# Linear Classification Methods

## Determining the solidness of borrowers via credit scoring

Thomas Bosch

02.12.2008

# ABSTRACT

The task of the author of this research paper was to select a specific linear classification method to create a rule to classify new customers into two mutually exclusive clusters. With this model, you can predict, if a customer of a bank will pay back a loan or not. In order to create such a rule, the author chose the linear discriminant analysis. The model consists of the formula of the discriminant function. To predict the credibility of a new customer, you have to insert the values of the variables in this discriminant function, and you will get the resulting classification. The quality of this assumed group assignment will be 76.50 percent.

# 1. Introduction

*Credit Scoring*

In credit business, banks are interested in information whether prospective consumers will pay back their credit or not. The aim of credit scoring is to model or predict the probability that a consumer with certain variables is to be considered as a potential risk.

*Purpose of the Research Paper*

The author of this research paper has to choose an appropriate linear classification method to create a rule to predict if a customer can pay back the loan or not and to select adequate variables for this classification. Another task of the research paper writer is to determine how much error is made applying this rule.

*Additional Information*

You can get a detailed description of the purpose of credit scoring, the dataset and the variables, and you also have the possibility to download the dataset by following the link http://www.stat.uni-muenchen.de/service/datenarchiv/kredit/kredit_e.html.

# 2. Description of the Data

First of all, the individuals of the dataset will be specified. Thereafter, the variables of the dataset will be described in detail.

## 2.1. Description of the Individuals

The dataset consists of 1000 consumer credits from a german bank. For each consumer the binary response variable "kredit" is available.

## 2.2. Description of the Variables

20 variables that are assumed to influence creditability were recorded. Overall, the description of the variable set includes 21 variables.

*Variable “kredit”*

The first variable, called “kredit” specifies the credibility of a customer and represents the assignment of a specific customer of a german bank to either group 1 or group two. If the value of this variable is 1, the customer is credit-worthy and if the value is 0, the customer is not credit-worthy.

*Detailed Description of the Variables*

The descriptions of all the 21 variables can be regarded in the appendix of this research paper.

The given score for the categorical variables is based on the assessment of experienced bank specialists dealing with credits.

## 3. Linear Discriminant Analysis

In order to solve the specific problem of this research paper, the linear discriminant analysis will be used. In the first subchapter, there the reader of this paper will get a description of this approach. After that, it will be explained, why the statistic software R will be utilized. Then you will read about the abstract methodology of the creation of the discriminant function, and when you know the abstract procedure, the concrete approach will be shown.

### 3.1. Description of the Linear Discriminant Analysis

The linear discriminant analysis is used to examine the difference between the two groups, representing the credibility of a customer, by concerning the variables of the dataset. It is assumed that the variables determine the group of a customer.

### 3.2. Statistic Software R

To solve the problem of this research paper, the linear discriminant analysis will be adopted. Since the application of the linear discriminant analysis is computationally extensive, the statistic software R will be used.

### 3.3. Methodology

*Data Import to R*

First, the writer of this paper imported the basic dataset into the statistic software R to solve the various problems.

*Discriminant Function*

In order to determine the discriminant function, you have to define the diverse parts of the formula of the discriminant function. You have to select the variables, which will be used to classify a specific customer. You also have to identify the coefficients of the discriminant function and the constant element.

*Additional Information*

The methodology can be read on the website of faes.de (2008).

### 3.4. Data Import to R

First, you have to import the data into the statistic program R. The dataset is given in an ASCII format. The following R statement will store the ASCII data in the variable called “data”.

```
data<-read.table("D:\\Studium\\Erasmus\\Courses\\Multivariate Techniques\\project 2\\kredit.asc",header=T)
```

The first parameter is assigned to the absolute path of the ASCII file. The second argument shows R that the first line in the data file contains the names of the variables.

### 3.5. Assumption of the Discriminant Function

On the basis of the groups, a discriminant function will be estimated by executing the linear discriminant analysis.

*Formula of Discriminant Function*

The form of this discriminant function is: $y = b_0 + b_1x_1 + b_2x_2 + \dots + b_jx_j$. y is the discriminant variable, $x_j$ is the variable with the index j, $b_j$ is the discriminant coefficient and $b_0$ is the constant element. The constant element $b_0$ will be selected in such way that the critical discriminant value will be zero. The critical discriminant value is that value, which will be considered to assign a new customer to one of the two groups.

*R Function lda*

The function “lda” of the package “MASS” will be used for the assumption of the discriminant function.

```
library(MASS)
lda(data[2:21], data[,1])
```

After loading the package “MASS”, while executing the first command in R, the function “lda” can be used. The function “lda” expects a dataset, containing both the groups and the variables. The first argument must be allocated with the variables and the second argument with the groups.

*“Prior probabilities of groups”*

To show the methodology, the author of this research paper regards in this first step all the possible variables. The outcome of the application of this R command “Prior probabilities of groups” shows that the grouping in the two groups 1 and 0 happened to 30 percent to the group 0 and to 70 percent to the group 1. This means that if you use this discriminant function, 700 of the 1000 customers will be allocated to group 1 and 300 to group 0. 700 customers will be credit-worthy and 300 will not be credit-worthy.

*“Group means”*

The group means of the variables are also presented on the screen.

*“Coefficients of linear discriminants”*

The assumed discriminant function can be formed, using the results in the shown section “Coefficients of linear discriminants”. These are the $b_j$ values from the formula of the discriminant function.

### 3.6. Determining the Quality of the Discriminant Function

Ahead, the author of this writing paper defined the discriminant function. Now, the next step is to detect the quality of the before defined discriminant function by utilizing the same function with the appended parameter “CV” assigned to “TRUE”.

```
lda(data[2:21], data[,1], CV = TRUE)
```

*“$posterior”*

The outcome of the previous R statement “$posterior” shows the assumption of the probable belonging to the group. For example, the value “0.57” means that this customer belongs with a likelihood of 57 percent to this group.

*“$class”*

The second result of the R command “$class” displays the assignment of a customer to a group according to the percentage values.

### 3.7. Selection of the variables

The main aim of this research paper is to determine that discriminant function, which maximizes the congruence between the group assignments, which can be extracted from the variable “kredit” of the basic dataset, and the assumed group assignments, which result after the application of the discriminant function on the same dataset.

The writer of this paper executed the following R commands. The second one was repeated with changing first formal parameter.

```
write.table(data$kredit, file = "excel import.csv")

write.table((lda(data[2:21], data[,1], CV = TRUE))$class, file = "excel import.csv")
```

The resulting vectors were exported to an excel spreadsheet. In this spreadsheet, the vector, containing the values of the “kredit” variable of the dataset was compared to the vector, which includes the assumed group assignments, which are the outcomes of the execution of the discriminant function. The code listing of the realization of this comparison, written in the programming language Visual Basic for Applications can be read in the appropriate section of the appendix. As you can see, the first argument of the prior second R statement determines the variables, which are used to create the formula of the discriminant function. This formula was changed step by step, in order to decrease the number and percentage of errors, which occurred at the comparison of the group allocations of the basic dataset in the variable “kredit” and the assumed group assignments. If you decrease the number and percentage of errors, you increase the percent congruence between these two groupings, at the same time. If the congruence is the maximum with the appropriate variables, you know that the quality of the discriminant function is very high. If you predict, if a new customers can pay back her or his loan, it will be more likely that this prediction will be verified.

The procedure is to determine a variable of the overall variable set of the dataset, which can be excluded from the construction of the discriminant function formula to decrement the number of errors and simultaneously to enhance the percentage of congruence between the group assignments by the variable “kredit” and the assumed one. This step was repeated until no improvement of the congruence could be detected. After that the final formula of the discriminant function could be determined and used for the predictions of new customers.

### 3.7.1. 1. Step: all Variables

Table 1

*Quality of the Discriminant Function – all Variables*

| Quality Measures | Value |
|---|---|
| Errors | 239 |
| Errors [%] | 23.90 |
| Congruence [%] | 76.10 |

The quality of the discriminant function is very high, if all the variables are considered in the process of creating the formula of the discriminant function, because the congruence in percent is 76.10. This means that 761 of the 1000 group assignments were assumed correctly by the discriminant function.

### 3.7.2. 2. Step: all Variables except “telef”

*Description of the variable “telef”*

The variable “telef” says, if the customer has a telephone connection. This seems to be an information without any effect on the model to evaluate the credibility of a customer.

*R statements*

The next R statements forms a discriminant function without the variable “telef” and exports the resulting groupings vector to an excel spreadsheet. In this spreadsheet, the comparison between the group assignments in the dataset and the group allocations, assumed by the discriminant function, will be done.

```
dataQualityTest <- cbind(data[1:19], data[21])
write.table((lda(dataQualityTest[2:20], dataQualityTest[,1], CV = TRUE))$class, file =
"excel import.csv")
```

In the next steps, there will not be any code listings, because they will change just in detail. The interested reader has the possibility to consult the appropriate section in the appendix.

*Conclusion*

Table 2

*Quality of the Discriminant Function – all Variables except “telef”*

| Quality Measures | Value |
|---|---|
| Errors | 239 |
| Errors [%] | 23.90 |
| Congruence [%] | 76.10 |

As you can see in table 2, the value of the congruence did not change. You can conclude that the variable “telef” has no impact on the determining, if a specific customer is credit-worthy or not.

### 3.7.3. 3. Step: all Variables except “telef” and “verw”

*Description of the variable “verw”*

The variable “verf” indicates the purpose of the credit. The author guesses that this variable does not affect the grouping decision.

*Conclusion*

Table 3

*Quality of the Discriminant Function – all Variables except “telef” and “verw”*

| Quality Measures | Value |
|---|---|
| Errors | 241 |
| Errors [%] | 24.10 |
| Congruence [%] | 75.90 |

The number and the percentage of errors increased. Two more customers were grouped not correctly. Accordingly, the congruence is less. In the next steps the variable “verw” will be included in the formula of the discriminant function.

### 3.7.4. 4. Step: all Variables except “telef” and “wohnzeit”

*Description of the variable “wohnzeit”*

The variable “wohnzeit” shows how long the customer lived in the current household. This seems not to be significant for the grouping decision.

*Conclusion*

Table 4

*Quality of the Discriminant Function – all Variables except "telef" and "wohnzeit"*

| Quality Measures | Value |
|---|---|
| Errors | 238 |
| Errors [%] | 23.80 |
| Congruence [%] | 76.20 |

The suspicion confirmed that the variable "wohnzeit" has no impact on the classification. In fact, the congruence increased and the error rate is reduced. In the next steps, this variable will not be included in the creation of the discriminant function.

### 3.7.5. 5. Step: all Variables except "telef", "wohnzeit" and "pers"

*Description of the variable "pers"*

"pers" signalized the number of persons entitled to maintenance. You will see, if the group allocations will be more exactly.

*Conclusion*

Table 5

*Quality of the Discriminant Function – all Variables except "telef", "wohnzeit" and "pers"*

| Quality Measures | Value |
|---|---|
| Errors | 238 |
| Errors [%] | 23.80 |
| Congruence [%] | 76.20 |

The results of the variable "kredit" in the original dataset and the assumptions of the application of the discriminant function are to 76.20 percent the same. The absence of the variable "pers" did not influence the finding. The author of this research paper did not consider this variable anymore in the process of generating the dicriminant function formula.

### 3.7.6. 6. Step: all Variables except “telef”, “wohnzeit”, “pers” and “gastarb”

*Description of the variable “gastarb”*

This variable indicates, if the customer of the german bank is a foreign worker.

*Conclusion*

Table 6

*Quality of the Discriminant Function – all Variables except “telef”, “wohnzeit”, “pers” and “gastarb”*

| Quality Measures | Value |
|---|---|
| Errors | 241 |
| Errors [%] | 24.10 |
| Congruence [%] | 75.90 |

The error rate increased and because of this, this variable will not be regarded in the next steps.

### 3.7.7. 7. Step: all Variables except “telef”, “wohnzeit”, “pers” and “wohn”

*Description of the variable “wohn”*

The variable “wohn” gives the type of the apartment, in which the customer lives.

*Conclusion*

Table 7

*Quality of the Discriminant Function – all Variables except “telef”, “wohnzeit”, “pers” and “wohn”*

| Quality Measures | Value |
|---|---|
| Errors | 239 |
| Errors [%] | 23.90 |
| Congruence [%] | 76.10 |

The congruence will also be less. Therefore, the variable “wohn” will stay part of the formula of the discriminant function.

### 3.7.8. 8. Step: all Variables except “telef”, “wohnzeit”, “pers” and “laufkont”

*Description of the variable “laufkont”*

“laufkont” stands for the balance of the current amount.

*Conclusion*

Table 8

*Quality of the Discriminant Function – all Variables except “telef”, “wohnzeit”, “pers” and “laufkont”*

| Quality Measures | Value |
|---|---|
| Errors | 257 |
| Errors [%] | 25.70 |
| Congruence [%] | 74.30 |

The congruence decrease with 1.9 percent from the previous maximum value. This has to be an important variable to evaluate the cluster of a specific customer of the german bank. This variable must be a part of the formula of the discriminant function.

### 3.7.9. 9. Step: all Variables except “telef”, “wohnzeit”, “pers” and “famges”

*Description of the variable “famges”*

The variable “famges” indicated both the marital status and the sex.

*Conclusion*

Table 9

*Quality of the Discriminant Function – all Variables except "telef", "wohnzeit", "pers" and "famges"*

| Quality Measures | Value |
|---|---|
| Errors | 240 |
| Errors [%] | 24.00 |
| Congruence [%] | 76.00 |

The congruence is less than 76.20. Because of this, the variable must be part of the formula of the discriminant function.

### 3.7.10. 10. Step: all Variables except "telef", "wohnzeit", "pers" and "alter"

*Description of the variable "alter"*

"alter" stands for the age in years.

*Conclusion*

Table 10

*Quality of the Discriminant Function – all Variables except "telef", "wohnzeit", "pers" and "alter"*

| Quality Measures | Value |
|---|---|
| Errors | 237 |
| Errors [%] | 23.70 |
| Congruence [%] | 76.30 |

The percentage value of the errors decreased with 0.1 percent. This variable will be excluded in the next steps from the generating of the discriminant function.

### 3.7.11. 11. Step: all Variables except “telef”, “wohnzeit”, “pers”, “alter” and “verm”

*Description of the variable “verm”*

In the variable “verm” for every customer of the german bank is stored the most valuable available assets.

*Conclusion*

Table 11

*Quality of the Discriminant Function – all Variables except “telef”, “wohnzeit”, “pers”, “alter” and “verm”*

| Quality Measures | Value |
|---|---|
| Errors | 246 |
| Errors [%] | 24.60 |
| Congruence [%] | 75.40 |

The congruence did not get better. Even the opposite happened. This variable has to be considered, if you want to cluster new customers according to the credibility.

### 3.7.12. 12. Step: all Variables except “telef”, “wohnzeit”, “pers”, “alter” and “laufzeit”

*Description of the variable “laufzeit”*

The duration in month is meant with “laufzeit”.

*Conclusion*

Table 12

*Quality of the Discriminant Function – all Variables except “telef”, “wohnzeit”, “pers”, “alter” and “laufzeit”*

| Quality Measures | Value |
|---|---|
| Errors | 247 |
| Errors [%] | 24.70 |
| Congruence [%] | 75.30 |

The congruence is less than 76.30. If you delete this variable from the formula of the discriminant function, the congruence level will be decreased.

### 3.7.13. 13. Step: all Variables except “telef”, “wohnzeit”, “pers”, “alter” and “moral”

*Description of the variable “moral”*

How can be described the payment of previous credits?

*Conclusion*

Table 13

*Quality of the Discriminant Function – all Variables except “telef”, “wohnzeit”, “pers”, “alter” and “moral”*

| Quality Measures | Value |
|---|---|
| Errors | 239 |
| Errors [%] | 23.90 |
| Congruence [%] | 76.10 |

The variable moral must be evaluated to get the assumed group assignment.

### 3.7.14. 14. Step: all Variables except “telef”, “wohnzeit”, “pers”, “alter” and “hoehe”

*Description of the variable “hoehe”*

The variable “hoehe” shows the reader the amount of credit in “Deutsche Mark”.

*Conclusion*

Table 14

*Quality of the Discriminant Function – all Variables except "telef", "wohnzeit", "pers", "alter" and "hoehe"*

| Quality Measures | Value |
|---|---|
| Errors | 236 |
| Errors [%] | 23.60 |
| Congruence [%] | 76.40 |

The number of errors decreased and so the percent congruence increased by 0.1, if the variable "hoehe" will not be part of the discriminant function formula.

### 3.7.15. 15. Step: all Variables except "telef", "wohnzeit", "pers", "alter", "hoehe" and "sparkont"

*Description of the variable "sparkont"*

"sparkont" determines the value of savings or stocks.

*Conclusion*

Table 15

*Quality of the Discriminant Function – all Variables except "telef", "wohnzeit", "pers", "alter", "hoehe" and "sparkont"*

| Quality Measures | Value |
|---|---|
| Errors | 245 |
| Errors [%] | 24.50 |
| Congruence [%] | 75.50 |

The congruence between the real grouping decisions and the assumed ones did not improve.

### 3.7.16. 16. Step: all Variables except “telef”, “wohnzeit”, “pers”, “alter”, “hoehe” and “beszeit”

*Description of the variable “beszeit”*

The “bestzeit” indicates the amount of years, the customer has bee employed by the current employer.

*Conclusion*

Table 16

*Quality of the Discriminant Function – all Variables except “telef”, “wohnzeit”, “pers”, “alter”, “hoehe” and “beszeit”*

| Quality Measures | Value |
|---|---|
| Errors | 239 |
| Errors [%] | 23.90 |
| Congruence [%] | 76.10 |

The formula of the discriminant function stays the same.

### 3.7.17. 17. Step: all Variables except “telef”, “wohnzeit”, “pers”, “alter”, “hoehe” and “rate”

*Description of the variable “rate”*

With the variable called “rate” you can get the installment in percent of the available income.

*Conclusion*

Table 17

*Quality of the Discriminant Function – all Variables except "telef", "wohnzeit", "pers", "alter", "hoehe" and "rate"*

| Quality Measures | Value |
|---|---|
| Errors | 240 |
| Errors [%] | 24.00 |
| Congruence [%] | 76.00 |

The variable "rate" is also important for getting the maximum of correct clustering results.

### 3.7.18. 18. Step: all Variables except "telef", "wohnzeit", "pers", "alter", "hoehe" and "buerge"

*Description of the variable "buerge"*

Are there further debtors, guarantors?

*Conclusion*

Table 18

*Quality of the Discriminant Function – all Variables except "telef", "wohnzeit", "pers", "alter , "hoehe" and "buerge"*

| Quality Measures | Value |
|---|---|
| Errors | 241 |
| Errors [%] | 24.10 |
| Congruence [%] | 75.90 |

You can not exclude this variable.

### 3.7.19. 19. Step: all Variables except "telef", "wohnzeit", "pers", "alter", "hoehe" and "weitkred"

*Description of the variable "weitkred"*

Are there further running credits?

*Conclusion*

Table 19

*Quality of the Discriminant Function – all Variables except "telef", "wohnzeit", "pers", "alter", "hoehe" and "weitkred"*

| Quality Measures | Value |
|---|---|
| Errors | 235 |
| Errors [%] | 23.50 |
| Congruence [%] | 76.50 |

The congruence improved. Because of this, the variable "weitkred" will not be part of the discriminant function anymore. It does not matter, if there are further running credits.

### 3.7.20. 20. Step: all Variables except "telef", "wohnzeit", "pers", "alter", "hoehe", weitkred" and "bishkred"

*Description of the variable "bishkred"*

The variable "biskred" indicates the number of previous credits at this bank including the running one.

*Conclusion*

Table 20

*Quality of the Discriminant Function – all Variables except “telef”, “wohnzeit”, “pers”, “alter”, “hoehe”, “weitkred” and “bishkred”*

| Quality Measures | Value |
|---|---|
| Errors | 236 |
| Errors [%] | 23.60 |
| Congruence [%] | 76.40 |

The congruence changed to the worse by 0.1 percent.

### 3.7.21. 21. Step: all Variables except “telef”, “wohnzeit”, “pers”, “alter”, “hoehe”, weitkred” and “beruf”

*Description of the variable “beruf”*

What kind of profession does the customer has?

*Conclusion*

Table 21

*Quality of the Discriminant Function – all Variables except “telef”, “wohnzeit”, “pers”, “alter”, “hoehe”, “weitkred” and “beruf”*

| Quality Measures | Value |
|---|---|
| Errors | 235 |
| Errors [%] | 23.50 |
| Congruence [%] | 76.50 |

The congruence is the same than the best, so this variable is not significant to group a customer.

The formula of the discriminant function is complete.

## 3.8. Formula of the Discriminant Function

*Abstract Form of the Discriminant Function*

The abstract form of the discriminant function is $y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_jx_j$. y is the discriminant variable, $x_j$ is the variable with the index j, $b_j$ is the discriminant coefficient and $b_0$ is the constant element. In order to get the formula of the discriminant function, you will have to determine each part of this abstract form.

*Variables*

The worst value for the congruence was 74.30 and the best value 76.50. The author of this research paper will exclude these variables, which have no impact on the creation of the two clusters “credible” and “not credible”. The variables “telef”, “wohnzeit”, “pers”, “alter”, “hoehe”, “weitkred” and “beruf” will not be part of the formula of the discriminant function.

*Coefficients*

By executing the “lda” function in the statistic program R you will get the coefficients of the discriminant function in the section “Coefficients of the linear discriminants” of the results. Now you can insert these coefficients into the formula of the discriminant function.

Table 22

*Coefficients of the discriminant function*

| Variable | Coefficient |
|---|---|
| laufkont | 0.52599678 |
| laufzeit | -0.03313090 |
| moral | 0.38067069 |
| verw | 0.02482113 |
| sparkont | 0.17534482 |
| beszeit | 0.14020532 |

| | |
|---|---|
| rate | -0.17357140 |
| famges | 0.21299822 |
| buerge | 0.28030022 |
| verm | -0.18938713 |
| wohn | 0.25803123 |
| bishkred | -0.26618071 |
| gastarb | 0.50817751 |

*Constant Element*

The constant element $b_0$ will be selected in such way that the critical discriminant value will be zero. The critical discriminant value is that value, which will be considered to assign a new customer to one of the two groups. The formula of the discriminant function without the constant element will be as follows:

$$y = 0.52599678 * \text{laufkont} + (-0.03313090) * \text{laufzeit} + 0.38067069 * \text{moral} + 0.02482113 * \text{verw} + 0.17534482 * \text{sparkont} + 0.14020532 * \text{beszeit} + (-0.17357140) * \text{rate} + 0.21299822 * \text{famges} + 0.28030022 * \text{buerge} + (-0.18938713) * \text{verm} + 0.25803123 * \text{wohn} + (-0.26618071) * \text{bishkred} + 0.50817751 * \text{gastarb}$$

To determine the constant element, you have to compare the discriminant value y of a specific customer (without consideration of the constant element) with the value in the result section “$x” after the application of the function “predict”, which is the discriminant value y of that specific customer (with consideration of the constant element). The function “predict” will be executed in this form:

```
dataQualityTest <- cbind(data[1:5], data[7:11], data[13], data[16:17], data[21])
predict(lda(dataQualityTest[2:14], dataQualityTest[,1]), dataQualityTest[2:14])
```

For the first customer the discriminant value without regarding the constant element is 2.09098589. The value for this customer in the “$x” section of the result of the function “predict” is -1.0340796936. The difference will be the searched constant element with the value of -3.12506556. To avoid errors, this was executed also with the second customer.

*Formula of the Discriminant Function*

After the identification of the different parts of the formula of the discriminant function, the whole discriminant function formula can be generated:

y = -3.12506556 + 0.52599678 * laufkont + (-0.03313090) * laufzeit + 0.38067069 * moral + 0.02482113 * verw + 0.17534482 * sparkont + 0.14020532 * beszeit + (-0.17357140) * rate + 0.21299822 * famges + 0.28030022 * buerge + (-0.18938713) * verm + 0.25803123 * wohn + (-0.26618071) * bishkred + 0.50817751 * gastarb

## 3.9. Classification of new Customers

After the creation of the formula of the discriminant function, you can classify new customers. The discriminant value y will be assumed and on the basis of that the new grouping will be executed. The author of this paper will assume on the basis of the training data and then you will see, how high the congruence will be between the original dataset and the assumed data. You can say that you assume the data against itself.

*“predict”*

Therefore, you have to use the R function “predict”. The call has to be in the form predict(model, dataset). The model is an assumption of the discriminant function. The dataset will be the same data, used to determine the formula of the discriminant function. The following R statement was executed.

```
dataQualityTest <- cbind(data[1:5], data[7:11], data[13], data[16:17], data[21])
predict(lda(dataQualityTest[2:14], dataQualityTest[,1]), dataQualityTest[2:14])
```

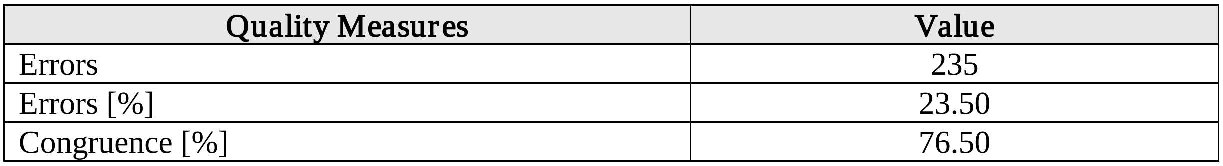
*“$class”*

The outome “$class” shows the assumed assignment to a group.

*Quality of the discriminant function to classify new customers*

Table 22

*Quality of the Discriminant Function to classify new customers*



| Quality Measures | Value |
|---|---|
| Errors | 235 |
| Errors [%] | 23.50 |
| Congruence [%] | 76.50 |

In 76.50 percent, the cluster in the variable “kredit” will be the same than with the discriminant function assumed cluster. This is a very high congruence in the opinion of the author of this research paper. This also means that, if you want to classify new customers of the german bank, you will be right in 76.50 percent, according to the training data of the 1000 customers.

## 4. Conclusion

*Purpose of the Research Paper*

The purpose of this writing paper was to choose an appropriate linear classification method to create a rule to predict if a customer can pay back the loan or not and to select adequate variables for this classification. Another task of the research paper was to determine how much error is made applying this rule.

*Methodology*

To achieve the goal of this paper, the authors used the linear discriminant analysis to predict, if new customers are credit-worthy or not.

*Formula of the Discriminant Function*

To generate the formula of the discriminant function, the variables had to be selected, the coefficients and the constant element had to determined.

$$y = -3.12506556 + 0.52599678 * \text{laufkont} + (-0.03313090) * \text{laufzeit} + 0.38067069 * \text{moral} + 0.02482113 * \text{verw} + 0.17534482 * \text{sparkont} + 0.14020532 * \text{beszeit} + (-0.17357140) * \text{rate} + 0.21299822 * \text{famges} + 0.28030022 * \text{buerge} + (-0.18938713) * \text{verm} + 0.25803123 * \text{wohn} + (-0.26618071) * \text{bishkred} + 0.50817751 * \text{gastarb}$$

*Quality of the discriminant function*

After the creation of the discriminant function, the quality of this function had to be determined. The values of the variable “kredit” of the basic dataset and the values, assumed by the discriminant function, had to be compared. The following table includes the number and percentage of the errors and the congruence in percent between these two values for all the 1000 customers of the german bank. These 1000 customers were used as training set to generate the discriminant function.

Table 23

*Quality of the Discriminant Function*

| Quality Measures | Value |
|---|---|
| Errors | 235 |
| Errors [%] | 23.50 |
| Congruence [%] | 76.50 |

You can see that 76.50 percent of the customers were allocated to the same cluster, which represents the credibility. If you want to classify new customers according to the training set used to generate the discriminant function, you will classify 76.50 percent of that new customers to the right cluster.

## ACKNOWLEDGEMENT

This paper was supported by Dimitris Karlis, Assistant Professor at the Department of Statistics, Athens University of Economics and Business.

# REFERENCES

Faes.de (2008). *Diskriminanzanalyse.* In: http://www.faes.de/Basis/Basis-Lexikon/Basis-Lexikon-Multivariate/Basis-Lexikon-Diskriminanz/basis-lexikon-diskriminanz.html#3GruppenBeispiel. Accessed on 27th of November, 2008.

# APPENDIX

## Description of the Variables

Table 24

Description of the Variables

| Variable | Description | Categories | Score | rel. frequency in % for | |
|---|---|---|---|---|---|
| | | | | good credits | bad credits |
| `kredit` | Creditability:<br>1 : credit-worthy<br>0 : not credit-worthy | | | | |
| `laufkont` | Balance of current account | no balance or debit | 2 | 35.00 | 23.43 |
| | | 0 <= ... < 200 DM | 3 | 4.67 | 7.00 |
| | | ... >= 200 DM or checking account for at least 1 year | 4 | 15.33 | 49.71 |
| | | no running account | 1 | 45.00 | 19.86 |
| `laufzeit` | Duration in months (metric) | | | | |
| `dlaufzeit` | Duration in months (categorized) | <=6 | 10 | 3.00 | 10.43 |
| | | 6 < ... <= 12 | 9 | 22.33 | 30.00 |
| | | 12 < ... <= 18 | 8 | 18.67 | 18.71 |
| | | 18 < ... <= 24 | 7 | 22.00 | 22.57 |
| | | 24 < ... <= 30 | 6 | 6.33 | 5.43 |
| | | 30 < ... <= 36 | 5 | 12.67 | 6.86 |
| | | 36 < ... <= 42 | 4 | 1.67 | 1.71 |
| | | 42 < ... <= 48 | 3 | 10.67 | 3.14 |
| | | 48 < ... <= 54 | 2 | 0.33 | 0.14 |
| | | > 54 | 1 | 2.33 | 1.00 |
| `moral` | Payment of previous credits | no previous credits / paid back all previous credits | 2 | 56.33 | 51.57 |
| | | paid back previous credits at this bank | 4 | 16.67 | 34.71 |
| | | no problems with current | 3 | 9.33 | 8.57 |

| | | | | | |
|---|---|---|---|---|---|
| | | credits at this bank | | | |
| | | hesitant payment of previous credits | 0 | 8.33 | 2.14 |
| | | problematic running account / there are further credits running but at other banks | 1 | 9.33 | 3.00 |
| verw | Purpose of credit | new car | 1 | 5.67 | 12.29 |
| | | used car | 2 | 19.33 | 17.57 |
| | | items of furniture | 3 | 20.67 | 31.14 |
| | | radio / television | 4 | 1.33 | 1.14 |
| | | household appliances | 5 | 2.67 | 2.00 |
| | | repair | 6 | 7.33 | 4.00 |
| | | education | 7 | 0.00 | 0.00 |
| | | vacation | 8 | 0.33 | 1.14 |
| | | retraining | 9 | 11.33 | 9.00 |
| | | business | 10 | 1.67 | 1.00 |
| | | other | 0 | 29.67 | 20.71 |
| hoehe | Amount of credit in "Deutsche Mark" (metric) | | | | |
| dhoehe | Amount of credit in DM (categorized) | <=500 | 10 | 1.00 | 2.14 |
| | | 500 < ... <= 1000 | 9 | 11.33 | 9.14 |
| | | 1000 < ... <= 1500 | 8 | 17.00 | 19.86 |
| | | 1500 < ... <= 2500 | 7 | 19.67 | 24.57 |
| | | 2500 < ... <= 5000 | 6 | 25.00 | 28.57 |
| | | 5000 < ... <= 7500 | 5 | 11.33 | 9.71 |
| | | 7500 < ... <= 10000 | 4 | 6.67 | 3.71 |
| | | 10000 < ... <= 15000 | 3 | 7.00 | 2.00 |
| | | 15000 < ... <= 20000 | 2 | 1.00 | 0.29 |
| | | > 20000 | 1 | 0.00 | 0.00 |
| sparkont | Value of savings or stocks | < 100,- DM | 2 | 11.33 | 9.86 |
| | | 100,- <= ... < 500,- DM | 3 | 3.67 | 7.43 |
| | | 500,- <= ... < 1000,- DM | 4 | 2.00 | 6.00 |
| | | >= 1000,- DM | 5 | 10.67 | 21.57 |

| | | | | | |
|---|---|---|---|---|---|
| | | not available / no savings | 1 | 72.33 | 55.14 |
| beszeit | Has been employed by current employer for | unemployed | 1 | 7.67 | 5.57 |
| | | <= 1 year | 2 | 23.33 | 14.57 |
| | | 1 <= ... < 4 years | 3 | 34.67 | 33.57 |
| | | 4 <= ... < 7 years | 4 | 13.00 | 19.29 |
| | | >= 7 years | 5 | 21.33 | 27.00 |
| rate | Installment in % of available income | >= 35 | 1 | 11.33 | 14.57 |
| | | 25 <= ... < 35 | 2 | 20.67 | 24.14 |
| | | 20 <= ... < 25 | 3 | 15.00 | 16.00 |
| | | < 20 | 4 | 53.00 | 45.29 |
| famges | Marital Status / Sex | male: divorced / living apart | 1 | 6.67 | 4.29 |
| | | female: divorced / living apart / married | 2 | 11.33 | 10.29 |
| | | male: single | 2 | 25.00 | 18.43 |
| | | male: married / widowed | 3 | 48.67 | 57.43 |
| | | female: single | 4 | 8.33 | 9.57 |
| buerge | Further debtors / Guarantors | none | 1 | 90.67 | 90.71 |
| | | Co-Applicant | 2 | 6.00 | 3.29 |
| | | Guarantor | 3 | 3.33 | 6.00 |
| wohnzeit | Living in current household for | < 1 year | 1 | 12.00 | 13.43 |
| | | 1 <= ... < 4 years | 2 | 32.33 | 30.14 |
| | | 4 <= ... < 7 years | 3 | 14.33 | 15.14 |
| | | >= 7 years | 4 | 41.33 | 41.29 |
| verm | Most valuable available assets | Ownership of house or land | 4 | 22.33 | 12.43 |
| | | Savings contract with a building society / Life insurance | 3 | 34.00 | 32.86 |
| | | Car / Other | 2 | 23.67 | 23.00 |
| | | not available / no assets | 1 | 20.00 | 31.71 |
| alter | Age in years (metric) | | | | |
| dalter | Age in years (categorized) | 0 <= ... <= 25 | 1 | 26.67 | 15.71 |
| | | 26 <= ... <= 39 | 2 | 47.33 | 52.72 |

| | | | | | |
|---|---|---|---|---|---|
| | | 40 <= ... <= 59 | 3 | 21.67 | 26.14 |
| | | 60 <= ... <= 64 | 5 | 2.33 | 3.00 |
| | | >= 65 | 4 | 2.00 | 2.43 |
| `weitkred` | Further running credits | at other banks | 1 | 19.00 | 11.71 |
| | | at department store or mail order house | 2 | 6.33 | 4.00 |
| | | no further running credits | 3 | 74.67 | 84.29 |
| `wohn` | Type of apartment | rented flat | 2 | 62.00 | 75.43 |
| | | owner-occupied flat | 3 | 14.67 | 9.14 |
| | | free apartment | 1 | 23.33 | 15.57 |
| `bishkred` | Number of previous credits at this bank (including the running one) | one | 1 | 66.67 | 61.86 |
| | | two or three | 2 | 30.67 | 34.43 |
| | | four or five | 3 | 2.00 | 3.14 |
| | | six or more | 4 | 0.67 | 0.57 |
| `beruf` | Occupation | unemployed / unskilled with no permanent residence | 1 | 2.33 | 2.14 |
| | | unskilled with permanent residence | 2 | 18.67 | 20.57 |
| | | skilled worker / skilled employee / minor civil servant | 3 | 62.00 | 63.43 |
| | | executive / self-employed / higher civil servant | 4 | 17.00 | 13.86 |
| `pers` | Number of persons entitled to maintenance | 0 to 2 | 2 | 84.67 | 84.43 |
| | | 3 and more | 1 | 15.33 | 15.57 |
| `telef` | Telephone | no | 1 | 62.33 | 58.43 |
| | | yes | 2 | 37.67 | 41.57 |
| `gastarb` | Foreign worker | yes | 1 | 1.33 | 4.71 |
| | | no | 2 | 98.67 | 95.29 |

**R - Code Listing**

```
data<-read.table("D:\\Studium\\Erasmus\\Courses\\Multivariate Techniques\\project 2\\kredit.asc",header=T)

library(MASS)

lda(data[2:21], data[,1])

lda(data[2:21], data[,1], CV = TRUE)

write.table(data$kredit, file = "excel import.csv")

1. step: all v

   write.table((lda(data[2:21], data[,1], CV = TRUE))$class, file = "excel import.csv")

2. step: all v - "telef"

   dataQualityTest <- cbind(data[1:19], data[21])
   dataQualityTest
   write.table((lda(dataQualityTest[2:20], dataQualityTest[,1], CV = TRUE))$class, file = "excel import.csv")

3. step: all v - "telef" and "verw"

   telef <- data[20]
   verw <- data[5]

   dataQualityTest <- cbind(data[1:4], data[6:19], data[21])
   dataQualityTest

   write.table((lda(dataQualityTest[2:19], dataQualityTest[,1], CV = TRUE))$class, file = "excel import.csv")

4. step: all v - "telef" and "wohnzeit"

   telef <- data[20]
   wohnzeit <- data[12]

   dataQualityTest <- cbind(data[1:11], data[13:19], data[21])
   dataQualityTest

   write.table((lda(dataQualityTest[2:19], dataQualityTest[,1], CV = TRUE))$class, file = "excel import.csv")
```

5. step: all v - "telef", "wohnzeit" and "pers"

```
telef <- data[20]
wohnzeit <- data[12]
pers <- data[19]

dataQualityTest <- cbind(data[1:11], data[13:18], data[21])
dataQualityTest

write.table((lda(dataQualityTest[2:18], dataQualityTest[,1], CV = TRUE))$class, file = "excel import.csv")
```

6. step: all v - "telef", "wohnzeit", "pers" and "gastarb"

```
telef <- data[20]
wohnzeit <- data[12]
pers <- data[19]
gastarb <- data[21]

dataQualityTest <- cbind(data[1:11], data[13:18])
dataQualityTest

write.table((lda(dataQualityTest[2:17], dataQualityTest[,1], CV = TRUE))$class, file = "excel import.csv")
```

7. step: all v - "telef", "wohnzeit", "pers" and "wohn"

```
telef <- data[20]
wohnzeit <- data[12]
pers <- data[19]
wohn <- data[16]

dataQualityTest <- cbind(data[1:11], data[13:15], data[17:18], data[21])
dataQualityTest

write.table((lda(dataQualityTest[2:17], dataQualityTest[,1], CV = TRUE))$class, file = "excel import.csv")
```

8. step: all v - "telef", "wohnzeit", "pers" and "laufkont"

```
telef <- data[20]
wohnzeit <- data[12]
pers <- data[19]
laufkont <- data[2]
```

```
dataQualityTest <- cbind(data[1], data[3:11], data[13:18], data[21])
dataQualityTest

write.table((lda(dataQualityTest[2:17], dataQualityTest[,1], CV = TRUE))$class, file = "excel import.csv")
```

9. step: all v - "telef", "wohnzeit", "pers" and "famges"

```
telef <- data[20]
wohnzeit <- data[12]
pers <- data[19]
famges <- data[10]

dataQualityTest <- cbind(data[1:9], data[11], data[13:18], data[21])
dataQualityTest

write.table((lda(dataQualityTest[2:17], dataQualityTest[,1], CV = TRUE))$class, file = "excel import.csv")
```

10. step: all v - "telef", "wohnzeit", "pers" and "alter"

```
telef <- data[20]
wohnzeit <- data[12]
pers <- data[19]
alter <- data[14]

dataQualityTest <- cbind(data[1:11], data[13], data[15:18], data[21])
dataQualityTest

write.table((lda(dataQualityTest[2:17], dataQualityTest[,1], CV = TRUE))$class, file = "excel import.csv")
```

11. step: all v - "telef", "wohnzeit", "pers", "alter" and "verm"

```
telef <- data[20]
wohnzeit <- data[12]
pers <- data[19]
alter <- data[14]
verm <- data[13]

dataQualityTest <- cbind(data[1:11], data[15:18], data[21])
dataQualityTest
```

```
write.table((lda(dataQualityTest[2:16], dataQualityTest[,1], CV = TRUE))$class, file = "excel
   import.csv")
```

12. step: all v - "telef", "wohnzeit", "pers", "alter" and "laufzeit"

```
telef <- data[20]
wohnzeit <- data[12]
pers <- data[19]
alter <- data[14]
laufzeit <- data[3]

dataQualityTest <- cbind(data[1:2], data[4:11],data[13], data[15:18], data[21])
dataQualityTest

write.table((lda(dataQualityTest[2:16], dataQualityTest[,1], CV = TRUE))$class, file = "excel
   import.csv")
```

13. step: all v - "telef", "wohnzeit", "pers", "alter" and "moral"

```
telef <- data[20]
wohnzeit <- data[12]
pers <- data[19]
alter <- data[14]
moral <- data[4]

dataQualityTest <- cbind(data[1:3], data[5:11],data[13], data[15:18], data[21])
dataQualityTest

write.table((lda(dataQualityTest[2:16], dataQualityTest[,1], CV = TRUE))$class, file = "excel
   import.csv")
```

14. step: all v - "telef", "wohnzeit", "pers", "alter" and "hoehe"

```
telef <- data[20]
wohnzeit <- data[12]
pers <- data[19]
alter <- data[14]
hoehe <- data[6]

dataQualityTest <- cbind(data[1:5], data[7:11],data[13], data[15:18], data[21])
dataQualityTest

write.table((lda(dataQualityTest[2:16], dataQualityTest[,1], CV = TRUE))$class, file = "excel
   import.csv")
```

15. step: all v - "telef", "wohnzeit", "pers", "alter", "hoehe" and "sparkont"

```
telef <- data[20]
wohnzeit <- data[12]
pers <- data[19]
alter <- data[14]
hoehe <- data[6]
sparkont <- data[7]

dataQualityTest <- cbind(data[1:5], data[8:11],data[13], data[15:18], data[21])
dataQualityTest

write.table((lda(dataQualityTest[2:15], dataQualityTest[,1], CV = TRUE))$class, file = "excel
   import.csv")
```

16. step: all v - "telef", "wohnzeit", "pers", "alter", "hoehe" and "beszeit"

```
telef <- data[20]
wohnzeit <- data[12]
pers <- data[19]
alter <- data[14]
hoehe <- data[6]
beszeit <- data[8]

dataQualityTest <- cbind(data[1:5],data[7], data[9:11],data[13], data[15:18], data[21])
dataQualityTest

write.table((lda(dataQualityTest[2:15], dataQualityTest[,1], CV = TRUE))$class, file = "excel
   import.csv")
```

17. step: all v - "telef", "wohnzeit", "pers", "alter", "hoehe" and "rate"

```
telef <- data[20]
wohnzeit <- data[12]
pers <- data[19]
alter <- data[14]
hoehe <- data[6]
rate <- data[9]

dataQualityTest <- cbind(data[1:5],data[7:8], data[10:11],data[13], data[15:18], data[21])
dataQualityTest

write.table((lda(dataQualityTest[2:15], dataQualityTest[,1], CV = TRUE))$class, file = "excel
   import.csv")
```

18. step: all v - "telef", "wohnzeit", "pers", "alter", "hoehe" and "buerge"

```
telef <- data[20]
wohnzeit <- data[12]
pers <- data[19]
alter <- data[14]
hoehe <- data[6]
buerge <- data[11]

dataQualityTest <- cbind(data[1:5],data[7:10],data[13], data[15:18], data[21])
dataQualityTest

write.table((lda(dataQualityTest[2:15], dataQualityTest[,1], CV = TRUE))$class, file = "excel
   import.csv")
```

19 step: all v - "telef", "wohnzeit", "pers", "alter", "hoehe" and "weitkred"

```
telef <- data[20]
wohnzeit <- data[12]
pers <- data[19]
alter <- data[14]
hoehe <- data[6]
weitkred <- data[15]

dataQualityTest <- cbind(data[1:5],data[7:11],data[13], data[16:18], data[21])
dataQualityTest

write.table((lda(dataQualityTest[2:15], dataQualityTest[,1], CV = TRUE))$class, file = "excel
   import.csv")
```

20 step: all v - "telef", "wohnzeit", "pers", "alter", "hoehe", "weitkred" and "bishkred"

```
telef <- data[20]
wohnzeit <- data[12]
pers <- data[19]
alter <- data[14]
hoehe <- data[6]
weitkred <- data[15]
bishkred <- data[17]

dataQualityTest <- cbind(data[1:5],data[7:11],data[13], data[16], data[18], data[21])
dataQualityTest

write.table((lda(dataQualityTest[2:14], dataQualityTest[,1], CV = TRUE))$class, file = "excel
   import.csv")
```

21 step: all v - "telef", "wohnzeit", "pers", "alter", "hoehe", "weitkred" and "beruf"

```
    telef <- data[20]
    wohnzeit <- data[12]
    pers <- data[19]
    alter <- data[14]
    hoehe <- data[6]
    weitkred <- data[15]
    beruf <- data[18]

    dataQualityTest <- cbind(data[1:5], data[7:11], data[13], data[16:17], data[21])
    dataQualityTest

    write.table((lda(dataQualityTest[2:14], dataQualityTest[,1], CV = TRUE))$class, file = "excel
       import.csv")

dataQualityTest <- cbind(data[1:5], data[7:11], data[13], data[16:17], data[21])
lda(dataQualityTest[2:14], dataQualityTest[,1])

predict(lda(dataQualityTest[2:14], dataQualityTest[,1]), dataQualityTest[2:14])
```

**VBA - Code Listing**

```
Option Explicit

Const firstRow As Integer = 6
Const numberCustomers As Integer = 1000
Const lastRow As Integer = numberCustomers + firstRow - 1
Const colKredit As Integer = 4
Const lastCol As Integer = 25

Private Sub btnCalc_Click()
   Dim c, r, errors As Integer
   Dim errorsPercent, congruence As Double

   '----- calculate errors and congruence -----
   For c = colKredit + 1 To lastCol
      For r = firstRow To lastRow
         If (Cells(r, colKredit).Value <> Cells(r, c).Value) Then
            errors = errors + 1
         End If
      Next
      Cells(2, c).Value = errors
```

```
        errorsPercent = errors / numberCustomers * 100
        Cells(3, c).Value = errorsPercent
        congruence = (numberCustomers - errors) * 0.1
        Cells(4, c).Value = congruence
        errors = 0
    Next

    '-----
End Sub
```