Vienna University of Technology

Diploma Thesis

Identification of Credit Default Drivers via Lasso Estimation in the Logistic Regression Model

Péter Blaskó, BSc.

Supervisor: Assoc. Prof. Mag. Ulrike Schneider, PhD.

Institut für Stochastik und Wirtschaftsmathematik

February 26, 2019

Contents

Abbreviations

AIC Akaike information criterion

AR accuracy ratio

AUC area under the curve

BCBS Basel Comittee on Banking Supervision

BER Bayesian error rate

BIC Bayesian information criterion

CART classification and regression tree

cdf cumulative distribution function

CRR Capital Requirements Regulation

CV cross-validation

EBA European Banking Authority

ECB European Central Bank

ER error rate

FNR false negative rate

FPR false positive rate

GLM generalized linear model

KLIC Kullback-Leibler information criterion

Lasso least absolute shrinkage and selection operator

LR likelihood ratio

ML maximum likelihood

OLS ordinary least squares

PD probability of default

pmf probability mass function

RF random forest

ROC receiver operating characteristics

SSM Single Supervisory Mechanism

TNR true negative rate

TPR true positive rate

Nomenclature

$\mathcal{A}$ active set

$\hat{\beta}^{A}_{\mathbf{Lasso}}$ adaptive Lasso

$\hat{\beta}_{\mathbf{Lasso}}$ Lasso

$L(\cdot)$ likelihood function

$\mathcal{L}(\cdot)$ objective function in an optimization problem

$l(\cdot)$ log-likelihood function

$\lambda_{\mathbf{max}}$ minimum value of λ which yields $\hat{\beta}_{\text{Lasso}}(\lambda) = 0$.

n number of observations

p number of covariates

$\hat{p}$ predicted probabilities

sgn sign function

$\hat{y}$ predicted values

Acknowledgement

In this place, I would like to thank all the people who made it possible for me to write this thesis.

First, especially my supervisor Dr. Ulrike Schneider, whose expertise in econometrics and especially Lasso regression awoke my interest for this topic in the first place. I would like to thank her for her guidance throughout the development of this work.

My parents, who supported me in every manner that they could, financially, morally, allowing me to enjoy such a level of high education.

My brother and sister, who always made me feel special with what I do.

And finally, all my class mates, who were always open for fascinating discussions.

Chapter 1

Introduction

Assessing the risk that a borrower will not be able to repay his debt or fail to meet his interest payment obligations is as old a task as money lending itself. The general interest in quantifying this risk by building internal probability of default (PD) models, however, only increased considerably when the Basel Comittee on Banking Supervision (BCBS) proposed its banking regulation framework Basel II in 2004, which allows banks to assess capital requirements according to the internally estimated credit risk of their counterparts. While discriminating between financially stable and unstable clients was the top priority for regulators during the time before the financial crisis 2008-09, recent developments shift their attention towards model calibration for performance in different times in the macroeconomic cycle. This has been influenced by the European Central Bank (ECB) and European Banking Authority (EBA) in the form of regular obligatory stress testing exercises for a large number of banks[1], which test banks on capital requirement sensitivity during strong downturns of the macroeconomic environment. In addition, the introduction of the new IFRS 9 accounting standard encourages many banks to incorporate macroeconomic variables into their PD model frameworks.

Due to these historical developments, drivers for credit defaults are frequently examined in a two-step approach. First, qualitative information on the borrower is used to create a rank – the credit score – according to characteristics of the individual. The current credit score should reflect the probability for the borrower to be unable to meet their payment obligation in the following year and therefore is the tool for measuring the average credit risk of the counterpart. In addition, there is clear evidence on the dependence of the PD on the macroeconomic environment, meaning that the probability to default during a macroeconomic downturn period is higher even if the credit scores do not change. This has not only been extensively tested in literature (for credit card holders e.g. in Bellotti & Crook, 2009), but also acknowledged as a fact by the EBA as well as the ECB and therefore part of the regular stress testing exercises

[1] All banks subject to the Single Supervisory Mechanism (SSM) are part of the stress testing exercises conducted by the EBA or the ECB.

for SSM-banks conducted by both EBA and ECB. Therefore, in the second step, PDs are assigned to the individuals depending on their current credit score as well as the macroeconomic environment.

Scoring models are based on the data available on the corresponding counterpart. For example, if the borrower is a private individual, this score can depend on socio-demographic (Oesterreichische Nationalbank, 2004, page 30) information such as age, profession or place of residence, but also on behavioral factors such as past due status and amounts thereof (Wang et al., 2015). The methodology to develop a scoring model reaches from expert judgment-based score cards to the usage of more complex forms of regression models, multivariate discriminant analysis, but also more computationally intensive models such as artificial neural networks. Based on the obtained credit score, borrowers are categorized into homogeneous buckets: the rating classes.

One of the industry standard statistical methods of quantifying default probabilities is logistic regression (Abdou & Pointon, 2011). Its advantage lies mostly in the straight-forwardness of the interpretation of its result – for each observation, the outcome can be easily explained by looking at the characteristics of the individual and the values of the corresponding coefficients in the regression model. Furthermore, from a more technical point of view, the maximum likelihood (ML) function can be easily derived, which is why the whole statistical toolkit for ML estimation can be applied. As a final point, it should not be left unmentioned that it is possible to incorporate macroeconomic influence as well as the drivers for a credit score at the same time.

However, the ML estimator in the logistic regression model struggles with the typical difficulties in ML estimation. First, there needs to be a procedure to establish model selection, since all explanatory variables included in the model are assigned a certain contribution to the result. This leads to a risk of overfitting the current data and, as a consequence, poor performance on test data sets. Typically, in order to solve this issue, best subset selection is used in order to select significant variables. However, best subset selection is computationally highly intensive and, additionally, academic research shows that it leads to extremely variable solutions (as stated e.g. in Zou, 2006). Second, multicollinearity in the explanatory variables leads to unreasonable statistical inference on the corresponding coefficients, since the estimation method cannot decide which one of the collinear variables explains the variance in the data.

A modern statistical estimation method, the least absolute shrinkage and selection operator (Lasso), attempts to tackle these issues. It was first introduced by Tibshirani (1996) for the ordinary least squares (OLS) regression model; since then, significant amount of research has been conducted on the Lasso due to its popularity resulting from its automatic model selection feature. The extension to general ML models is viewed as only one of the achievements of the past years. The Lasso incorporates a tuning parameter λ as well as a penalty on the size of the coefficients which shrinks some of them towards zero. In fact, some of the coefficient estimates then obtain the exact value of zero.

$$\hat{\beta}_{\text{Lasso}}(\lambda) = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left(f(\beta; \mathbf{X}, y) + \lambda \sum_{i=1}^{p} |\beta_i| \right) \tag{1.1}$$

The automatic model selection feature of the Lasso means that if the contribution of a specific factor to the dependent variable is neglectably small, the value of its estimated contribution is set to zero. The number of variables with contribution set to zero, however, depends on the size of the tuning parameter λ. The model selection procedure therefore breaks down to a one-dimensional problem of optimizing the tuning parameter to its optimal value.

It is shown in Zou (2006) that certain small adaptions to the Lasso – resulting in the adaptive Lasso – lead to a consistent estimation of the parameter in generalized linear models. Furthermore, the Lasso can always be computed, even if the column rank of the matrix does not equal the number of explanatory variables, e.g. when the number of variables included in the regression exceeds the number of observations. As a consequence, it is possible to include a high number of explanatory variables in the model – even if linear dependence between the explanatory variables is not even in this case truly advisable.

Fast algorithms have already been implemented to compute the Lasso path (1.1), which provide a possibility to conduct model selection even when the number of explanatory variables is large. Therefore, it is possible to include a high number of desired variables into the regression and detect influence of those possible explanatory variables on the dependent variable, which have not been considered influential before. The insignificant variables are ruled out during the model selection process.

The assumption of a linear influence of the explanatory variables on the log-odds of the probabilities is one of the greatest criticisms of logistic regression. Variable transformations may be used to tackle this issue, however, any function could be used for transforming the variables and therefore it is not clear, which transformation would yield the best results. Discretizing the covariates and applying the fused Lasso, a useful modification of the Lasso, allows an analysis of the influence. This avoids arbitrary transformations of the variables, conversely, the influence of the explanatory variables are tested and non-linearities captured.

This work uses credit bureau data from kaggle and attempts to find significant credit default drivers by using the logistic Lasso (Chapter 2). The data are transformed and discretized in order to check for non-linear influences (Chapter 4) using the fused Lasso (Section 2.3.3). Model selection is conducted via cross-validation using the AUC as the performance measure on out-of-sample data or via AIC/BIC on in-sample data (Chapter 3). Results of model selection and comparison between selected estimators are presented in Chapter 5.

Chapter 2

Estimating Credit Default Drivers via the Lasso

In its original form as introduced by Tibshirani (1996), the Lasso refers to a shrinkage estimator in the linear regression model. However, its application can be extended to more general types of models, e.g. all parametrized statistical models where parameter estimation is based on the optimization of a function. This includes for example generalized linear model (GLM) such as logistic regression as well as models descending from survival analysis techniques[1]. The methodology to use the Lasso within the logistic regression model will be outlined in Section 2.2.

2.1 Preliminary Work

Although Lasso regression is a relatively new method for estimating the coefficients in the linear regression model, it has already been thoroughly studied in hundreds of publications. Consequently, algorithms for the calculation of tuning parameter paths were developed for the logistic regression model (e.g. in Park & Hastie (2007) for several types of GLM). Applications can be found in many types of classification problems, encompassing a wide range of fields, such as medicine efficiency assessment, genome (Wu et al., 2009), image (Sun et al., 2014) or text classification (Genkin et al., 2007).

In credit scoring literature, classical logistic regression is probably the most wide-spread method for estimating models and may be considered as the industry standard (Thomas et al., 2002). However, logistic Lasso methods have only been sporadically applied in the context of estimating credit scoring models so far, mostly in the context of general machine learning algorithms and in comparison with other algorithms.

[1] For a textbook treatment of these methods, see e.g. Cameron & Trivedi (2005).

Wang et al. (2015) compare simple Lasso-logistic regression to an ensemble of Lasso-logistic base learners created by bagging (Breiman, 1996) and other tree-based algorithms like classification and regression tree (CART) and random forest (RF) using the same data set as this work. They find that the ensemble does not significantly outperform the single Lasso-logistic regression algorithm as well as the RF in terms of area under the curve (AUC), but outperforms them in terms of F-measure. Choi et al. (2015) use the fused Lasso (Tibshirani et al., 2005) on simulated data as well as on the German credit data available on the UCI machine learning repository especially in order to deal with categorical and ordinal variables. They find that the fused Lasso outperforms the Lasso in terms of misclassification rate.

2.2 Logistic Regression

2.2.1 Overview

From the perspective of the credit lender, at the beginning of a year, two possible outcomes may occur at the end of the same year: either, the client i is unable to repay his debt or else fails to meet his interest payment obligations, the default event, or he is able to do so, the non-default event. Considering the credit default as a random variable Y with two possible outcomes and coding it accordingly, i.e., $[Y = 1]$ corresponds to a default event and $[Y = 0]$ to a non-default event, it can be seen as a non-degenerate Bernoulli distributed variable with parameter $p := P(Y = 1) \in (0, 1)$.[2]

A regression model arises when the probability of a default event is set into relationship with a set of attributes $\mathbf{x} = (x_1, \ldots, x_p)' \in \mathbb{R}^p$ of the client, the covariates or explanatory variables. The link between the covariates $\mathbf{x}$ and the outcome Y is established through the probability p, which is connected to the set of covariates by using a link function $F : \mathbb{R} \to [0, 1]$ and an unknown parameter $\beta \in \mathbb{R}^p$:

$$p = F(\mathbf{x}'\beta) = F(x_1\beta_1 + \ldots + x_p\beta_p). \tag{2.1}$$

In the case of logistic regression, the link function is chosen as $F(z) = \Lambda(z) := \frac{1}{1+\exp(-z)}$ and is therefore strictly monotonically increasing. The choice of the link function is important for the interpretation of the result, which can be conducted by considering the marginal effects $\frac{\partial p}{\partial x_j}$. They measure the dependence of the probabilities on the covariates by capturing the change of probability, if the j-th explanatory variable is increased by a unit. For the logistic regression

[2] The assumption of a PD equal to zero is highly unrealistic – history has shown that even states, which are considered one of the safest investments, can default. In fact, even the EBA prescribes a minimum PD of 0.03% in the Capital Requirements Regulation (CRR). Conversely, application of a PD value of 1 is unnecessary; in this case, the client may as well be assumed as already in default, while we focus solely on the prediction of default probabilities for non-defaulted clients.

model, the effects are as follows:

$$\frac{\partial p}{\partial x_j} = \Lambda(\mathbf{x}'\beta)(1 - \Lambda(\mathbf{x}'\beta))\beta_j = p(1-p)\beta_j. \tag{2.2}$$

For low values of $p \approx 0$, as is typical in credit scoring, β_j can therefore be interpreted as the percentage change of p, if the j-th covariate x_j is increased by a unit.

$$\frac{\partial \log p}{\partial x_j} = \frac{1}{p}\frac{\partial p}{\partial x_j} = (1-p)\beta_j \approx \beta_j.$$

In case of logistic regression, the linear term $z := \mathbf{x}'\beta$ from Equation (2.1) corresponds to the log-odds between probabilities and counter-probabilities $\log(p/(1-p))$, because

$$\log \frac{p(z)}{1-p(z)} = -\log\left(\frac{\frac{1}{1+\exp(-z)}}{\frac{\exp(-z)}{1+\exp(-z)}}\right) = -\log(\exp(-z)) = z.$$

Therefore we can conclude that the log-odds are linearly dependent on the covariates and this term is given special attention when it comes to the analysis of results in logistic regression.

2.2.2 Parameter Estimation

If the parameter β is already known, the logistic regression model provides default probabilities for all counterparties with given covariates $\mathbf{x}$. However, in practice, this is not the case, which is why β needs to be estimated from default data $y = (y_1, \dots, y_n) \in \{0,1\}^n$ and their respective covariates $X = (\mathbf{x}_1, \dots, \mathbf{x}_n)' \in \mathbb{R}^{n \times p}$. For that, we need to assume that the data are realizations of independent Bernoulli distributed random variables $(Y_1, \dots, Y_n)$ with parameters $p_i := P(Y_i = 1|\mathbf{x}_i)$.

The most wide-spread method for parameter estimation in the logistic regression model is Maximum-Likelihood (ML). The ML estimator $\hat{\beta}_{\text{ML}}$ is thereby defined as the solution of the optimization problem of maximizing the likelihood function $L(\cdot)$.

$$\hat{\beta}_{\text{ML}} := \underset{\beta \in \mathbb{R}^p}{\arg\max}\, L(\beta; y, X)$$

The likelihood function $L : \mathbb{R}^p \to \mathbb{R}$ is defined as the joint probability mass function (pmf) of the random variables $(Y_1, \dots, Y_N)$ depending on the parameter β evaluated at the data y and X.

In the following, we will deduct L for the logistic regression model. Since the Y_i are independent, the joint pmf is equal to the product of the marginal pmfs of the Y_i, which are Bernoulli distributed by definition. The pmf of a Bernoulli distributed variable can be written as follows:

$$P(Y_i = y_i) = p_i \mathbb{1}_{\{y_i=1\}} + (1-p_i)\mathbb{1}_{\{y_i=0\}} = p_i^{y_i} \cdot (1-p_i)^{1-y_i}.$$

As a consequence, the likelihood function is as follows:

$$L(\beta; y, X) = P(Y = y|X) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} = \prod_{i=1}^{n} \Lambda(\mathbf{x}_i'\beta)^{y_i}(1-\Lambda(\mathbf{x}_i'\beta))^{1-y_i}.$$

However, for improved numerical conditioning, easier assessment of the scores (the first derivative) and because of its statistical properties, the log-likelihood function $l(\cdot)$ is usually considered instead of the likelihood function.

$$\begin{aligned} l(\beta; y, X) = \log L(\beta; y, X) &= \sum_{i=1}^{n} \{y_i \log \Lambda(\mathbf{x}_i'\beta) + (1-y_i)\log(1-\Lambda(\mathbf{x}_i'\beta)\} \\ &= \sum_{i=1}^{n} \left\{ y_i \log \frac{\Lambda(\mathbf{x}_i'\beta)}{1-\Lambda(\mathbf{x}_i'\beta)} + \log(1-\Lambda(\mathbf{x}_i'\beta) \right\} \\ &= \sum_{i=1}^{n} \{y_i \mathbf{x}_i'\beta - \log(1+\exp(\mathbf{x}_i'\beta)\} \end{aligned}$$

Existence and Uniqueness

If the ML estimator exists, i.e., there is $\beta \in \mathbb{R}^p$, which maximizes the likelihood function, uniqueness is given if the matrix X has full column rank, since the likelihood function is strictly concave in this case.

However, the existence of an optimal value is not easily guaranteed for any continuous and strictly monotonically increasing $F : \mathbb{R} \to (0,1)$. In fact, ML estimation in logistic regression is not possible if the respective covariates of the default and non-default observations can be discriminated with a linear hyperplane (a detailed discussion of this can be found in Albert & Anderson, 1984). In this case, the orthogonal vector v to the hyperplane fulfills that $v \cdot \mathbf{x}_i \leq 0$ for all i with $y_i = 0$ as well as $v \cdot \mathbf{x}_i \geq 0$ for all i with $y_i = 1$.[3] In this case, the value of the objective function can always be increased by moving into the direction of v increasing the scores $\mathbf{x}_i \cdot (\beta + v)$ for all observations with $y_i = 1$ and decreasing the scores for all observations with $y_i = 0$. Since F is strictly monotonically increasing, i.e., increasing scores always lead to increasing probabilities and vice versa, but never actually reaches the values of $\{0,1\}$, the ML estimator does not exist.

[3] If the model contains an intercept, the property of perfect linear discrimination needs to be tested only with the remaining variables and the hyperplane can be affine, i.e., zero in the inequalities above can be replaced by a constant $\eta \in \mathbb{R}$, because $(-\eta, v')' \cdot (1, \mathbf{x}_i')' = -\eta + v \cdot \mathbf{x}_i$.

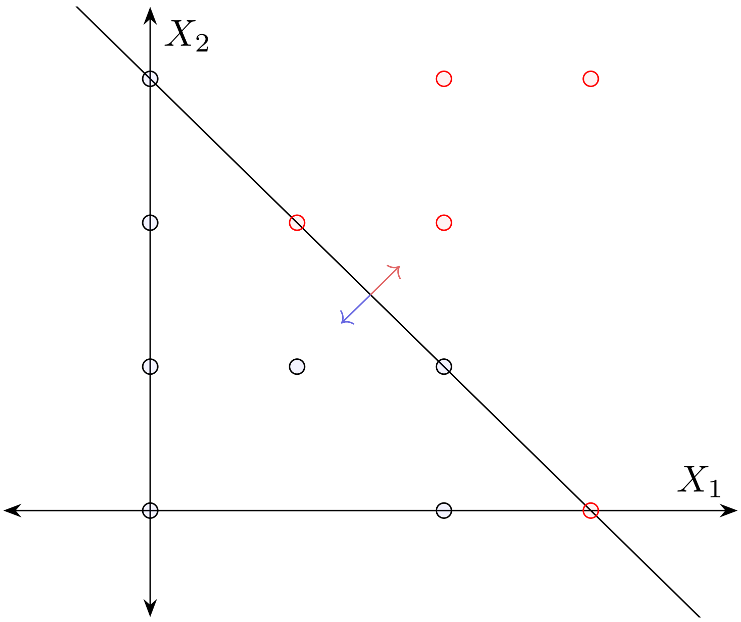

Figure 2.1: Example of a case where the ML estimator does not exist. The red dots denote the default observations, the blue dots the non-default observations. The model contains an intercept and v is the red arrow.

In Silvapulle (1981), a slightly different but equivalent formulation has been proven. It has been shown that the ML estimator exists if and only if the positive half cones spanned by the line vectors of all observations with $y_i = 1$ and those with observations $y_i = 0$ intersect. The following formalization holds true:

Theorem 1 (*Silvapulle, 1981*):

A: Let y, X be the realizations in a logistic regression model, $J := \{i \in \{1, \dots, n\} : y_i = 1\}$ the set of all indices i where $y_i = 1$, $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$ the i-th row vector of X, $S := \{\sum_{i \in J} k_i \mathbf{x}_i : k_i > 0\}$, $F := \{\sum_{i \in J^c} k_i \mathbf{x}_i : k_i > 0\}$ the positive half cones spanned by the row vectors of $X \in \mathbb{R}^{n \times p}$ with indices in J and J^c, respectively, and let $x_{i1} = 1, i = 1, \dots, n$, such that the model includes an intercept.

S: Then, $\hat{\beta}_{\text{ML}}$ exists and the set of optimal values is bounded if and only if $S \cap F \neq \emptyset$. Furthermore, $\hat{\beta}_{\text{ML}}$ is unique if $\text{rk}(X) = p$.

In this work, all variables are transformed into indicator variables (see Section 4.3). Therefore, the consequences of the theorem above will be analyzed for matrixes X consisting of an intercept and indicator variables $X = (\mathbf{1}, \boldsymbol{\delta}_2, \dots, \boldsymbol{\delta}_p), \boldsymbol{\delta}_j \in \{0, 1\}^n, j = 2, \dots, p$ in detail. The theorem in

Silvapulle (1981) implies that each level of each indicator variable necessarily needs to contain observations with $y = 0$ as well as $y = 1$. If this is not the case for any variable $\boldsymbol{\delta}_j, j = 1, \ldots, p$, e.g. $\boldsymbol{\delta}_j = 1$ implies that $y = 1$, then any increase in β_j implies an increase in the value of the likelihood function and the ML estimator does not exist.

However, the condition above is not sufficient for the existence of the ML estimator. For example, consider the case with $p = 3$ and $n = 6$ and the following data:

$$y = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

In this case, every level of each variable individually contains observations with both $y = 1$ and $y = 0$, however, $\hat{\beta}_{\mathrm{ML}}$ does not exist, since the respective sets $F = \{(a, b, c)' : a > c > b > 0\}$ and $S = \{(a, b, c)' : a > b > c > 0\}$ do not intersect.

Obviously, a sufficient condition for the existence of $\hat{\beta}_{\mathrm{ML}}$ is that for every row $\mathbf{x}_i'$ with $y_i = 1$ there exists a corresponding index $j(i)$ with $y_{j(i)} = 0$ and $\mathbf{x}_i = \mathbf{x}_{j(i)}$ and vice versa. However, this condition is not necessary, as it is very strong. A specific rule for the existence of the ML estimator could not be found in this work. The explanations in this chapter shall indicate that existence is not always guaranteed in logistic regression and should therefore always be individually checked for in the estimation process.

Consistency and Asymptotic Normality

The most important reason for the popularity of ML estimation is the availability of an extensive statistical toolkit that has been especially developed for these types of estimators. In particular, asymptotic properties have been of strong interest and therefore have been thoroughly studied e.g. in Amemiya (1985). It was shown that ML estimators require only mild regularity conditions in order to achieve consistency and asymptotic normality.

A sequence of estimators $\hat{\beta}_n$ is called consistent, if it converges in probability to the true parameter, i.e.,

$$\forall \varepsilon > 0 : \lim_{n \to \infty} P\left(\left\| \hat{\beta}_n - \beta \right\| > \varepsilon \right) = 0.$$

$\hat{\beta}_n$ is called asymptotically normally distributed if its difference to the true parameter multiplied by $\sqrt{n}$ converges in distribution to a non-degenerate normal distribution.

$$\lim_{n\to\infty} \sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow[n\to\infty]{d} N(0,\Omega), \quad \Omega > 0.$$

For the logistic regression model in particular, consistency and asymptotic normality are already achieved with restrictions made only on the regressor matrix X. The following formulation can be shown:

Theorem 2

A: In a logistic regression model, with the regressor matrix $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)' \in \mathbb{R}^{n\times p}$ having the following properties:

- $\mathrm{rk}(X) = p$, i.e., X has full column rank
- $\exists \eta > 0 : \|\mathbf{x}_i\| < \eta$ for all $i = 1, \ldots, n$, i.e., the $\mathbf{x}_i$ are uniformly bounded
- $\frac{1}{n}X'X \to C$ for a positive definite matrix $C \in \mathbb{R}^{p\times p}$
- $G_n(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,\mathbf{x}]}(\mathbf{x}_i)$ converges to a non-degenerate cumulative distribution function G.

S: Then, the ML estimator $\hat{\beta}_{\mathrm{ML}}$ is consistent and asymptotically normal.

Further Properties

If the ML estimator exists and is unique, the estimator can be derived from the scores $\frac{\partial l}{\partial \beta}$ of the log-likelihood function.

$$\frac{\partial l(\beta; y, X)}{\partial \beta} = \sum_{i=1}^{n} \left\{ y_i \mathbf{x}_i - \frac{\exp(\mathbf{x}_i'\beta)}{1+\exp(\mathbf{x}_i'\beta)}\mathbf{x}_i \right\} = \sum_{i=1}^{n} \left(y_i - \Lambda(\mathbf{x}_i'\beta)\right)\mathbf{x}_i = \mathbf{0}. \tag{2.3}$$

In the logistic regression model, the predicted probabilities $\hat{p}$ are equal to $\Lambda(\mathbf{x}_i'\hat{\beta})$. Therefore, the scores in Equation (2.3) can be interpreted as a weighted average of residuals. If the model includes an intercept (i.e., $x_{i1} = 1, i = 1, \ldots, n$), the residuals sum up to zero, which results in the well-known fact that the average of predicted probabilities is equal to the average default rate in the data.

$$\frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} \hat{p}_i. \tag{2.4}$$

Equation (2.4) is one of the reasons to prefer the logistic link function to any other distribution function, where this identity does not necessarily hold true.

2.3 The Lasso Estimator

2.3.1 Lasso Estimation in Generalized Linear Models (GLMs)

In its original form, the Lasso estimator was introduced for the linear regression model. Its general idea is to shrink the parameters $\beta_1, \dots, \beta_p$ by introducing a constraint directly on their size. Ideally, the coefficients of non-significant[4] variables obtain the exact value of zero, which excludes them from the model. In order to reach this, the size of the parameter vector is measured in its l1-norm

$$\|\beta\|_1 := \sum_{j=1}^{p} |\beta_j| \,.$$

For estimating the parameter in GLM, pairs of data $(y_i, \mathbf{x}_i')' \in \mathbb{R}^{p+1}, i = 1, \dots, n$ are given, where the y_i are viewed as realizations of random variables Y_i having the same conditional distribution as a random variable $Y \in \Theta_y \subseteq \mathbb{R}$ and $\mathbf{x}_i \in \mathbb{R}^p$, where the following relationship is assumed between the random variable and the predictors:

$$g\left(\mathbb{E}[Y_i|\mathbf{x}_i]\right) = \mathbf{x}_i'\beta, \quad i = 1, \dots, n. \tag{2.5}$$

$g : \mathbb{R} \to \mathbb{R}$ is the so-called link function between the mean of the random variable Y and the predictors. The distribution of Y is assumed to be part of the exponential family. Therefore, its density function, or pmf, can be expressed as

$$f_Y(y|\theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right).$$

Parameter estimation in its most wide-spread form is performed by maximizing an objective function $-\mathcal{L}(\cdot)$, which in many cases corresponds to the log-likelihood function (e.g. in the logistic regression model) or partial likelihood functions (e.g. in the Cox regression model), which yield the ML estimator $\hat{\beta}_{\text{ML}}$.

$$\hat{\beta}_{\text{ML}} = \underset{\beta \in \mathbb{R}^p}{\arg\max} -\mathcal{L}(\beta) = \underset{\beta \in \mathbb{R}^p}{\arg\min}\, \mathcal{L}(\beta). \tag{2.6}$$

Similarly to the ML estimator, the Lasso-estimator is the solution of an optimization problem using the same objective function $\mathcal{L}(\cdot)$. However, as opposed to ML-estimation, a constraint on the parameter size is imposed. The Lasso estimate $\hat{\beta}_{\text{Lasso}}$ is therefore defined as the optimal solution of the optimization problem

$$\hat{\beta}_{\text{Lasso}}(t) := \underset{\beta \in \mathbb{R}^p}{\arg\min}\, \mathcal{L}(\beta), \quad \text{s.t.} \ \ \|\beta\|_1 \le t. \tag{2.7}$$

The introduction of the tuning parameter $t \geq 0$ is convenient, since it generates a transition

[4] In this context, "significant" is not to be interpreted in a classical statistical way, but in the sense that it is significantly improving the value of an objective function.

from $\hat{\beta}_{\text{Lasso}}(0) = \mathbf{0} \in \mathbb{R}^p$ to the ML estimator $\hat{\beta}_{\text{Lasso}}(t) = \hat{\beta}_{\text{ML}}$, which is true for any $t \geq \left\| \hat{\beta}_{\text{ML}} \right\|_1$. This is obvious, since for any $t \geq \left\| \hat{\beta}_{\text{ML}} \right\|_1$, the global solution of problem (2.6) lies within the bounds imposed in problem (2.7).

Tibshirani (1996) lists two main reasons to use the Lasso instead of the OLS estimator in the linear regression model. First, prediction accuracy of the coefficients can be improved by increasing the bias, but decreasing the estimator variance at the same time. Second, the Lasso can set the value of rather insiginificant coefficients to zero (as illustrated in Figure 2.2[5]), which results in a smaller subset of variables included in the model and makes it easier to interpret.

The set of variables with coefficients set to a value different from zero is called the active set and grows with t towards the set of all explanatory variables. It can be shown that the transition from 0 to the OLS estimator is piecewise linear in t (Rosset & Zhu, 2007, page 1017). More specifically, on intervals of t, on which the active set does not change, $\hat{\beta}_{\text{Lasso}}(t)$ is a linear function in t.

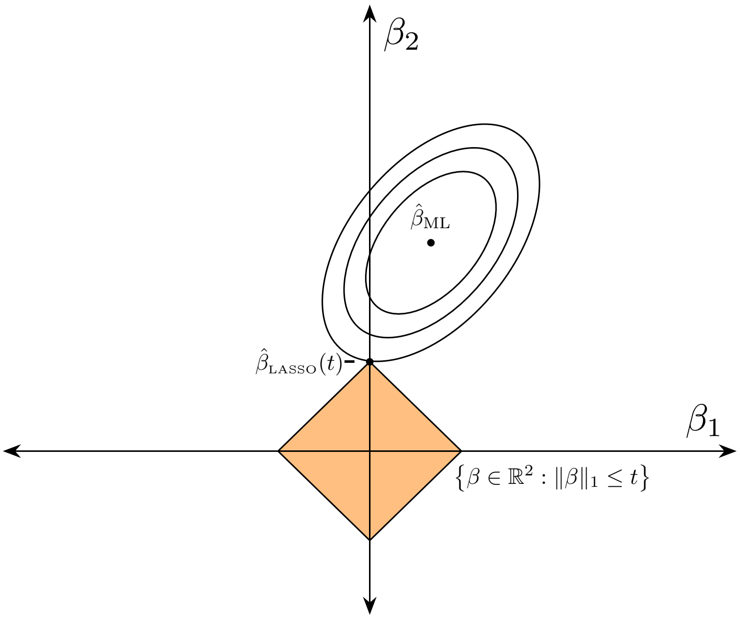
β_2
$\hat{\beta}_{\text{ML}}$
$\hat{\beta}_{\text{LASSO}}(t)$
β_1
$\{\beta \in \mathbb{R}^2 : \|\beta\|_1 \leq t\}$

Figure 2.2: Illustration of the Lasso solution for a given t. In this case, $\hat{\beta}_1 = 0$

[5] In fact, $\hat{\beta}_1$ obtains a value of exactly zero, in this example. This is possible, because the set of admissible values (the shaded area) has non-differentiable corners at the values of zero. The ellipses around the OLS estimator represent sets of parameters β, where the value of the objective function in problem (2.7) is constant.

Typically, the Lasso is analyzed by using its reformulation by its Kuhn-Tucker conditions (see e.g. Rockafellar, 1996, Theorem 28.1). In this case, the Lagrange function $\tilde{\mathcal{L}}$ is written as $\tilde{\mathcal{L}} : \mathbb{R}^p \times \mathbb{R}_+ \to \mathbb{R}, (\beta, \lambda) \mapsto \mathcal{L}(\beta) + \lambda \left\|\beta\right\|_1$. The Lasso can therefore also be viewed as a function of the tuning parameter λ.

$$\hat{\beta}_{\text{Lasso}}(\lambda) := \underset{\beta\in\mathbb{R}^p}{\arg\min}\, \tilde{\mathcal{L}}(\beta, \lambda) = \underset{\beta\in\mathbb{R}^p}{\arg\min} \left\{\mathcal{L}(\beta) + \lambda \left\|\beta\right\|_1\right\}. \tag{2.8}$$

The notation $\hat{\beta}_{\text{Lasso}}(\lambda)$ will also be used to indicate the use of the formulation in (2.8).

As expected, there are no huge differences in these fomulations. In fact, the following theorem can be shown:

Theorem 3 (*Osborne et al., 2000*):

A: In a logistic regression model, denote with $\hat{\beta}_A(t)$ the solution of problem (2.7) and with $\hat{\beta}_B(\lambda)$ the solution of problem (2.8).

S: Then these problems are equivalent in the following sense:

- For any $t < \left\|\hat{\beta}_{\text{LS}}\right\|_1$, there is $\lambda_0 > 0$ such that $\hat{\beta}_B(\lambda_0) = \hat{\beta}_A(t)$.
- For any $\lambda > 0$, there is $t_0 = \left\|\hat{\beta}(\lambda)\right\|_1$ such that $\hat{\beta}_A(t_0) = \hat{\beta}_B(\lambda)$.

We can therefore use both formulations in order to analyze the properties of the Lasso solution path.

Since the objective function of problem (2.8) is only piecewise differentiable, it is useful to reformulate it by introducing new variables $\beta^+, \beta^- \geq \mathbf{0}$ and replacing β with the term $\beta^+ - \beta^-$. This increases the dimension of the optimization problem and the number of constraints[6], but makes the problem globally differentiable such that equalities hold in the Kuhn-Tucker conditions. The transformed problem yields the following Lagrange function $\tilde{\mathcal{L}}$:

$$\tilde{\mathcal{L}}(\beta^+ - \beta^-, \lambda) = \mathcal{L}(\beta^+ - \beta^-) + \lambda\mathbf{1} \cdot (\beta^+ + \beta^-) - \lambda^+\beta^+ - \lambda^-\beta^-. \tag{2.9}$$

The Kuhn-Tucker conditions of problem (2.9) are as follows:

$$s(\beta) := \frac{\partial\mathcal{L}}{\partial\beta} = -\lambda + \lambda^+ \geq -\lambda \tag{2.10a}$$

$$s(\beta) := \frac{\partial\mathcal{L}}{\partial\beta} = \lambda - \lambda^- \quad \leq \lambda \tag{2.10b}$$

[6] For the replacement of β to make sense, β^+ and β^- need to be componentwise nonnegative. In order to omit the constraints, additional Lagrange parameter vectors λ^+ and λ^- are introduced, respectively.

$$\lambda^+, \lambda^- \geq 0, \;\; \lambda^+\beta^+ = \lambda^-\beta^- = 0. \tag{2.11}$$

The inequalities in Equation (2.10) hold componentwise and hold true because of the inequalities in Equation (2.11). The conditions yield that the scores $s(\beta)$ therefore fulfill the following constraint:

$$|s(\beta)| \leq \lambda, \tag{2.12}$$

where the absolute value is again to be understood componentwise. Equality in (2.12) holds for the j-th component if $|\beta_j| > 0$ and the score $s(\beta)_j$ is positive if and only if $\beta_j < 0$. The set $\mathcal{A}$ of all indices j such that $|\beta_j| > 0$ is called the active set. Its components naturally depend on the value of λ – the notation $\beta_{\mathcal{A}}(\lambda)$ will therefore be used to indicate those indices of β that belong to $\mathcal{A}$ at λ. For $\mathcal{A}$, Equation (2.10) can be rewritten as follows:

$$s(\beta)_{\mathcal{A}} = \frac{\partial \mathcal{L}}{\partial \beta_{\mathcal{A}}} = -\operatorname{sgn}(\beta_{\mathcal{A}})\lambda. \tag{2.13}$$

The sign function sgn is again to be understood componentwise.

If we interpret β in Equation (2.13) as an implicitly defined function of λ, we can determine the first derivative $\beta'(\lambda)$ as

$$\frac{\partial^2 \mathcal{L}}{\partial \beta_{\mathcal{A}} \partial \beta'_{\mathcal{A}}} \beta'_{\mathcal{A}}(\lambda) = -\operatorname{sgn}(\beta_{\mathcal{A}}) \implies \beta'_{\mathcal{A}}(\lambda) = \left(\frac{\partial^2 \mathcal{L}}{\partial \beta_{\mathcal{A}} \partial \beta'_{\mathcal{A}}}\right)^{-1} \operatorname{sgn}(\beta_{\mathcal{A}}). \tag{2.14}$$

From Equation (2.14), we can deduce that $\beta'_{\mathcal{A}}(\lambda)$ is constant as long as the active set or the sign of the parameter does not change if the objective function is quadratic, as for the OLS estimator in the linear regression model.

This is a result from Rosset & Zhu (2007), who found that solution paths for Lagrange functions $\tilde{\mathcal{L}}$ having the form $\mathcal{L} + \lambda \cdot P_\lambda$ are piecewise constant if and only if $\mathcal{L}$ is piecewise quadratic and the penalty function P_λ – which in the case of the Lasso equals $\lambda \left\|\beta\right\|_1$ – is piecewise linear in the components of β.

2.3.2 Adaptive Lasso

From formulation (2.8), it can be deduced that the Lasso is not independent from the scale of the explanatory variables. A sensible estimator in a model fulfills the condition that the rescaling of any single variable (i.e., the multiplication of the variable by a constant $c \neq 0$) yields the same predicted values $\hat{y}$. This is fulfilled for any estimator that is defined as a solution of an optimization problem, where the objective function is not explicitly based on β, but on $X\beta$[7],

[7] In this case, the notation $\mathcal{L}(X\beta)$ will be used.

because

$$X\beta = \underbrace{XD}_{=:\tilde{X}}\underbrace{D^{-1}\beta}_{=:\tilde{\beta}} = \tilde{X}\tilde{\beta}, \tag{2.15}$$

where $D = \operatorname{diag}(d_1, \ldots, d_p)$ is any diagonal matrix with entries $d_i \neq 0$. As can be seen from Equation (2.15), the scaling of X results solely in a reparametrization of the optimization problem and therefore does not change the objective function and thus the result.

For the Lasso, however, the objective function does not solely depend on $X\beta$, but also explicitly on β. This can be observed in Equation (2.8), where a multiplicative transformation of X relates to the original problem as follows:

$$\mathcal{L}(\tilde{X}\tilde{\beta}) + \lambda \left\|\tilde{\beta}\right\|_1 = \mathcal{L}(X\beta) + \lambda \sum_{j=1}^{p} |d_j|^{-1} |\beta_j| \,. \tag{2.16}$$

Problems (2.8) and (2.16) therefore differ and in general also result in two different solutions. In essence, the Lasso is not a scale-independent estimator if used like in Equation (2.8). In order to show the effects of this property, assume that the X_i are centered and have unit variance, such that the coefficients $|d_i|$ can be interpreted as the standard deviations in $\tilde{X}_i$. From Equation (2.16), we can deduce that variables with high standard deviations are less penalized than those with relatively small ones.

In order to correct that, two solutions have been proposed. Both solutions introduce adaptive weights $w_j \geq 0,\ j = 1, \ldots, p$, which scale the penalty on the coefficient differently for every coefficient.

$$\hat{\beta}^{A}_{\text{Lasso}}(\lambda) := \left\{ \underset{\beta \in \mathbb{R}^p}{\arg\min}\, \mathcal{L}(\beta) + \lambda \sum_{j=1}^{p} w_j \, |\beta_j| \right\}. \tag{2.17}$$

The obvious choice for the weights w are the respective standard deviations of the variables X_i. Zou (2006) proposed to use the inverse absolute values of the OLS estimator as weights and named $\hat{\beta}^{A}_{\text{Lasso}}$ the adaptive Lasso. He proved the following asymptotic statement for the adaptive Lasso using any root-n-consistent[8] estimator as weights.

[8] A root-n-consistent estimator is defined as a consistent and asymptotically normal distributed estimator with convergence factor $\sqrt{n}$, i.e., $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow[n\to\infty]{d} N(0, \Sigma)$ for a positive definite matrix Σ.

Theorem 4 (*Special case of Zou (2006, Theorem 2) with* $\gamma = 1$):

A: Let $\hat{\beta}$ a root-n-consistent estimator for β in the linear regression model, assume that $X'X \xrightarrow{n\to\infty} C$, where C is a positive definite matrix (i.e., $C > 0$) and choose a sequence λ_n with the properties $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n \to \infty$. Define $\mathcal{A}$ as the set of indices with true coefficient different from zero and $\mathcal{A}_n^*$ the indices with Lasso estimates different from zero.

S: Then the adaptive Lasso estimates $\hat{\beta}^{*(n)} := \hat{\beta}^{\mathcal{A}}_{\text{Lasso}}(\lambda_n)$ with weights $1/\left|\hat{\beta}_i\right|$ satisfy:

1. $\lim_{n\to\infty} P(\mathcal{A}_n^* = \mathcal{A}) = 1$
2. $\sqrt{n}(\hat{\beta}^{*(n)}_{\mathcal{A}} - \beta^*_{\mathcal{A}}) \xrightarrow{d} N(0, C_{11})$

for a positive definite matrix $C_{11} \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{A}|}$.

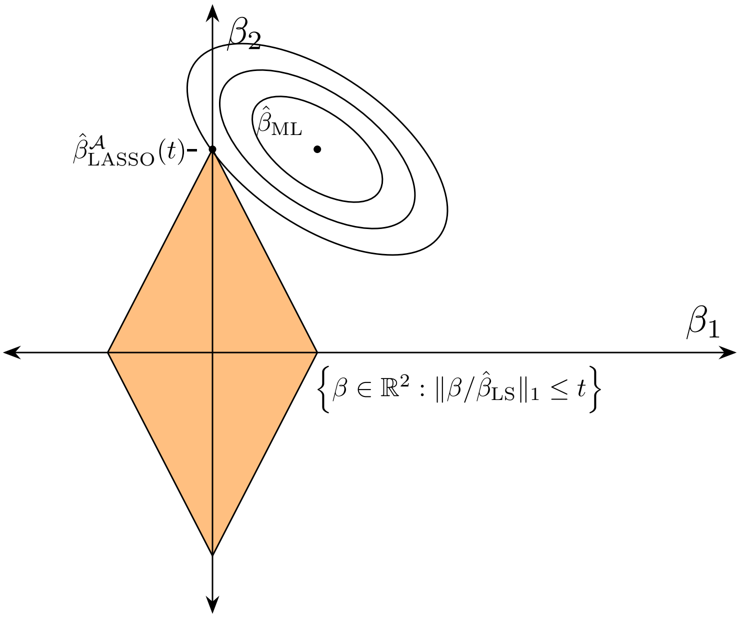
β_2
$\hat{\beta}_{\text{ML}}$
$\hat{\beta}^{\mathcal{A}}_{\text{LASSO}}(t)$
β_1
$\left\{\beta \in \mathbb{R}^2 : \|\beta/\hat{\beta}_{\text{LS}}\|_1 \leq t\right\}$

Figure 2.3: Illustration of the adaptive Lasso solution for a given t. In this case, $t = 1$ and $\hat{\beta}_1^{\mathcal{A}} = 0$
.

2.3.3 Fused Lasso

GLMs are designed to be applied using a set of numerical, especially continuous covariates $X_1, \dots, X_p$, since Equation (2.5) is undefined for non-numerical variables. In order to include a non-numerical, especially categorical explanatory variable X_j in the regression, as an obvious choice, indicator variables are created for each level $\{l_1, \ldots, l_k\}$ of X_j and included in the regression instead of X_j. While this is a valid approach, its usage has a major limitation: some of the categories might have similar influence on the dependent variable, but the Lasso can only identify that if this influence is zero.

As for ordinal variables, similar drawbacks can be observed: While it is possible to treat them as continuous variables, it is not advisable, since their influence might be highly non-linear, which leads to bad estimation results. Another solution is to treat them as categorical variables, accepting the drawback that the approach does not account for the order of the categories. In order to eliminate both limitations, Tibshirani et al. (2005) propose the fused Lasso, which is a slight modification of the Lasso that is designed to include ordinal variables into the model.

In order to introduce the fused Lasso for a single ordinal explanatory variable X with levels $1, \ldots, k$, create variables $X_1, \ldots, X_k$ with $X_j = \mathbb{1}_{\{X=j\}}$ and define $\beta_0 := 0$. In this case, the fused Lasso estimator is defined by

$$\hat{\beta}_{\text{Lasso}}(\lambda) := \underset{\beta \in \mathbb{R}^k}{\arg\min} \left\{ \mathcal{L}(X\beta) + \lambda \sum_{i=1}^{k} |\beta_i - \beta_{i-1}| \right\}. \tag{2.18}$$

Note that this penalty can also be reached with the regular Lasso optimization algorithm by using a transformation of the variables $X_1, \ldots, X_k$. Define $\tilde{\beta}_i := \beta_i - \beta_{i-1},\ i = 1, \ldots, k$. Then, $\tilde{\beta}$ is a linear transformation of β; in matrix notation, $\tilde{\beta} = U\beta$, with

$$U = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & -1 & 1 \end{pmatrix}, \qquad U^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 1 & \cdots & 1 & 1 \end{pmatrix},$$

such that the optimization problem in (2.18) can be viewed as a regular Lasso problem with parameter $\tilde{\beta}$ and variables $\tilde{X}_j := \sum_{i=j}^{k} X_i = \mathbb{1}_{\{X \geq j\}}$.

Finally, it is important to state that the fused Lasso is also appropriate to study non-linear dependencies on continuous variables. For such a variable X, define k breakpoints[9] $x_i,\ i = 1, \ldots, k$ and define variables $X_i := \mathbb{1}_{\{X \geq x_i\}}$. Then the resulting coefficient sequence β_i can be interpreted as the values of a piecewise constant function on the increments $[x_i, x_{i+1})$.

[9] For a detailed description of possible approaches to define breakpoints see Section 4.3.1.

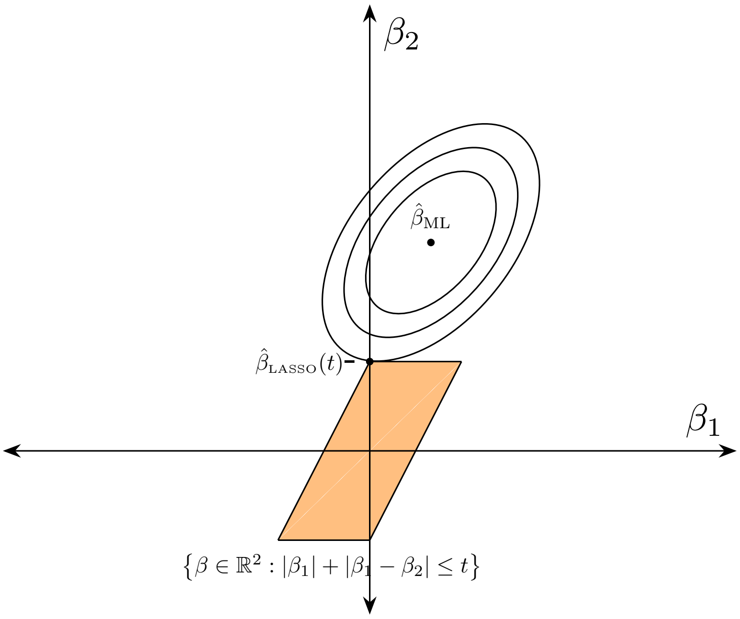
β_2
$\hat{\beta}_{\mathrm{ML}}$
$\hat{\beta}_{\mathrm{LASSO}}(t)$
β_1
$\{\beta \in \mathbb{R}^2 : |\beta_1| + |\beta_1 - \beta_2| \leq t\}$

Figure 2.4: Illustration of the fused Lasso solution for a given t. In this case, $\hat{\beta}_1 = 0$
.

2.4 Lasso in the Logistic Regression Model

After introducing the Lasso for the linear regression model, significant research was conducted on its introduction for GLMs. Earlier works include Lokhorst (1999), Roth (2004) and Shevade & Keerthi (2003). Since then, faster and more efficient algorithms were developed for calculating the Lasso solution path.

2.4.1 Definitions

The Lasso in the logistic regression model is a penalized ML estimator, such that – continuing with the notation from Section 2.2 and 2.3 – $\mathcal{L}(\beta) = -l(\beta; y, X)$ and

$$\hat{\beta}_{\text{Lasso}}(\lambda) = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ -l(\beta; y, X) + \lambda \|\beta\|_1 \right\}. \tag{2.19}$$

In logistic regression, the intercept is approximately equal to the logit transformed average probability to default, if the matrix X has standardized columns. Since standardization of y is not possible, an intercept is therefore typically included in the regression, but excluded from penalization.

2.4.2 Properties

As opposed to the Lasso in the linear regression model, the solution paths $\hat{\beta}_{\text{Lasso}}(\lambda)$ are not piecewise linear in general. This can be directly deduced from formula (2.14), since the Hessian matrix of the log-likelihood function is not independent of the parameter β.

$$\frac{\partial^2 \mathcal{L}}{\partial \beta \partial \beta'} = \sum_{i=1}^{n} \Lambda(\mathbf{x}_i \beta)(1 - \Lambda(\mathbf{x}_i \beta))\mathbf{x}_i \mathbf{x}_i'.$$

In Section 2.2.2, we could see that the ML estimator has the property that the average probabilities equal the average observations. This property remains intact if the intercept is unpenalized. This can be seen by a combination of equations (2.3) and (2.10), using that the first score is unpenalized and is therefore equal to zero.

Furthermore, existence of the Lasso estimator is provided easier than in the case of standard ML estimation. This is because the objective function is convex[10] and the set of admissible values is a convex and compact subset of $\mathbb{R}^p$. However, this justification does not hold true for the intercept, since it is unpenalized and therefore has any admissible value in $\mathbb{R}$. It has been shown in Meier et al. (2008, page 55), however, that the Lasso exists and is unique if there are i, j with $y_i = 0$ and $y_j = 1$ and $\lambda > 0$ and if X has full column rank.

Finally, asymptotic normality and consistency have also been proved for the adaptive Lasso in the logistic regression model in Zou (2006, Theorem 4). They can be achieved with almost the same assumptions as for the linear regression model (see Section 2.3.2). The only additionally needed assumptions are mild regularity conditions, which are merely required for the asymptotic normality of the ML estimator.

[10] The objective function is of course strictly convex if X has full column rank.

Chapter 3

Model Performance Criteria

The measurement of model performance is an integral part in choosing a suitable model for estimating credit scores. Especially in the assessment of the optimal value of the tuning parameter λ and hence the selection of the model, the methodology to measure the performance decides which explanatory variables will be included in the final model.

At this point, it is important to emphasize that there is a huge difference, whether performance is analyzed on data that the model was built upon (development sample) or data that it was not (validation sample). A model might fit the development sample well (i.e. have good in-sample performance) and at the same time estimate strongly biased scores for the validation sample (i.e. poor out-of-sample performance). This happens especially in case when the contribution of a certain explanatory variable is overestimated during model development due to a random variation in the development sample (i.e. the data are overfitted), which is not present in the validation sample.

In Chapter 2 we saw that estimation of the parameter β is based on the maximization of the log-likelihood function restricting the parameter space to a subset of $\mathbb{R}^p$, see Equation (2.7). Considering the log-likelihood function as an in-sample performance criterion is therefore not senseful as such – since its restrictions are stronger, the smaller one of two nested models[1] will always yield worse performance on data that it is estimated upon than the other. In reality, the difference in performance may not be statistically significant; For the likelihood-function, criteria based on information theory (e.g. AIC, BIC, see Section 3.1) attempt to quantify when this difference becomes statistically significant by penalizing model size (i.e. the number of included variables) and thus make it suitable as in-sample performance measure.

For binary target variables, there might be more suitable conditions for measuring performance

[1] Two models are nested, if and only if one model is a special case of the other. This is especially the case, if the set of variables in one model is a subset of the variables included in the larger model.

than the value of the log-likelihood function. Typically, these cannot be adjusted easily in order to provide meaningful in-sample measures as the AIC and the BIC. However, any criterion that indicates good model performance may be applied on the validation sample, i.e. as out-of-sample performance measure. As the most straight-forward criterion, the value of the log-likelihood function already provides an appropriate measure. For binary models, however, more sophisticated methods have been developed using measures of discriminatory power. In particular, the ROC and modifications thereof (see Section 3.2) established themselves in credit scoring literature.

One of the most important decisions in this work is whether to use an in-sample or an out-of-sample performance measure in order to select the optimal value of the tuning parameter. In literature, both approaches are established and neither is preferred to the other in terms of number of appearances: Zhang et al. (2010); Fan & Tang (2013); Choi et al. (2015) use information theory-based criteria, while Genkin et al. (2007); Meier et al. (2008); Wu et al. (2009) use cross-validation (see Section 3.3.1) to create validation samples with different criteria to choose an optimal model.

3.1 In-Sample Performance

As mentioned in the introduction of this chapter, in-sample performance is traditionally measured using information theory. The most widely used criteria are the Akaike information criterion (AIC) proposed by Akaike (1973) and the Bayesian information criterion (BIC) (Schwarz, 1978). They will be further discussed in Section 3.1.1.

In order to test significant differences in model performance of two nested models, the likelihood ratio test will be introduced in Section 3.1.2.

3.1.1 Akaike and Bayesian Criteria

The field of information theory in general considers available data as realizations of a random vector and attempts to extract information on the distribution of the vector by using the data. In order to measure the distance between the true distribution $f(\mathbf{x})$ and the estimated distribution $\hat{f}(\mathbf{x})$, the Kullback-Leibler information criterion (KLIC) has been introduced:

$$\text{KLIC} = \mathbb{E}\left[\log\left(\frac{f(\mathbf{x})}{\hat{f}(\mathbf{x})}\right)\right]. \tag{3.1}$$

If the family of possible distributions $\hat{f}$ is parametrized, i.e., $\hat{f}(x) = g(x|\theta)$ and the expected value in Equation (3.1) is omitted, the minimization of the KLIC is equivalent to the maximization of the log-likelihood function, yielding the ML estimator.

One of the more famous derivations of the AIC is attributed to Amemiya (1980). Attempting to improve the approximation of the KLIC, Amemiya (1980) conducted a Taylor-expansion of the KLIC around the ML estimator $\hat{\theta}_{\text{ML}}$ and showed asymptotically, that under general assumptions, the average improvement attainable in terms of model fit measured by the value of the log-likelihood function by using K_1 additional variables with true parameter 0 (i.e., not belonging to the true model) amounts to exactly K_1. For p covariates, the value of the criterion is therefore defined as

$$\text{AIC} = -2\left(\log L - p\right).$$

Using a Bayesian approach, Schwarz (1978) proposed the BIC under the assumption that the random variable belongs to the exponential family[2]. Under general assumptions of prior distributions, Schwarz (1978) arrives at the result that, as the number of observations $n \to \infty$, selecting the model that is a posteriori most probable is equivalent to maximizing the following criterion:

$$\text{BIC} = -2\left(\log L - p \cdot \frac{\log n}{2}\right).$$

The difference between AIC and BIC is solely the size of the penalties: The BIC also depends on the number of observations n used for estimating the models, imposing a less conservative penalty on the number of parameters included in the model.

3.1.2 Likelihood Ratio (LR) Test

In the case of testing significant fit differences in two nested models, ML theory provides a useful test. As the name suggests, the likelihood ratio (LR) test is based on the ratio of the values of the two models' respective likelihood functions and uses this ratio to evaluate statistically significant differences.

Formalizing this concept, we have $\hat{f}(x) = g(x|\theta)$ with $\theta \in \Theta$, where Θ is a convex subset of $\mathbb{R}^p$ and write $\theta = (\theta_1', \theta_2')'$ with $\theta_1' \in \mathbb{R}^{p_1}$. In the context of Lasso estimation, the LR test is an asymptotic test for the following hypothesis:

$$\begin{matrix} \mathcal{H}_0 : \theta_2 = 0 \\ \mathcal{H}_1 : \theta_2 \neq 0 \end{matrix}, \qquad \hat{T} = -2\log\left(\frac{\sup_{\theta\in\Theta}\{L(\theta; y, X);\ \theta_2 = 0\}}{L(\hat{\theta}_{\text{ML}}; y, X)}\right).$$

It can be shown that, under $\mathcal{H}_0$ and certain assumptions, the likelihood ratio converges to a $\chi^2_{p_2}$ distribution, where $p_2 = p - p_1$.

Since the LR test provides a useful method to evaluate significant differences in model fits, it is possible to use it for model selection. Denoting two models M_1, M_2 with the respective

[2] as e.g. the dependent variable in the logistic regression model

maximum values of the likelihood functions L_1 and L_2 having p_1 and p_2 covariates, with $p_1 > p_2$ without loss of generality, the test statistic $\hat{T}$ suggests that M_1 is preferred over M_2, if

$$\hat{T} = 2 \log L_1 - 2 \log L_2 > \chi^2_{p_1 - p_2}(0.95).$$

The statistic $\hat{T}$ makes use of the log-likelihood functions $l_i := \log L_i, i = 1, 2$, similar to the AIC and the BIC. In fact, $\hat{T}$ can be expressed as a function of both criteria.

$$\hat{T} = 2(l_1 - p_1) - 2(l_2 - p_2) + 2(p_1 - p_2) = \mathrm{AIC}(M_1) - \mathrm{AIC}(M_2) + 2(p_1 - p_2).$$

For the BIC, similar deductions can be made:

$$\hat{T} = \mathrm{BIC}(M_1) - \mathrm{BIC}(M_2) + \log(n)(p_1 - p_2).$$

We can see that all three criteria of choosing M_1 over M_2 can be reduced to the value of T as well as differences in model size. In fact, preferring M_1 over M_2 using any of these criteria is equivalent to $\hat{T}$ exceeding the following thresholds:

$$\begin{aligned} \mathrm{LR}: \quad & \hat{T} > \chi^2_{p_1 - p_2}(1 - \alpha) \\ \mathrm{AIC}: \quad & \hat{T} > 2(p_1 - p_2) \\ \mathrm{BIC}: \quad & \hat{T} > \log(n)(p_1 - p_2). \end{aligned}$$

3.2 Measures of Discriminatory Power

In particular in the context of the development and validation of rating models, discriminatory measures are used in order to examine the ability of the model to distinguish between clients who are able to meet their payment obligations and those who are not[3].

3.2.1 Definitions

All of the discriminatory measures discussed in this work are based on the so-called classification table, which describes the relationship between predicted and observed defaults[4].

In the following, using the terminology from Table 3.1, some important terms will be defined.

The true positive rate (TPR), also referred to as “sensitivity”, is defined as the share of predicted good customers in all non-defaulting customers, i.e., the correctly positively classified customers.

[3] For a detailed discussion of rating model validation see e.g. Chapter 6 of Oesterreichische Nationalbank (2004)

[4] In Table 3.1 the defaulted customers are classified as “bad”, non-defaults are depicted as “good”

	observed good	**observed bad**
predicted good	*true positive (TP)*	*false positive (FP)*
predicted bad	*false negative (FN)*	*true negative (TN)*
	Sensitivity TP / (TP+FN)	**Specificity** TN / (TN+FP)

Table 3.1: Confusion matrix, terminology from Siddiqi (2006, Exhibit 6.19)

The false negative rate (FNR), or "false alarm ratio", which is equal to 1-TPR, is defined as the predicted bad customers as a share of all good customers.

The true negative rate (TNR) or "specificity", is defined as the share of predicted bad customers in all defaulting customers, i.e., the ratio of correctly specified bad customers.

The false positive rate (FPR) (1-TNR) is defined as the share of predicted good, but defaulting customers, i.e., the wrongly classified bad customers.

Finally, the accuracy ratio (AR) is defined as the share of all correctly classified customers, and the error rate (ER) equals 1-AR.

Logistic regression does not classify the customers as "good" or "bad" as in Table 3.1, but assigns probabilities to each customer. In order to achieve a classification into these classes, a cutoff point is chosen – customers having default probabilities above the cutoff point are regarded as "predicted bad" and, conversely, customers with default probabilities below the cutoff point are regarded as "predicted good". Typically, the selected cutoff point is 0.5. More sophisticated methods will be outlined in Section 3.2.2.

3.2.2 Receiver Operating Characteristics Curve

The market-standard method (Oesterreichische Nationalbank, 2004) for analyzing the discriminatory power of a rating system is the receiver operating characteristics (ROC) curve[5]. The idea behind the ROC curve is that choosing a single probability cutoff point for achieving a confusion matrix as in Table 3.1 provides a very limited view on the true discriminatory power of the rating system. Therefore, the ROC curve depicts the relationship between Sensitivity and Specificity as a function of the cutoff point c.

In order to formally define the ROC curve, it is necessary to formalize some concepts. From

[5] The name of this curve can be traced back to World War II, where it was used to quantify the ability of a radar receiver operator to correctly detect Japanese aircraft from their radar signal (Green & Swets, 1966).

Section 2.2, we know that $Y_i|X_i = \mathbf{x} \sim B(p(\mathbf{x}))$, where $p(\mathbf{x}) = F(\mathbf{x}\beta)$ and $S := \mathbf{x}\beta$ is called the credit score of customer i. This implies that customers with high credit score have a high PD. We define F_N as the distribution function of the scores of all “observed good” customers, and F_D as the distribution function of the scores of all “observed bad” customers.

For some theoretical concepts in this section, we have to assume that the theoretical distributions of the scores F_N and F_D are continuous. In order to use the ROC curve on real data, we introduce the empirical counterparts of the distribution functions as well:

$$\begin{aligned} \hat{F}_N : \mathbb{R} &\to [0,1] \\ s &\mapsto \frac{\sum_{i=1}^n \mathbb{1}_{\{Y_i=0, S_i \le s\}}}{\sum_{i=1}^n \mathbb{1}_{\{Y_i=0\}}}, \\ \hat{F}_D : \mathbb{R} &\to [0,1] \\ s &\mapsto \frac{\sum_{i=1}^n \mathbb{1}_{\{Y_i=1, S_i \le s\}}}{\sum_{i=1}^n \mathbb{1}_{\{Y_i=1\}}}. \end{aligned}$$

Denoting the quantile functions of the theoretical distribution functions with F_N^{-1} and F_D^{-1},

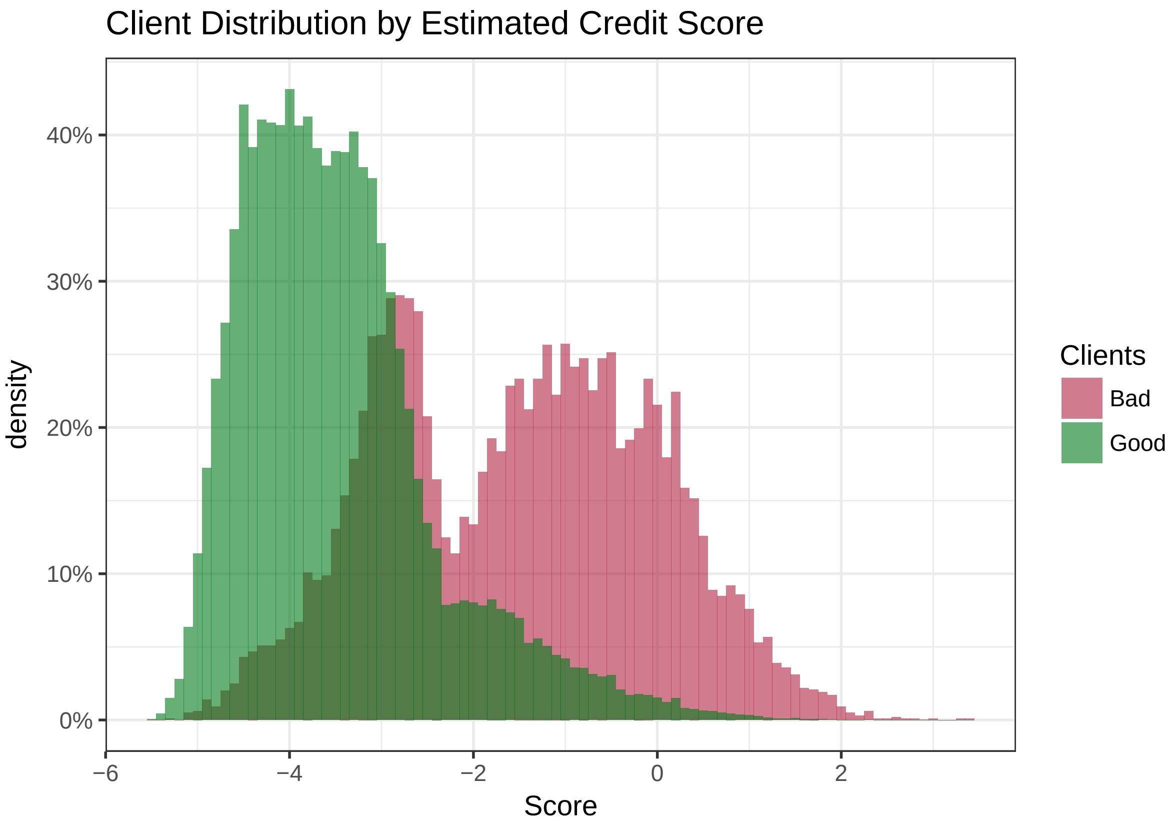

Figure 3.1: Client distributions by estimated credit score

respectively, the ROC curve is defined as follows:

$$\begin{aligned} \mathrm{ROC} : [0,1] &\rightarrow [0,1] \\ p &\mapsto F_N(F_D^{-1}(p)). \end{aligned}$$

The value $s := F_D^{-1}(p)$ can be interpreted as one of the cutoff-points[6] used to create a confusion matrix as in Table 3.1. The value $p = F_D(F_D^{-1}(p)) = F_D(s)$ is then the probability that a defaulting customer's credit score is smaller than s, which is the theoretical counterpart of the false positive rate. Finally, $F_N(s)$ is the probability of non-defaulting customers to have credit scores smaller than s, i.e., the counterpart of the true positive rate (sensitivity).

The ROC curve is typically plotted as in Figure 3.2, where the cumulative good cases (sensitivity) are plotted on the y axis and the cumulative bad cases[7] (1 − specificity) are plotted on the x axis. One point (x, y) on the ROC curve can be interpreted as follows: if the cutoff-value corresponding to this point is chosen, a share x of the bad cases will not be identified as bad and the share $1-y$ of good cases will not be identified as good. Therefore, the ideal point would be $(0, 1)$, which of course cannot be reached in realistic cases.

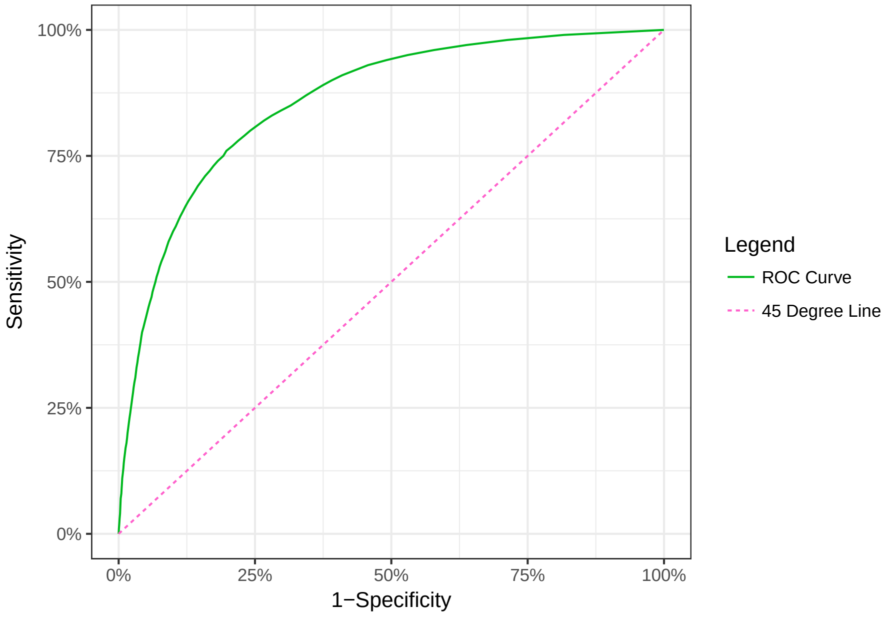

Figure 3.2: Example of an ROC curve

The ROC curve also provides a suitable method to analyze the relationship between the credit

[6] Earlier, we defined cutoff-points as probabilities p. In case the population distribution is continuous, this is equivalent to the concept of using scores s as cutoff-values, because $p = F_D(s)$ is fulfilled in this case.

[7] The cumulative good and bad cases are to be understood as the share of the respective observed cases positively predicted for the same cutoff point.

score and the PD. In the logistic regression model, this relationship is implicitly assumed to be strictly monotonically increasing, since $p = \Lambda(s)$. However, this is not necessarily true for any resulting model and can be tested in the following way for continuous populations: If the relationship is strictly monotonous, the function $\mathrm{PD}(s) := P(Y = 1|S = s)$ is strictly monotonous in s. For this analysis, we need to assume that S is a continuous random variable. Using Bayes' theorem for densities and defining $\mathrm{PD} := P(Y = 1)$, we can see that

$$P(Y = 1|S = s) = \frac{f_D(s)P(Y = 1)}{f_{N+D}(s)} = \frac{f_D\,\mathrm{PD}}{f_D\,\mathrm{PD} + (1-\mathrm{PD})f_N} = \frac{1}{1 + \frac{1-\mathrm{PD}}{\mathrm{PD}}\frac{f_N}{f_D}}.$$

Since $x \mapsto \left(1 + x\frac{1-\mathrm{PD}}{\mathrm{PD}}\right)^{-1}$ is a strictly monotonically decreasing function in x, $\mathrm{PD}(s)$ is strictly monotonically increasing if and only if $s \mapsto f_D(s)/f_N(s)$ is a strictly monotonically increasing function.

On the other hand, the first derivative of the ROC curve is as follows:

$$\mathrm{ROC}'(u) = \frac{d}{du}F_N(F_D^{-1}(u)) = f_N(F_D^{-1}(u))\frac{d}{du}F_D^{-1}(u) = \frac{f_N(F_D^{-1}(u))}{f_D(F_D^{-1}(u))}.$$

Since $x \mapsto F_D^{-1}(x)$ is a strictly monotonically increasing function, we can see that the first derivative of the ROC curve is therefore strictly monotonically decreasing, if and only if $s \mapsto f_D(s)/f_N(s)$ is a strictly monotonically increasing function in s. This is the case, if and only if the ROC curve is strictly concave.

Summarizing the explanation above, we can conclude that the relationship between the credit scores and the PD is strictly monotonically increasing, if and only if the ROC curve is strictly concave. This is of course a result only applicable for infinite observations, since it was conducted for densities[8].

A further interpretation of the ROC curve can be provided in terms of depicting the relationship of α- and β-errors. For this, let S_i be the credit score of an individual i, for whom it is unknown whether he belongs to the population of the defaulted or the non-defaulted. Given the test

$$\begin{array}{l}\mathcal{H}_0 : S_i \in F_N \\ \mathcal{H}_1 : S_i \in F_D\end{array}, \qquad \hat{T}(s) = \begin{cases} 0 & S_i < s \\ 1 & \text{else} \end{cases} \tag{3.2}$$

then the α-error is defined as $P(\hat{T}(s) = 1|\mathcal{H}_0)$ and the β-error equals, conversely, $P(\hat{T}(s) =$

[8] Remark: Strict concavity of the ROC curve as a definition only makes sense for densities, since otherwise the ROC curve is an empirical distribution function, which is piecewise constant. In real applications, a grid of ordered quantiles $q_1, \ldots, q_m$ is defined to check the assumption of concavity.

$0|\mathcal{H}_1)$. Transforming these expressions leads to

$$\alpha(s) = P(\hat{T}(s) = 1|\mathcal{H}_0) = P(S \geq s|\mathcal{H}_0) = 1 - F_N(s)$$
$$\beta(s) = P(\hat{T}(s) = 0|\mathcal{H}_1) = P(S < s|\mathcal{H}_1) = F_D(s).$$

We can therefore conclude that the ROC curve depicts the relationship between β-error and $(1-\alpha)$-error.

In case the credit score of an individual is used in order to decide whether a loan will be granted to him based on a decision rule as in (3.2), the α-error can be interpreted as the proportion of lost business to the bank, while the β-error is the proportion of defaulting cases. In such sense, the β-error is much more costly for the bank.

3.2.3 The Area Under the Curve (AUC)

As the most wide-spread measure for discriminatory power, the area under the ROC curve – the AUC – is used. It measures the probability that a randomly chosen good customer has a better score than a randomly chosen bad customer. For the proof, let $N \sim F_N$ and $D \sim F_D$ be two independent random variables on a probability space $(\Omega, \mathfrak{A}, P)$[9], then

$$\text{AUC} = \int_0^1 \text{ROC}(u)du = \int_0^1 F_N(F_D^{-1}(u))du. \tag{3.3}$$

We can see that AUC can be interpreted as an expected value of a uniformly distributed variable. Let $U \sim U_{[0,1]}$, then $F_D^{-1}(U) \sim F_D$, which leads to

$$\text{AUC} = \mathbb{E}\left[F_N(F_D^{-1}(U))\right] = \mathbb{E}\left[F_N(D)\right] = \int_\Omega F_N(D)dP = \int_{D(\Omega)} F_N(s_d)dP^D(s_d). \tag{3.4}$$

Now, we have to expand $F_N(s)$ in order to evaluate the value of the integral above.

$$F_N(s_d) = P(N \leq s_d) = \int_\Omega \mathbb{1}_{\{N \leq s_d\}}dP = \int_{N(\Omega)} \mathbb{1}_{\{s_n \leq s_d\}}dP^N(s_n). \tag{3.5}$$

Combining equations (3.4) and (3.5) as well as taking into account that X and Y are independent leads to the final result

$$\text{AUC} = \int\int \mathbb{1}_{\{s_n \leq s_d\}}dP^N dP^D = \int_{D(\Omega)\times N(\Omega)} \mathbb{1}_{\{s_n \leq s_d\}}d(P^D \otimes P^N) = P(D \leq N). \tag{3.6}$$

Using the result from Equation (3.6), we can conclude that the following properties hold true:

- In case the model cannot separate well between defaulted and non-defaulted cases, i.e.,

[9] Remark: For this proof, F_N and F_D do not have to be continuous.

there is a significant overlap between the credit score distributions of D and N, the model will provide an AUC close to 50%.

- In case the model separates well between defaulted and non-defaulted cases, the AUC will be close to 100%.[10]

In this sense, the AUC measures the level of discrimination between the score distributions of D and N.

Another interpretation of the AUC is provided using the term $F_N(s_d) = P(N \leq s_d)$ from Equation (3.4): s_d can be interpreted as a cutoff-value, below which clients are predicted good and above which they are predicted bad. Fixing this point, we observe that the term $F_N(s_d)$ denotes the proportion of good clients classified as good by the model. For the AUC, the cutoff-point is varied and weighted by the probabilities of a bad client to be above the cutoff-point, i.e., classified as bad by the model. In such a sense, the AUC is a probability-weighted average over a goodness-of-fit measure depending on cutoff-points.

While the AUC is probably established as the most important discriminatory measure in rating validation, as its use is e.g. recommended by Oesterreichische Nationalbank (2004), sole reliance on it should be viewed with caution, since it does not account for the shape of the ROC curve. For example, as already discussed, it does not measure, whether the monotony in the relationship between the credit scores and the PD is intact.

3.2.4 Further Discriminatory Measures

All of the discriminatory measures described in this work are based on the ROC curve. Three further measures will be discussed in this section because of their economic relevance and applications in fields other than statistical modeling of credit risk. These measures are all summarized from Oesterreichische Nationalbank (2004, Chapter 6.2).

First, the Gini Coefficient – which is also referred to as the Accuracy Ratio (AR) – is equivalent to the AUC and can be calculated as

$$\mathrm{AR} = 2 \cdot \mathrm{AUC} - 1. \tag{3.7}$$

It is therefore twice the area between the ROC curve and the 45 degree line as shown in Figure 3.2. The Gini Coefficient is also used in economics to analyze differences in the distributions of wealth. For this, the values of every individual's assets $a_1, \dots, a_n$ are ordered such that $a_i \leq a_{i+1}$ for all $i = 1, \dots, n-1$. Furthermore, two distributions on the interval $[0, 1]$ are generated: F_1 describes the empirical distribution function of $\{\frac{1}{n}, \dots, \frac{n}{n}\}$, whereas F_2 describes the

[10] According to Oesterreichische Nationalbank (2004, chart 60), most credit rating systems relying on multivariate binary logit models reach values between 0.8 and 0.9.

empirical distribution function of $\{\frac{a_1}{a_n}, \dots, \frac{a_n}{a_n}\}$. The Gini coefficient in economics is then defined by using Formula (3.7) and the ROC curve generated by these two distribution functions, i.e., $\mathrm{ROC}(u) = F_2(F_1^{-1}(u))$.

Another important measure is the Pietra Index[11], which is based on the ROC curve as well. It measures the maximum distance between the ROC curve and the 45 Degree line, or, equivalently, the minimum sum of α and β-errors generated by the model.

$$\begin{aligned}\text{Pietra} &:= \max_{x\in[0,1]} (\mathrm{ROC}(x) - x) = \max_{x\in[0,1]} (F_N(F_D^{-1}(x)) - x) \\ &= \max_{s\in\mathbb{R}} (F_N(s) - F_D(s)) = \max_{s\in\mathbb{R}} 1 - \alpha(s) - \beta(s). \end{aligned} \tag{3.8}$$

A slightly more sophisticated measure is the Bayesian error rate (BER), which is the minimum of the weighted averages of the α and β errors, respectively.

$$\text{BER} := \min_{s\in\mathbb{R}} \left[(1-p)\cdot\alpha(s) + p\cdot\beta(s)\right]. \tag{3.9}$$

By defining p as the overall default rate within the sample, the BER can be interpreted as follows: For any threshold s, the term $(1-p)\alpha(s)$ denotes the proportion of individuals within the sample who do not default but are classified as bad by the model. In addition, the term $p\beta(s)$ is the share of individuals within the sample that do default but are classified as good by the model. The BER therefore minimizes the share of individuals within the sample that is misclassified by the model by application of a decision rule as in (3.2).

3.3 Out-of-sample Performance

As mentioned in the introduction of this chapter, it is not advisable to use information theory for measuring out-of-sample performance. However, any other performance measure may be used. For binary logit models, the most wide spread measure for out-of-sample performance is the AUC (see Section 3.2.3). In literature, those models are regarded as useful, which maximize the out-of-sample AUC. However, in order to measure out-of-sample performance, a data set with observations not contained in the original data set would be needed. One solution could be to use only a subset of the original data for estimating the model and validate the model based on the remaining data points. However, in this case, the final model does not incorporate the whole information attainable from the data set, which is a drawback for the validity of the model.

Therefore, more ways of generating out-of-sample data sets were developed. The most common method is Cross-validation, which will be discussed in Section 3.3.1.

[11] In economics, the Pietra Index is sometimes also referred to as the Robin Hood Index.

3.3.1 Cross-Validation (CV)

Avoiding the implementation of a model that is suitable for the selected data subset but works poor on test data sets, on the basis of which the model was not developed on, is one of the key challenges in model selection. A well-established statistical technique to avoid such biases is cross-validation (CV). CV is a random sampling technique used in order to assess the out-of-sample predictive power of a model. It is commonly used due to its flexibility, since any types of models can be compared in terms of performance in case a suitable performance measure for the dependent variable's type has been defined.

The most wide-spread type of CV is the so-called k-fold cross-validation. It essentially consists of randomly splitting the data set into k approximately equally sized parts (folds) and using all folds except for one for estimating a model. The data which the model was estimated on is called the training data, the remaining part is called the test data. The model generated by the training data provides predictions for the test data set. The predictions are then compared to the outcome and performance is measured choosing any suitable performance measure. This iteration is repeated until all folds have been declared as the test data set. The cross-validated model performance is then calculated as

$$P_{\text{CV}} = \frac{1}{k} \sum_{i=1}^{k} P(y^{\text{test}(i)}, \hat{y}^{\text{test}(i)}). \tag{3.10}$$

In case of logistic regression, P is any performance measure discussed in Section 3.2. Furthermore, specifically for tuning parameter selection in Lasso problems, performance will be evaluated for a sequence of tuning parameter values such that P_{CV} is a function of λ. As already mentioned, in literature, the AUC is mostly used as the performance measure. However, any type of measurement is applicable and Formula (3.10) will not change.

Chapter 4

Data Description & Software

The analyses were conducted on data available on kaggle's Competition "GiveMeSomeCredit"[1] that was launched on 19 September 2011. The competition consisted of modeling two-year default probabilities of retail customers with different types of loans (mortgage loans, credit cards and other revolving loans). The provided explanatory dataset consists of 150.000 observations containing socio-demographic variables such as age and number of dependents as well as quantitative variables such as debt and income, and finally qualitative variables like the number of times that the customer was past due in the past. The complete variable list can be found in the Appendix in Table A.1.

4.1 Detailed Data Description

In this section, a detailed description of each variable and the manner of its treatment is provided.

4.1.1 The Dependent Variable

We model predictors of default events and their measured influence on the invocation of such an event. However, the definition of a serious delinquency can be formulated in different ways and has therefore not been unique throughout the history of regulatory requirements. This is because the failure of a bank's client to keep up with his redemption or interest payment until the due date does not always imply that he is not able to do so, and, conversely, a bank may not always be able to identify whether a creditor is unable to meet his payment obligations.

In fact, the definition of the default of a counterparty has recently been updated and implemented in EU law within Article 178 of Regulation (EU) No 575/2013. Within this regulation and especially concerning retail lending, the established criterion in declaring a client as defaulted is the number of days he is past due with his interest payments or his redemption plan,

[1] This data set has also been analyzed in Wang et al. (2015).

where the cutoff-date is defined after 90 days[2]. The kaggle data also refers to the same definition of default, i.e., if a client is past due more than 90 days, he is considered as defaulted in the data set.

The data per se is imbalanced concerning the distribution of the clients in the defaulted and non-defaulted classes, since only 6.68% of the clients experienced a serious delinquency event in the past two years – this is typical in the case of modeling credit defaults. In such a case, the discrimination between good and bad cases is conducted in the tails of the distribution function Λ, which however is more linear[3] in that area – this might lead to worse discrimination between good and bad cases. Therefore, defaulted observations will be weighted $(1-6.68\%)/6.68\% \approx 14$ times higher than non-defaulted observations. This leads to a miscalibration of probabilities[4], but makes sure that those properties that are leading to default events are considered equally strongly in the course of the selection of variables.

4.1.2 The Predictors

For easier readability, the variables present in the data are categorized into three different buckets in the following sub-section: sociodemographic variables, characteristics of the loan or loans and behavioral variables. These buckets are designed and ordered according to the likeliness for these properties to change throughout the observation period, i.e., sociodemographic variables such as age and income are less likely to change than properties of the client reflecting his behavior in the recent months. Choosing explanatory variables that are stable throughout time is desirable for a credit analyst, since it makes the credit score for a selected client and thus his rating more stable by design. In this work, however, this differentiation is not made within the estimation.

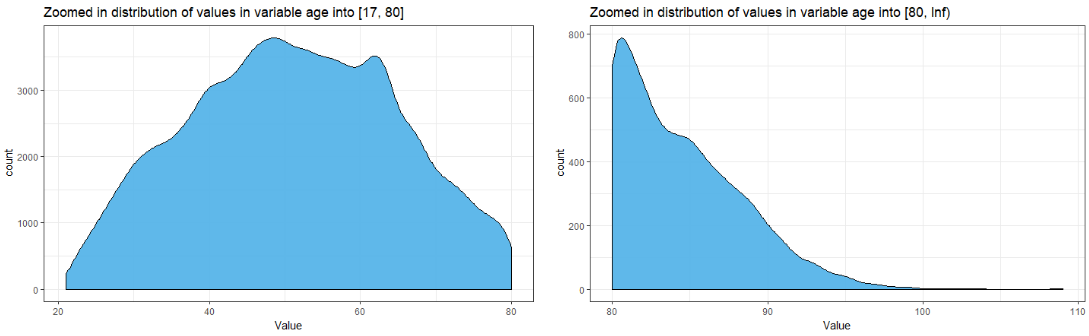

Figure 4.1: Age distribution in the dataset split at the age of 80.

[2] Here, materiality thresholds are defined, meaning that if the client missed a payment that is immaterial, he is not considered as defaulted.

[3] Meaning that the second derivative $\Lambda'' = \Lambda(1-\Lambda)(1-2\Lambda)$ is closer to zero.

[4] On average, the model will assign PDs of 50%, while the average estimated PD should be equal to the average default rate in the data set, which is 6.68%.

Sociodemographic Variables

The kaggle data contains information on retail clients such as the age (variable name: `age`) as well as the gross income (`MonthlyIncome`) of the client and the number of their dependents (`NumberOfDependents`), which refers to all members of the family excluding themselves, who the borrower has to provide for.

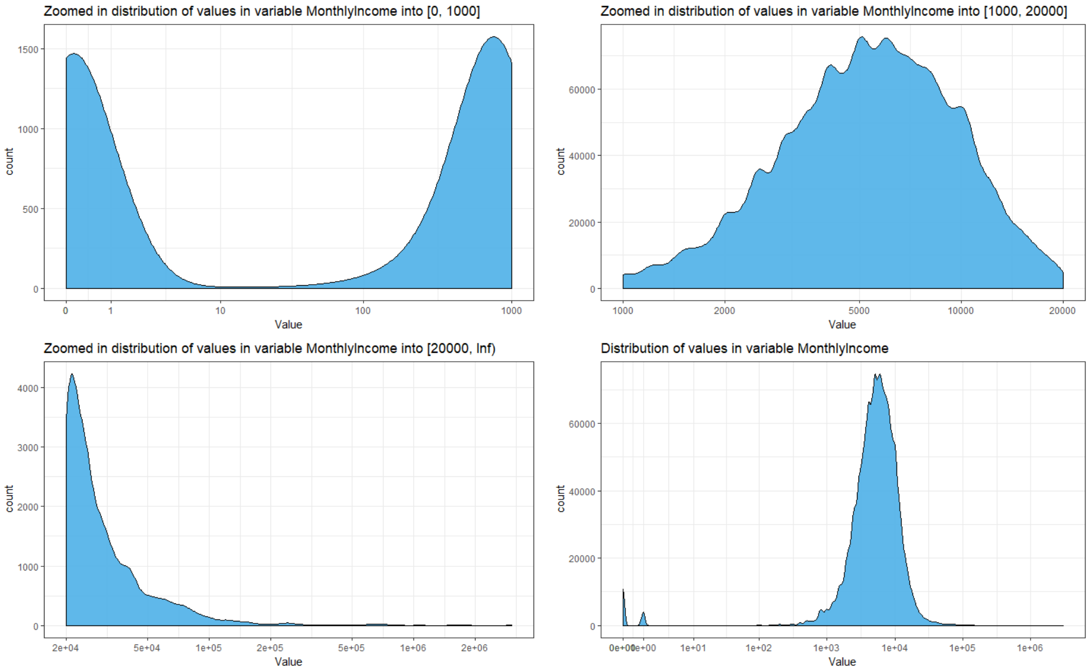

Figure 4.2: Income distribution in the dataset split into three income regions (low, middle, high, total).

Concerning `age`, the majority of the observations are located between ages 21 and 80, even though still a significant number of observations can be found in the upper tail from 80 to 110 (see Figure 4.1).

`MonthlyIncome` is present mostly in the interval [1.000, 20.000] (see Figure 4.2), while extreme values can be found with a low of zero and highs up to income of 2 Million USD. Unknown values will be replaced with the value of zero and a separate indicator variable will be introduced for those (see Section 4.2.1).

The majority of clients in the data set has no dependents, and if so, only a negligible number of observations have more than five. There is, however, a considerable number of clients, where this figure is unknown – they will again be replaced with the standard value of zero. The values of 13 and 20, while certainly almost unrealistically high, will not be interpreted as missing, since there is no evidence to use them as such.

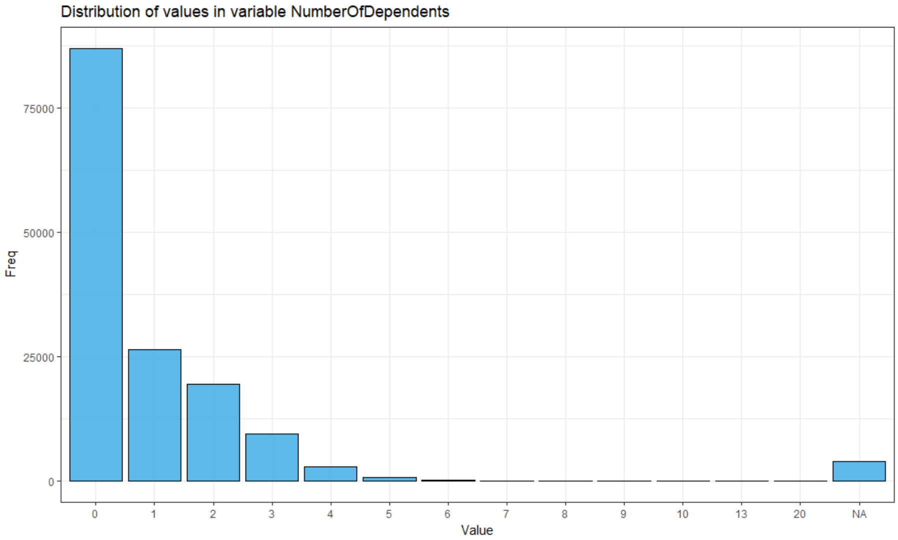

Figure 4.3: Distribution of variable `NumberOfDependents`.

Loan Characteristics

The data contains variables on the debt to income ratio of the client (`DebtRatio`, see Figure 4.4), which in fact does not solely refer to the debt to the bank, but adds up monthly debt payments, alimony and living costs as a share of the monthly gross income of the borrower. Additionally, the data contains information on the total number of credit loans or lines (`NumberOfOpenCreditLinesAndLoans`, Figure 4.5) as well as the number of real estate loans or lines (`NumberRealEstateLoansOrLines`, Figure 4.6), which is a subset thereof.

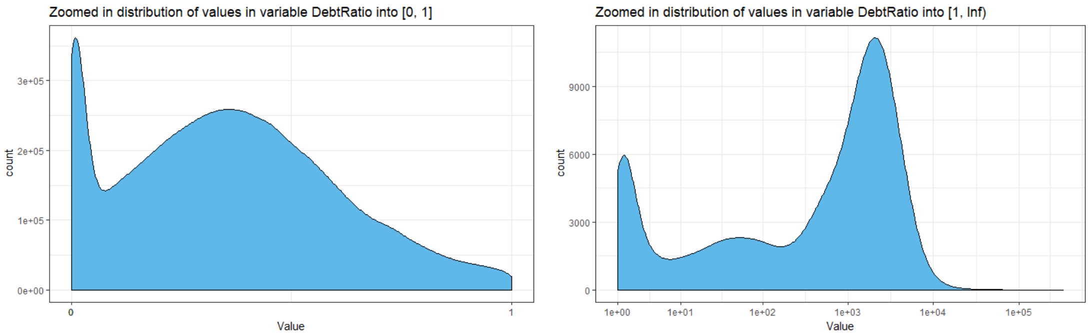

Figure 4.4: Distribution of variable `DebtRatio` (low, high).

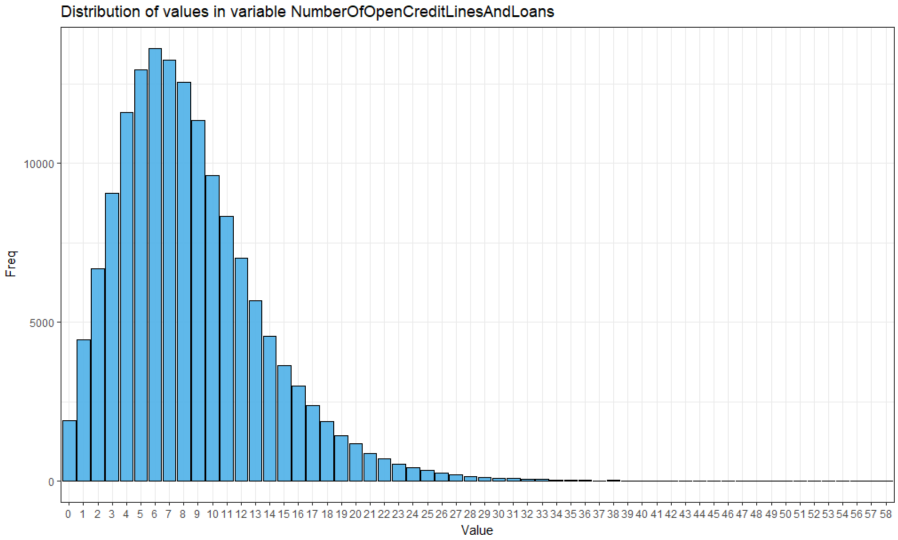

Figure 4.5: Distribution of variable `NumberOfOpenCreditLinesAndLoans`.

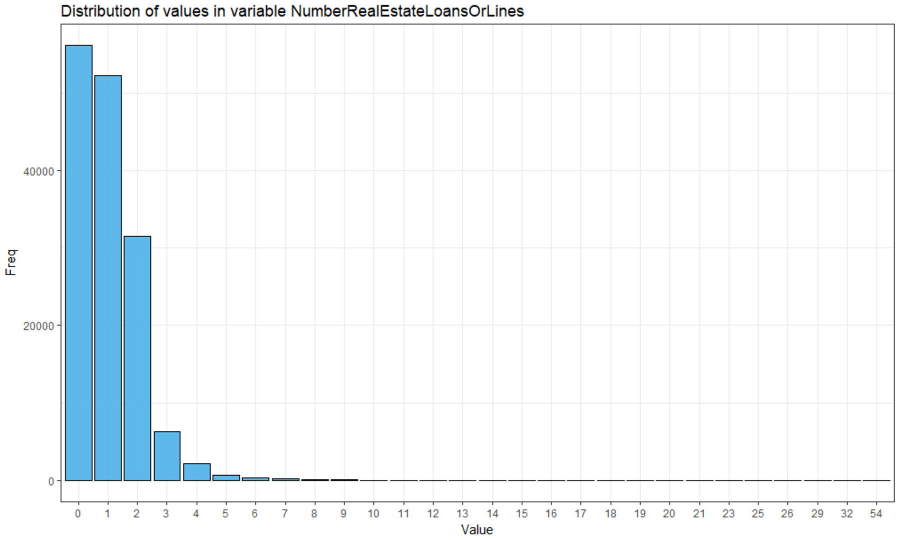

Figure 4.6: Distribution of variable `NumberRealEstateLoansOrLines`.

As expected, the majority of observations in the variable `DebtRatio` are located in the interval [0,1], however, quite a significant number of observations are located above 1. The missing values are replaced with zero.

As can be deduced from Figure 4.5 and Figure 4.6, the high number of loans and lines are mainly driven by non-real estate covered loans. These variables also do not contain missing values.

Behavioral Variables

Outstanding predictors of default are variables which depict the client's behavior in the past. In this context, this especially means payment behavior, i.e., the past due status of these clients in the past, but also variables which depict usage of unsecured credit lines[5].

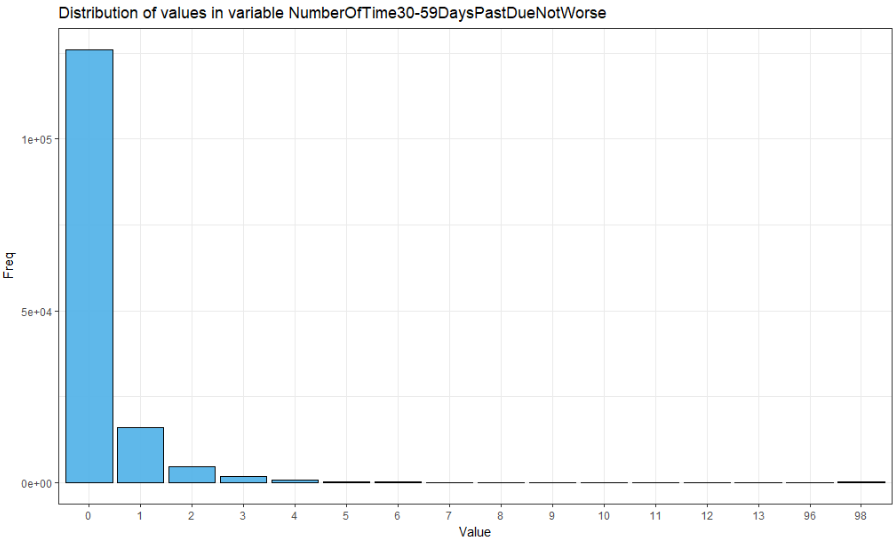

Figure 4.7: Distribution of variable `NumberOfTime30-59DaysPastDueNotWorse`.

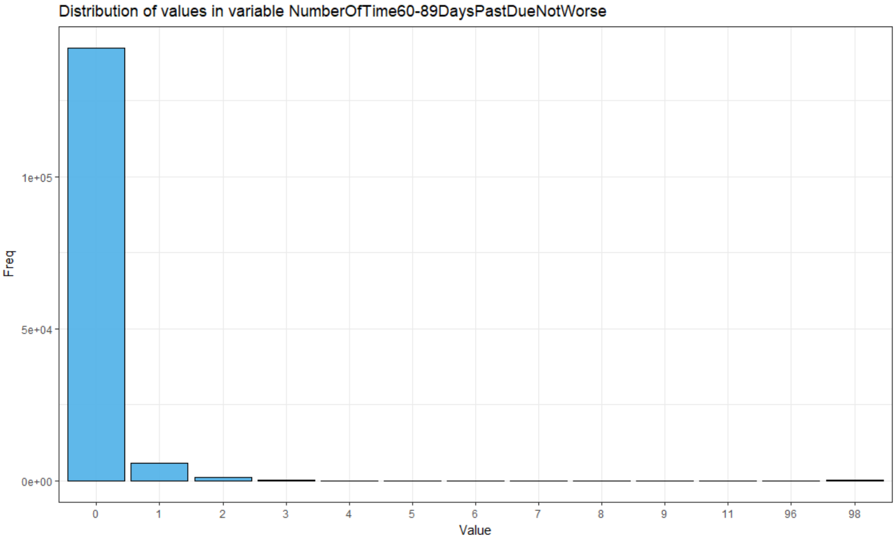

Figure 4.8: Distribution of variable `NumberOfTime60-89DaysPastDueNotWorse`.

In the kaggle data, especially payment behavioral variables are present with the number of times

[5] Credit lines are products, which allow the borrower to withdraw money up to a certain threshold – comparable to loans, they can also be secured by collateral. The most wide-spread form of credit lines for retail customers are credit cards.

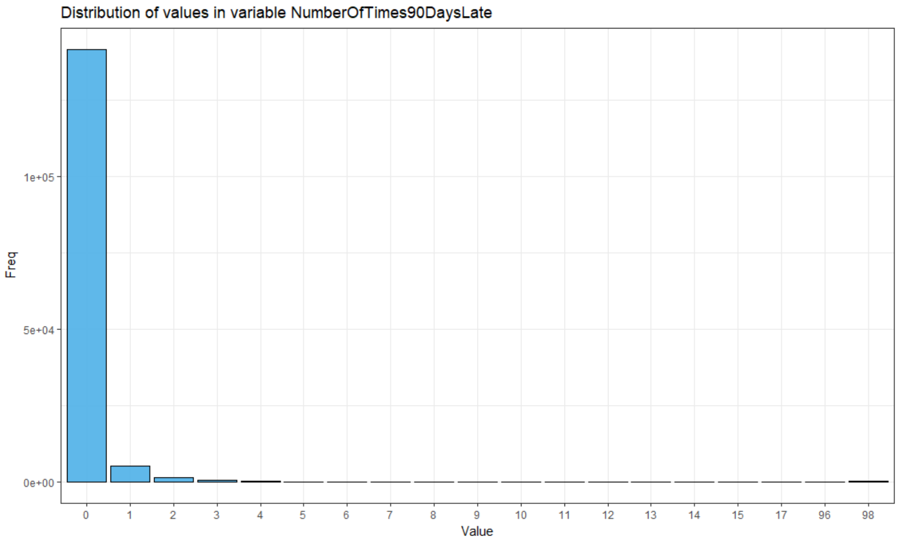

Figure 4.9: Distribution of variable `NumberOfTimes90DaysLate`.

past due for a certain number of days but not worse. More precisely, the number of times the client has been past due 30-59 days (`NumberOfTime30-59DaysPastDueNotWorse`), 60-89 days (`NumberOfTime60-89DaysPastDueNotWorse`) as well as more than 90 days (`NumberOfTimes90DaysLate`) are present in the data set.

We can see in Figures 4.7, 4.8 and 4.9 that while no missing observations are present, all of these variables contain values of 96 and 98, which appear to be a code for missing values. Therefore, they are replaced by the standard value of zero and treated as missing once again.

Additionally, line usage is also present (`RevolvingUtilizationOfUnsecuredLines`), which uses the total balance on all credit cards and other personal lines of credit excluding real estate loans and installment debt as a share of the sum of total credit limits.

4.2 Data Cleansing

4.2.1 Missing Value Treatment

Although the dependent variable `SeriousDlqin2yrs` is complete, some of the explanatory variables used in the data set contain missing values. In practical applications, if the model is already developed, some of the information will again not be available, but a credit score will have to be assigned to the client, nevertheless. Therefore, it is necessary to find a possibility to also include observations into the estimation process, even if there are missing values in some explanatory variables.

In the literature, different ways are proposed with different drawbacks: One possibility is to

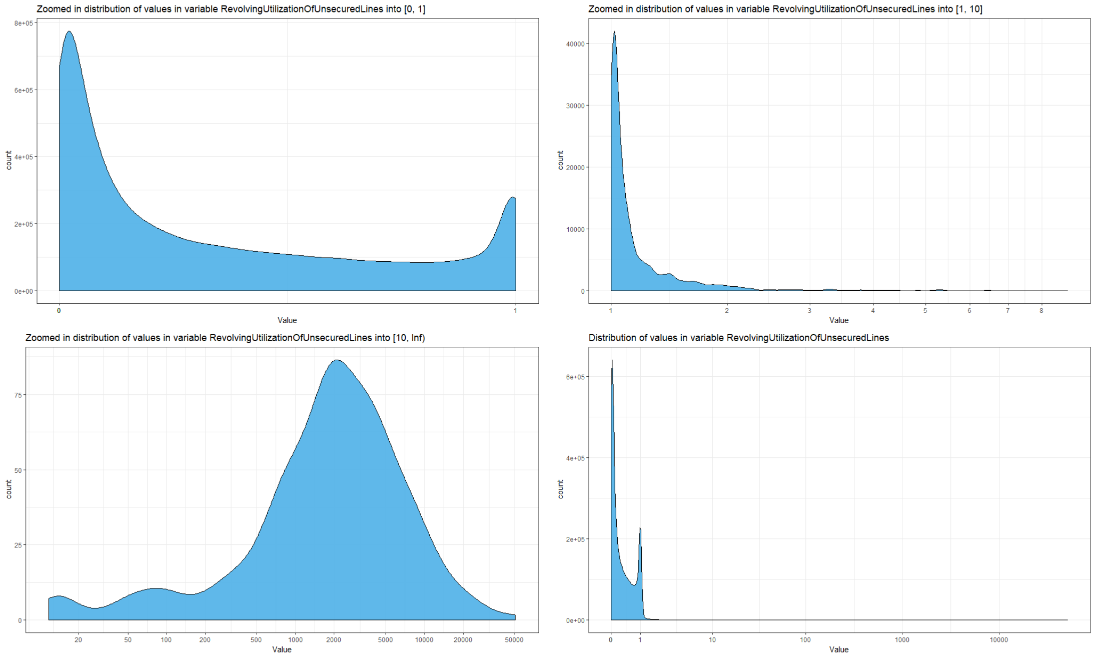

Figure 4.10: Distribution of variable `RevolvingUtilizationOfUnsecuredLines` (low, middle, high) – total in the bottom right corner.

replace the missing values in the explanatory variables with some average calculated by the remaining (non-missing) values. However, additional information will thus be ignored, which may result in a worse model fit and therefore validity.

As an alternative, standard values can be defined variable by variable and missing values can be replaced by these. Additionally, new indicator variables are generated for missing values in the respective variable. This procedure may, however, lead to an overfit and in some cases it may be the reason that the ML estimator does not exist (see Section 2.2.2).

Since this work uses the Lasso, however, overfitting and existence of the estimator is not an issue in the logistic regression model (see Section 2.4.2). Therefore, the second possibility is adopted. This is also the more convenient procedure for a regular application of the model, since the standard values are not data dependent in this case as opposed to the first possibility.

4.2.2 Coded Values

The behavioral variables denoting the number of times that a client was past due in the past two years contain coded values $\{96, 98\}$, which cannot actually denote the number of times past due, because there simply is not enough time to be 30 to 90 days past due that often in two years. Therefore, it is assumed to be an indicator of missing values and is treated as such.

Additionally, the variable `RevolvingUtilizationOfUnsecuredLines` contains a high number of observations with a value of "0.999999". Since it is unrealistic that such a high number of clients obtain this particular value it is again treated as missing and, as a consequence, an indicator variable is generated for its treatment.

4.3 Variable Transformation & Generation

The reasons for transforming existing variables or generating additional ones and including them in the regression are wide-ranged.

The assumption of a linear relationship between the explanatory variables and the log-odds of default probabilities is quite restrictive and hence one of the greatest criticisms of logistic regression. Therefore, continuous variables were transformed into ordinal ones and the result can be viewed as a piecewise constant function f of the original variable. This will be outlined in Section 4.3.1.

In Section 4.3.2, we provide an overview of further possible reasons to include additional variables. For example, simple economic reasoning leads to the replacement of the monthly income by the monthly disposable income, which might explain default risk more accurately.

Finally, as already explained in Section 4.2, the data consists of variables containing missing and coded values that cannot be included without treatment into the regression.

4.3.1 Variable Discretization

The treatment of explanatory variable types only distinguishes between ordinal and continuous variables.

Ordinal variables

Ordinal variables v are characterized by the property that they can be ordered in a certain sense, while the difference between the values does not provide information on the true difference between the values[6]. They are treated using the fused Lasso estimator (see Section 2.3.3) in the following way: The variables are coded according to their sizes with levels $\{1, \ldots, m\}$ and additional (cumulative) variables $\tilde{v}_2, \ldots, \tilde{v}_m$ are created, which are then included into the regression instead of the initial variable v. The variables are generated according to the following

[6] A typical example for an ordinal variable is the rating, which can be ordered in a certain sense. For example, the Standard & Poor's rating AAA is better than AA as well as AA is better than A, but the difference between AAA and AA in terms of PD may be much smaller than the difference between AA and A.

formula:

$$\tilde{v}_j = \begin{cases} 1 & v \geq j \\ 0 & \text{else} \end{cases}, \qquad j = 2, \dots, m. \tag{4.1}$$

Note that only $m-1$ variables are generated, because, if $\tilde{v}_1$ was generated according to formula (4.1), it would be equal to the intercept, the columns of the design matrix X would be linearly dependent and the Lasso therefore may not be unique (see e.g. Section 2.4.2).

This approach ensures that the influence of the original variable v on the dependent variable is piecewise constant and therefore allows for a full flexibility of modeling possible nonlinearities. The relationship can then be analyzed using an illustration as in Figure 4.11. This Figure depicts the coefficients $\beta_2, \dots, \beta_m$ resulting from the fused Lasso estimates $\tilde{\beta}_2, \dots, \tilde{\beta}_m$ discussed in Section 2.3.3 according to the formula $\beta_j = \sum_{i=2}^{j} \tilde{\beta}_j$ and $j = 2, \dots, m$.

The value of the coefficient describes the increase in log-odds compared to the base value of 1. For example, in Figure 4.11 we could derive that the PD of clients with 13 or more open credit lines and loans is 0.28 higher in log-odds than the PD of those with only one credit line. As already discussed in Section 2.2.1, this can be approximately viewed as a 28%-increase in PD, which is a good approximation for PD ≈ 0.

If regarded useful and economically reasonable, the resulting piecewise linear function can still be approximated and replaced by a nonlinear function of the levels of the original variable, excluding the newly generated variables $\tilde{v}_2, \dots, \tilde{v}_m$ and once again reducing the number of variables.

Continuous variables

The differences in the values of continuous variables v have a significance and therefore it is appropriate to include variables of this type in the regression as is. However, if one does not believe in a linear influence of the variable on the log-odds ratio, the variable may be discretized by defining m ordered intervals $(x_i, x_{i+1}]$ for $i = 0, \dots, m$ with $x_0 := -\infty$ and $x_{m+1} := \infty$ and defining an ordinal variable v^o using the formula

$$v^o := \sum_{i=1}^{m} i \mathbb{1}_{(x_i, x_{i+1}]}(v).$$

The variable v^o is then treated as an ordinal variable and is analyzed with the methods according to this variable type.

The only question remaining in this Section is how the intervals $(x_i, x_{i+1}]$ can be defined. In

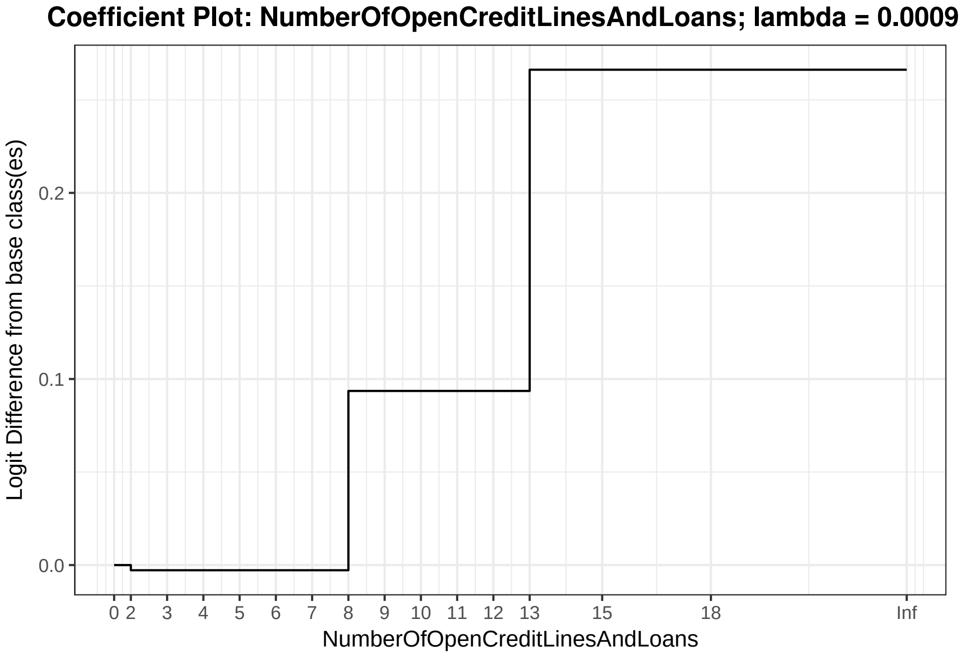

Figure 4.11: Example coefficient plot of an ordinal variable

this work, two possibilities have been tested.

First, variables may be split by empirical quantiles of the distribution of the continuous explanatory variable. This is a very stable method, since even for high m, it is guaranteed with high probability that in all of the buckets, there exist observations i, j with $v_i, v_j \in (x_i, x_{i+1}]$ and $y_i = 1$ as well as $y_j = 0$. This is important for the existence of the ML estimator, as shown in Section 2.2.2[7]. However, this approach of course highly depends on the empirical distribution of the variable and therefore is highly data-driven.

The second approach would be a split-up defined by pre-defined intervals. The intervals are defined in an equidistant manner, i.e., interval borders of the form $x_{i+1} = x_i + c$ are generated. If appropriate, a logarithmic transformation is imposed on the variable, such that the interval borders follow the recursive rule $x_{i+1} = ax_i$ for a constant $a > 1$.

Whichever approach may be chosen, sparsely populated intervals may always be a risk for the ML estimator not to exist. As a solution, interval borders x_{i+1} are only used for the generation of the ordinal variable v^o if two conditions are met. First, a minimum number of values is defined, i.e., at least a minimum number of observations need to be contained in the interval $(x_i, x_{i+1}]$. Additionally, both default and non-default observations need to be contained in the

[7] The Lasso estimator still exists for these variables, but for $\lambda \to 0$, the coefficient converges to $\pm\infty$, depending on whether all $y_i = 1$ or all $y_i = 0$.

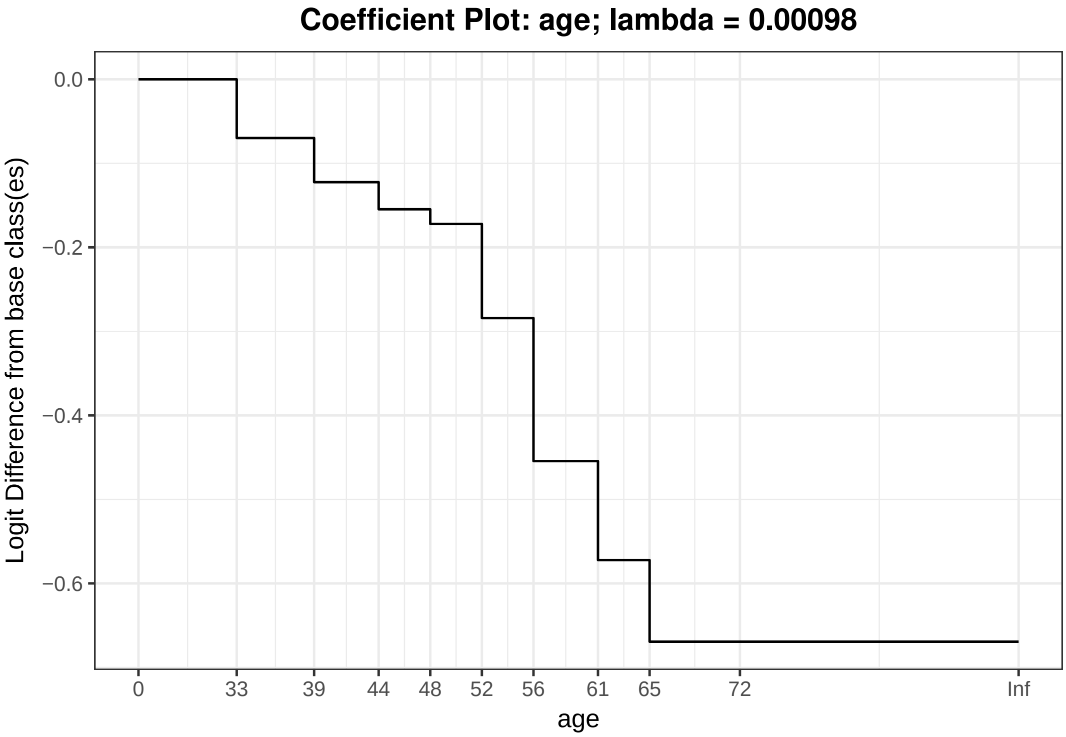

Figure 4.12: Example coefficient plot of a discretized continuous variable

interval in order to guarantee existence of the ML estimator as discussed in Section 2.2.2. If both conditions are met, x_{i+1} will be used as an interval border; otherwise, x_{i+2} will be tested for compliance with the conditions above and if it passes, the interval $(x_i, x_{i+2}]$ is used as the next ordinal level of the variable.

4.3.2 Creation of Further Variables

The methods applied in Section 4.3.1 aim to tackle the issue that due to the resulting model, the marginal effects of the explanatory variables on the PD depend on the shape of the logistic cumulative distribution function (cdf) – the increase in number of regressors is only a side-effect that is automatically resolved by the variable selection property of the Lasso estimator. However, it can still be senseful to create additional variables in order to improve the model by considering the following possible effects:

1. Nonlinear effects of correlations between regressors on the PD
2. Transformations of variables might be more meaningful from economical interpretation perspective
3. The shape of the empirical distribution function of the explanatory variables might affect that an approach as in Section 4.3.1 is not sensible.

In this work, these points are treated exemplarily for some variables – they are described in the subsequent paragraphs.

The first improvement displayed in the list above is referring to the possibility that the simultaneous occurence of two variables increases the PD more than the additive marginal effect implicitly assumed in the logistic regression model (see Section 2.2.1). For a full assessment, every pairwise correlation between all variables would need to be analyzed – in this work, this is treated with the inclusion of additional indicator variables for two specific combinations which are assessed with economic reasoning.

- Well-behaving clients within the last two years, which were never past due in the last two years of the observation period, i.e.,

$$\begin{aligned}\texttt{NeverPastDue} := \ & \mathbb{1}_{\{\texttt{NumberOfTime30-59DaysPastDueNotWorse}=0\}} \\ & \times \mathbb{1}_{\{\texttt{NumberOfTime60-89DaysPastDueNotWorse}=0\}} \\ & \times \mathbb{1}_{\{\texttt{NumberOfTimes90DaysLate}=0\}}.\end{aligned} \tag{4.2}$$

- Clients, which have a combination of high income and debt ratio over one:

$$\texttt{HighDebt\_HighIncome} := \mathbb{1}_{\{\texttt{MonthlyIncome}>2000\}} \times \mathbb{1}_{\{\texttt{DebtRatio}>1\}} \tag{4.3}$$

The second improvement can be made on the definition of economically more meaningful variables. This work includes two examples, where this has been executed:

- The disposable income was included, which – opposed to the gross income – takes the payments that the client has to make per month into account.

$$\texttt{DisposableIncome} := \max(\texttt{Income} - \texttt{Debt},\ 0) = \texttt{Income} \cdot \max(1 - \texttt{DebtRatio},\ 0) \tag{4.4}$$

- Another adjustment to the problem includes separating the number of loans by loan type. In the data set, the number of real estate loans as well as total loans are available – we extract the non-real estate loans from the data as follows:

$$\begin{aligned}\texttt{NumberNonRELoansLines} := \ & \texttt{NumberOfOpenCreditLinesAndLoans} \\ & - \texttt{NumberRealEstateLoansOrLines}\end{aligned} \tag{4.5}$$

- Finally, low income clients are also added an indicator variable:

$$\texttt{LowIncome} := \mathbb{1}_{\{\texttt{MonthlyIncome}\leq 1000\}} \tag{4.6}$$

Finally, it might happen that neither of the discretization algorithms are suitable for the full distribution of the explanatory variable.

- For `DebtRatio`, we can see in Figure 4.4 that while the majority of observations are between $[0, 1]$, there is still a significant portion of observations between $(1, 100.000]$. While the discretization could be performed by using the quantile method as described in Section 4.3.1, it is advantageous to employ a separation of the methodologies within the intervals referred to above. For this, the variable is simply split into two separate parts:

$$\texttt{DebtRatio\_Low} := \min(\texttt{DebtRatio}, 1) \tag{4.7}$$

$$\texttt{DebtRatio\_High} := \max(\texttt{DebtRatio}, 1). \tag{4.8}$$

- Similar observations can be made for `RevolvingUtilizationOfUnsecuredLines` (see Figure 4.10). Here, in addition to the example above, there are many clients with no utilization of their credit lines, which results in a heavy weight at zero. Therefore, three additional variables are introduced:

$$\texttt{No\_Line\_Utilization} := \mathbb{1}_{\{\texttt{RevolvingUtilizationOfUnsecuredLines}=0\}} \tag{4.9}$$

$$\texttt{LowUtilization} := \min(\texttt{RevolvingUtilizationOfUnsecuredLines},\ 1) \tag{4.10}$$

$$\texttt{HighUtilization} := \max(\texttt{RevolvingUtilizationOfUnsecuredLines},\ 1). \tag{4.11}$$

4.4 Algorithms & Software

Since the introduction of the Lasso, various algorithms have been developed for calculating the coefficient paths $\hat{\beta}_{\text{Lasso}}(\lambda)$ and implemented in R.

Friedman et al. (2010) implemented the package glmnet, which relies on a coordinate-descent algorithm and uses a pre-defined sequence of λ-values for which it calculates the Lasso solution path. The pre-defined sequence is based on the minimum value $\lambda_{\max}$ of tuning parameters λ for which all coefficients are set to zero.

$$\lambda_{\max} = \min\{\lambda \in \mathbb{R}^+ : \hat{\beta}_{\text{Lasso}}(\lambda) = \mathbf{0}\}.$$

The existence of this value was proven in Theorem 4[8]. In applications, it can be calculated using the results from Section 2.3. From Equation (2.13), we conclude that as long as the absolute value of a given variable's score at the current estimate is smaller than λ, it will not be included in the active set $\mathcal{A}$, i.e.,

$$\left| s\left(\hat{\beta}_{\text{Lasso}}(\lambda)\right)_j \right| < \lambda \Longleftrightarrow j \notin \mathcal{A}. \tag{4.12}$$

[8] Setting $t = 0$ in Theorem 4 (i.e., $\mathbf{0} \in \mathbb{R}^p$ is the only admissible value in the optimization problem), we can see that there is $\lambda_0 > 0$ such that $\hat{\beta}_{\text{Lasso}}(\lambda_0) = \mathbf{0}$. We conclude that $\lambda > \lambda_0 \Longrightarrow \hat{\beta}_{\text{Lasso}}(\lambda) = \mathbf{0}$ accordingly.

For $\lambda > \lambda_{\max}$, we know that $\hat{\beta}_{\text{Lasso}}(\lambda) = 0$ and $\mathcal{A} = \emptyset$. For $\lambda < \lambda_{\max}$, there is $j \in \mathcal{A}$, such that Equation (4.12) gives

$$\left| s\left(\hat{\beta}_{\text{Lasso}}(\lambda)\right)_j \right| = \lambda.$$

The continuity of this optimization problem finally provides that there is $j \in \{1, \ldots, p\}$ fulfilling

$$\left| s\left(\hat{\beta}_{\text{Lasso}}(\lambda_{\max})\right)_j \right| = |s(\mathbf{0})_j| = \lambda_{\max}. \tag{4.13}$$

Finally, combining formula (4.12) and Equation (4.13) yields the following formula for calculating $\lambda_{\max}$:

$$\lambda_{\max} = \max_{j=1,\ldots,p} |s(\mathbf{0})_j|\,.$$

Based on $\lambda_{\max}$, a log-scaled sequence for λ is built and the exact values of the solution path at those values are calculated using coordinate descent.

Further algorithms have been developed e.g. by Park & Hastie (2007) and implemented in the R package glmpath. Starting with $\lambda_{\max}$, they approximate the values, where the active set $\mathcal{A}$ changes, and linearly interpolate the path at those values. Therefore, they use a slight modification of the LARS algorithm developed by Efron et al. (2004), which yields the exact solution path for the linear regression model.

Chapter 5

Results

We now turn to presenting the results of the methods introduced in Sections 2 and 3 applied on the data set described in detail in Section 4.

As discussed in aforementioned chapters, a logistic Lasso path was estimated to predict default probabilities of bank's clients using the adaptive Lasso estimator. The optimal value of the tuning parameter λ was determined by using three different methods of model selection presented in Chapter 3, AIC, BIC as well as cross-validation (using $K = 10$) with the AUC as measure for model performance. In order to test if cross-validation results are stable, it was repeated 10 times with independent samples. Results will be displayed for all three methods of tuning parameter selection throughout this chapter.

We find that cross-validation provides the highest value for lambda, such that, for all repetitions,

$$\lambda(\text{AIC}) < \lambda(\text{BIC}) < \lambda(\text{CV}).$$

In terms of AUC and in-sample vs. (cross-validated) out-of-sample performance, we find the following results:

	λ	**In-sample AUC**	**Out-of-sample AUC**
AIC	0.0021	86.50%	86.38%
BIC	0.0032	86.47%	86.38%
CV	0.0049	86.42%	86.39%

Table 5.1: Result Overview

We can deduct from Table 5.1 that AIC provides the best in-sample fit, while its cross-validated AUC is lowest compared to BIC or CV[1]. Even though CV provides the highest value of λ within

[1] We will denote the three optimal values of λ with $\lambda(\text{AIC}), \lambda(\text{BIC})$ and $\lambda(\text{CV})$, respectively.

all 10 repetitions of cross-validation, the selection of the tuning parameter seems not to have a huge influence on the model's out-of-sample performance – the difference is just slightly more than one basis point.

As discussed in Chapter 4, several variable transformations, generations and treatments were conducted. The full list of final variables is displayed in the appendix in Table A.2.

5.1 Multivariate Variable Strength

The measurement of multivariate variable strength will be conducted on two dimensions: The first dimension is related to the first entry point of the variable x_i into the Lasso path, i.e., the minimum $\lambda > 0$, for which $\hat{\beta}^i_{\text{Lasso}}(\lambda) \neq 0$, while the second dimension is related to the size of the parameter at the chosen points $\lambda(\text{AIC}), \lambda(\text{BIC})$ and $\lambda(\text{CV})$. As for the first entrance point, an overview is provided in the appendix in Table A.3. Note that discretized variables might have more than one entry point depending on the number of components within the model. As for the parameter size, we will differentiate between discretized and dummy variables as follows:

For dummy variables, we will use the absolute value of their coefficients, as they can already be interpreted as the percentage change of PD when the dummy switches from zero to one.

For discretized variables, this procedure is not conducive, because each variable is represented by a set of parameters (see Section 4.3.1). This set, however, does not depend on the size of the original variables, as they are discretized by using dummy variables in the first place, see Equation (4.1). For such a variable it is useful to know how well it differentiates between the best and worst clients, i.e., which range of differentiation it covers. Therefore, the measure applied for these variables – using explicitly the coefficients $\beta_2, \ldots, \beta_m$ from Section 4.3.1 – is the following:

$$\max_{i,j=2,\ldots,m} |\beta_i - \beta_j| \,. \tag{5.1}$$

5.2 Dummy Variables

In order to measure the strength of a dummy variable, we draw the Lasso path of the coefficients on a log-scale (see Figure 5.1). We do this only for these variables, since a representation of all variables in a single picture would be unclear for the reader.

Measured by the first time that the coefficient is different from zero, we can see that the most important variable is `NeverPastDue`, the indicator representing clients that have never been

Variable	AIC	BIC	CV
LowUtilization	2.12	2.09	2.04
NumberOfTimes90DaysLate	1.53	1.49	1.45
TimesLateCoded9698	1.26	1.21	1.15
NumberOfTime60-89DaysPastDueNotWorse	1.11	1.07	1.02
NeverPastDue	1.06	1.07	1.08
NumberRealEstateLoansOrLines	1.01	1.00	0.97
age	0.80	0.74	0.68
HighUtilization	0.73	0.56	0.29
NumberOfTime30-59DaysPastDueNotWorse	0.73	0.70	0.65
DisposableIncome	0.59	0.53	0.50
NumberNonRELoansLines	0.59	0.53	0.47
LowIncome	0.52	0.53	0.51
Coded_Utilization	0.45	0.42	0.37
No_Line_Utilization	0.42	0.40	0.35
DebtRatio_Low	0.33	0.26	0.22
HighDebt_HighIncome	0.24	0.22	0.22
DebtRatio_High	0.10	0.08	0.05
NumberOfDependents	0.04	0.02	0.00
Income_Unknown	0.00	0.00	0.00
NumberOfDependents_Unknown	0.00	0.00	0.00

Table 5.2: Range of variable coefficients as in Equation (5.1), evaluated at different estimators ordered by strength of AIC estimator. For continuous and dummy variables, the absolute value of the coefficient is displayed.

past due in the past (decreasing the default probability; –).

Measured by the size of the coefficient, all three model selection methods further identify `TimesLateCoded9698`, the coded values of the times past due (+) as similarly important. Weakly important are also `HighDebt_HighIncome` (+), `No_Line_Utilization` (+), `Coded_Utilization` (–) as well as `LowIncome` (–).

In the whole coefficient path, `NumberOfDependents_Unknown`, the indicator for unknown number of dependents, does not appear at all. This may have two possible reasons: Either the number of unknown dependents is too small in the data set, or the categorization to the group of clients with no dependents is already accurate. Since there are almost 4.000 observations having unknown number of dependents (out of 150.000, ≈ 2.5%), the first reason may be ruled out. Therefore, it can be safely assumed that the categorization is accurate in terms of influence on default probabilities. This claim is also supported by the fact that default probabilities do not substantially depend on the number of dependents – this will be discussed in Section 5.3.1 in more details.

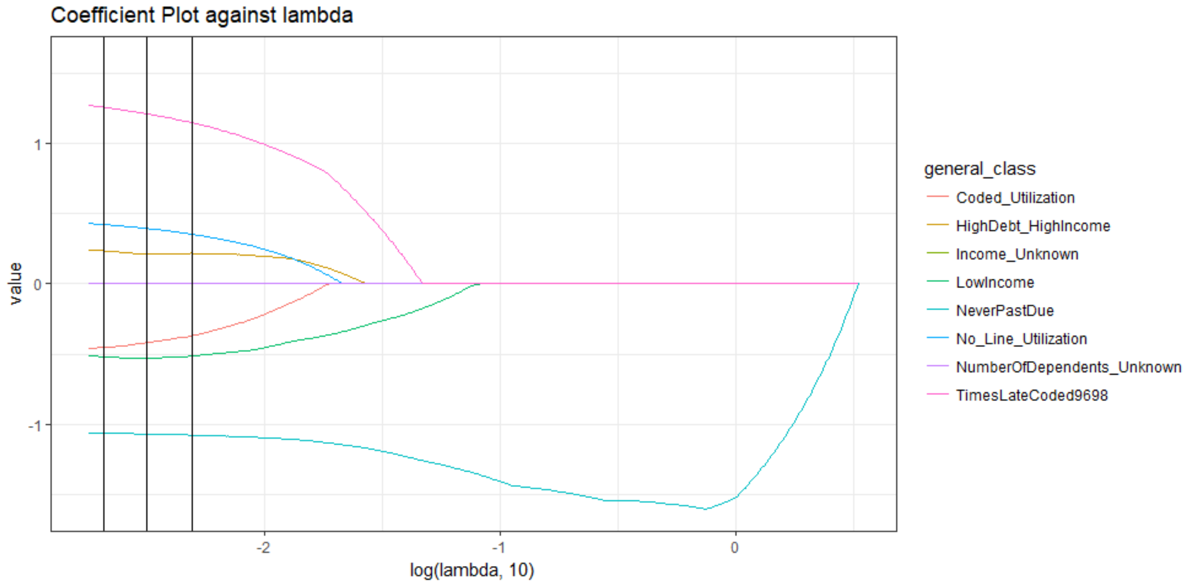

Figure 5.1: The Lasso path for all continuous and dummy variables. The three bold vertical lines denote the places for the optimal lambda values achieved by the three different model selection techniques.

5.3 Discretized Variables

Due to the chosen methods of including continuous and ordinal variables into the analysis, a discretized variable is represented by a set of parameters in the Lasso path (see e.g. Section 4.3.1). This means that at every fixed value of λ, a single variable is represented by a set of parameters displaying the shape of dependence of default probabilities on this variable (see e.g. Figure 4.12). Therefore, it is not conducive to analyze the full coefficient path for this variable type – conversely, results for discretized variables will be shown only for the three different values of lambda achieved by the respective tuning parameter selection methods (see Table 5.1).

5.3.1 Sociodemographic Variables

Age

We can deduce from Figure 5.2 that the age is a significant predictor for the default of a client. All three model selection techniques predict the relationship to be monotonically decreasing – the lowest probabilities to default seem to be starting at the age of 66. The AIC and BIC further discriminates at the ages starting at 81, where the PD starts to rise again.

Since the reasons for the defaults are not available in this data set, it is not possible to provide a perfect explanation for the outcomes. However, this result seems to correspond to basic intuition, since it seems plausible that people tend to seek more stability the older they get and

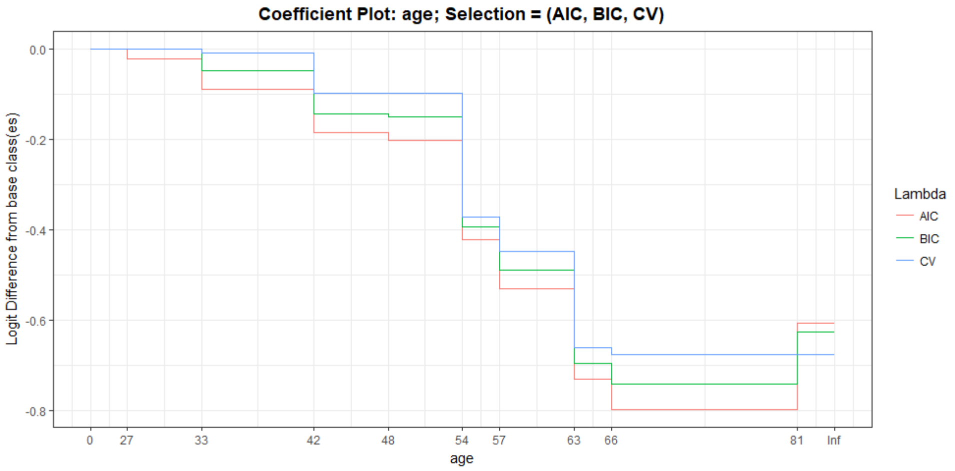

Figure 5.2: Estimated influence of the client's age on the default risk

therefore are less under risk to losing their job and thus ceasing to be able to repay their debt. Another possible explanation could be that banks are less keen to sell risky products to older clients the closer they get to retirement age. Finally, the increase in PD after the age of 81 could possibly be explained by the rise in death probability. Since this increase is not reflected by the Lasso estimator chosen by cross-validation, it might appear solely due to the variation in the data. However, no further analyses were conducted to support this claim.

Disposable Income

The economic intuition behind the inverse relationship between the PD of a client and his disposable income seems clear – the more income is at the disposal of the client, the less will he be likely to default. This intuition is mostly confirmed by Figure 5.3, except for the existence of a slight increase at extremely small levels of income within the dependence structure evaluated at λ(AIC). This small peak between 316$ and 619$ cannot be explained by economic reasoning, therefore we have to assume that this happens due to variation in the data and possibly by an overfit by λ(AIC).

Otherwise, the dependence structures for all three estimators look similar – from graphical inspection, it seems quite linear, i.e., the log-odds of the PD is linearly dependent on the amount of disposable income. Finally, we can see that the level of discrimination in terms of range between clients with different income levels is highest for λ(AIC), see also Table 5.2.

Dependents

The a priori economic intuition concerning the number of dependents would dictate that a higher number of dependents would lead to an increase in PD, since the probability that costs rise need to include the probability that they rise for every single dependent of the client.

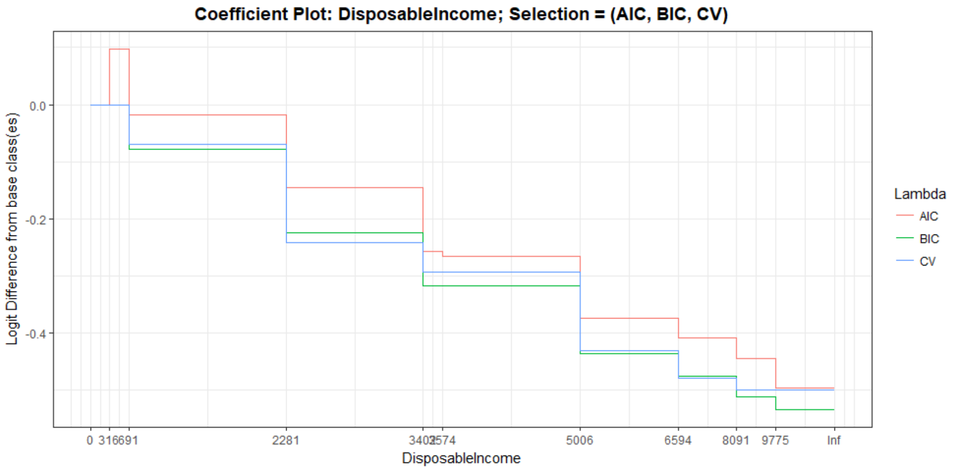

Figure 5.3: Estimated influence of the client's disposable income on the default risk

However, we see none of this in the data. The estimator achieved by cross-validation completely excludes the variable, while the AIC and BIC assign negligibly small contributions with a positive sign – e.g. λ(BIC) assigns a relative difference in PD of 2% to clients with at least one dependent.

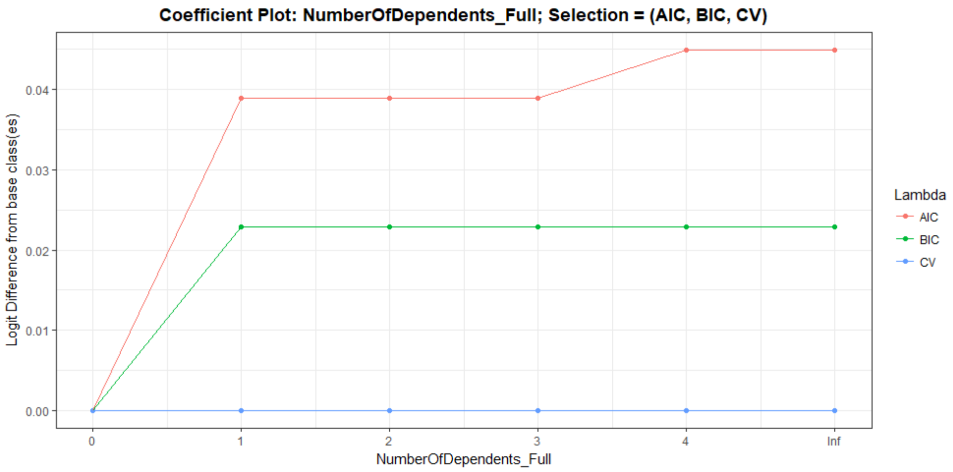

Figure 5.4: Estimated influence of the number of the client's dependents on the default risk

5.3.2 Loan Characteristics

Debt to Income Ratio

Economic intuition suggests that a high debt to income ratio – which is probably intuitively the best proxy for the financial burden of an individual – should lead to higher PDs. This is clearly fulfilled for the interval $[0, 1]$ (see Figure 5.5), where the PD shows a monotonically increasing

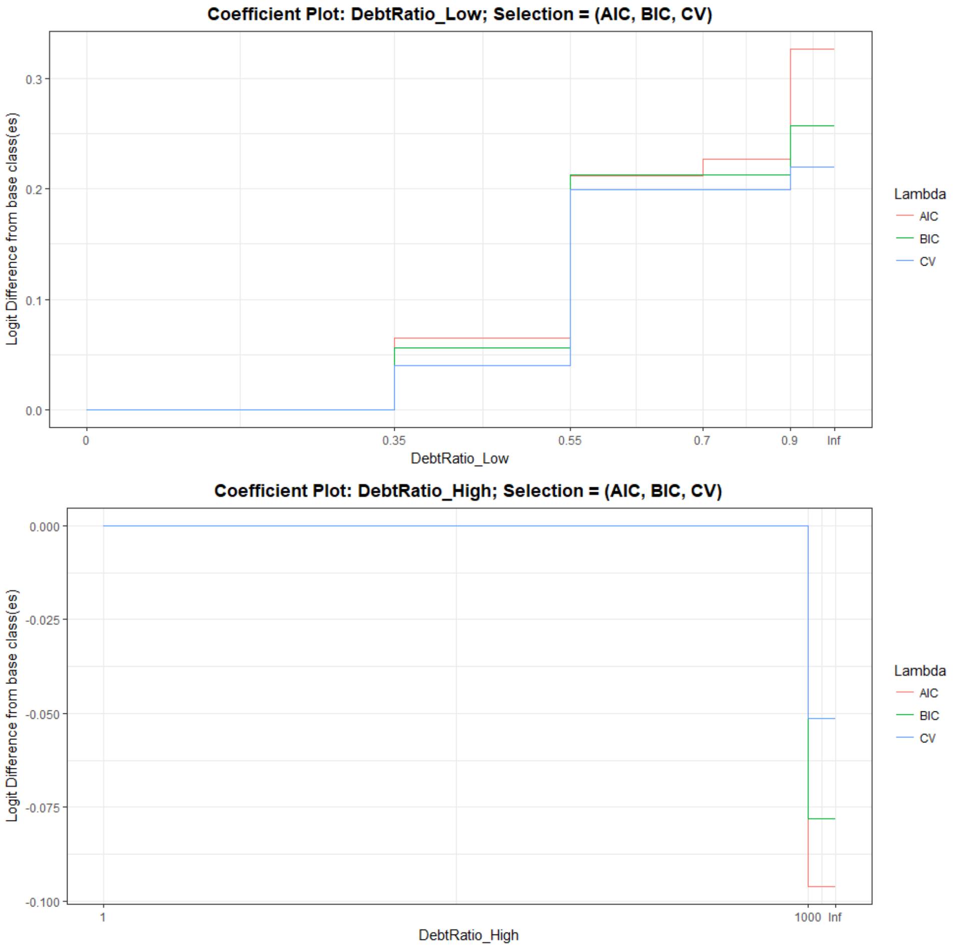

Figure 5.5: Estimated influence of the client's debt to income ratio on the default risk.

form. A little surprisingly, however, PDs cease to rise for debt to income ratios higher than one[2] and, additionally, decrease starting at the value of 1000. This decrease is quite insignificant, however, due to the low absolute values of the coefficient.

In order to get a complete picture of the dependency of the PD on the debt to income ratio, Figure 5.5 has to be viewed in conjunction with the coefficient of the dummy variable `HightDebt_HighIncome`. As can be seen in Figure 5.1, the coefficient of this dummy variable is assigned a value of around 0.2 for all methods of tuning parameter selection. This means that individuals with higher debt than one and income over 2000$ are ceteris paribus 20% more probable to default within two years than others. We can see that the income condition was necessary to include, because apparently it is much more likely that high debt ratios affect the PD, when debt is compared to a high enough value of income, such that this ratio becomes

[2] Notice that due to the design of the newly introduced variables, Figure 5.5 needs to be interpreted cumulatively, such that a client with a debt to income ratio in the interval $(1, 1000)$ would have the same default risk if he had a debt to income ratio of 1.

meaningful.

This result is somewhat surprising – economic intuition would suggest that high debt payments are only truly significant in case the income of the client is low or does not exist (e.g. due to unexpected unemployment). However, it seems that either data quality is poor for clients with income reported in a low range or, more probably, they have a different source of income that is not captured in the data – e.g. a spouse that is a co-debtor on the loan. It might also be the case that these clients only have small revolving loans (e.g. credit cards) that are cancelled by the bank after cutoff-date of the data set.

Number of Loans and Credit Lines

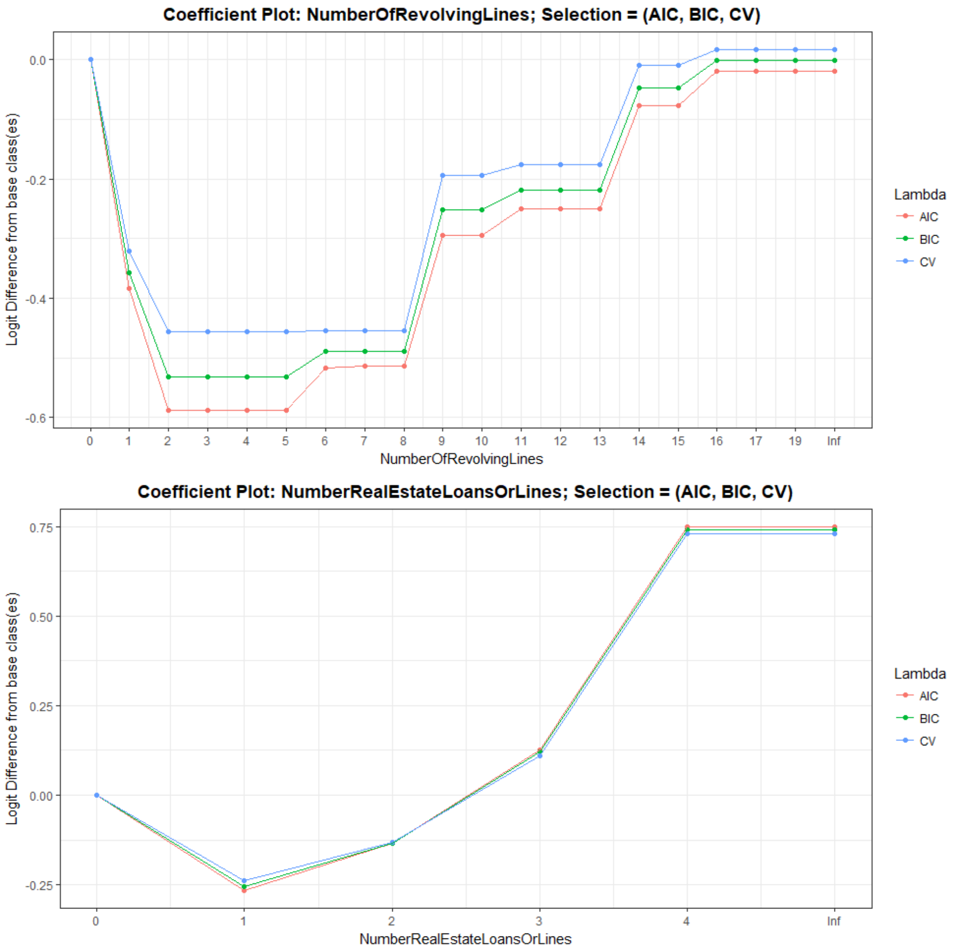

Figure 5.6: Estimated influence of the number of lines the client holds on the default risk.

Both the number of non-real estate loans[3] as well as the number of real estate loans show similar shapes. Creditors without any of those products have a relatively high default risk. Possessing

[3] Real estate loans (or mortgages) are loans that are secured by real estate collateral, i.e., in case the client defaults, the bank is allowed to claim or sell the collateral up to the notional value of the loan.

a small number of loans (this number is between 1 and 8 for non-real estate loans and between 1 and 2 for real estate loans) decreases the PD, which sharply increases for clients who purchased more products. For example, the default risk of a creditor may double if he increases the number of real estate secured loans from one to 4.

The shape of dependence between the number of loans and the default risk measured by the model seems to be highly nonlinear but surprisingly stable to which selection method is used. Creditors with a low number of loans seem to be less probable to default than those without any loans, while at some point, default risk rises again. Probably the most suprising result is the fact that default risk seems to be high in case the client does not have any loans. Observing this, the question arises on how the client actually goes into default, if he does not have any loan to default upon. Based on the data this question cannot be evaluated – there might be some unknown underlying processes (e.g. these clients might have had a revolving loan in the past that has already been cancelled by the bank) that cause this behavior.

5.3.3 Behavioral Variables

Days Past Due

Quite unsurprisingly, the variables indicating payment behavior in the past are strong predictors for credit defaults as indicated in Figure 5.7. If a client was at least twice 90 days past due in the past two years, his two-year PD will increase by up to 150%. Although the less serious delinquency events do not leave space for such a strong increase in default risk, however, twice experiencing a 30 days past due event will already increase the PD of the client by 50%.

Understandably, these variables are frequently used by banks to discriminate between “good” and “bad” customers. However, relying on payment behavioral variables also has some drawbacks:

On the one hand, this information is not available for clients without loans, e.g. those who only keep their salary account at the bank. While this is not a problem for scoring existing customers, it becomes problematic in case the bank wants to grant a loan to a new customer, who it will not be able to score, since no past behavior of that client will be available.

On the other hand, credit scores might become unstable as the number of past due events in a certain time period can increase or decrease strongly over time. In addition, delinquency counters are bound to some processes, e.g. in many banks counters are not even triggered in case the overdue amount is below a certain threshold. Any process change (e.g. changing the threshold) can distort the meaning of the counter over time.

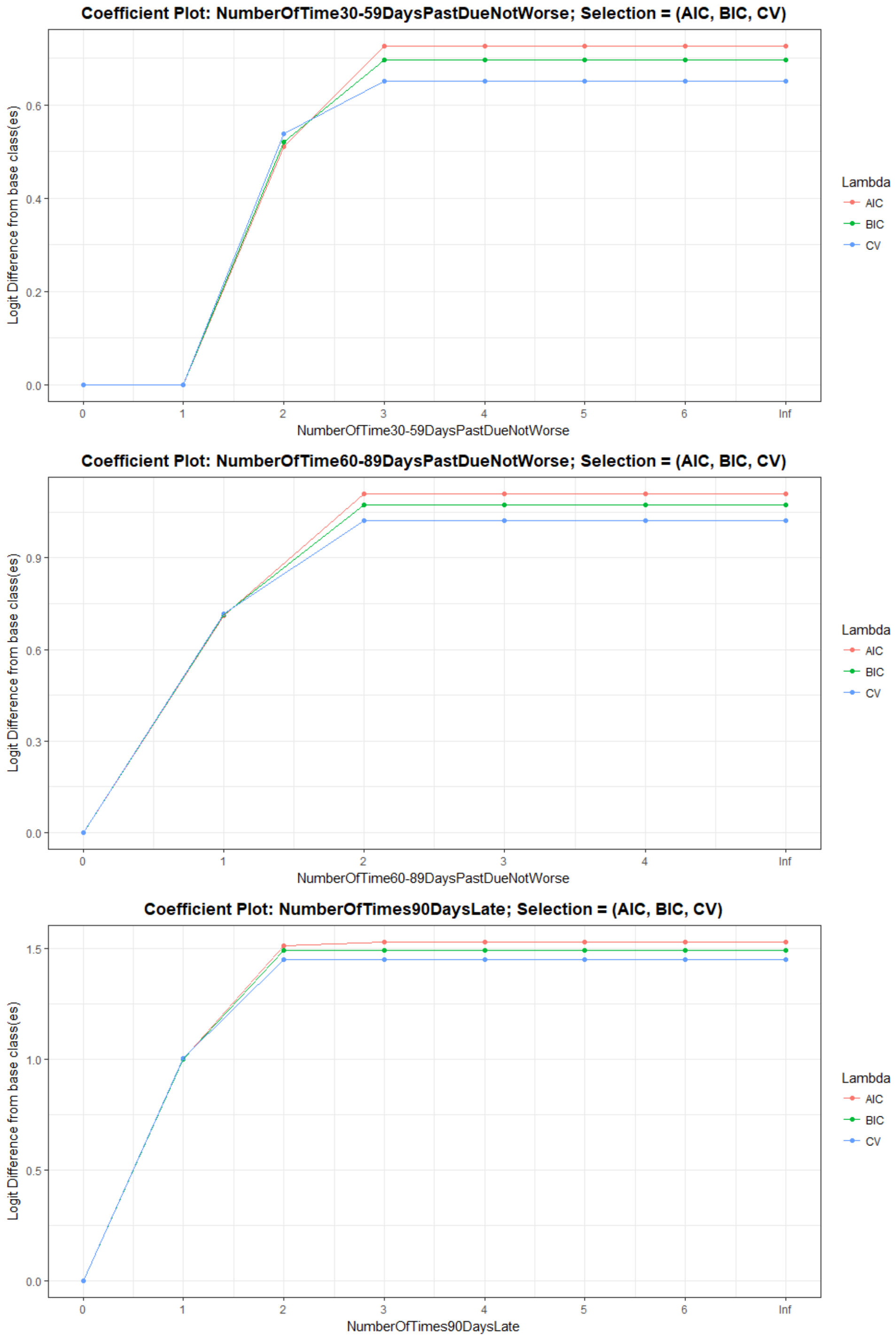

Figure 5.7: Influence of the client's past due behavior on the PD

Credit Line Utilization

Credit cards or other credit lines are products, which usually charge relatively high interest rates. Therefore, it is not advantageous for a bank's client to extensively use his line and it would be expected that he only does this in case he is financially unstable, such that he has troubles paying his bills using his current account. As a consequence, the modeller would expect that higher credit line usage is an indication for a higher probability to default.

Conversely, clients who do have a credit line at their disposal but did not make use of it in the past are more probable to not making use of it in the future as well and are therefore less probable to go into default at all on these products. Consequently, economic intuition dictates that lower credit line usage leads to lower probability to default.

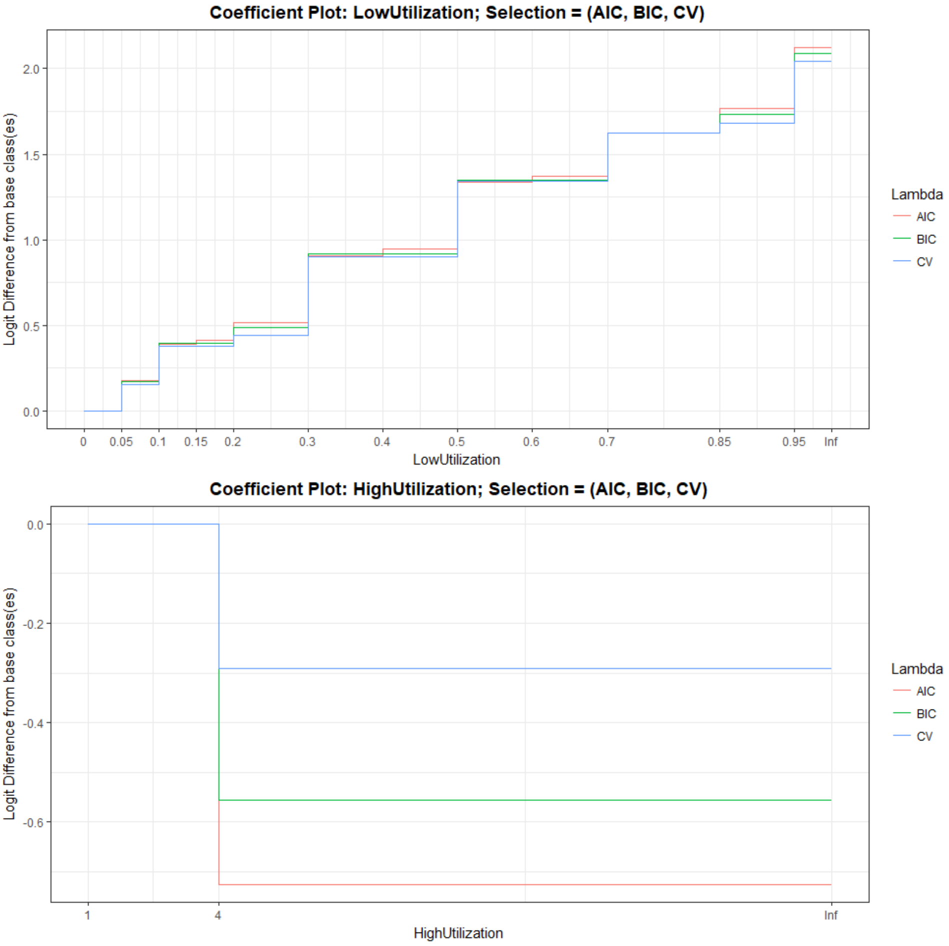

Figure 5.8: Estimated influence of the client's credit line utilization behavior on the default risk

This intuition is fully reflected in the data as we can see in Table 5.2 – credit line usage is not only strongly positively correlated with default probabilities, but also is the strongest driver in terms of coefficient range. More specifically, we can deduce from Figure 5.8 that the default risk of a client increases by 200% if his credit line usage reaches 100%.

Although unclear how it is possible, exceeding this threshold leads to a decrease of default risk – this is at least reflected in values starting at a utilization of 400%, where the PD bounces back to the level, which can be observed at the utilization level of 50-70%. As there are no more details at disposal, the author cannot analyze whether this is a data quality issue or if there might be a reasoning behind these results. In general, this contradicts economic intuition and cannot be explained based on the author's experience.

5.4 Score Distributions

One capability of a well-performing model has to be that as many good customers as possible are attached a good credit score based on their characteristics, while as many bad customers as possible should receive a bad score. As already discussed in Section 3.2, it is not straightforward to define the threshold, below which credit scores are good and above which scores are bad. An average over different thresholds is provided by the AUC (in the sense as discussed in Section 3.2.3), for which results have already been presented in the beginning of this Section.

Performance Measure	**AIC**	**BIC**	**CV**
Pietra-Index	0.5715	0.5717	0.5718
cutoff-value s	-0.1876	-0.1119	-0.1410
fnr $(\alpha(s))$	0.2237	0.2086	0.2158
fpr $(\beta(s))$	0.2048	0.2197	0.2124
Bayesian Error Rate	0.0634	0.0635	0.0636
cutoff-value s	2.8956	2.8533	2.8081
fnr $(\alpha(s))$	0.0085	0.0087	0.0089
fpr $(\beta(s))$	0.8293	0.8283	0.8266
Zero-Cutoff	–	–	–
cutoff-value s	0.0000	0.0000	0.0000
fnr $(\alpha(s))$	0.1866	0.1864	0.1872
fpr $(\beta(s))$	0.2445	0.2441	0.2438

Table 5.3: ROC performance measures of the model – each measure defines a cutoff point s, for which we display FNR and FPR (α- and β-errors).

If the modeller wants to analyze a specific cutoff-value, he has more natural choices available. Specifically for this work, $s = 0$ already provides a natural choice, because it was made sure that the average in-sample score is zero. Otherwise, the performance measures introduced in Section 3.2.4, the Pietra-Index, see Equation (3.8), as well as the Bayesian Error Rate, see Equation (3.9), are both based on a single specific cutoff-value s derived by the solution of an optimization problem.

Table 5.3 provides an overview over these values and illustrates the values of the α- and β-errors at the cutoff-points. For example, we can deduce that $s = 0$ leads to a misclassification of 24.5% of bad cases (score below zero) and 18.7% of the good cases (score above zero) in case we use the model selected by AIC.

Based on this table, we can also deduce that Pietra-Index puts the cutoff-value rather precautiously, i.e., classifies a big proportion of clients as bad, while the BER cuts off at a point, where almost only defaulted clients are in the sample. This is due to the fact that in this sample, BER weights α-errors much stronger than β-errors, because the a-priori default rate is equal to 6.68%. However, any α-error is much more costly for a bank than a β-error[4], while BER weights any event equally, because it only cares about misclassification in the full sample. For example, it is easy to see that plugging in $p = 50\%$ into Equation (3.9) results in the same cutoff-value as the one resulting by the Pietra-Index – in our sample this would correspond to a $(1 - 6.68\%)/6.68\% \approx 14$ times higher weight for a false positive than a false negative event[5].

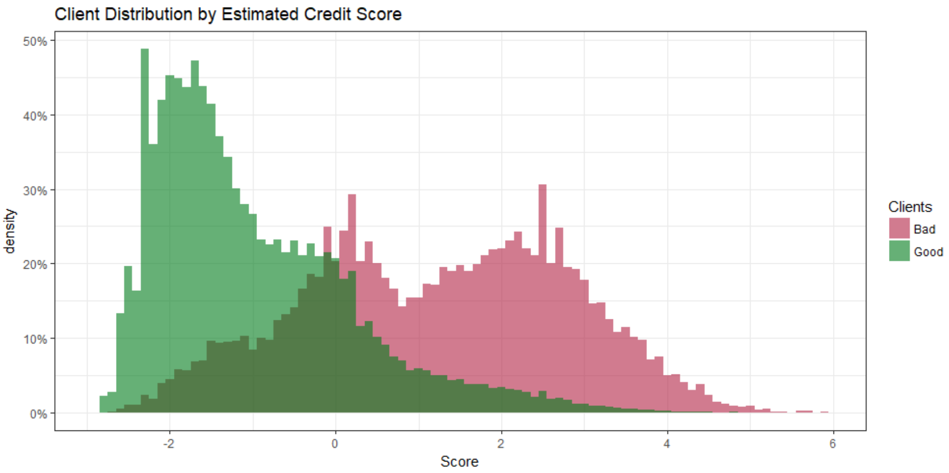

Figure 5.9: Credit score distributions using the AIC as measure for model selection.

In Figure 5.9 we show the credit score distributions from the defaulted (Bad) and the non-defaulted (Good) sub-sample using λ(AIC). The histogram confirms our insights from Table 5.3 – the majority of good cases $(1 - 24.45\% = 75.55\%)$ is attached a credit score below zero, while the majority of bad cases $(1 - 18.66\% = 81.34\%)$ receives a credit score above zero. We can also see that the means of the distributions are visibly different from each other. Concerning

[4] see interpretation of α and β-errors in Section 3.2.2

[5] It is not straightforward to see, whether this value is realistic or not. Cost of default and cost of missed new business depend on highly non-linear factors such as national legislation, interest rates or banks processes. However, some banks do monitor which weight to use in order to determine a suitable cutoff-value.

in particular the distribution of bad cases, we can see that it is bi-modal – this is caused by the predictor `NeverPastDue`, which separates the bad cases into those never past due more than 30 days before their default (i.e., the extremely well-behaving clients) and those, which have already been in delay with their payments.

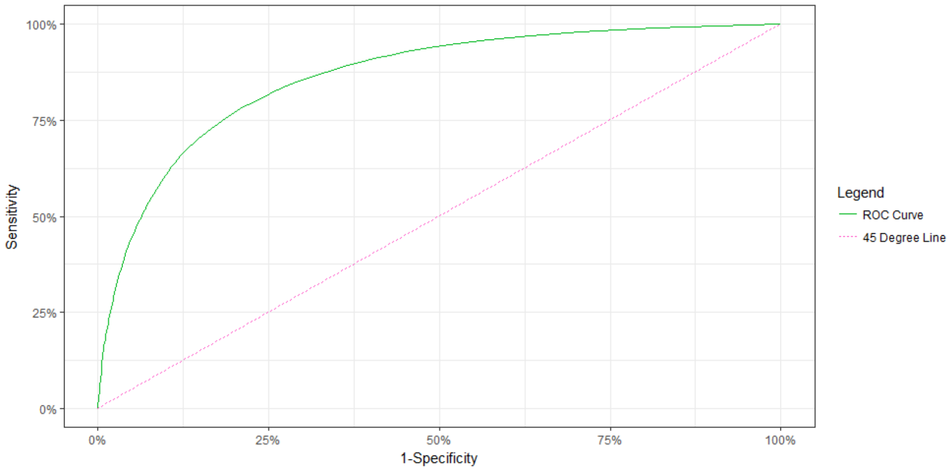

Figure 5.10: Final result of the ROC Curve using the AIC as measure for model selection.

Finally, we display the ROC curve of the model chosen by $\lambda(\text{AIC})$. As already shown in Table 5.1, the area between the ROC curve and the x-axis makes up to over 86% of the total area. This excellent result is in line with our expectations from Section 3.2.3, where we stated that binary logit credit score models reach values between $80 - 90\%$. By visual inspection we can also deduct that the ROC curve is concave such that the PD is a strictly monotonic function of the credit score (see Section 3.2.2).

5.5 Discussion of Results

At this point we turn to summarizing the results displayed in the previous sections.

In general, we can state that the shape of dependence between the PD and most variables selected by all three estimators follows economic intuition. Exceptions can appear within small, poorly populated intervals of some variables, where the estimator chosen by AIC measures some significant jumps in directions that are economically unexpected, e.g. it measures a drop in default risk after a certain level of credit line utilization. These inconsistencies are most probably caused by the variation within the data.

As discussed in Section 5.3.2, the shape of dependence between the number of credit lines and the probability to default is economically difficult to interpret. A priori, the influence could be thought positive in case the portfolio reflects rather risky clients, who might be more keen to accept many loans because they are in financial difficulties and use them for redemption or interest payment of other loans. The influence might also be negative, as a client with a huge number of loans is most probably wealthy, otherwise the bank would not have granted him this number of loans. However, as there are no more details within the dataset, this cannot be properly justified.

In many applications, modelers would discard a variable having such a shape, because it introduces another level of complication into the model, even though there might be a reason behind the shape that is not obvious to the author. In order to analyze whether this is a suitable variable, the modeler would need to evaluate whether its influence shape is stable over time, which is, however, not possible using this data set.

Based on the results, we can deduct that the strongest variables are those which reflect the past behavior of the client, i.e., past due status or utilization of credit line. The absolute strongest variable is `NeverPastDue`, since it is the first one entering the Lasso path and also provides a very good discrimination between good bad clients based on its strongly negative coefficient of -1.

While a good separation between strong and weak clients is the goal of any model measuring default probabilities, it is risky to base a model solely on behavioral variables, because it might cause strong variation in the estimate of this risk over time. Therefore, it is favorable to include relatively stable variables such as income or age, which are solid risk drivers, but by far not as strong as behavioral variables. In addition, it introduces the problem that the model cannot assign default probabilities to new clients of the bank, about whom no behavioral information is available.

Chapter 6

Conclusion

In this work, a binary logistic regression model for two-year default probabilities has been fitted on a data set containing information on 150.000 clients available on kaggle's competition "GiveMeSomeCredit". We have selected the optimal model by choosing a subset of continuous, categorical and ordinal variables reflecting sociodemographic and behavioral properties of the client as well as characteristics of their loans using the Lasso estimator. We have tackled the issue of non-linear dependence of default probabilities on the regressors by their discretization and graphical evaluation in a multivariate environment.

We find that the model provides an excellent fit of the data by reaching an average out-of-sample AUC of over 86%, independent of the model selection criterion (AIC, BIC or CV), which lies in the upper range of the industry standard and in range of more complicated modeling approaches such as in Wang et al. (2015). We see that the estimator gives the strongest weights to behavioral variables such as past due status and limit utilization, while sociodemographic variables and loan properties are much less significant. This is problematic for the application of the model on credit decisions for counterparts that are not yet clients of the bank, where no behavioral component can be observed.

The biggest caveat of the model is that only one time snapshot is available and thus no stability analyses can be conducted over time. This is especially important in order to analyze whether the variables in the model also perform equally well during an economic crisis.

Further research on this data set may focus on testing interactions between the variables in order to further discriminate between good and bad clients. Especially, as can be seen in the bi-modal distribution in Figure 5.9, it might be advantageous to differentiate between well and worse-behaving clients by separating the data set into clients which were never past due and the rest and test whether the same coefficients result in these subsegments.

Bibliography

ABDOU, H. & POINTON, J. (2011). Credit scoring, statistical techniques and evaluation criteria: A review of the literature. *Intelligent Systems in Accounting, Finance and Management* **18**, 59–88.

AKAIKE, H. (1973). *Information Theory and an Extension of the Maximum Likelihood Principle*. Budapest: Akademiai Kiado, pp. 267–281.

ALBERT, A. & ANDERSON, J. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71**, 1–10.

AMEMIYA, T. (1980). Selection of regressors. *International Economic Review* **21**, 331–354.

AMEMIYA, T. (1985). *Advanced Econometrics*. Massachusetts: Cambridge University Press.

BELLOTTI, T. & CROOK, J. (2009). Credit scoring with macroeconomic variables using survival analysis. *The Journal of the Operational Research Society* **60**, 1699–1707.

BREIMAN, L. (1996). *Bagging Predictors*, vol. 24. Kluwer Academic Publishers, pp. 123–140.

CAMERON, A. C. & TRIVEDI, P. K. (2005). *Microeconometrics - Methods and Applications*. Cambridge University Press.

CHOI, H., KOO, J.-Y. & PARK, C. (2015). Fused least absolute shrinkage and selection operator for credit scoring. *Journal of Statistical Computation and Simulation* **85:11**, 2135–2147.

EFRON, B., HASTIE, T., JOHNSTONE, I. & TIBSHIRANI, R. (2004). Least angle regression. *The Annals of Statistics* **32**, 407–499.

FAN, Y. & TANG, C. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society* **75**, 531–552.

FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1–22.

GENKIN, A., LEWIS, D. & MADIGAN, D. (2007). Large-scale bayesian logistic regression for text categorization. *Technometrics* **49**, 291–304.

Green, D. M. & Swets, J. A. (1966). *Signal detection theory and psychophysics.* John Wiley and Sons Inc.

Lokhorst, J. (1999). The lasso and generalised linear models. Honors project, University of Adelaide, Adelaide.

Meier, L., van de Geer, S. & Brühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society* **70**, 53–71.

Oesterreichische Nationalbank (2004). *Guidelines on Credit Risk Management – Rating Models and Validation.* Oesterreichische Nationalbank.

Osborne, M., Presnell, B. & Turlach, B. (2000). On the lasso and its dual. *Journal of Computational and Graphical Statistics* **9**, 319–337.

Park, M. & Hastie, T. (2007). L_1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society* **69**, 659–677.

R (1992). The r project for statistical computing. `http://www.r-project.org`. Accessed: 31 October 2017.

Rockafellar, R. T. (1996). *Convex analysis.* Princeton Landmarks in Mathematics and Physics. Princeton University Press.

Rosset, S. & Zhu, J. (2007). Piecewise linear regularized solution paths. *The Annals of Statistics* **35**, 1012–1030.

Roth, V. (2004). The generalized lasso. *IEEE Transactions on Neural Networks* **15**, 16–28.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

Shevade, S. & Keerthi, S. (2003). A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics* **19**, 2246–2253.

Siddiqi, N. (2006). *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Silvapulle, M. (1981). On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society* **43**, 310–313.

Sun, X., Qu, Q., Nasrabadi, N. & Tran, T. (2014). Structured priors for sparse-representation-based hyperspectral image classification. *CoRR* **abs/1401.3818**.

Thomas, L., Edelman, D. & Crook, J. (2002). *Credit Scoring and Its Applications.* Philadelphia, USA: SIAM Monographs on Mathematical Modeling and Computation.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society* **58**, 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society* **67**, 91–108.

Wang, H., Xu, Q. & Zhou, L. (2015). Large unbalanced credit scoring using lasso-logistic regression ensemble. *PLoS One* **10**, 1–20.

Wu, T., Chen, Y., Hastie, T., Sobel, E. & Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **25**, 714–721.

Zhang, Y., Li, R. & Tsai, C. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association* **105**, 312–323.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

Chapter A

Appendix

Variable Name	Step	# Components	Selection
`(Intercept)`	1	-	
`NeverPastDue`	2	1	
`NumberOfTimes90DaysLate`	3	1	
`LowUtilization`	4	1	
`LowUtilization`	5	2	
`NumberOfTime60-89DaysPastDueNotWorse`	6	1	
`age`	7	1	
`NumberOfTime30-59DaysPastDueNotWorse`	7	1	
`NumberRealEstateLoans`	8	1	
`LowUtilization`	9	3	
`LowIncome`	10	1	
`DisposableIncome`	11	1	
`NumberNonRELoansLines`	12	1	
`TimesLateCoded9698`	13	1	
`LowUtilization`	14	4	
`DisposableIncome`	15	2	
`age`	16	2	
`LowUtilization`	17	5	
`NumberRealEstateLoans`	18	2	
`NumberRealEstateLoans`	19	3	
`DebtRatio_Low`	20	1	
`NumberOfTimes90DaysLate`	21	2	
`HighDebt_HighIncome`	22	1	
`LowUtilization`	23	6	
`No_Line_Utilization`	24	1	
`NumberNonRELoansLines`	25	2	
`Coded_Utilization`	26	1	
`NumberNonRELoansLines`	27	3	
`NumberRealEstateLoans`	28	4	
`DisposableIncome`	29	3	
`NumberOfTime60-89DaysPastDueNotWorse`	30	2	

Variable Name	Step	# Components	Selection
DisposableIncome	31	4	
age	32	3	
NumberNonRELoansLines	32	4	
age	33	4	
DebtRatio_High	34	1	
LowUtilization	34	7	
DebtRatio_Low	35	2	
NumberOfTime30-59DaysPastDueNotWorse	36	2	
NumberNonRELoansLines	37	5	
DisposableIncome	38	5	
LowUtilization	39	8	
NumberNonRELoansLines	39	6	
DebtRatio_Low	40	3	
HighUtilization	41	1	
age	42	5	
DisposableIncome	43	6	
age	44	6	
NumberNonRELoansLines	45	7	CV
age	46	7	
NumberOfDependents	47	1	
DisposableIncome	48	7	
age	49	8	
LowUtilization	50	9	BIC
LowUtilization	51	10	
DisposableIncome	52	8	
LowUtilization	53	11	
age	54	9	
DebtRatio_Low	55	4	
DisposableIncome	56	9	
NumberOfTimes90DaysLate	57	3	
NumberNonRELoansLines	57	8	
NumberOfDependents	58	2	
DisposableIncome	59	10	AIC

Table A.3: (Discretized) variables entering the lasso path. *Variable Name:* Name of variable; *Step:* Current step in lasso path; *# Components:* Number of (discrete) components in the model at current step; *Selection:* Selected estimator at current step.

Variable Name	Description	Type
`SeriousDlqin2yrs`	*Target variable* Person experienced 90 days past due delinquency or worse	Y/N
`RevolvingUtilizationOfUnsecuredLines`	Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits	percentage
`age`	Age of borrower in years	integer
`NumberOfTime30-59DaysPastDueNotWorse`	Number of times borrower has been 30-59 days past due but no worse in the last 2 years.	integer
`DebtRatio`	Monthly debt payments, alimony,living costs divided by monthy gross income	percentage
`MonthlyIncome`	Monthly income	real
`NumberOfOpenCreditLinesAndLoans`	Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards)	integer
`NumberOfTimes90DaysLate`	Number of times borrower has been 90 days or more past due.	integer
`NumberRealEstateLoansOrLines`	Number of mortgage and real estate loans including home equity lines of credit	integer
`NumberOfTime60-89DaysPastDueNotWorse`	Number of times borrower has been 60-89 days past due but no worse in the last 2 years.	integer
`NumberOfDependents`	Number of dependents in family excluding themselves (spouse, children etc.)	integer

Table A.1: Original variables contained in the data set

Name	Variable Name	Original	Formula	NA Value	Var Type	Bucket Formula
x_1	age	Yes			Continuous	$x_{i+1} = x_i + 3$
x_2	DebtRatio_Low		(4.7)		Continuous	$x_{i+1} = x_i + 0.05$
x_3	DebtRatio_High		(4.8)		Continuous	$x_{i+1} = 10 \cdot x_i$
x_4	DisposableIncome		(4.4)	0	Continuous	$x_i = q_{[i/k]}$
x_5	LowIncome		(4.6)	0	Indicator	
x_6	NumberOfDependents	Yes		0	Ordinal	
x_7	HighDebt_HighIncome		(4.3)	0	Indicator	
x_8	LowUtilization		(4.10)		Continuous	$x_{i+1} = x_i + 0.05$
x_9	HighUtilization		(4.11)		Continuous	$x_{i+1} = 2 \cdot x_i$
x_{10}	No_Line_Utilization		(4.9)		Indicator	
x_{11}	NumberOfTime30-59 DaysPastDueNotWorse	Yes		0	Ordinal	
x_{12}	NumberOfTime60-89 DaysPastDueNotWorse	Yes		0	Ordinal	
x_{13}	NumberOfTimes90DaysLate	Yes		0	Ordinal	
x_{14}	NumberRealEstateLoansOrLines	Yes			Ordinal	
x_{15}	NumberNonRELoansLines		(4.5)		Ordinal	
x_{16}	Coded_Utilization		$x_8 = 0.99999$		Indicator	
x_{17}	NeverPastDue		(4.2)		Indicator	
x_{18}	Income_Unknown		$x_4 = \text{NA}$		Indicator	
x_{19}	NumberOfDependents_Unknown		$x_6 = \text{NA}$		Indicator	
x_{20}	TimesLateCoded9698		$x_{13} \in \{96, 98\}$		Indicator	

Table A.2: Full list of variables after all transformations. *Name:* Short name of variable; *Variable Name:* Final name of variable; *Original:* Yes if contained in original data set; *Formula:* Reference to formula if Original $\neq$ Yes; *NA Value:* Treatment of NA value if available; *Var Type:* Variable type (continuous, ordinal or indicator variable); *Bucket Formula:* If Var Type = Continuous, recursive formula to determine buckets