Humboldt University Berlin

Term paper

Credit Default Prediction

Valeryia Mosinzova, Michail Psarakis, Thomas Siskos

Statistical Programming Languages

supervised by
Alla Petukhina

March 16, 2018

Contents

List of Tables

List of Figures

List of Algorithms

1 Introduction

The term default probability is a financial term which describes the likelihood that a borrower will fail to make scheduled repayments over a specific time horizon, which is usually a year. The effective estimation and prediction of corporate defaults are crucial for asset pricing, credit risk assessment of loan portfolios as well as the valuation of other financial products exposed to corporate defaults (Miao et al., 2018). This issue has been considered in several studies. There are two basic approaches to deal with default risk analysis: the market-based model (a.k.a. structural model) and the statistical approach, determined through the empirical analysis of historical data like accounting data (Härdle et al., 2012).

In our paper we aim to perform default prediction analysis by following the methodologies of Härdle et al. (2012), Chen et al. (2011) and Zhang & Härdle (2010) We seek to replicate some of their results with the help of the R programming language. Our analysis is based on the Creditreform-database, which provides several financial statement variables of German firms. In our work we perform linear and nonlinear techniques for predicting bankruptcy based on financial ratios. We compare the predictive performance between following models: logit, probit, CART (Classification and Regression Tree), random forest and LDA (Linear Discriminant Analysis).

The rest of the paper is organized as follows. In the next section we describe the creditreform-dataset. In Section 3 we explain the data preparation procedures as well as the variables and ratios used in this paper. Section 4 contains a brief introduction to Binary Response Models, with a custom implementation and an application of a Logit model to the creditreform-dataset. In Section 5 we apply Classification and Regression Trees (CART), which is followed by a Section about Random Forests (RF). Linear Discriminant Analysis (LDA) is briefly explained and applied in Section 6. In Section 7, we present the steps we take to evaluate predictions of the selected models and the suitable performance measures. Section 8 summarises the conclusions we draw from our analysis.

2 The dataset

We used the creditreform database obtained from the Laboratory for Empirical and Quantitative Research (LEQR) of the Humboldt University of Berlin (leqr.wiwi.hu-berlin.de) to analyse a sample of 20,000 solvent and 1,000 insolvent German firms from the period of 1996-2007. Due to incomplete data from 2003 onwards and missing data for insolvent firms in 1996 we will focus our analysis on the data of the period 1997 to 2002. About half of the data refers to the years 2001 and 2002. The majority of firms appear several times in different years in the dataset, while the data for the insolvent firms was collected two years prior to the default (Chen et al. (2011)). Each firm is described by several financial statement variables as those in balance sheets and income statements. A complete list of all variables of the Creditreform database as well as their descriptions is provided in table 5 in the Appendix.

The firms are divided into the sectors of construction (39.7%), manufacturing (25.7%), wholesale and retail trade (20.1%), real estate (9.4%) and others (5.1%) (including agriculture, mining, electricity, gas and water supply, hotels and restaurants, transport and communication, financial intermediation and social service activities. The respective composition of solvent firms is manufacturing (27.4%), wholesale and retail trade(24.8%), real estate (16.9%), construction (13.9%) and others (17.1%), including publishing, administration and defence, education and health (Chen et al. 2011). In our project we focus on the four largest industry sectors.

3 Data preparation

We used the free programming language R for our analysis. For cleaning the data we use primarily the dplyr package, which allows us to manipulate data easier than with base R allowing for less verbose, i.e. more easily readable code.

In order to clean the data we load the Creditreform dataset into R with the read.csv() command and store it as "data". Using dplyrs's filter() command we choose only those observations from the years between 1997-2002 and store them as data1.

```
data= read.csv("Data/SPL_data.csv", sep = ";", dec = '.', header = TRUE,
               stringsAsFactors = TRUE)

# Due to missing insolvencies in 1996 and missing data from 2003 onwards,
# we choose only the data of the period 1997-2002
#data1 = data 1997-2002
data1 = filter(data, JAHR>= 1997 & JAHR<=2002)
```

Then we want to choose only those observations belonging to the four most prevalent industry sectors in both solvent and insolvent firms, i.e manufacturing, wholesale and retail trade, real estate and construction. We extract the industry class of firms by using the substring() command and save it in a new column of data1 as follows:

```
data1$Ind.Klasse = substring(data1$VAR26, 1, 2)
```

These two digits contained in the variable "Ind.Klasse" are then used to identify in which of the 17 broad industry sectors, defined in the German Classification of Economic Activities Edition 1993 (WZ93), issued by the German Federal Statistical Office (destatis.de), each firm belongs. If the value of the variable is in the range 15-37 then the firm belongs to the manufacturing sector. Accordingly the ranges 50-52 and 70-74 correspond to "Wholesale and Retail Trade" and "Real Estate" respectively, while the value 45 corresponds to "Construction". We create four subsets of data1 for each of the above mentioned sectors by using the filter() command again and remove data and data1 with the rm() command as we do not need them anymore. The 4 subsets are then bound with rbind() into a new dataset with the name "data".

```
Man = filter(data1, Ind.Klasse %in% as.character(15:37))
Man$Ind.Klasse = "Man"

WaR = filter(data1, Ind.Klasse %in% as.character (50:52))
WaR$Ind.Klasse = "WaR"

Con = filter(data1, Ind.Klasse == "45")
Con$Ind.Klasse = "Con"

RE = filter(data1, Ind.Klasse %in% as.character(70:74))
RE$Ind.Klasse = "RE"

# Remove data and data1 & bind the above subsets to get one dataset containing
# only companies of interest

rm(data, data1)

data = rbind(Man, WaR, Con, RE)
```

Then we turn our interest on the size of the companies. Specifically the distribution of total assets which can be considered to be representative of the distribution of the companies' size (Chen et al., 2011). Following the methodology of Zhang & Härdle (2010), we keep only firms with total assets

Table 1: Number of solvent and insolvent companies per year in the dataset

Year	Solvent	Insolvent
1997	1084	126
1998	1175	114
1999	1277	147
2000	1592	135
2001	1920	132
2002	2543	129

(VAR6 in the dataset) in the range of $10^5 - 10^8$ Euros, since the credit quality of small firms often depends mostly on the finances of a key individual (e.g. the owner).

Finally, we eliminate observations with zero values in variables used as denominators in the calculation of the financial ratios that will be used for the classification of the companies (lines 60-67 in the code) and save the result as "data_clean".

```
60 data_clean = data %>% filter(VAR6 != 0,
61                              VAR16 != 0,
62                              VAR1 != 0,
63                              VAR2 != 0,
64                              VAR12 != 0,
65                              VAR12 + VAR13 != 0,
66                             VAR6 - VAR5 - VAR1 - VAR8 != 0,
67                             VAR19 != 0)
```

We end up with 9591 solvent and 783 insolvent firms, which is a similar result as in Chen et al. (2010).

3.1 Financial Ratios

The Creditreform database contains many financial statement variables for each company. Such statements are often used by investors to evaluate firms in two basic ways. A firm is either compared with itself by analysing how it has changed over time, or the firm is compared to other similar firms by means of a common set of financial ratios (Berk & DeMarzo, 2016). We follow the methodology of Chen et al. (2011) and use 23 financial statement variables to create 28 financial ratios to be used in classification. The variables used in the creation of the financial ratios are summarized in table 1. These financial ratios can be divided into six main groups (risk factors): profitability, leverage, liquidity, activity, firm size and percentage change for some variables.

Table 1 presents all the financial ratios used in the current study, the formulas used for their calculation and their category. In our code, the calculation procedure of the financial ratios can be found in lines 92-121 where we add the ratios for each firm to the dataset by using the `dplyr`'s `mutate()`- command. Then we create a dataset (test_data_rel) where only relevant variables are kept, i.e. ID of the firm, solvency status, year and the 28 financial ratios.

```
test_data = data_clean %>%
  mutate(x1 = VAR22/VAR6,
         x2 = VAR22/VAR16,
         x3 = VAR21/VAR6,
         x4 = VAR21/VAR16,
         x5 = VAR20/VAR6,
         x6 = (VAR20+VAR18)/VAR6,
         x7 = VAR20/VAR16,
         x8 = VAR9/VAR6,
         x9 = (VAR9-VAR5)/(VAR6-VAR5-VAR1-VAR8),
         x10 = VAR12/VAR6,
         x11 = (VAR12-VAR1)/VAR6,
         x12 = (VAR12+VAR13)/VAR6,
         x13 = VAR14/VAR6,
         x14 = VAR20/VAR19,
         x15 = VAR1/VAR6,
         x16 = VAR1/VAR12,
         x17 = (VAR3-VAR2)/VAR12,
         x18 = VAR3/VAR12,
         x19 = (VAR3-VAR12)/VAR6,
         x20 = VAR12/(VAR12+VAR13),
         x21 = VAR6/VAR16,
         x22 = VAR2/VAR16,
         x23 = VAR7/VAR16,
         x24 = VAR15/VAR16,
         x25 = log(VAR6),
         x26 = VAR23/VAR2,
         x27 = VAR24/(VAR12+VAR13),
         x28 = VAR25/VAR1)
```

Profitability ratios have appeared in many studies to be strong predictors for bankruptcy (Chen et al., 2011). They measure the ability of a firm to generate revenue relative to its costs over a specific time period. We calculate 7 ratios belonging to this group (ratios x1-x7). The return on assets ratio (x1), for example, provides information on how effective a firm is in making use of its assets to create income. A higher ratio signals that a firm is able to earn more money on less investment (Chen et al., 2011).

Among the profitability ratios, the net profit margin ratio (x2) shows the percentage of sales which the firm keeps in earnings. A high ratio corresponds to a firm with more profitability and better control over its costs (Chen et al., 2011).

Another important factor of risk measurement is leverage. It refers to the extent that a firm

Table 2: Variables used for the calculation of financial ratios and their description

Variable	Description	Variable	Description
VAR1	Cash and cash equivalents	VAR14	Bank debt
VAR2	Inventories	VAR15	Accounts payable
VAR3	Current assets	VAR16	Sales
VAR5	Intangible assets	VAR18	Amortization and depreciation
VAR6	Total assets	VAR19	Interest expenses
VAR3 - VAR2	Quick assets	VAR20	EBIT
VAR7	Accounts receivable	VAR21	Operating income
VAR8	Lands and buildings	VAR22	Net income
VAR9	Equity (own funds)	VAR23	Increase (decrease) inventories
VAR12	Total current liabilities	VAR24	Increase (decrease) liabilities
VAR12 + VAR13	Total liabilities	VAR25	Increase (decrease) cash
VAR3 - VAR12	Working capital		

Table 3: Definitions of financial ratios

Ratio No.	Formula	Ratio	Category
x1	VAR22/VAR6	Return on assets (ROA)	Profitability
x2	VAR22/VAR16	Net profit margin	Profitability
x3	VAR21/VAR6		Profitability
x4	VAR21/VAR16	Operating profit margin	Profitability
x5	VAR20/VAR6		Profitability
x6	(VAR20+VAR18)/VAR6	EBITDA	Profitability
x7	VAR20/VAR16		Profitability
x8	VAR9/VAR6	Own funds ratio (simple)	Leverage
x9	(VAR9-VAR5)/(VAR6-VAR5-VAR1-VAR8)	Own funds ratio (adjusted)	Leverage
x10	VAR12/VAR6		Leverage
x11	(VAR12-VAR1)/VAR6	Net indebtedness	Leverage
x12	(VAR12+VAR13)/VAR6		Leverage
x13	VAR14/VAR6	Debt ratio	Leverage
x14	VAR20/VAR19	Interest coverage ratio	Leverage
x15	VAR1/VAR6		Liquidity
x16	VAR1/VAR12	Cash ratio	Liquidity
x17	(VAR3-VAR2)/VAR12	Quick ratio	Liquidity
x18	VAR3/VAR12	Current ratio	Liquidity
x19	(VAR3-VAR12)/VAR6		Liquidity
x20	VAR12/(VAR12+VAR13)		Liquidity
x21	VAR6/VAR16	Asset turnover	Activity
x22	VAR2/VAR16	Inventory turnover	Activity
x23	VAR7/VAR16	Accounts receivable turnover	Activity
x24	VAR15/VAR16	Accounts payable turnover	Activity
x25	log(VAR6)		Size
x26	VAR23/VAR2	Percentage of incremental inventories	Percentage
x27	VAR24/(VAR12+VAR13)	Percentage of incremental liabilities	Percentage
x28	VAR25/VAR1	Percentage of incremental cash flow	Percentage

relies on debt as a source of financing (Berk & DeMarzo, 2016). As firms combine debt and equity to finance their operations, leverage ratios are useful in evaluating a firm's ability to meet its financial obligations. We calculate 7 ratios belonging to this group (ratios x8-x14). An example of a leverage ratio is the net indebtedness (x11) which measures the level of short term liabilities not covered by the firm's most liquid assets as a proportion to the firm's total assets. Except from measuring the short term leverage of a firm, this ratio provides a measure of liquidity as well. Another popular leverage ratio is the debt ratio (x13), which is defined as the debt of a company divided by its total assets. While this ratio performs well for public firms, it performs considerably worse for private firms compared to the total liabilities to total assets ratio (x12). The reason for that is that liabilities is a more inclusive term which includes debt, deferred taxes, minority interest, accounts payable and other liabilities (Chen et al.,2011).

The next six financial ratios we calculate belong to the family of liquidity ratios (ratios x15-x20). Liquidity is a common variable in many credit decisions and represents a firm's ability to convert an asset into cash quickly (Chen et al.,2011). Liquidity ratios are important indicators of a firm's health as they assess its ability to meet its debt obligations. Chen et al. (2011) note that the cash to total assets ratio (x15) is the most important single variable relative to default in the private dataset. The quick ratio (x17) is an indicator used to assess if a firm has adequate liquidity to meet short term needs. A higher quick ratio indicates that a cash shortfall of the firm is less likely to occur in the near future (Berk & DeMarzo, 2016).

Another type of ratios which deliver important information on insolvency are the activity ratios (x21-x24). They measure the efficiency of a firm in using its own resources to generate cash and revenue. The asset turnover ratio (x21) measures the ability of a company to produce sales from its assets by comparing sales to its asset base. Accounts receivable (x23) and accounts payable (x24) turnover ratios are powerful predictors (Chen et al., 2011).

We also compute a risk indicator accorting to the size of each firm, which is defined as the

Table 4: Three number summary of the financial ratios for solvent and insolvent firms.

	Insolvent			Solvent		
Ratio	$q_{0.05}$	Median	$q_{0.95}$	$q_{0.05}$	Median	$q_{0.95}$
x1	-0.19	0.00	0.09	-0.09	0.02	0.19
x2	-0.15	0.00	0.06	-0.07	0.01	0.09
x3	-0.22	0.00	0.10	-0.11	0.03	0.27
x4	-0.16	0.00	0.06	-0.08	0.02	0.13
x5	-0.09	0.02	0.13	-0.09	0.05	0.27
x6	-0.13	0.07	0.21	-0.04	0.11	0.35
x7	-0.14	0.01	0.10	-0.07	0.02	0.14
x8	0.00	0.05	0.40	0.00	0.14	0.60
x9	-0.01	0.05	0.56	0.00	0.16	0.96
x10	0.18	0.52	0.91	0.09	0.42	0.88
x11	0.12	0.49	0.89	-0.05	0.36	0.83
x12	0.29	0.76	0.98	0.16	0.65	0.96
x13	0.00	0.21	0.61	0.00	0.15	0.59
x14	-7.75	1.05	7.19	-6.76	2.16	74.37
x15	0.00	0.02	0.16	0.00	0.03	0.32
x16	0.00	0.03	0.43	0.00	0.08	1.41
x17	0.18	0.68	1.88	0.24	0.94	4.55
x18	0.57	1.26	3.72	0.64	1.58	7.15
x19	-0.32	0.15	0.63	-0.22	0.25	0.73
x20	0.34	0.84	1.00	0.22	0.86	1.00
x21	0.24	0.61	2.31	0.16	0.48	2.01
x22	0.02	0.16	0.88	0.01	0.11	0.56
x23	0.02	0.12	0.33	0.00	0.09	0.25
x24	0.03	0.14	0.36	0.01	0.07	0.23
x25	13.01	14.87	17.16	12.82	15.41	17.95
x26	-1.20	0.00	0.74	-0.81	0.00	0.57
x27	-0.44	0.00	0.47	-0.53	0.00	0.94
x28	-12.17	0.00	0.94	-7.03	0.00	0.91

logarithm of total assets (x25) in order to study the insolvency risk of small, medium and large firms (Chen et al., 2011). Finally, we calculate the ratios of the percentage change of incremental inventories, liabilities and cash flow (x26-x28). As the increased cash flow is the additional operating cash flow that an entity receives from taking on a new project, a positive incremental cash flow means that the firm's cash flow will increase with the acceptance of a project, the ratio of which indicates that the firm should invest time and money in the project (Chen et al., 2011).

In order to avoid sensitivity to outliers in applying the Random Forest, CART and the logit model, we follow the methodogy of Chen et al. (2011) and we replace extreme ratio values according to the following rule: For i = 1,...,28, if $x_i < q_{0.05}(x_i)$, then $x_i = q_{0.05}(x_i)$, and if $x_i > q_{0.95}(x_i)$, then $x_i = q_{0.95}(x_i)$, where $q_{0.05}(x_i)$ and $q_{0.95}(x_i)$ refer to the 0.05 and 0.95 quantiles of the ratio x_i respectively. This will make our results robust and insensitive to outliers. For that purpose we create the function "replace_extreme_values()" (lines 83-91 of the "data.preparation" quantlet), which is then separately applied to the subsets of solvent and insolvent companies (lines 134-145). The lower and upper quantile as well as the median of the financial ratios for solvent and insolvent firms are presented in table 4. Our results coincide almost entirely with those of Chen et al. (2011). The final clean dataset to be used in further analysis is then created by binding the two subsets of solvent and insolvent firms and is saved as "data_clean".

4 Binary Response Models

4.1 Theory

Logit and Probit models are the most important and probably most used member of the class of models called Binary Response or Generalized Linear Models. These parametric models are applied to estimate the probability of an observation belonging to a particular class while using observed data of known outcomes to infer models which predict the outcome of future observations.

The outcome of interest in the Creditreform-dataset is of binary nature. Firms can either go bankrupt or not. This behavior is commonly modeled by Binary Response Models like the probit or the logit model. Binary response variables follow a Bernoulli probability function

$$f(y|x) = P(y=1|x)^y \left(1 - P(y=1|x)\right)^{1-y}, \qquad y \in \{0,1\},\ x \in \mathbb{R}^d, d \in \mathbb{N}, \tag{1}$$

where $P(y=1|x)$ stands for the conditional probability of observing $y=1$ given x. Both probit and logit models have in common that $P(y=1|x)$ is modeled by a monotonic transformation of a linear function

$$P(y=1|x) = G(x\prime\beta), \qquad \beta \in \mathbb{R}^d, \tag{2}$$

where $x\prime\beta$ is the scalar product of x and β. Additionally we require that $0 \leq G(x\prime\beta) \leq 1$, since it denotes a probability.

For the probit model G will be the cumulative density function of the normal distribution

$$P(y=1|x) = G(x\prime\beta) = \Phi(x\prime\beta) = \int_{-\inf}^{x\prime\beta} \frac{1}{\sqrt{2\pi}} exp\left[-(\frac{t^2}{2})\right] dt. \tag{3}$$

Here $\Phi(x\prime\beta)$ stands for the cumulative density function of the normal distribution.

For the logit model G will be replaced by the cumulative density function of the logistic distribution $\Lambda(x\prime\beta)$:

$$P(y=1|x) = G(x\prime\beta) = \Lambda(x\prime\beta) = \frac{exp(x\prime\beta)}{1+exp(x\prime\beta)} \tag{4}$$

The parameter vector β is obtained by the Maximum-Likelihood method. Given independent and identically distributed samples, the Likelihood function can be written as

$$\begin{aligned} L(\beta; y, x) &= \prod_{i=1}^{n} f(y_i|x_i) = \prod_{i=1}^{n} P(y_i = 1|x_i)_i^y \left(1 - P(y_i=1|x_i)\right)^{1-y_i} \\ &= \prod_{i=1}^{n} G(x_i\prime\beta)_i^y \left(1 - G(x_i\prime\beta)\right)^{1-y_i} \end{aligned} \tag{5}$$

where we just take the product over all individual Bernoulli-functions.

The log-Likelihood can thus be written as

$$l = \log L(\beta; yx) = \sum_{i=1}^{n} y_i \log G(x_i\prime\beta) + (1-y_i)\log(1 - G(x_i\prime\beta)) \tag{6}$$

The Maximum-Likelihood estimators $\beta_M L$ are calculated as

$$\beta_M L = argmax(l) \tag{7}$$

and solve the first order conditions for a maximum.

$$\frac{\partial l}{\partial \beta} \stackrel{!}{=} 0 \tag{8}$$

In general, the resulting system of equations has no closed-form solution for $\beta_M L$ and numerical solutions are needed which can be obtained by iterative optimization techniques, one we will implement in the next section.

4.2 Implementation

The architecture of the `brm`-class follows the general structure outlined in the chapter before. First, it generates a log-Likelihood-function, which it then optimizes using a Gradient-Descent-Algorithm. After training the model it is possible for the user to very easily obtain predictions by invoking the `predict()`-function, which has been augmented with a method for the `brm`-class.

4.2.1 Obtaining the Likelihood

The first task is to define a function that accepts a distribution and yet undefined data as it's input and first extracts all suitable variables, then expresses the Likelihood-function from 6 and finally returns another function which depends only on the weights β. This task is performed by the `get_likehood()`-function. We outline the function pass of `get_likehood()` first in pseudo-code followed by a look on the implementation in the `R` language.

Algorithm 1 get_likehood()

procedure SET UP AUXILIARY VARIABLES
  $grp \leftarrow$ unique labels
  $nums \leftarrow$ extract numeric columns
  $Xy\_mat \leftarrow$ bind numeric variables as a matrix
  $y\_pos \leftarrow$ cache position of the outcome variable
procedure SET UP LOG-LIKELIHOOD
  $l \leftarrow 0$
  for: x_i, y_iin data :
    calculate : $j = y_i \log G(x_i\prime\beta) + (1 - y_i)\log(1 - G(x_i\prime\beta)$
    update : $l \leftarrow l + j$
  return : $l(\beta)$

The heavy lifting in this function is done by this `R`-snippet:

```r
l = function(x){
  - sum(apply(Xy_mat, 1,
              function(X){
                X[y_pos] * distr(t(X[-y_pos]) %*% x,
                                 lower.tail = TRUE,
                                 log.p = TRUE) +
                  (1-X[y_pos]) * distr(t(X[-y_pos]) %*% x,
                                       lower.tail = FALSE,
                                       log.p = TRUE)}))}
# Return loglikelihood as a function of x (here 'x' stands for the weights)
return(l)
```

The `for`-loop from the pseudo code is implemented as an `apply`-call to `Xy_mat`, which in turn is a matrix of numeric columns. The `apply`-function initially selects each row in `Xy_mat` and extracts the outcome-variable `Xy_mat[y_pos]` which corresponds to y_i from 6. It then computes the scalar product between the regressors of `Xy_mat`'s row, which plays the role of $x_i'\beta$ in 6. The scalar product is wrapped in `dist` which represents the cumulative distribution function $G(x'\beta)$ and is one of the arguments to `get_loglikelihood()`.

```
  distr = switch (mode,
                  "logit" = plogis,
                  "probit" = pnorm
  )
```

This way `dist` will point to the built-in functions for computing probabilities, depending if the user wishes to train a logit or a probit model. The arguments `lower.tail=TRUE` and `lower.tail=FALSE` stand for $G(x'\beta)$ and $G(x'\beta) = 1 - G(x'\beta))$ respectively. Finally the resulting vector is summed up and multiplied by -1. This is done because of the way we implemented the Gradient Descent algorithm. Currently `gradientDescentMinimizer()` can only find minima. However, Maximum-Likelihood estimation poses a maximization problem. Luckily, we can transform any maximization problem into a minimization problem by multiplying with minus one.

4.2.2 Gradient Descent

Since we now have a log-Likelihood function, the next step is to optimize it. To this effect we deploy a Gradient Descent algorithm. Theory tells us that in order to reach the minimum of a function $f(x)$ starting at a particular $x \in \mathbb{R}^d, d \in \mathbb{N}$ one needs to follow the negative gradient $\nabla f(x)$ of f evaluated at x. This leads to the iterative rule we can exploit

$$x_{t+1} = x_t - \eta \cdot \nabla f(x_t), \qquad t \in \mathbb{N}, \eta \in \mathbb{R}^+ \tag{9}$$

where η is the learning rate. To this standard method of performing a Gradient Descent routine we will also make some minor modifications. First of all, we will approximate the gradients by taking finite differences, which is easier to implement albeit computationally inefficient. Finite differences are computed by

$$\nabla f(x_t) \approx \frac{f(x_{t+1}) - f(x_t)}{\epsilon}, \qquad \epsilon > 0 \tag{10}$$

Secondly, we will before we initialize the algorithm, try a set of random points and chose the one that provides the lowest value of the objective function as a starting point for the Gradient Descent Routine. This also ensures to an extend that the algorithm, if it reaches convergence, finds the global minimum. The final modification will be to prune the gradients. When computing gradients using finite differences, it may happen that the gradient's values can become extremely high for large denominators and very small ϵ. In fact, they can become high enough for `R` to treat them as `Inf` which results in the gradients being treated as `NaN` (not a number). To counteract that, we will limit the gradients to the interval $[-100, 100]$.

The `gradientDescentMinimizer()`-function accepts following arguments:

1. `obj`: an objective function, that accepts exactly one argument called 'x'.
2. `n_pars`: an integer specifying the dimensions of the objective.
3. `epsilon_step`: a float defining the stepwidth used for computing the finite differences.
4. `max_iter`: an integer for the maximum number of iteration before the algorithm aborts.
5. `precision`: a float defining the precision of the solution. All elements of the gradient have to be absolutely lower than `precision` for the algorithm to converge.

Algorithm 2 gradientDescentMinimizer()

procedure SET UP AUXILIARY VARIABLES
    learn_rates ← descending sequence from `learn` to 0
    a ← matrix of 1000 randomly initialized points
    f_a ← vector of function values for each element of a
    update:a ← argmin(f_a)
    gradient ← **compute gradient evaluated at a**
    $i \leftarrow 0$
procedure PERFORM GRADIENT DESCENT
    $l \leftarrow 0$
    while: $i \leq max\_iter$ and any element of gradient > 0 :
        update : $a = a -$ `learn_rates[i]` $\cdot$ *gradient*
        calculate : *gradient* = **calculate gradient**
    if i = max_iter **then raise** warning
    return : a

6. `learn`: a positive float representing the learning rate.

7. `verbose`: a boolean indicating if additional information during training is desired. The default is `FALSE`

8. `report_freq`: If `verbose` is `TRUE`, define how often to print the logstring. The default is 10 which corresponds to a console output being printed every 10 steps.

```
143  a = matrix(data = runif(1000 * n_pars,
144                           min = -10,
145                           max = 10),
146              ncol = n_pars)
147  f_a = apply(a, 1, obj)
148  a = a[which.min(f_a), ]
```

We begin by filling the matrix `a` with 1000 `n_pars`-dimensional points which we draw from the uniform distribution, making use of `R`'s built-in `runif`-function. We draw random numbers within the range of $[-100, 100]$ to cover a wide part of the objective function's domain.

```
get_gradient = function(x, d = n_pars,
                        objective = obj,
                        epsilon = epsilon_step){
  init = matrix(data = x, nrow = d, ncol = d, byrow = TRUE)
  steps = init + diag(x = epsilon, ncol = d, nrow = d)
  f_steps = apply(steps, 1, objective)
  f_comp =  apply(init, 1, objective)
  D = (f_steps - f_comp) / epsilon
  D_trimmed = ifelse(abs(D) <= 100, abs(D), 100) * sign(D)
  return(D_trimmed)}
```

The workhorse in this routine is the `get_gradient()`-function, which computes the finite differences. First we need to compute the values of the objective function at the current and next step (lines 146 and 147). Then we can apply the current and next step as inputs to the objective function and compute the difference $f(x_{t+1}) - f(x_t)$. The current step is provided as the `x` argument to the function call. The next steps need to be inferred by the function. If $f(x_t)$ is multidimensional we need to perform an ϵ-step in each dimension of the vector, since we want to approximate the partial derivatives of $f(x)$ evaluated at x_t. I.e. first we want to increment just the first element of x and store the result, then just the second element, and repeat the process until we reach the last element. If we stack these vectors, we get a matrix of one-directional ϵ-steps that are essentially updates of the starting point x_t with which it is easy to compute the gradients as their element-wise difference, normalized by `epsilon_step`. The gradients are finally trimmed if necessary and returned.

Algorithm 3 get_gradient()

procedure SET UP AUXILIARY VARIABLES

$$init \leftarrow \begin{bmatrix} x_1 & x_2 & \cdots \\ x_1 & x_2 & \cdots \\ x_1 & x_2 & \cdots \\ \vdots & \ddots & \cdots \end{bmatrix}$$

$$steps \leftarrow \begin{bmatrix} x_1 + \epsilon & x_2 & \cdots \\ x_1 & x_2 + \epsilon & \cdots \\ x_1 & x_2 & \cdots \\ \vdots & \ddots & \cdots \end{bmatrix}$$

  $f\_comp \leftarrow$ **apply row-wise:** objective function to *init*

  $f\_steps \leftarrow$ **apply row-wise:** objective function to *steps*

procedure COMPUTE FINITE DIFFERENCES

$$D \leftarrow \frac{f\_steps - f\_comp}{\epsilon}$$

  if $any d \in D \notin [-100; 100]$ **then replace** d by $100 \cdot \text{sign}(\text{d})$

```r
while(any(abs(gradient) >= precision) & i <= max_iter){
  if(i %% report_freq == 0 & verbose) {
    cat("\nStep:\t\t", i,
        "\nx:\t\t", a,
        "\ngradient:\t", gradient,
        "\nlearn:\t", learn_rates[i],
        "\n------------------------------------")}

  i = i + 1
  a = a - learn_rates[i] * gradient
  gradient = get_gradient(a)
}

cat("\nResults\n",
    "\nIteration:\t", i,
    "\nx:\t\t", a,
    "\nf(x):\t\t", obj(a),
    "\ndf(x):\t\t", gradient,
    "\n")

if(i >= max_iter){
  warning("Maximum number of iterations reached.")}
return(a)
```

In each iteration the `while`-loop ensures that convergence has not been reached. This is implemented by a call to `any` wrapped around a vector of logical expressions. If any element of the gradient is still greater than the specified precision, the call to `any` will evaluate to `TRUE`. The second breaking criterion is a safeguard for the loop not to run infinite times. If the current iteration is larger than `max_iter` the algorithm will break and the user will receive a warning (lines 175-176). If the user wishes to receive information about the status of the algorithm during runtime, the optional argument `verbose` can be set to `TRUE` which will print a logstring to the console in regular intervals (lines 156-161).

4.2.3 Predictions

In order to facilitate making predictions based on the `brm`-class we augmented the built-in function `predict()` with a method that works on our custom class in a predefined way.

```
61 predict.brm = function(model, data){
62   Xb = as.matrix(cbind(constant = 1, data)) %*% model$weights
63   predictions = sapply(X = Xb,
64                        FUN = model$distribution,
65                        log.p = FALSE,
66                        lower.tail = TRUE)
67   return(predictions)
68 }
```

A call to `predict()` on a `brm`-model will add a column of ones to the provided `data` and multiply the matrix with the weights calculated during training of the model (line 62). Finally these scores of the index-function $X\beta$ will be applied to the correct distribution. The distribution is stored inside `model$distribution` which points to `pnorm` in case of `brm`-model of `mode` "probit" and a pointer to `plogis` if the `mode` is equal to "logit".

4.3 Using the Logit-Model on the creditreform Dataset

We use the `caret` package to build a logistic regression model. This package provides a `train()` method for fitting the data for various algorithms. We select the algorithm via the parameters supplied to the `train()`-method. The `method` parameter specifies the type of model which is applied to predict the target variable (Kuhn, et al., 2017).

```
  glm_mod=train(as.factor(status)~.,data=training_complete, method="glm", family="
    binomial")
```

We apply the `glm` method. The function glm(), and the predict method for glm objects, reflect a classical approach to statistical inference. It only handles binary classification. Thus, we define our target variable `status` as categorical by using`as.factor()` command (Stephens, 2016). All the other financial ratios are independent variables that influence the binary outcome of a target variable `status`. One important argument is `familyb̄inomial`. Family() function specifies the assumed distribution of the dependent variable `status`. In our case, we model a target variable as a binomial distribution. The binomial family accepts the links such as logit, probit, cauchit, log and cloglog. The link function `links` the output to a linear model. It passes dependent variable through the link function, and then models the resulting value as a linear function of the values of independent variables. Different combinations of family functions and link functions lead to different kinds of generalized linear models (for example, Poisson, or probit). Without an explicit link argument, the link function defaults to standard `logit` (Zumel & Mount, 2014). We store the resulting model in the object texttglm_mod and use `predict()` function to predict the insolvency of the German firms in the years 2000 until 2002. To this end, we pass the trained model `glm_mod` and the test set features without the actual labels to the predict() function. We specify the type of prediction as `type=prob` to receive the probability for a firm's insolvency. The predicted probabilities are passed into `evaluate_prediction` function to estimate the overall prediction power over models.

```
pred_logit=predict(glm_mod, newdata=validierung_logit,type="prob")
```

The practical advantage of logit model is that it provides insight into the impact of each predictor variable on the response variable. None of the nonparametric methods really provide this information. The coefficients of a logistic regression model can be thus treated as advice and may serve as orientation for decision making (Zumel & Mount, 2014). Yet, the calculation and interpretation of coefficients exceeds the scope of this project.

The logit model has some limitations. For example, if we include the wrong independent variables, the model will have little to no predictive value. It requires each data point be independent of all other data points. If observations are related to one another, then the model will tend to overweight the significance of those observations (Robinson, 2018). Further the method is searching for a single linear decision boundary that separates insolvent and solvent firms. This works well if the data are linearly separable. A linear separating boundary is, however, not suitable if there is doubt that the separation mechanism is of a nonlinear kind (Chen, et al., 2011). Therefore, it is always a good idea to try different models and choose the one which performs best on a test data set. Tree-based models for instance, divide the feature space into half-spaces using axis-aligned linear decision boundaries. CART model, which is introduced next, is nice in that it is easy to implement and easy to explain.

5 Classification and Regression Trees (CART)

Tree-based methods, or decision tree methods, are used for two broad types of problems - classification and regression. These methods are appropriate by extensive datasets. Its strength is that, in large data sets, it has the potential to reflect relatively complex forms of data structures, which may be hard to detect with conventional regression modelling. Additionally, the methodology is relatively easy to use and can be applied to a wide class of problems (Maindonald & Braun, 2010). Once again, we use the `caret` package to build a CART model. We pass the `rpart` method into train() function to fit the data (Kuhn, et al., 2017). There is a package with same name `rpart`, which is specifically available for decision tree implementation. Caret links its train function with other packages simplifying the work process significantly.

```
  modfit=train(status~.,method="rpart",data=training_complete)

  #calculate predictions and store the results in confusion matrix:
  pred.cart=predict(modfit,newdata=validierung_cart)
  conf_mat_cart=table(pred.cart,validierung_cart$status)
  TN=TN+conf_mat_cart[1,1]
  FP=FP+conf_mat_cart[1,2]
  TP=TP+conf_mat_cart[2,2]
  FN=FN+conf_mat_cart[2,1]
}
#calculate correlation matrix
correlations=round(cor(training_complete[, -1]),2)
pdf("CorrPlot.pdf")
corrplot(correlations, order = "hclust")
dev.off()
#perform variable importance
pdf("varImp.pdf")
plot(varImp(modfit), main="Variable Importance", top=10)
dev.off()

# use rpart for plotting fancy decision tree and predictions
mod_fit=rpart(as.factor(status)~.,data=training_complete)

pdf("fancyRpartPlot.pdf")
```

```
fancyRpartPlot(mod_fit)
dev.off()

pred_cart=predict(mod_fit,newdata=validierung_cart,type="prob")
write.csv(cbind(label=as.numeric(as.character(validierung_cart$status)),pred_cart),
   file="CART/cart_pred.csv")
```

The basic way to visualize classification or regression tree built with R's `rpart()` function is to call `plot()`. The call typically produces a couple of black clouds of overlaid text. Therefore, the traditional representation of the CART model is not graphically appealing. There are better ways to plot `rpart()` trees for example by using the `prp()` function from `rpart.plot package` or the `fancyRpartPlot()` from `rattle` package (Milborrow, 2016). We have chosen the `fancyRpartPlot()` function to get a legible structure of our decision tree.

fancyRpartPlot() builds more elaborate and clean trees, which can be easily interpreted. Each node box displays the classification, the probability of each class at that specific node and the percentage of observations that arrived at each node (Rickert, 2013).

Once we have built the model, we are ready to validate it on a separate data set - a test set. We store the results in the object `mod_fit` and apply `predict()` function to calculate the probabilities for further analysis.

5.1 Variable Importance

It might be interesting to acquire information regarding which financial statement variables are most important in predicting the outcome. There are practical reasons for doing so. For example, we want to rank-order financial ratios in regard to their relative importance. This information can provide insights about the most important causes of insolvency, allow a deeper understanding of financial health of the companies and enable investors more effective decision making. If predictor variables were uncorrelated, this would be a simple task. We would rank-order the predictor variables by their correlation with the response variable. In our case though, the financial ratios variables are correlated (as shown in the chart below), and this complicates the task significantly.

We used `corrplot` package to graphical display a correlation matrix. Positive correlations are displayed in blue and negative correlations in red colour. Colour intensity and the size of the circle are proportional to the correlation coefficients. Every correlation coefficient ranges from 1 to -1, where 1 is a prefect positive correlation, -1 a perfect negative correlation, and 0 indicates no correlation between two features.

`caret` package includes a command to perform variable importance. The varImp() function calculate important features of almost all models (Kuhn, et al., 2017). We use a plot method to visualize the results for importance scores generated from varImp() function. We apply the `top=10` option to make the chart more readable and select the ten most important variables.

We can see that the varImp function considers Accounts Payable Turnover (AP/SALE, x24) to be the most important feature to predict German default firms, followed by Interest Coverage Ratio Leverage (EBIT/INTE, x14), Quick Ratio (QA/CL, x17), Operating Income/Total Assets (OI/TA, x3), Net Indebtedness Leverage ((CL-CASH)/TA, x11).

In the whole, decision trees are an attractive method since they work with any type of data, numerical or categorical, without any distributional assumptions and without pre-processing. Furthermore, most R's implementations handle missing data and this method is also robust to redundant and nonlinear data. The algorithm is easy to use, and the output (the tree) is relatively easy to understand. On the other hand, decision trees have some shortcomings. They tend to overfit, especially without pruning, because the decision tree can `memorize` the training set. The model then become very specific to the training data and consequently has less ability to classify the unobserved data correctly. Besides, they have high training variance, which means that samples drawn from the same

Figure 1: Decision Tree

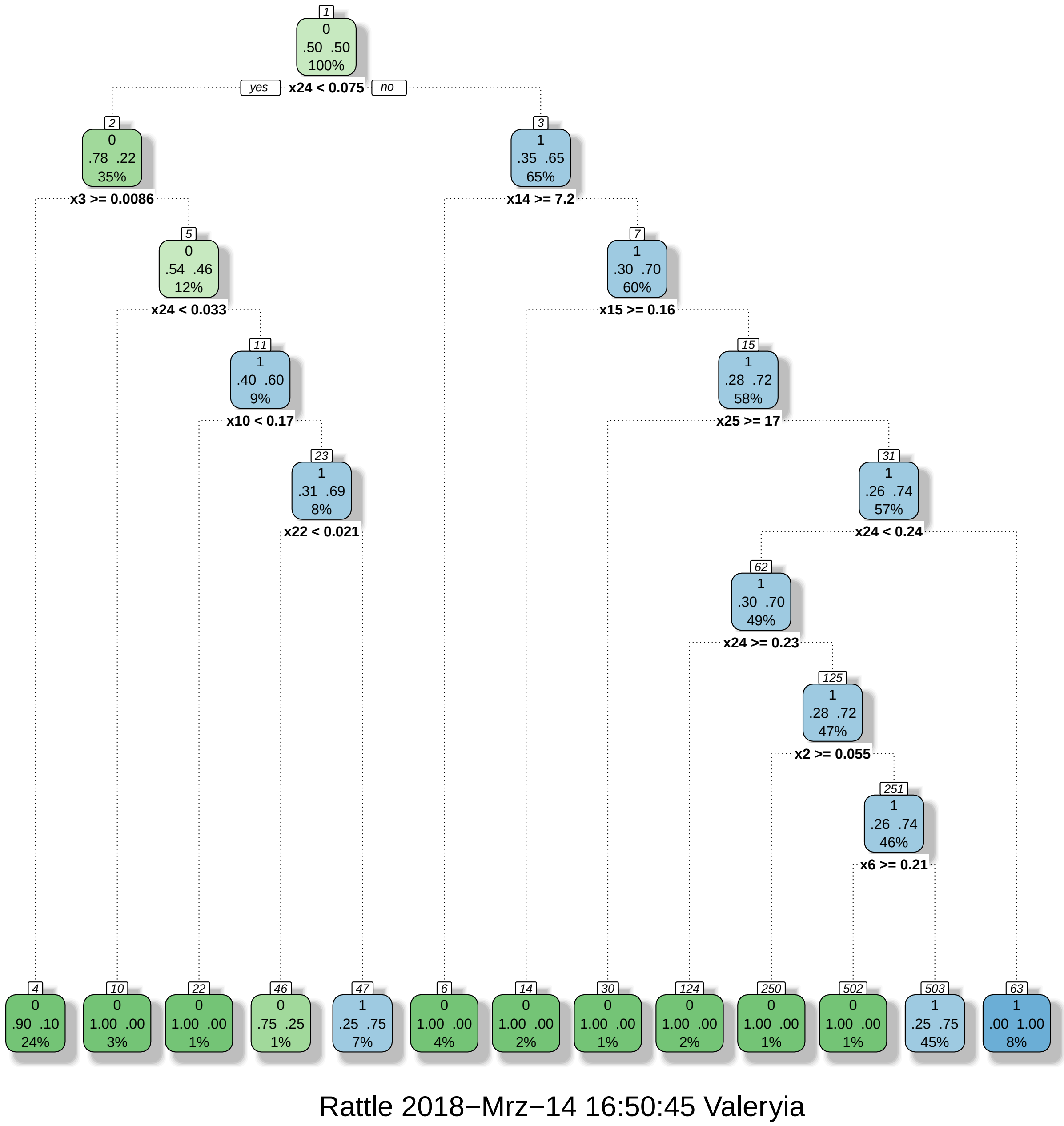


Figure 2: Correlation Matrix

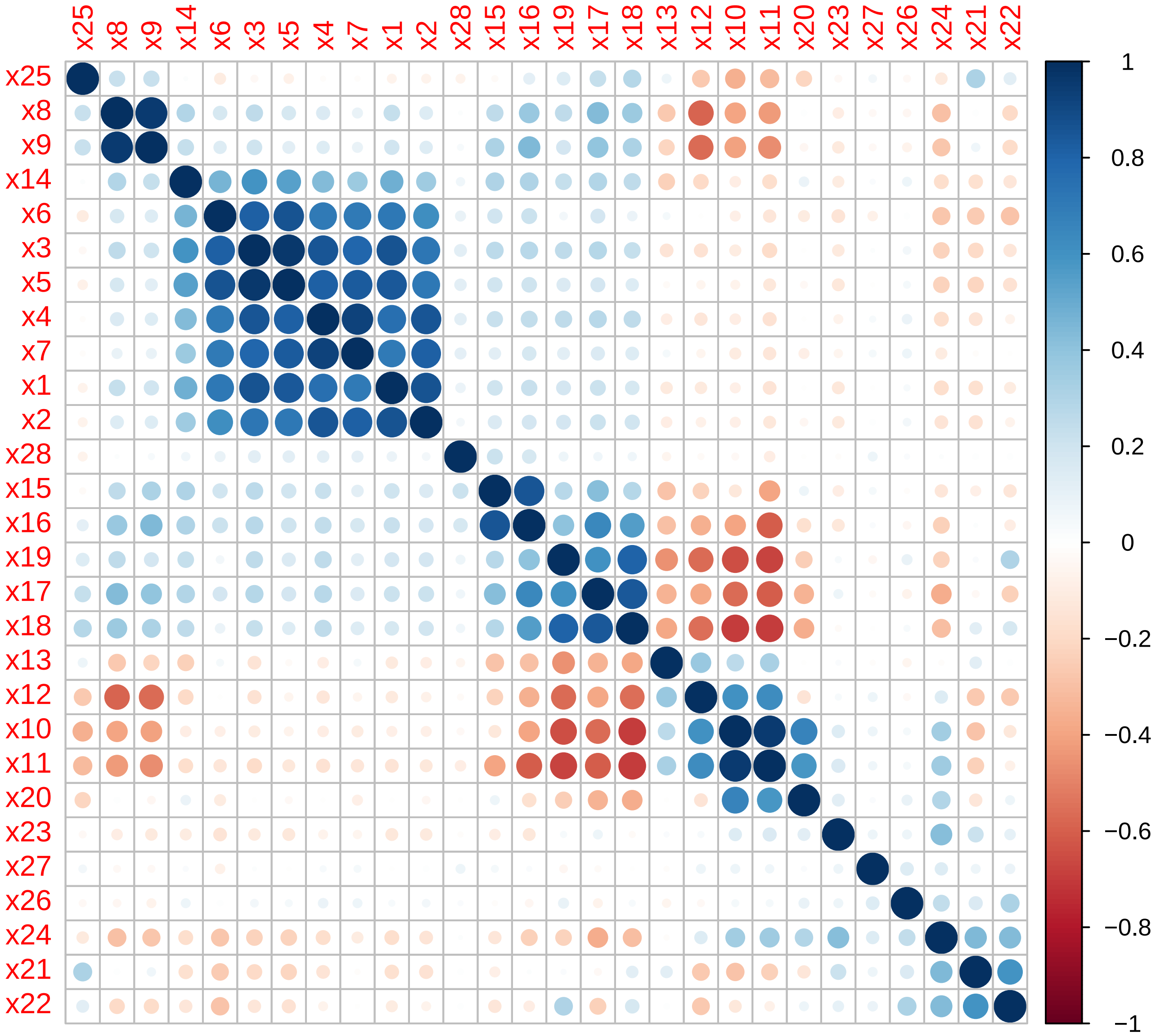
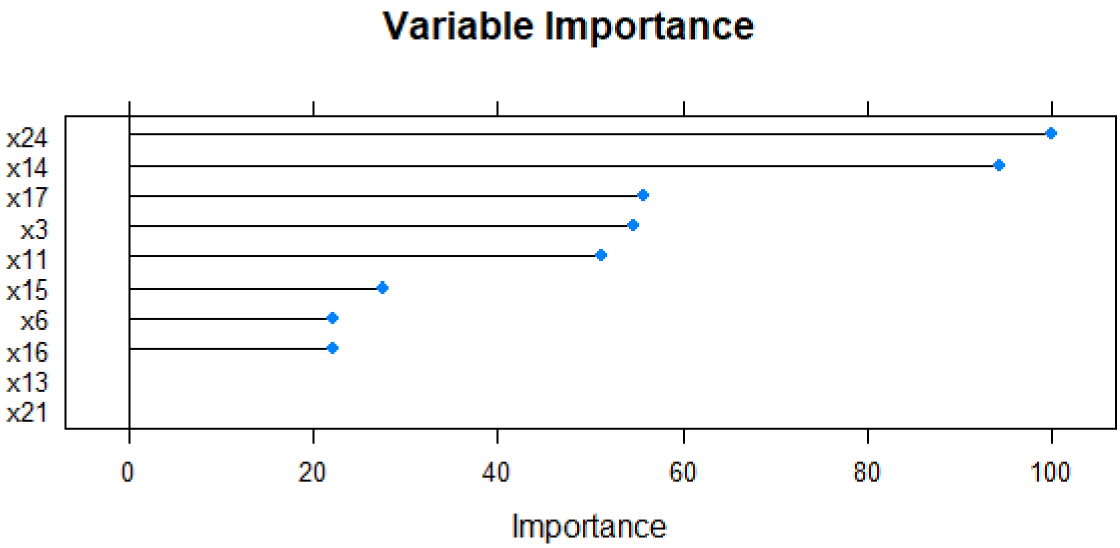


Figure 3: Variable Importance - varImp() function



population can produce trees with different structures and different prediction accuracy. Additionally, a prediction accuracy can be eventually low, compared to other methods. For these reasons a technique called bagging is often used to improve decision tree models. Random forest is a more specialized approach, which directly combines decision trees with bagging and yields markedly better results (Zumel & Mount, 2014).

6 Random Forest

Random forest approach often gives an improved predictive accuracy for relatively complex tree models. This ensemble method combines several individual classification trees in the following way: from the original sample several bootstrap samples are drawn, and an unpruned classification tree is fitted to each bootstrap sample. The variable selection for each split in the classification tree is conducted only from a small random subset of predictor variables. From the complete forest the status of the response variable is predicted as an average or majority vote of the predictions of all trees (Maindonald & Braun, 2010).

The function `randomForest()` in the `randomForest` package is an attractive alternative to rpart() (Breiman, et al., 2015). The following syntax uses randomForest() function applied to our dataset:

```
mod_forest_I=randomForest(as.factor(status)~.,data=training_complete,importance=T
,
                                ntree = 2000, maxnodes= 100, norm.votes = F)
```

We pass several arguments in to the `randomForest()` function. First of all, since it is a classification problem we define our target variable `status` as categorical by using `as.factor()` (Stephens, 2016). In comparison to CART models, trees are grown independently to their maximum extent, limited however by nodesize (minimum number of trees at a node). This parameter implicitly sets the depth of generated trees. Additionally, maxnodes can be used to limit the number of nodes (Maindonald & Braun, 2010).

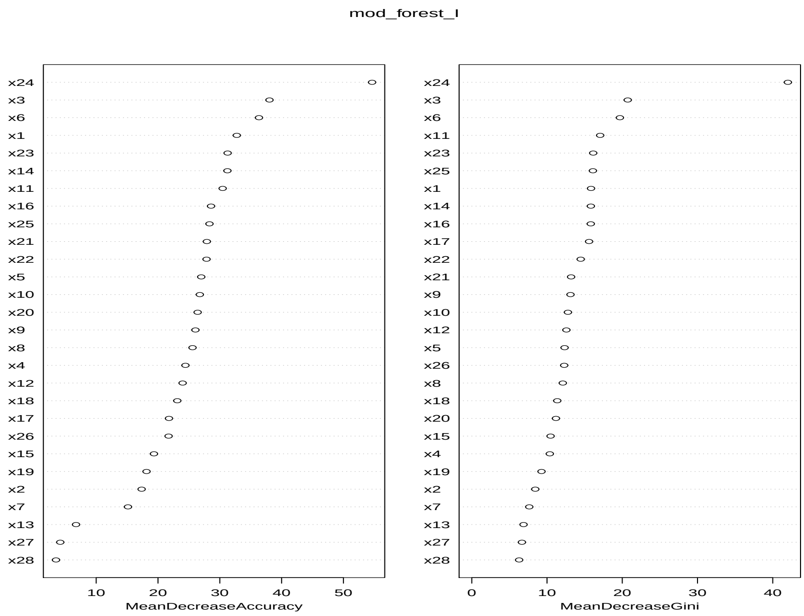
The main tuning parameter is the number `mtry` of variables that are randomly sampled at each split. By default, the `randomForest()` function in R draws square root of the total number of variables at each node for classification trees. Essentially, `mtry` controls the trade-off between the amount of information in each individual tree, and the correlation between trees. Smaller `mtry` values will grow the trees faster. However, in our case using a larger `mtry` is better, since we have a large number of variables to choose from, of which only a small fraction is actually useful. By doing so, we are more likely to draw some useful variables at every step of the tree-growing procedure (Zumel & Mount, 2014).

Further, the `ntree` argument specifies how many trees we want to grow. By tuning this parameter, the final accuracy of a model might increase up to some point. Also, limiting the number of trees restrict the complexity and computing time (Maindonald & Braun, 2010).

Random forests highly increase the prediction accuracy as compared to individual classification trees. However, the interpretability of a random forest is not as straightforward as that of an individual classification tree, where the influence of a predictor variable directly corresponds to its position in the tree. Thus, alternative measures for variable importance are required for the interpretation of random forests (Strobl, et al., 2007).

We can calculate the variable importance by specifying `importance¯TRUE` in the `randomForest()` function, and then calling the functions `importance()` and `varImpPlot()`. Two measures of importance for each variable are then calculated: `MeanDecreaseAccuracy` and `MeanDecreaseGini`. The first measure is a `leave-one-out` type assessment of variable's contribution to prediction accuracy. Thereby the variable's values are randomly permuted in the out-of-bag samples, and the corresponding decrease in each tree's accuracy is estimated. If the average decrease over all the trees is large, then the variable is considered important - its value makes a big difference in predicting the outcome. If the average decrease is small, then the variable does not make much difference to the outcome. The second measure shows how each variable affects the quality of the tree by representing the decrease in node purity that occurs from splitting on a permuted variable (Maindonald & Braun, 2010).

Figure 4: Variable Importance - `importance()`-function



Both measures rate equally the first three and the fifth variables: Accounts Payable Turnover (AP/SALE, x24), Operating Income/Total Assets (OI/TA, x3), EBITDA ((EBITA+AD)/TA, x6) and Account Receivable Turnover Activity (AR/SALE, x23). Interestingly, this ranking order of `importance` differs from the results obtained with `varImp()` function applied for the CART model. Knowing which variables are most important or at least, which variables contribute the most to the structure of the underlying decision trees, can help us with variable reduction. This is useful not only for building smaller, faster trees, but for choosing variables to be used by another modelling algorithm, if that is desired.

The `random forest` package is robust and very user friendly. However, it is important to note that it has a drawback that it is invariably biased towards features with many cut points. There is a package `party` which provides the cforest() command to build a forest of conditional inference trees. The basic construction of each tree is fairly similar to a random forest. Hence the decisions are made in slightly different ways, using a statistical test rather than a purity measure. The `party` package is better than the `randomForest` package in terms of accuracy, since cforest uses the weighted average of the trees to get the final ensemble. However, it is computationally more expensive (Modi, 2016). In our dataset we do not have features which have many categories and therefore we favour the `random forest` package to calculate the insolvency predictions. Additionally, it is computationally more efficient.

7 Linear Discriminant Analysis

7.1 Theory

Linear Discriminant Analysis (LDA) is a technique for dimensionality reduction that encorporates information on class-labels of the different observations. In contrast to Principal Component Analysis, which is a unsupervised dimensionality reduction technique, it finds the rotation that ensures the highest separability between classes. It accomplishes this goal by trying to maximize between class variance while simultaneously minimizing within class variance.

$$max\ J_b(w) = w\prime S_b w, \qquad w \in \mathbb{R}^d, d \in \mathbb{N} \tag{11}$$

$$min\ J_w(w) = w\prime S_w w \tag{12}$$

This is done by maximizing the so called Rayleigh coefficient

$$maxJ = \frac{J_b(w)}{J_w(w)} = \frac{w\prime S_b w}{w\prime S_w w}. \tag{13}$$

The matrices for between and within class variance are defined as

$$S_b = \sum_{c=1}^{C} (\mu_c - \mu)(\mu_c - \mu)\prime \tag{14}$$

$$S_w = \sum_{c=1}^{C} \sum_{i \in c} (x_i - \mu_c)(x_i - \mu_c)\prime \tag{15}$$

where C is the number of classes, μ_c is the vector of sample means for each class respectively and μ is the vector of sample means for the full dataset. For identification purposes we can always chose weights w such that $w\prime S_w w = 1$, since J is constant with regards to rescalings. We can therefore replace w by αw which will result in the constant α canceling out. This way the initial optimization problem can be formulated as

$$\arg\min_{w} -\frac{1}{2} w\prime S_b w \quad s.t. \quad w\prime S_w w = 1 \tag{16}$$

with the lagrangian being

$$\mathcal{L} = -\frac{1}{2} w\prime S_b w + \frac{1}{2}\lambda \left(w\prime S_w w - 1 \right). \tag{17}$$

The halves are added for more convenient matrix derivatives. The Karush-Kuhn-Tucker conditions imply that the solution to this maximization problem and subsequently the vector of weights we want to find needs to fulfill

$$S_b w = \lambda S_w w. \tag{18}$$

This is a generalized eigenvalue problem for which there exists a convenient `R`-solution in the form of the `geigen`-package.

7.2 Implementation

The result from 18, which is essentially the rotation of the underlying data's column space that ensures the highest separability, is implemented in the `lda`-class for the two-class case. The `lda()` function accepts two arguments:

1. `data`: a `data.frame` containing at least one column of factors indicating the class label.
2. `by`: a `character`-string equal to the column's name containing the class labels. Note that all other non-numerical columns will be ignored by the function, since LDA is only meaningful for continuous variables.

The function in a first step extracts the useable columns and the number of classes provided in `data`. It then performs a quick check if the prerequisites are met and then continues with the calculation of the class-means μ_1 and mu_2, the overall mean μ, as well as the scatter-matrices S_b and S_w. With these we can solve the generalized eigenvalue problem from 18.

The first task of the `lda()`-function is to determine, if the prerequisites for further computation are met. To this avail it first asserts that the user provided a dataframe that contains at least one numeric column and a column that contains exactly two distinct class labels (lines 11-17). In a second step it calculates all required variables such as the class means μ_1 and μ_2, the overall mean μ and the covariance matrices for each group (lines 18-19). The class specific metrics are computed by calls to custom made functions defined in the `utils.R`-file of the `LDA`-directory. Essentially these functions are wrappers for subsetting a provided dataframe by a provided key, here this is the class-label, and and an `apply` call looping over the respective subset's columns. The `get_class_means()` calculates the mean inside the `apply`-function whereas `get_class_cov()` uses a call to the built-in `cov`-function. However note that `cov` calculates

$$\hat{cov}(X) = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})\prime.$$

Unfortunately this is not exactly what we want for the scatter matrix defined in 15. We need to multiply the resulting list of covariance matrices elementwise by the number of observations in each class minus one. We correct this before calculating the within class scatter matrix on line 31.

Algorithm 4 `lda()`

procedure SET UP AUXILIARY VARIABLES
  num ← **extract numeric columns**
  classes ← **extract class labels**
  mu ← **calculate class means**
  x_bar ← **calculate overall means**
  S ← **calculate variance per group**
  S_b1 & S_b2 ← **get summands for between class scatter matrix**
procedure CALCULATE SCATTER MATRICES
  S_b ← `S_b1 + S_b2`
  S_w ← `S[[1]] + S[[2]]`
function SOLVE GENERALIZED EIGENVALUE PROBLEM
  V ← **calculate Eigenvectors and Eigenvalues**
  v ← **extract 2 Eigenvectors associated to largest Eigenvalues**
  lda1 & lda2 ← **calculate first two LDA-Components**
  inertia ← **calculate percentage of explained variance**
  return: `lda1, lda2, classes, mu, v, inertia`

```
lda = function(data, by){
  # Closed form solution of LDA:
  # w = S_b^-0.5 * largest_eigenvector(S_b^0.5 * S_w^-1 * S_b^0.5)

  # Check prerequisites
  num = get_numeric_cols(data = data)
  classes = unique(data[, by])

  stopifnot(
    length(num) > 0,
    length(classes) == 2)

  mu = get_class_means(Data = data, By = by, na.rm = TRUE)
  S = get_class_cov(Data = data, By = by, use = "complete.obs")
  n = sapply(classes, function(x){sum(data[, by] == x)})
  print(n)
  # Compute overall mean:
  x_bar = colMeans(data[, num], na.rm = TRUE)

  # Compute between class scatter matrix:
  Sb1 = (mu[, 1] - x_bar) %*% t(mu[, 1] - x_bar)
  Sb2 = (mu[, 2] - x_bar) %*% t(mu[, 2] - x_bar)
  S_b = Sb1 + Sb2

  # Compute within class scatter matrix:
  S_w = (n[1]-1)*S[[1]] + (n[2]-1)*S[[2]]
```

With the `matrix`-object `S_w` storing the within-scatter-matrix S_w and `S_b` storing S_b we can continue by solving the generalized Eigenvalue problem posed in 18. This is done by the `geigen()` function from the `geigen` package.

```
  V = geigen(S_b, S_w, symmetric = TRUE)

  # Extract (absolutely) largest eigenvalue
  ev_order = order(abs(V[["values"]]), decreasing = TRUE)
  v = V[["vectors"]][ev_order[1:2], ]
```

```
41
42  # Percent of variance explained:
43  inertia = (V[["values"]][ev_order[1:2]])^2 / sum(V[["values"]]^2)
```

We extract the eigenvectors and eigenvalues from `V` (line 36) in form of a list. However, caution is required because the eigenvectors are not ordered, which is different to the implementation of eigenvalues in the `base`-function `eigen()`. Therefore we rearrange the eigenvectors according to their absolute eigenvalues (line 39). We extract only the first two eigenvectors, because we require the most informative rotations in order to produce two-dimensional plots. Additionally we compute the percentage of explained variance as

$$intertia_i = \frac{\lambda_i^2}{\sum_{j=1}^{d} \lambda_j^2}. \tag{19}$$

The rotations `lda1` and `lda2` are computed by an `apply`-call (lines 45-46) to an anonymous function which emulates a scalar product by multiplying the two vectors `v[, i]`, which is the i-th eigenvector $i \in \{1, 2\}$, with each observation in the provided dataset. Computationally this is equivalent to the matrix multiplication Xv_i, where X is the data matrix and v_i is the eigenvector.

Finally we gather the results into a `list` and sets it's class to `flda`. This allows us to augment pre-existing functions with a custom method for predicting and plotting for future objects of the`flda`-class in an object-oriented fashion.

7.2.1 Plotting

With a trained model of the `flda`-class it may be of interest to plot the rotations in order to get a visualisable idea of the separability of the two classes. To make this as easy as possible for the end-user we implemented a `plot` method for each instance of a `flda`-class. This means that a user can simply store a call to the `lda`-function inside a variable, called `model` for example. To plot the rotations one can simply invoke `plot` on `model` as easy as with any other model class by typing `plot(model)`.

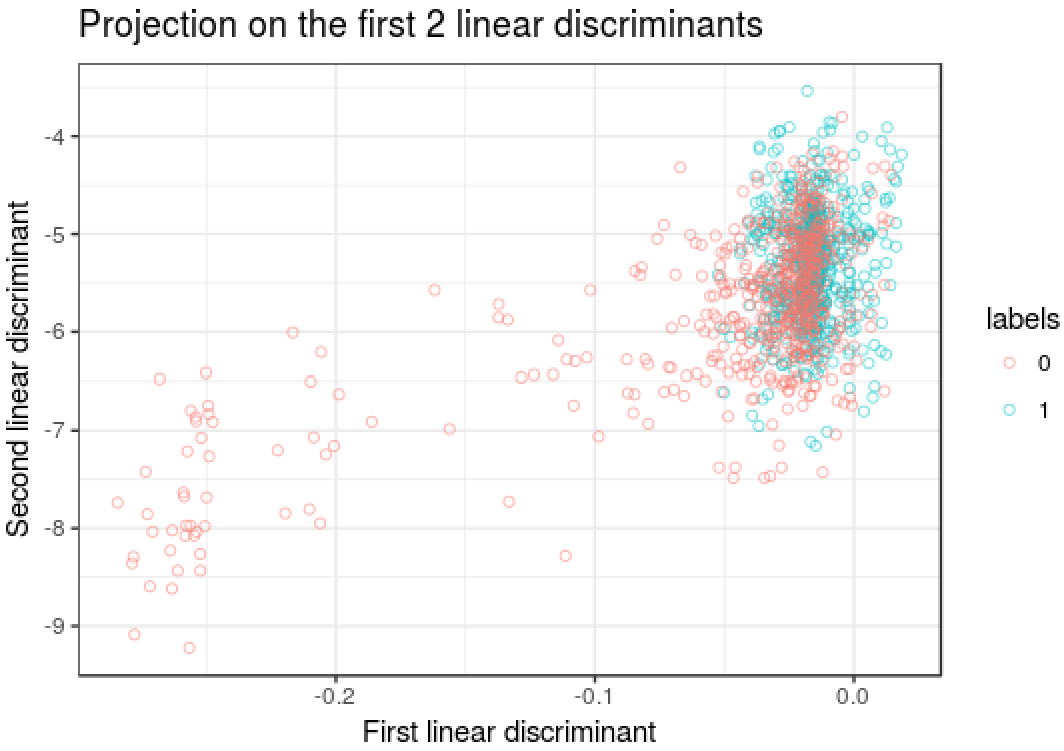
This is done by augmenting `R`'s `plot`-function using the `plot.__className__` syntax. Here we replace `__className__` by `flda`

```
plot.flda = function(model){
  p = ggplot2::ggplot(data = model$X, ggplot2::aes_string(x = "lda1", y = "lda2",
    color = "labels")) +
    ggplot2::geom_point(alpha = 0.5, shape=21) +
    ggplot2::ggtitle("Projection on the first 2 linear discriminants") +
    ggplot2::xlab("First linear discriminant") +
    ggplot2::ylab("Second linear discriminant") +
    ggplot2::theme_bw()
  return(p)
}
```

We overwrite the behavior of `plot` in order to accomodate `flda`-objects. Here we pass the model's precomputed rotations to the `ggplot2` function. The `ggplot2::__function__` makes it explicit that we want the `__function__` from the `ggplot2` package. Another benefit of this syntax is that this way we do not implicitly change a users namespace, which may potentially hide some functions by overwriting them with `ggplot2` routines and therefore may result in unexpected behavior.

Looking at the creditreform dataset we can apply our custom routines by creating a train-set first and then visualizing the results. This reveals already why LDA's predictive results are of poor quality. The feature space spanned by the financial ratios is simply not linearly separable.

Figure 5: Using Linear Discriminant Analysis on the Creditreform Database




```
creditLDA = lda(data = train,
                by = "T2")
plot(creditLDA)
```

7.2.2 Predictions

To make predicting with the `flda`-class as easy as plotting we augment the standard `predict` function in a similar fashion. We expect the user to have already trained an `flda`-model to which she or he wishes to apply new data and obtain predictions on the class labels.

```
predict.flda = function(model, data){
  v = model$scalings[1, ]
  num = get_numeric_cols(data)

  Data = data[, num]
  dims = dim(Data)
  stopifnot(dims[2] == length(v))

  # Compute discriminant as the dot product of every observation
  # with the scaling vector v:
  discr = apply(X = Data, MARGIN = 1, FUN = function(x){sum(x * v)})

  # Compute cutoff threshold:
  c = 0.5 * t(v) %*% (rowSums(model$class_means))
  predictions = ifelse(discr <= as.numeric(c), model$classes[1], model$classes[2])
```

```
86
87     return(predictions)
88 }
```

The provided new data should have a similar structure to the training data. This means that especially the order of the columns should be the same. We check for obviously non-conforming data in line 77 but we do not perform explicit tests on the order of columns. If the data has the required shape we calculate the discriminant function for each observation, which is the scalar product of the observation's data vector and the first eigenvector from the trained model. We then proceed calculating a threshold c for being able to discriminate between the two groups by following the rule

$$c = \frac{1}{2} v\prime \left(\mu_1 + \mu_2 \right) \tag{20}$$

predictions are made according to the threshold, if the value of the discriminant function is smaller than the threshold we assign the label of the first class and the second class otherwise.

8 Evaluation of predictions

In this section we explain the steps followed in the "evaluate_predictions" quantlet. Purpose of this quantlet is to create a function that will take labels (actual solvency status) and predictions about the solvency status of firms and will return the confusion matrix, the ROC curve and the AUC as well as some evaluation metrics like sensitivity, specificity, precision and accuracy.

The confusion matrix, also known as error matrix, is a matrix that illustrates the performance of a classification model on a set of test data with known true values (labels). Each row of the matrix corresponds to a case of a predicted class while each column corresponds to a case of an actual class. It is called confusion matrix because it depicts whether the classifier is confusing the two classes, i.e. if it is wrongly labelling one class as another. It is a special case of contingency table, with two dimensions ("actual" and "predicted") and two "classes" (solvent and insolvent in our case) in each dimension. The cells of our confusion matrix will present the number of:

1. True Positives (TP): firms who are correctly predicted to be insolvent (hits).
2. False Positives (FP): solvent firms were wrongly predicted to be insolvent (false alarm or Type I error)
3. False Negatives (FN): insolvent firms were wrongly predicted to be solvent (mass or Type II error)
4. True Negatives (TN): firms were correctly predicted to be solvent (correct rejection)

These values will then be used to get some important metrics which are described next.

Sensitivity and specificity are assessment metrics of the discriminative power of classification methods (Härdle et al., 2012). Sensitivity (also known as the true positive rate) measures the proportion of positives (defaulted firms in our case) that are correctly identified as such (Fawcett, 2006). It is defined as

$$TPR = \frac{TP}{TP + FN}.$$

Specificity (or true negative rate) measures the proportion of negatives (solvent firms in our case) that are correctly identified as such (Fawcett, 2006). It is defined as

$$TNR = \frac{TN}{TN + FP}.$$

Precision is analogous to the positive predictive value (PPV) and is a measure of exactness (Härdle et al., 2012). It is defined as

$$PPV = \frac{TP}{TP + FP}.$$

Finally accuracy measures the fraction of correct predictions and is defined as

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

The values of specificity and sensitivity will allow us to plot the ROC curve. The ROC curve (Receiver Operating Characteristic curve) is a graphical plot which depicts the diagnostic ability of a binary classification model for different values of its discrimination threshold. In our case, this threshold refers to the probability value above of which a firm is predicted to be insolvent. It is created by plotting the sensitivity values(y axis) against their corresponding $1 - specifity$ values (x axis) as the threshold varies (Fawcett, 2006). The area under the ROC curve (AUC) can be interpreted as the average power of the test on default or non-default corresponding to all discrimination thresholds (Härdle et al.(2012)). A larger AUC corresponds to a better classification result. A model with perfect discriminative power will have an AUC value of 1, while a random model without discriminative power will have an AUC value of 0.5, i.e. its ROC curve will be the 45 degree line. Thus, any reasonable rating model is expected to have an AUC value above 0.5, while AUC values close to 1 will indicate models with high diagnostic ability.

The first function we created in this quantlet is the get_prediction() function (lines $19 - 22$). This function takes as inputs the fitted probabilities for insolvency and the threshold above of which a firm would be predicted to be defaulted. Using the ifelse() function in the body of this function, get_prediction() will return predictions for the solvency status of each firm. We construct then the evaluate_predictions() function (lines $24 - 47$). Inputs of this function are the actual solvency status of each company (labels), the corresponding predictions about the status and the "verbose" parameter which is equal to FALSE by default. In the body of the function we create first the previously mentioned confusion matrix which we expect to be a 2x2 matrix. If it is not the case, the test in line 30 will show us a warning message. Then we get the values of TP, TN, FP and FN and we compute the values of sensitivity, specificity, precision and accuracy, which are then saved as a list in "reports". The function will also print a data frame with those reports if we change the logical value of "verbose" to TRUE. The two previously mentioned functions are used in the body of the last function which completes the task of the quantlet.

The final function created in this quantlet is the evaluate_model() function (lines $49 - 101$), which takes the fitted probabilities for bankruptcy and the actual status (labels) as inputs. At first we create a threshold list in the body of the function with threshold values varying from 0 to 1 by tiny steps. Then we apply the get_prediction() function to this list in order to get a list of predictions for each threshold value which we store as "pred_list". Afterwards we apply the evaluate_predictions() function to pred_list and we get a list of reports. We use then the values of sensitivity and specificity from reports in order to compute the ROC curve and the area under it (AUC) (lines $58 - 70$). For calculating the AUC we create the function get_auc() which takes the values of $1 - specificities$ and sensitivities as inputs. This function uses the trapezoidal method to approximate the AUC (solution found in stackoverflow.com). For each difference in $1 - specificities$ there is one rectangle which underestimates the area under curve (corresponding to the "left" value of sensitivities) and one rectangle that overestimates it (corresponding to the "right" value of sensitivities). Therefore the function approximates the AUC by taking the average of the two rectangles.

We are next interested in finding the values of sensitivities and specificities for which the distance between the ROC curve and the 45-degree line (i.e. the ROC curve of a model with no discriminative power) is maximized. We use the index of these values in order to find the optimal values of sensitivity, specificity and the threshold, which is then used to compute the optimal predictions. These predictions are then used to create the optimal confusion matrix and get the optimal accuracy (lines 73 – 84). In the final part of the evaluate_model() function the ROC curve is plotted and the function returns a list with the optimal measures calculated above.

9 Appendix

9.1 Unit Tests

9.1.1 LDA

In order to test the `lda`-function we take a dataset from which we know that it is linearly separable already and see if we can reproduce the results. Since Linear Discriminant Analysis was first proposed by Sir Ronald A. Fisher it is only fitting that we test it on his famous `iris` dataset. The `iris` dataset contains measurements for the sepal and petal length and width of three different species of iris flowers. We know that the sepal length and the petal length are sufficient variables to create a obviously visible separation line in figure 9.1.1 which discriminates the Setosa-species almost perfectly from the rest.

Should the `lda`-function work as intended we would expect this result only to become better, i.e. the points being perfectly separable. We first import the `iris` dataset and transform the Species column in order to fit our binary task, that is we create a dichotomous variable indicating if the particular flower is of the Setosa species or not. We can then directly use the `lda`-function.

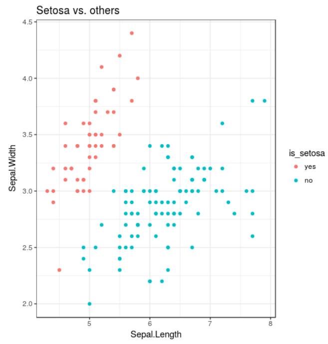


Figure 6: The Setosa-Species is already almost separable, even in the normal space.

```
irisLDA = lda(iris_test, "is_setosa")
plot(irisLDA)
```

9.1.2 Binary Response Models

9.1.3 Evaluation

9.2 Tables

Table 5: Variables of the Creditreform database

Column	Value
ID	ID of each company
T2	Indicator of solvency status (solvent=0, insolvent=1)
Jahr	Year
VAR1	Cash and cash equivalents
VAR2	Inventories
VAR3	Current assets
VAR4	Tangible assets
VAR5	Intangible assets
VAR6	Total assets
VAR7	Accounts receivable
VAR29	Accounts receivable against affiliated companies
VAR8	Lands and buildings
VAR9	Equity (own funds)
VAR10	Shareholder loan
VAR11	Accrual for pension liabilities
VAR12	Total current liabilities
VAR13	Total longterm liabilities
VAR14	Bank debt
VAR15	Accounts payable
VAR30	Accounts payable against affiliated companies
VAR16	Sales
VAR17	Administrative expenses
VAR18	Amortization and depreciation
VAR19	Interest expenses
VAR20	Earnings before interest and taxes (EBIT)
VAR21	Operating income
VAR22	Net income
VAR23	Increase (decrease) in inventories
VAR24	Increase (decrease) in liabilities
VAR25	Increase (decrease) in cash
VAR26	Industry classification code
VAR27	Legal form
VAR28	Number of employees
Rechtskreis	Accounting principle
Abschlussart	Type of account

References

[1] Winkelmann, R., Boes, S. (2009): *"Analysis of Microdata"*, 2nd edition.

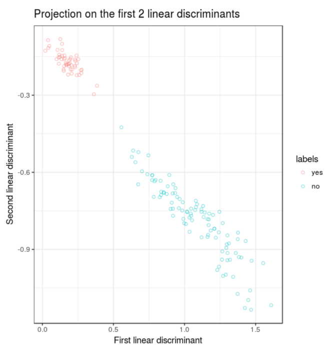


Figure 7: We already see that the first LDA-component is enough for a good classification.

[2] Berk, J., DeMarzo, P. (2016), Corporate Finance, 4th edition, Harlow, Pearson Education

[3] Chen, S., Härdle, W.K., Moro, A.R (2011), Modeling default risk with support vector machines, Quantitative Finance, 11(1), 135-154

[4] Fawcett, T. (2006), An introduction to ROC analysis, Pattern Recognition Letters 27, 861-874

[5] Härdle, W.K., Prastyo, D.D., Hafner, C. (2012), Support vector machines with evolutionary model selection for default prediction, SFB 649 Discussion Paper 2012-030

[6] Miao, H., Ramchander, S., Ryan, P., Wang, T. (2018), Default prediction models: The role of forward-looking measures of returns and volatility, Journal of Empirical Finance, 46, 146-162

[7] Zhang, J.L., Härdle, W.K. (2010). The Bayesian Additive Classification Tree Applied to Credit Risk Modelling. Computational Statistics and Data Analysis, 54: 1197-1205