: Junior High	34,282	22.9%	2.19%	M_1
Male: Senior High	29,000	19.3%	1.82%	M_2
Male: Technical High	30,972	20.6%	1.54%	M_3
Male: Junior College	11,110	7.4%	1.28%	M_3
Male: Bachelor	5,339	3.6%	0.67%	M_4 & F
Male: Master and Above	100	0.1%	2.00%	M_4 & F
Female: Elementary	136	0.1%	0.00%	M_4 & F
Female: Junior High	9,244	6.2%	0.89%	M_4 & F
Female: Senior High	9,268	6.2%	0.93%	M_4 & F
Female: Technical High	10,873	7.2%	0.79%	M_4 & F
Female: Junior College	5,949	4.0%	0.42%	M_4 & F
Female: Bachelor	3,041	2.0%	0.16%	M_4 & F
Female: Master and Above	48	0.0%	0.00%	M_4 & F
Total	150,000	100.0%	1.50%	4 Categories

Table 4.6: Variable combination example – Sex & Education (Gini 22.24%).

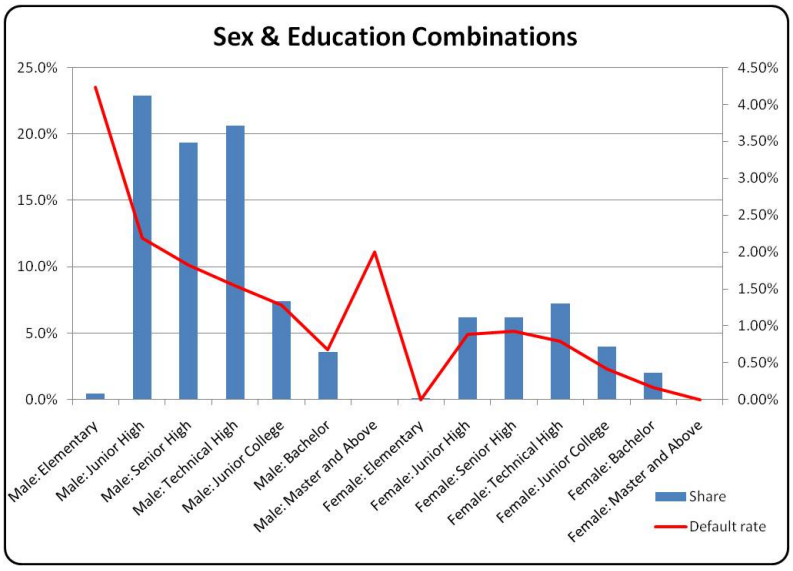

Figure 4.6: Variable combination example – Sex & Education (Gini 22.24%).

Under this condition the performance of a numerical predictor can be sometimes better than the performance of a corresponding categorized predictor. Moreover, for small scorecards (i.e. developed scoring function formulas) the presence of a numerical predictor can help to extend the number of potential score values (i.e. decrease the number of ties) – and the scoring function is then easier to implement in production.[11]

[11] With a small scorecard (e.g. with only 5 parameters of categorized predictors) it can happen

Sex & Education	Population	Share	Default rate
Male: Elementary & Junior High	34,920	23.3%	2.23%
Male: Senior High	29,000	19.3%	1.82%
Male: Technical High & Junior College	42,082	28.1%	1.47%
Male: University & Female	43,998	29.3%	0.73%
Total	150,000	100.0%	1.50%

Table 4.7: Variable combination example – Sex & Education in 4 categories (Gini 21.12%).

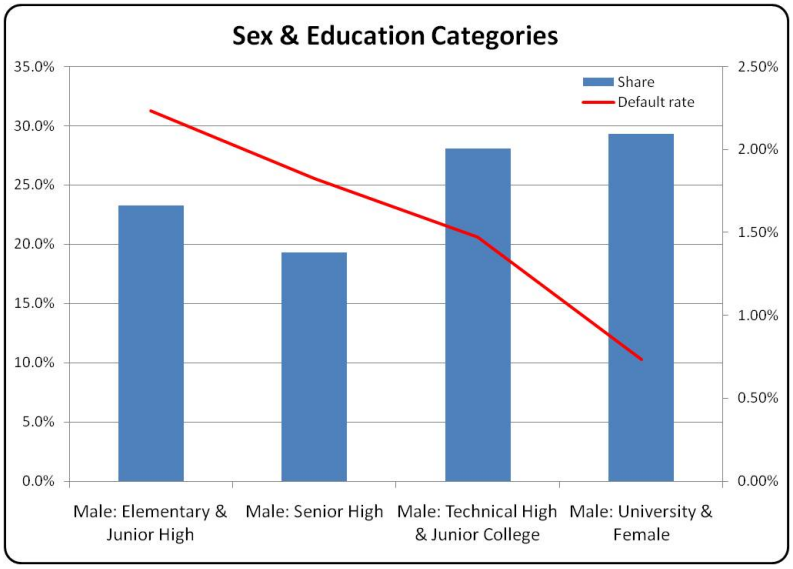

Figure 4.7: Variable combination example – Sex & Education in 4 categories (Gini 21.12%).

Example 4.5. *As an example of a suitable numerical predictor we can take the down payment rate, i.e. the percentage of the product price paid by the client as an initial payment at the moment of approving the loan. If we create 10 equal size groups according to the dawn payment values (i.e. deciles), we can see from Table 4.8 and Figure 4.8 that the default performance is really monotonous along the down payment range (the higher down payment the lower risk). Therefore, we can try to use this predictor (or its transformation) as numerical. At the same time, we categorize this predictor and let the model select the better one.*

that there are many clients with the same score and it might be difficult to setup who should be approved and who should be rejected. A numerical predictor can help to spread the score values for easier diversification.

Down payment	Population	Share	Default rate
1st decile	15,000	10.0%	2.36%
2nd decile	15,000	10.0%	1.84%
3th decile	15,000	10.0%	1.84%
4th decile	15,000	10.0%	1.81%
5th decile	15,000	10.0%	1.80%
6th decile	15,000	10.0%	1.48%
7th decile	15,000	10.0%	1.41%
8th decile	15,000	10.0%	1.19%
9th decile	15,000	10.0%	0.85%
10th decile	15,000	10.0%	0.42%
Total	150,000	100.0%	1.50%

Table 4.8: Numerical variable example – Down payment in 10 categories (Gini 19.75%).

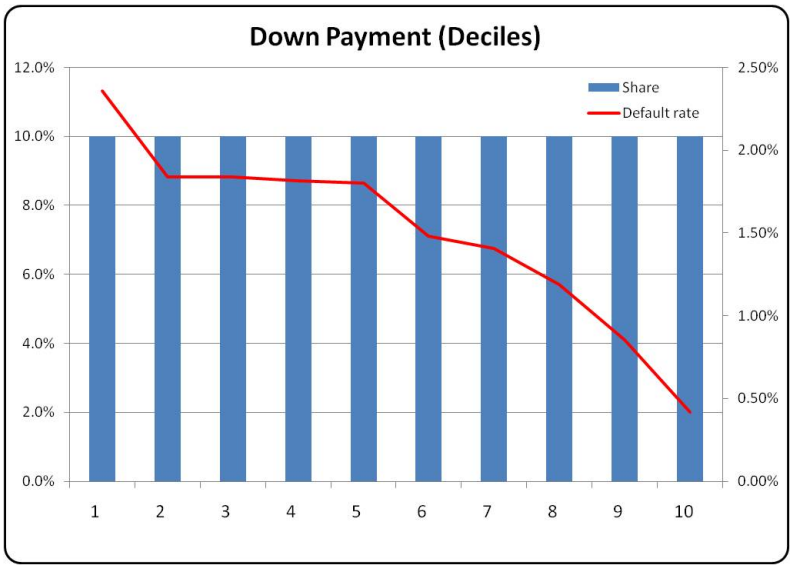

Figure 4.8: Numerical variable example – Down payment in 10 categories (Gini 19.75%).

4.3 Scoring Modeling

Once we have the initial categorization done, we can start with the actual scoring modeling. For this we can use any type of statistical modeling software like SAS, SPSS, Eviews or R.

4.3.1 Initial Models

Now we are in the situation when we have a great amount of different predictors and their categorizations (usually more than 100) and we have to select the most important ones for future development. In this case it is usually advised to start with an automatic selection process for some initial exploration. By different combinations of the forward selection, backward selection and stepwise selection, we can distinguish the more and the less valuable predictors.

- *Forward selection* – helps to identify the most powerful predictors (these are selected first).
- *Backward selection* – helps to exclude the less powerful or correlated (duplicate) predictors.
- *Stepwise selection* – we can use the stepwise selection to build the first initial models from a selected set of predictors.
- *Gini information* – the information from the variable exploration and categorization step can also help with predictors evaluation.
- *Expert expectation* – often also an experience and expectation can help to select good predictors for future modeling.

After this step we have a smaller set of potential predictors and several initial models for future development. At this moment we observe the most important characteristics.

- *Training Gini* – for the initial models it is useful to observe the value of training Gini. Usually the initial models have more degrees of freedom (and thus these are not that stable), and their training Gini then stands for some target value we are trying to achieve with the smaller and more stable models.
- *Testing Gini* – provides us the information about the out of sample performance of the model. From the perspective of stability and future implementation, this characteristic is even more important than the training Gini. If the value of the testing Gini is close to the value of training Gini, then we can say that the model is not overfitted.
- *Degrees of freedom* – the number of degrees of freedom is the total number of parameters estimated. Generally, the fewer degrees of freedom the more stable model we get.

Example 4.6. *We get an initial model with 10 predictors and 21 degrees of freedom. Then this model has the training Gini value of 44.1% and the testing Gini of 35.3%. This model is then not very stable and we try to decrease the number of degrees of freedom by excluding, combining or changing its predictors.*

4.3.2 Correlation Structure

Another important characteristic of a model is the correlation structure of its predictors' categories. Most statistical software enables to compute the estimate of the correlation matrix for individual parameters (see an example in Table 4.11). If there is a high correlation coefficient (e.g. higher than 0.1), we should try to understand the reasons, and (if possible) remove correlation from the model. This can be achieved by two main approaches.

- *Remove one of the predictors* – first we can try to remove one of the correlated predictors. If we see that the Gini of the model decreased a lot, we try to find another convenient predictor to be used instead.
- *Combine the predictors* – if we know the reason of correlation, we can combine the correlated predictors (in the terms of Paragraph 4.2.3).[12]

Example 4.7. *We face the problem of correlation between the predictor document (i.e. the type of the second document provided) and working industry. This is an example of an understandable correlation because most students (special industry type) provide their student card as the second document, whereas most of manufacturing workers provide their employer card. Therefore, in this case we can suggest combining of these two predictors.*

4.3.3 Final Candidates

After having the initial models, there is a long process of manual optimization in order to achieve a stable and well performing model. The process contains mainly the following activities.

- *Manual adding and removing of predictors* – we try to decrease the number of degrees of freedom by removing the least significant predictors. After every change of the model, we try to add another predictor and watch the sensitivity of Gini.
- *Different categorizations of the same predictors* – for some predictors we prepared more different categorizations. In different models different categorizations can perform better.
- *Combinations of predictors based on correlation matrix* – we try to replace or combine the correlated predictors.

[12] Combined predictors are often used in scoring models; however, it is better to avoid combinations of more than 2 (exceptionally 3) different predictors into one.

This way we try different combinations of predictors and closely watch the key performance indicators for the final model candidate.

- *Training and testing Gini* – the aim is to get the testing Gini value close to the training Gini (for stability) and both of them reasonably high (for performance). The target level can be partially given by the original training Gini for the initial models.
- *Significance of parameters* – all parameters of the model should be significant.
- *Correlation matrix* – the correlation matrix should not contain high correlations between different predictors (except for a correlation with intercept and correlations among different categories of the same predictor).
- *Degrees of freedom* – the lower number of degrees of freedom (with the same performance otherwise) the simpler and more stable model.

Example 4.8. *After the optimization process we can get a candidate model with 7 predictors, 15 degrees of freedom, training Gini 43.2%, testing Gini 42.8% and acceptable correlation matrix. See the variable selection in Table 4.9, the parameter estimates and category description*[13] *in Table 4.10 and the correlation matrix in Table 4.11.*

Predictor	Description	Categories	DF	Test Value	P-value
x_1	Retailer Type	2	1	88.7	< 0.0001
x_2	Sex & Education	4	3	165.2	< 0.0001
x_3	Industry & Document	4	3	88.0	< 0.0001
x_4	Goods & Loan	3	2	34.3	< 0.0001
x_5	Submitting Time	3	2	28.2	< 0.0001
x_6	Client Age	3	2	76.5	< 0.0001
x_7	Down Payment	(numerical)	1	241.7	< 0.0001

Table 4.9: Variable selection example.

4.4 Stability Testing and Validation

After having several candidate models (usually 2–3) we can test the time stability of predictors, performance on a new sample and average score precision for some important segments – to help us choose the most suitable model.

[13]Some of the category descriptions are not specific to protect the anonymity and know-how of the data provider.

Parameter	Predictor Description	Category Description	Estimate
β_0	Intercept	Intercept	-3.5266
β_{10}	Seller Type	Contractor	0
β_{11}	Seller Type	Employee	-0.5009
β_{20}	Sex & Education	Male: University & Female	0
β_{21}	Sex & Education	Male: Senior High	0.8288
β_{22}	Sex & Education	Male: Technical High & Junior College	0.7408
β_{23}	Sex & Education	Male: Elementary & Junior High	1.0403
β_{30}	Industry & Document	Combination 1	0
β_{31}	Industry & Document	Combination 2	-1.1173
β_{32}	Industry & Document	Combination 3	-0.2312
β_{33}	Industry & Document	Combination 4	-0.4311
β_{40}	Goods & Loan	Cheap Others	0
β_{41}	Goods & Loan	Cheap Mobile & Expensive Others	0.4336
β_{42}	Goods & Loan	Expensive Mobile	0.6738
β_{50}	Submitting Time	Weekend	0
β_{51}	Submitting Time	Weekday: 14–18 & 21–23	0.3254
β_{52}	Submitting Time	Weekday: Others	0.1498
β_{60}	Client Age	28+	0
β_{61}	Client Age	18–23	-0.5249
β_{62}	Client Age	24–27	-0.2248
β_7	Down Payment	Initial Payment Percentage	-4.4932

Table 4.10: Parameters estimates example.

4.4.1 Stability Testing

First we look at the time stability of the used predictors. We divide the development time period into smaller segments (e.g. months or weeks) and compare the performance on these segments. Ideally all categories should have a stable share in time and stable default rates.

- *Stable share* – instability of share (and especially a significant trend) on the development sample suggests an instability of future predictors. This can lead to a changing distribution of score.[14]
- *Stable default rate* – this is manly to ensure that there are no changes in the population behavior for the selected predictors, i.e. that the group order (according the default rate) keeps stable in time.

Example 4.9. *As an example, we take the combined predictor of sex and education and show the stability on 8 time segments (i.e. half a month for each segment).*

[14]Stability of the score distribution is an assumption of a stable underwriting model. Otherwise, for fixed score cutoffs (i.e. values of score which divide the applicants into different underwriting strategies) we would get a changing distribution of underwriting strategies shares, and thus instable approval rate.

Variable	x_0	x_{11}	x_{21}	x_{22}	x_{23}	x_{31}	x_{32}	x_{33}	x_{41}	x_{42}	x_{51}	x_{52}	x_{61}	x_{62}	x_7
x_0	1.00	−0.21	−0.30	−0.27	−0.32	−0.22	−0.32	−0.25	−0.70	−0.71	−0.20	−0.17	−0.24	−0.18	−0.42
x_{11}	−0.21	1.00	0.04	−0.03	0.11	−0.04	0.00	−0.07	0.05	0.03	0.01	0.02	0.00	0.00	0.09
x_{21}	−0.30	0.04	1.00	0.63	0.66	0.04	−0.01	−0.04	0.01	0.01	0.00	0.00	0.07	0.01	−0.03
x_{22}	−0.27	−0.03	0.63	1.00	0.67	0.00	−0.02	−0.06	0.00	−0.01	0.01	0.00	0.01	−0.02	−0.04
x_{23}	−0.32	0.11	0.66	0.67	1.00	0.03	−0.05	−0.06	−0.01	0.01	0.00	0.00	0.08	0.03	−0.04
x_{31}	−0.22	−0.04	0.04	0.00	0.03	1.00	0.47	0.39	0.02	0.03	0.04	0.01	−0.04	0.00	0.01
x_{32}	−0.32	0.00	−0.01	−0.02	−0.05	0.47	1.00	0.65	0.01	0.02	0.05	0.01	−0.02	−0.01	−0.03
x_{33}	−0.25	−0.07	−0.04	−0.06	−0.06	0.39	0.65	1.00	0.01	0.04	0.06	0.01	−0.07	−0.03	0.02
x_{41}	−0.70	0.05	0.01	0.00	−0.01	0.02	0.01	0.01	1.00	0.91	−0.01	−0.01	0.01	0.01	0.08
x_{42}	−0.71	0.03	0.01	−0.01	0.01	0.03	0.02	0.04	0.91	1.00	−0.02	−0.02	0.00	0.00	0.16
x_{51}	−0.20	0.01	0.00	0.01	0.00	0.04	0.05	0.06	−0.01	−0.02	1.00	0.53	0.00	0.00	−0.02
x_{52}	−0.17	0.02	0.00	0.00	0.00	0.01	0.01	0.01	−0.01	−0.02	0.53	1.00	0.02	0.01	0.00
x_{61}	−0.24	0.00	0.07	0.01	0.08	−0.04	−0.02	−0.07	0.01	0.00	0.00	0.02	1.00	0.53	0.04
x_{62}	−0.18	0.00	0.01	−0.02	0.03	0.00	−0.01	−0.03	0.01	0.00	0.00	0.01	0.53	1.00	0.00
x_7	−0.42	0.09	−0.03	−0.04	−0.04	0.01	−0.03	0.02	0.08	0.16	−0.02	0.00	0.04	0.00	1.00

Table 4.11: Correlation structure example.

From Figure 4.9 we can see that the shares are quite stable in time. The default rate stability in Figure 4.10 is not perfect but still the risk order of the selected categories is consistent.

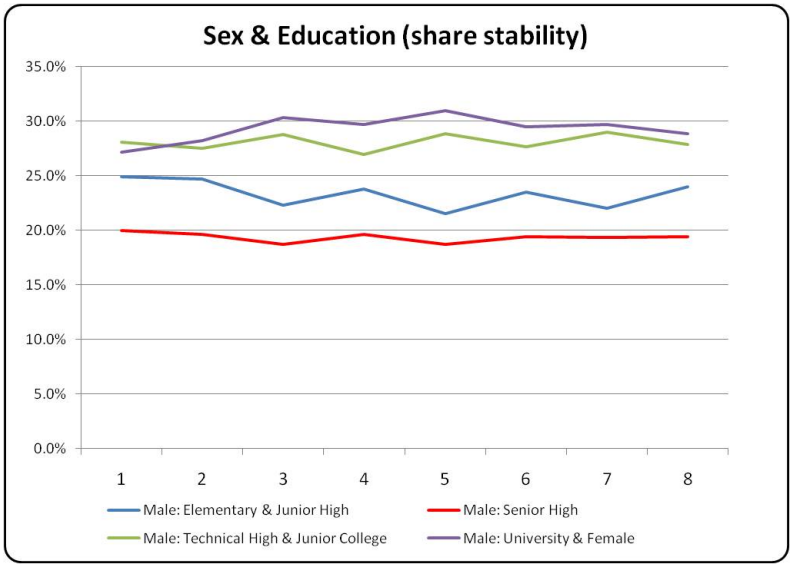

Figure 4.9: Share stability testing example – Sex & Education.

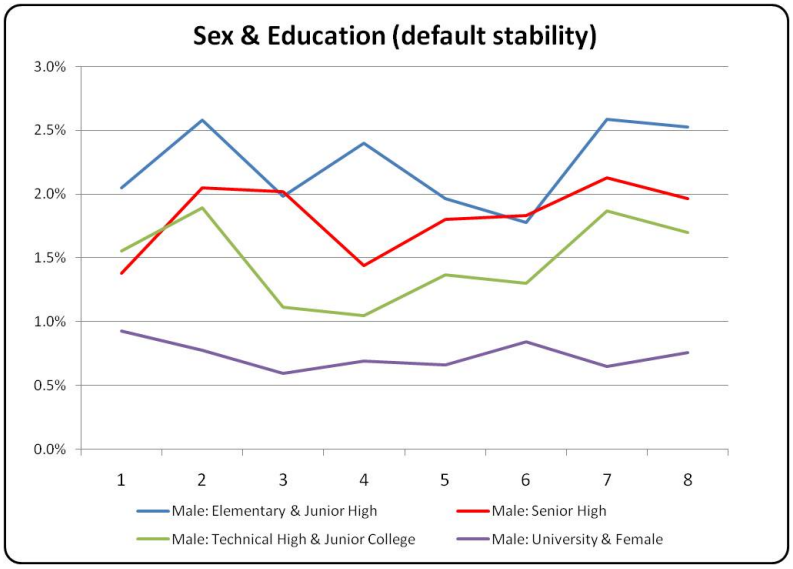

Figure 4.10: Default stability testing example – Sex & Education.

4.4.2 Testing on Validation Sample

The next step of the candidate model testing process is the validation sample testing. As introduced in Paragraph 4.1.3, a validation sample consists of the data

from a later time period than the development sample data. Therefore, this is a kind of ex ante prediction performance testing.

Same as for the testing sample, we first prepare all the necessary predictors and their categorizations used for the candidate models, and then compute the score for every client using the estimated parameters. Finally, we compute the validation Gini and compare with the training and testing Gini. Small difference of training and validation Gini suggests a good time stability of the model.

Example 4.10. *We test our candidate model of Example 4.8 on a one month validation sample data. Comparing the training Gini 43.2% and validation Gini 41.9%, we can conclude that this model is quite stable in time.*[15]

4.4.3 Testing of Average Score

The final step of the pre-selection testing is the testing of average score. As we know from Chapter 1, score represents the probability of default; and therefore, we would like to know how much are the average score values consistent with the average default rate on some specific segments.

First, we can concentrate on the consistence for different values of score. Here it is important to know whether high or low values of score overestimate or underestimate the original default rates. For this we can divide the data sample according to score into deciles – to compare the average score and default rate.

Example 4.11. *For our candidate model of Example 4.8 we divide the data into deciles according to score, and plot the corresponding values of average score and default rate into a scatter plot. Ideally, all the points should lay on the diagonal of the square. From Figure 4.11 we can see that the score values are quite consistent with the default rates and there is no systematical overestimating or underestimating.*

Secondly, we can be interested in the consistence for specific segments of the data. This is important mainly in the case where the segments are particularly important for our business strategies – and in this case we want to well understand the score value meaning (i.e. as the expected default rate).

Example 4.12. *Again we take as example our candidate model of Example 4.8. In Table 4.12 we see the average scores and original default rates for all different commodity types in our data sample.*

[15] Of course, we can expect a different Gini value after implementation to the real production (because of the sample inconsistence described in Paragraph 4.1.4), and this value decreasing in time due to changing population behavior; however, with this performance on the validation sample we can believe that the decrease will not be that rapid.

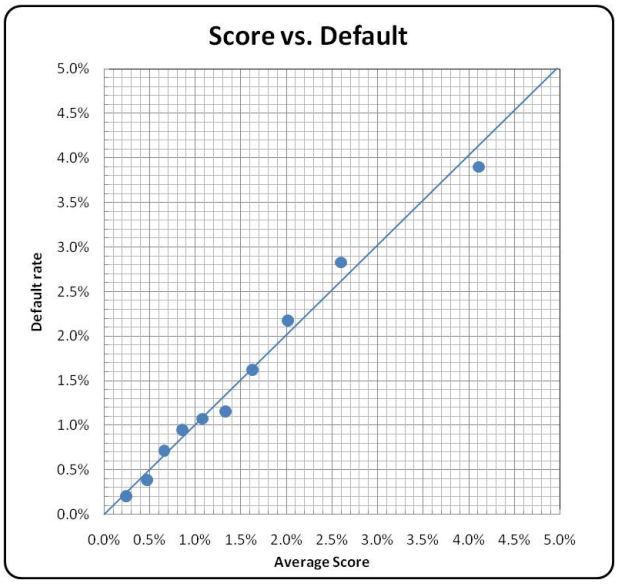

Figure 4.11: Example: average score and default rate comparison.

Segment	Average score	Default rate
Commodity 1	1.64 %	1.62 %
Commodity 2	0.87 %	0.95 %
Commodity 3	1.44 %	1.35 %
Commodity 4	0.81 %	0.73 %

Table 4.12: Example: average score and default rate comparison – Commodity segment.

4.5 Final Model Selection

Finally, after performing all the necessary tests (some of them described in Section 4.4; others suggested by the developer based on the actual needs) we can select the final model, which best fits our requirements.

After the final model is selected, we can consider the final re-computation of its parameters on the full available data - i.e. on the data covering all training, testing and validation samples. Especially for small development samples, this can help to strengthen the stability of the model.

Example 4.13. *After re-computation of the parameters of our final model (i.e. the candidate model from Example 4.8), we get the new set of parameters estimates in Table 4.13. If we compare the results with the original training estimates in Table 4.10, we can see that there is a visible difference. However, the main difference is in decreasing the standard error of these parameters and thus in reducing the confidence interval size (see Table 4.14).*

Parameter	Predictor Description	Category Description	Estimate
β_0	Intercept	Intercept	−3.729
β_{10}	Seller Type	Contractor	0
β_{11}	Seller Type	Employee	−0.5162
β_{20}	Sex & Education	Male: University & Female	0
β_{21}	Sex & Education	Male: Senior High	0.8525
β_{22}	Sex & Education	Male: Technical High & Junior College	0.7629
β_{23}	Sex & Education	Male: Elementary & Junior High	1.01
β_{30}	Industry & Document	Combination 1	0
β_{31}	Industry & Document	Combination 2	−0.9452
β_{32}	Industry & Document	Combination 3	−0.2622
β_{33}	Industry & Document	Combination 4	−0.4171
β_{40}	Goods & Loan	Cheap Others	0
β_{41}	Goods & Loan	Cheap Mobile & Expensive Others	0.5148
β_{42}	Goods & Loan	Expensive Mobile	0.7759
β_{50}	Submitting Time	Weekend	0
β_{51}	Submitting Time	Weekday: 14–18 & 21–23	0.3168
β_{52}	Submitting Time	Weekday: Others	0.1987
β_{60}	Client Age	28+	0
β_{61}	Client Age	18–23	−0.4915
β_{62}	Client Age	24–27	−0.2042
β_7	Down Payment	Initial Payment Percentage	−3.9964

Table 4.13: Parameters estimates example – full model.

4.6 Monitoring in Production

In this section we add one last comment to the monitoring process after the final model is implemented in the real production. Again, it is very important to understand that the development sample is different from the original (real) population; and therefore, we have to closely monitor the scoring performance in production as well. We should focus mainly on the following characteristics.

- *Monitoring of Gini performance* – it is useful to monitor the Gini performance in time after implementation to the real production. This way we can observe the possibly decreasing trend of the scorecard's diversification power. Moreover, it is recommended to observe the Gini performance on different target variables (i.e. first payment default, second payment default etc.) and on different segments (i.e. regions or commodities) to see the sensitivity of the function. In addition to that, we can also monitor the lift development in time.

- *Monitoring of predictor shares and default rates* – this is the same monitoring as in Paragraph 4.4.1. Here we can notice the changing share or default rate performance of different predictors.

Parameter	Training Estimate	Standard Error	Final Estimate	Standard Error
β_0	-3.5266	0.1826	-3.729	0.1385
β_{10}	0	–	0	–
β_{11}	-0.5009	0.0532	-0.5162	0.0398
β_{20}	0	–	0	–
β_{21}	0.8288	0.0868	0.8525	0.0646
β_{22}	0.7408	0.0840	0.7629	0.0626
β_{23}	1.0403	0.0814	1.01	0.0612
β_{30}	0	–	0	–
β_{31}	-1.1173	0.1247	-0.9452	0.0879
β_{32}	-0.2312	0.0736	-0.2622	0.0551
β_{33}	-0.4311	0.0912	-0.4171	0.0678
β_{40}	0	–	0	–
β_{41}	0.4336	0.1288	0.5148	0.0990
β_{42}	0.6738	0.1330	0.7759	0.1021
β_{50}	0	–	0	–
β_{51}	0.3254	0.0618	0.3168	0.0464
β_{52}	0.1498	0.0669	0.1987	0.0497
β_{60}	0	–	0	–
β_{61}	-0.5249	0.0613	-0.4915	0.0461
β_{62}	-0.2248	0.0739	-0.2042	0.0555
β_7	-4.4932	0.2890	-3.9964	0.2119

Table 4.14: Parameters estimates example – standard error comparison.

- *Monitoring of average score* – this monitoring is again similar to the tests in Paragraph 4.4.3. However, the main added value is in the context of the whole underwriting process. If we have for example different underwriting strategies for different score segments after the first scoring (i.e. it is possible to reject the application later based on some additional information), we can observe the difference between average score and average default rate as the result of this additional underwriting strategy performance.[16]

[16]This can be understood in the following way – we see that the expected default rate of a particular segment (i.e. the average score) is higher, but it is the result of the additional strategy that the worst cases of this segment are rejected, and thus the the real default rate is lower than "expected" by observing the average score.

Conclusions

The aim of this work was to describe the practical application of the logistic regression model in the field of credit scoring. Also the scoring development process was introduced and illustrated on real data examples from a financial market. This paper can be thus used as a basic guideline for beginning underwriting analysts in different financial companies as well as for anybody interested in this field.

Although the risk-based underwriting process is probably the most common approach currently used by banks and other financial institutions, we can think of some generalizations in the terms of expected profit calculation and dynamic scoring systems. In that case clients are evaluated not solely based on their risk characteristics but mainly based on their expected profitability for the loan providing company – thus it can happen that a client having higher probability of default gets higher interest rate on the loan to balance the risk. This approach then enables an advanced underwriting management to maximize the company's profit.

All the models in this thesis are oriented to estimate the credit risk of one single customer. However, in the terms of portfolio management and portfolio risk evaluation it is important to consider also the fact that the customers' defaults can be dependent (e.g. a bankruptcy of one employer can cause a default of many customers). Therefore, the dependence structure of the portfolio has to be described and used in the portfolio modeling. One of the most famous (and also most criticized) solutions introduced in Li (2001) uses copula functions to determine the correlation structure. For more details about portfolio credit risk, copulas and correlation structure we refer to McNeil et al. (2005) and Rychnovský (2010).

For wider context of credit risk modeling we can recommend for example Hull (2009) or Witzany (2010). Here we can find the connection of the probability of default modeling to the capital requirements given by Basel II (2001) as well as some more references to the portfolio credit risk management.

Bibliography

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

Aspey, J., Hinder, J., and Lucas, A. (2003). Rhino Risk Mission Statement [online]. http://www.crc.man.ed.ac.uk/conference/archive/2003/presentations/lucas.pdf (2011-08-05).

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).

Collett, D. (2003). *Modelling survival data in medical research*. CRC press. ISBN 978-15-848-8325-8.

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.

Cox, D.R. (1975). Partial likelihood. *Biometrika*, volume 62, no. 2, pages 269–276. ISSN 0006-3444.

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.

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

Hull, J. (2009). *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.

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

Li, D. (2001). On default correlation: A copula function approach. *Journal of fixed income*, volume 9, pages 43–54. ISSN 1059-8596.

Lucas, A. (2004). Gini coeficient [online]. Rhino Risk Ltd., http://www.rhinorisk.com/Publications/Gini Coefficients.pdf (2011-08-05).

McNeil, A.J., Frey, R., and Embrechts, P. (2005). *Quantitative risk management: Concepts, techniques and tools*. Princeton University Press. ISBN 0-691-12255-5.

Merton, R. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of finance*, volume 29, no. 2, pages 449–470. ISSN 0022-1082.

Rao, C.R. (1973). *Linear statistical inference and its applications*. Wiley Series in Probability and Statistics. ISBN 978-04-717-0823-0.

Rychnovský, M. (2008). *Postupná výstavba modelů ohodnocení kreditního rizika (Step by step credit risk model construction)*. Bachelor Thesis, Faculty of Mathematics and Physics, Charles University, Prague.

Rychnovský, M. (2010). *Portfolio credit risk models*. Master Thesis, Faculty of Sciences, Vrije Universiteit Amsterdam.

Schönbucher, P. (2003). *Credit derivatives pricing models: Models, pricing and implementation*. John Wiley & Sons, Ltd. (UK). ISBN 978-04-708-4291-1.

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.

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

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

Appendix A

Different Approaches to Scoring Modeling

Although this paper is mainly dedicated to the standard logistic regression scoring models, we would like to mention two different approaches to the probability of default modeling. This chapter is based on Rychnovský (2010) and original sources.

A.1 Dynamic Models

Whereas the scoring models estimate only the probability of default based on defaults occurring in some predefined horizon, dynamic models understand the *time until default* as a random variable. One of the most common dynamic approaches is based on survival analysis. This section is based mainly on Kalbfleisch et al. (1980) and Collett (2003).

A.1.1 Survival Analysis

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. In the case of probability of default modeling we model the time until default of the client with given characteristics. First we introduce several terms of survival analysis.

Assume that X is an absolutely continuous nonnegative random variable representing the time to default of a client. Denote F the distribution function and f the

density of X. Then we define a *hazard function* (or *intensity*) of the client as

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

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

$$\lambda(t) = \lim_{h \to 0+} \frac{F(t+h) - F(t)}{h} \frac{1}{S(t)} = \frac{f(t)}{S(t)} = -\frac{\mathrm{d}}{\mathrm{d}t} \ln S(t). \tag{A.2}$$

From (A.2) we get also a converse relation

$$S(t) = \exp\left[- \int_0^t \lambda(u) \mathrm{d}u \right]. \tag{A.3}$$

Finally, we define a *cumulative hazard function* as

$$\Lambda(t) = \int_0^t \lambda(u) \mathrm{d}u = -\ln S(t). \tag{A.4}$$

A.1.2 Cox Model

D. R. Cox in Cox (1972) assumed the hazard function of subject k at time t in the form

$$\lambda(t; \boldsymbol{x}_k) = \lambda_0(t) \exp(\boldsymbol{x}_k' \boldsymbol{\beta}), \tag{A.5}$$

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

$$\frac{\lambda(t; \boldsymbol{x}_k)}{\lambda(t; \boldsymbol{x}_l)} = \frac{\exp(\boldsymbol{x}_k' \boldsymbol{\beta})}{\exp(\boldsymbol{x}_l' \boldsymbol{\beta})}$$

depends only on clients' characteristics, the Cox model is often called the *proportional hazards model*.

In Cox (1975) Cox introduced a generalization of (A.5) by implementing time dependent explanatory characteristics $\boldsymbol{x}_k(t)$. This model then assumes the hazard function in the form

$$\lambda(t, \boldsymbol{x}_k(t)) = \lambda_0(t) \exp(\boldsymbol{x}_k(t)' \boldsymbol{\beta}). \tag{A.6}$$

The corresponding survivor function is then

$$S(t, \boldsymbol{x}_k(t)) = P(T > t | \boldsymbol{x}_k(t)) = \exp\left[- \int_0^t \lambda_0(u) \exp(\boldsymbol{x}_k(u)' \boldsymbol{\beta}) \mathrm{d}u \right].$$

Then for the case of discrete time and no multiple defaults at any time, we can derive the Breslow-Crowley maximum likelihood estimator of the baseline hazard function. If n is the number of clients in our database, $Y_k(t)$ is an indicator that client k has not defaulted until time t, and $\mathrm{d}N_k(t)$ is an indicator that client k defaulted in the time interval $(t-1, t]$, we can estimate the baseline hazard function as

$$\widehat{\lambda}_0(t) = \frac{\sum_{k=1}^{n} \mathrm{d}N_k(t)}{\sum_{k=1}^{n} \exp\big(\boldsymbol{x}_k'(t)\boldsymbol{\beta}\big) Y_k(t)}. \tag{A.7}$$

For $\boldsymbol{\beta}$ we then substitute an estimate $\widehat{\boldsymbol{\beta}}$. If m is the number of defaulted clients and $t_1 < \cdots < t_m$ are the observed default times, we can define auxiliary functions $\mathcal{M}$ and $\mathcal{N}$ as

$$\mathcal{M}(\boldsymbol{\beta}, t_k) = \frac{\exp(\boldsymbol{x}_k(t_k)'\boldsymbol{\beta})}{\sum_{i=1}^{n} Y_k(t_k) \exp(\boldsymbol{x}_k(t_k)'\boldsymbol{\beta})}$$

and

$$\mathcal{N}(\boldsymbol{\beta}, t_k) = \sum_{i=1}^{m} Y_k(t_k) \boldsymbol{x}_k(t_k) \mathcal{M}(\boldsymbol{\beta}, t_k).$$

Using this notation, the estimate of $\boldsymbol{\beta}$ can be computed from the expression for the partial likelihood function by solving the equation

$$\sum_{k=1}^{m} [\boldsymbol{x}_k(t_k) - \mathcal{N}(\boldsymbol{\beta}, t_k)] = 0. \tag{A.8}$$

The t-year probability of default can be in terms of (A.3) and (A.6) then estimated as

$$\pi(\boldsymbol{x}, t) = 1 - \exp\left[- \int_0^t \widehat{\lambda}_0(t) \exp(\boldsymbol{x}'\widehat{\boldsymbol{\beta}}) \mathrm{d}u \right], \tag{A.9}$$

where $\boldsymbol{x}$ is the vector of clients characteristics.

For further details as well as formulas for multiple defaults we refer to Collett (2003). As a parametric alternative to the Cox model also the *Accelerated Failure Time (AFT) model* can be used (see e.g. Kalbfleisch et al. (1980)).

A.2 Structural Models

Both the scoring models and and dynamic models belong to the class of so called *reduced form models*. These are easily calibrated to estimate the probability of default but give no information about the circumstances of the default. This is on the other hand the main aim of so called *structural* or *firm-value models*. These approaches are designed to model the underlying structure of the firm's value in time. In the following text we describe the Merton model as the original concept for many of the present structural models.

A.2.1 Merton Model

The original model was introduced in Merton (1974). Consider a firm with a stochastic value process V_t. Assume that the value V_t of the firm's assets at time t consists of its equity value S_t and its debt value B_t (the value in time t of a single debt obligation with maturity T and face value B). Thus,

$$V_t = S_t + B_t \quad \text{for} \quad t \in [0, T]. \tag{A.10}$$

At time T two situations may occur.

1. $V_T > B$. In this case the value of the firm's assets exceeds the value of the debt. Here the debt is fully recovered and the shareholders get the residual value. Then $B_T = B$ and $S_T = V_T - B$.

2. $V_T \leq B$. In this case the value of the firm's assets is less than its liabilities, and the firm falls into default. Here all the value of the firm's assets is paid to the bondholders. Then $B_T = V_T$ and $S_T = 0$.

Summarizing these two situations, we get similar expressions as are known from derivatives pricing models,

$$S_T = (V_T - B)^+ \tag{A.11}$$
$$B_T = B - (B - V_T)^+. \tag{A.12}$$

Therefore, to develop the Black-Scholes-type pricing model, Merton (1974) makes the following assumptions:

1. There are no transactions costs, taxes, or problems with indivisibilities of assets.

2. There is a sufficient number of investors with comparable wealth levels so that each investor believes that he can buy and sell as much of an asset as he wants at the market price.

3. There exists an exchange market for borrowing and lending at the same rate of interest.

4. Short-sales of all assets, with full use of the proceeds, is allowed.

5. Trading in assets takes place continuously in time.

6. The Modigliani-Miller theorem that the value of the firm is invariant to its capital structure obtains.

7. The term-structure is "flat" and known with certainty. I.e., the price of a riskless discount bond which promises a payment of one dollar at time T in the future is $P(T) = \exp[-rT]$ where r is the (instantaneous) riskless rate of interest, the same for all time.

8. The dynamics for the value of the firm, V_t, through time can be described by a diffusion-type stochastic process with stochastic differential equation

$$dV_t = (\alpha V_t - C)dt + \sigma V_t \mathrm{d}W_t, \tag{A.13}$$

where α is the instantaneous expected rate of return on the firm per unit time, C is the total dollar payouts by the firm per unit time to either its shareholders or liabilities-holders (e.g., dividends or interest payments) if positive, and it is the net dollars received by the firm from new financing if negative; σ^2 is the instantaneous variance of the return on the firm per unit time; $\mathrm{d}W_t$ is a standard Wiener process.

According to (A.11), the equity value at the terminal time T corresponds to a European call option on V_t with strike price B and maturity T. Then the value of the equity today can be expressed in the Black-Scholes-type formula,

$$S_0 = V_0\Phi(d_1) - Be^{-rT}\Phi(d_2), \tag{A.14}$$

where Φ is the cumulative distribution function of the standard normal distribution, and

$$d_1 = \frac{\ln\frac{V_0}{B} + (r + \frac{1}{2}\sigma^2)T}{\sigma\sqrt{T}}, \tag{A.15}$$

$$d_2 = d_1 - \sigma\sqrt{T}. \tag{A.16}$$

Moreover, under the risk neutral measure $\mathbb{Q}$ we have

$$\ln V_T \sim N\left(\ln V_0 + (r - \frac{1}{2}\sigma^2)T, \sigma^2 T\right).$$

And thus at time $t = 0$ we get the probability of default as

$$\pi = \mathbb{Q}(V_T \leq B) = 1 - \Phi\left(\frac{\ln\frac{V_0}{B} + (r + \frac{1}{2}\sigma^2)T}{\sigma\sqrt{T}} - \sigma\sqrt{T}\right), \tag{A.17}$$

i.e. in the form $\pi = 1 - \Phi(d_2)$.

For more information about structural models together with some applications we refer to McNeil et al. (2005) or Schönbucher (2003).