Multivariate Statistics Group Project

Natalja Nešumajeva, Alexander Bakalkin

14/12/2020

Contents

1 Research Problem

Businesses are enterprises which produce goods or render services for profit motive. To be able to predict the financial soundness of a business has led to many research works. Financial ratios are a key indicator of financial soundness of a business. Financial ratios are a tool to determine the operational & financial efficiency of business undertakings. There exist a large number of ratios propounded by various authors. Altman developed a z-score model using ratios as its foundation. With the help of the Z-Score model, Altman could predict financial efficiency /Bankruptcy up to 2-3 years in advance (Anjum 2012).

But is it possible to apply the Altman Z-Score model on Estonian companies, and to predict the financial efficiency/Bankruptcy up to 2-3 years in advance only relying on the data possessed in Estonian e-Business Register (research question nr 1), or would instead a grouping technique be more effective for bankruptcy identification (research question nr 2)?

First part of the research work question is predictive, where the second part of the questions is exploratory. Respectively the Linear Discriminant Analysis (LDA) and Clyster analysis and Multidimensional scaling (MDS) is going to be used in the group project for investigating the research question. In its turn MDS is used to prove or refute the answer on the research question nr1. given by LDA application (research question nr 3).

Altman Z-Score. In 1968, Edward Altman published what has become the best known predictor of bankruptcy. This predictor is a statistical model that combines five financial ratios to produce a product called a Z-score. The model has proven to be a dependable instrument in forecasting failure in a diverse mix of business entities (Anjum 2012):

In 1983, Altman developed a revised Z-score model for privately held firms (the original model is only applicable to publicly traded entities). The revised Z-scores substitute the book value of equity for the market value (Anjum 2012).

The revised Z-score model is the model that bests suits the data that was used in the given research paper, the data of the privately held companies.

The Z-score model ratios are following:

- X1= Working capital/total assets
- X2= Retained earnings/total assets
- X3= EBIT/total assets
- X4= Net Worth (book value)/total liabilities
- X5= Sales/total assets

And the Z-score formula is following:

$$Z = 0.717(X1) + 0.847(X2) + 3.107(X3) + 0.420(X4) + 0.998(X5) \quad (1)$$

Cut off scores of <1.23 indicate bankrupt firms and scores of >2.90 are indicators of non bankrupt firms. Firms with scores between 1.23 and 2.90 are determined to exist in the grey area or zone of ignorance (Anjum 2012).

Altman's Z-score model produces results of 90.9% accuracy in bankruptcy forecasting at least one year prior to actual failure. Firms with scores over 2.90 have a 97% chance of continuing operations with financial health (Anjum 2012).

2 Data

The data for the research work was obtained from the Estonian e-Business Register. Totally were collected the information from the financial reports of the 165 companies, among them are 81 companies in bankruptcy.

Along with the data from the financial reports, the additional informative data was added as well. Such as number of employees, size of the company, age, dissolved year, last report year, EMTAK code, location of the company. Based on the financial data from the reports financial ratios of the Z-score model were calculated. Afterwards the Z-score itself as well.

The exact name of the variables used in the research work is presented in the Appendix 1, the short name of the variables can be seen in Table 1.

Table 1 provides an overview of the data descriptive statistics used in the research work. Based on coefficient of variation (CV),it can be stated that the ZscoreIIndModel, RE_TA ratio, EBIT_TA ration and RE vary the most from all the variables. The skewness of the variables is also the largest among others. The ratio and financial data variables have very high Excess Kurtosis, what means that they are not normally distributed.

Table 1: A Summary statistics

Variable	Mean	Median	Minimum	Maximum	Std. Dev.	C.V.	Skew.	Ex. Kurt.
ZscoreIIndModel	12.0	3.2	-1'129	1'006	131.0	11.0	-0.7	54.9
Type	2.2	3.0	1	3	0.9	0.4	-0.4	-1.6
Bankrupt	0.5	0.0	0	1	0.5	1.0	0.0	-2.0
WC_TA	0.2	0.2	-13	1	1.2	5.2	-9.0	98.5
RE_TA	-1.0	0.2	-179	1	14.0	14.4	-12.6	157.4
EBIT_TA	-1.9	0.0	-325	1	25.3	13.1	-12.7	159.9
BVE_TL	38.6	0.4	-1	2'393	228.2	5.9	8.2	74.1
TS_TA	2.4	1.1	0	32	4.1	1.7	5.0	31.8
CA	1'069'500.0	77'537.0	0	40'822'000	4'321'200.0	4.0	6.7	49.8
TA	3'592'400.0	139'520.0	0	262'790'000	21'881'000.0	6.1	10.5	119.2
CL	770'920.0	51'822.0	0	20'166'000	2'698'500.0	3.5	5.5	31.3
TL	1'604'400.0	79'259.0	0	39'087'000	5'320'700.0	3.3	4.6	22.7
RE	1'657'200.0	11'906.0	-3'665'600	214'370'000	16'866'000.0	10.2	12.3	152.3
EBIT	347'720.0	3'580.0	-1'791'500	38'381'000	3'109'600.0	8.9	11.4	134.8
BE	1'988'100.0	29'696.0	-5'352'100	233'270'000	18'547'000.0	9.3	12.0	145.6
TS	2'573'100.0	164'620.0	0	85'307'000	9'668'800.0	3.8	6.3	43.7
Employees	15.3	2.0	0	797	65.5	4.3	10.5	121.7
Size	1.9	2.0	1	3	0.6	0.3	0.0	-0.4
Age	11.3	10.0	1	27	6.6	0.6	0.5	-0.6
Dissolved_year	2'019.4	2'020.0	2'017	2'020	1.1	0.0	-1.5	0.5
Last_Report	2'017.5	2'019.0	2'012	2'019	2.0	0.0	-1.2	0.3

The correlation matrix in Figure 1 below show that financial data of the companies have positive correlations and exists a strong correlations for a lot of financial data. In correlation matrix of financial ratios (see Figure 2) exists a negative correlation for some variables. The strongest correlation is between EBIT_TA and RE_TA.

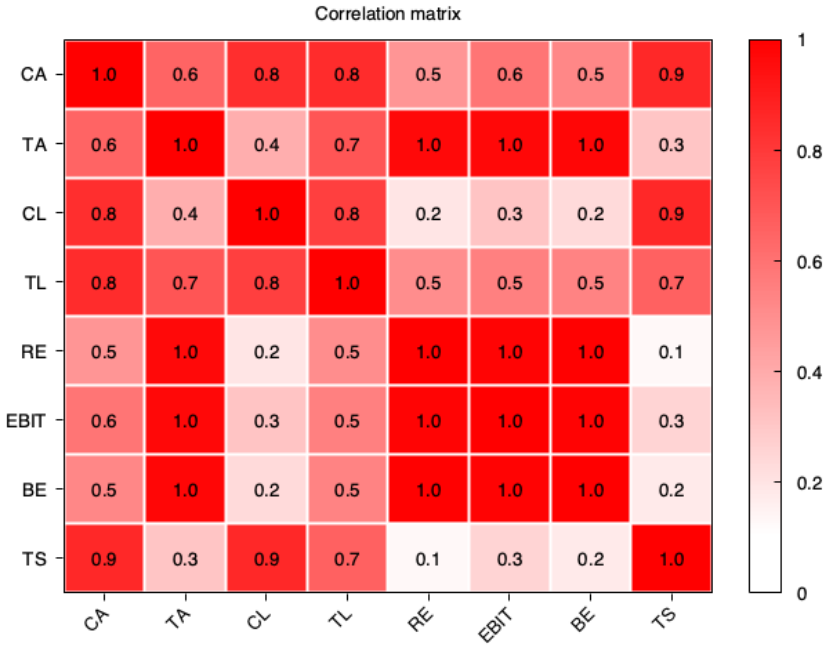

Figure 1: Correlation matrix of financial data

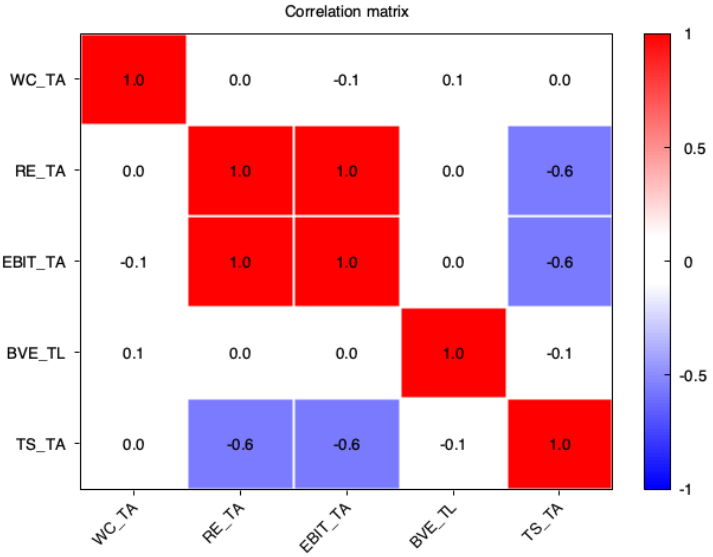

Figure 2: Correlation matrix of financial ratios

3 Methodology

The investigation of the research problem is based on the three multivariate methods. Research question nr 1 is investigated using Linear Discriminant Analysis and research question nr 2 is explored with the help of Cluster analysis. Linear Discriminant Analysis (LDA) is a linear transformation techniques used for dimensionality reduction. LDA is "supervised" algorithm and computes the directions ("linear discriminants") that will represent the axes that maximize the separation between classes. Clyster analysis is an exploratory technique. The Analysis searches for patterns in data by grouping observations into clusters. In Clyster Analysis nothing is known: nor the numbers of clusters/groups nor the assignment of groups to observations.

Multidimensional scaling (MDS) is an exploratory technique as well. MDS is an addition to research question nr 1. MDS is visualizing the degree of resemblance of a dataset's individual instances. It refers to a collection of similar ordination techniques used in the visualization of information, in particular to display the data stored in a matrix of distances. The goal of an MDS algorithm is to position each object in N-dimensional space such that the distances between objects are kept as best as possible. In each of the N dimensions, any object is then given co-ordinates. The number of dimensions can exceed 2 for an MDS plot N and is defined a priori. For a two-dimensional scatterplot, choosing N=2 optimizes the object positions

Excel was used for initial data preparation. Data needed to be appropriate for the import into R and Gretl. Gretl was used for retrieving summary statistics and correlation matrices of the data.

Linear Discriminant Analysis and Cluster Analysis were performed in R. The 165 observations were divided into two groups: bankrupt (as 1) or non-bankrupt type (0). In bankruptcy group there were 81 companies. For Linear Discriminant Analysis the lda function in R was used. Along with the separation of the observations into 2 groups, the LDA analysis was also used for reviewing whether the Z-score model cut off scores introduced by Edward Altman are applicable to a financial data of companies presented in the Estonian e-Business Register. The LDA analysis was used for separating data into 3 groups according to the Z-score values calculated with the formula (1) and according cut off points.

MDS was also performed in R. The aim was to find out the similarity of results with the Z-Score coefficient. During MDS was computed the distance matrix of our standardized variables and pasted the specimen of first 5 observations from our distance matrix. During the process we were to find the correlation between fitted and observed distances.

4 Results

4.1 Linear Discriminant Analysis (research question nr 1)

For the first part of the research question: "whether it is possible to apply the Altman Z-Score model on Estonian companies, and to predict the financial efficiency/Bankruptcy up to 2-3 years in advance only relying on the data possessed in Estonian e-Business Register" answer was searched with the help of LDA. Was estimated the linear discriminant function using the lda function in R. For the analyse were taken the variables listed in Table 2. Coefficients for the first discriminant function are also listed in Table 2

The Linear discriminant function is as following:

$$LD1 = 1.372(X1)+0.910(X2)-0.513(X3)-0.002(X4)-0.009(X5)-0,002(X6)+0.669(X7)+0.012(X8)+0.157(X9) \quad (2)$$

For detailed description of X's please refer to Appendix 1.

Table 2: LDA scores for the first LDA function

	x
WC_TA	-1.3725
RE_TA	0.9098
EBIT_TA	-0.5129
BVE_TL	-0.0018
TS_TA	0.0089
Employees	-0.0019
Size	0.6687
Age	0.0120
Tallinn	0.1569

The accuracy of the prediction is shown in Table 3 and in Table 4. In Table 3 ia presented the prediction in quantities and in Table 4 the prediction in percentage for each category of Bankrupt. The total percentage of the correct prediction is 72.73%.

Table 3: Assessment of the prediction accuracy

	0	1
0	60	24
1	21	60

Table 4: Percent correct for each category of Bankrupt

	x
0	0.71
1	0.74

In Figure 3 are presented the LDA histograms based on discriminant scores. The LDA histograms show that there exists an overlapping within the groups and there is no quite distinct separation of the groups.

In Figure 4 is presented a Scattergram with LDA scores. The Scattergram once again helps to view that the predictions are not quite exact, and there exists some wrong predictions among the correct ones.

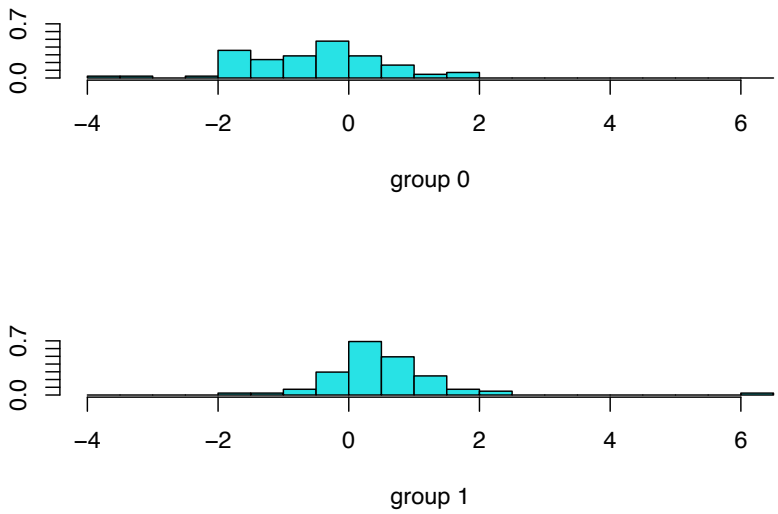

Figure 3: LDA histograms based on discriminant scores

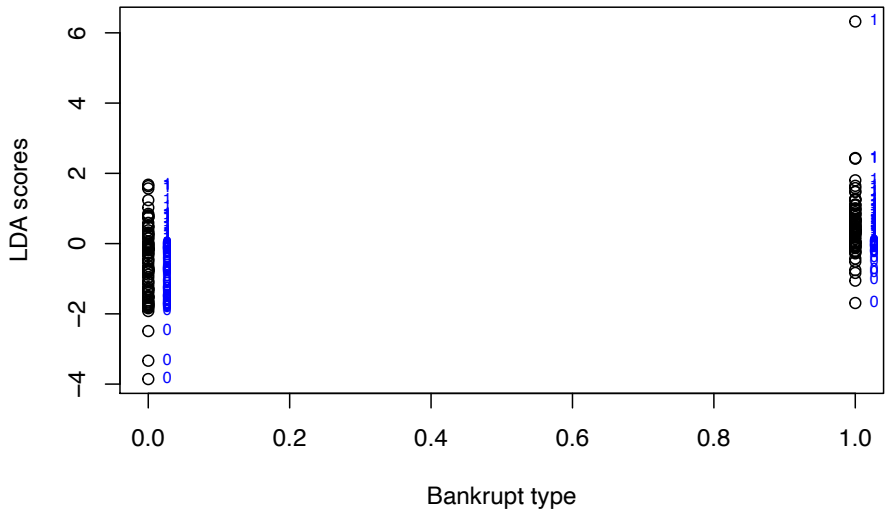

Figure 4: Scattergram with LDA score

Significance test and measure of Discriminant Power are presented in Table 5. As can be seen from the table, the "Employees" variable is not significant in the current LDA model (but the variable has been kept in the model: the exclusion from the model has not contributed significantly to the raising of the model prediction accuracy). "WC_TA", "BVE_TL", "Size" and "Age" variables are all significant at 10% significance level.

Table 5: Significance test and measure of Discriminant Power of explanatory variables

	correl_ratio	wilks_lambda	F_statistic	p_value
WC_TA	0.09	0.91	16.87	0.00
RE_TA	0.01	0.99	1.45	0.23
EBIT_TA	0.01	0.99	1.09	0.30
BVE_TL	0.02	0.98	4.14	0.04
TS_TA	0.01	0.99	1.46	0.23
Employees	0.00	1.00	0.11	0.74
Size	0.06	0.94	11.11	0.00
Age	0.02	0.98	3.01	0.08
Tallinn	0.01	0.99	1.40	0.24

4.2 Linear Discriminant Analysis for 3 groups, employing Altman Z-score model cut off scores of Formula (1)

The Scatterplot of Discriminant Function for 3 groups is presented in Figure 5. Based on the plot it can be concluded that Altman Z-score model cut off scores does not quite good separate the groups. Type 2 is overlapping with Type 1 and 3, on the other hand separation between Type 1 and 3 is more readable.

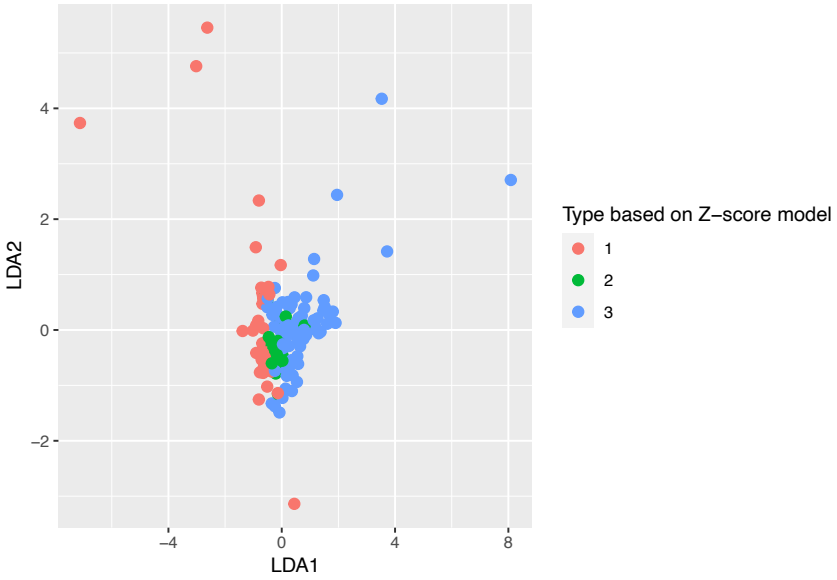

Figure 5: Scatterplot of Discriminant Function

4.3 Clyster Analysis (research question nr 2)

With Clyster Analysis was searched the answer to the question whether a grouping technique, a Clyster Analysis, can be more effective for bankruptcy identification (research question nr 2) in comparison to LDA.

The optimal number of clusters in presented in Figure 6. According to it, the optimal number of clusters is 2, what corresponds to the total number of Bankruptcy type (2).

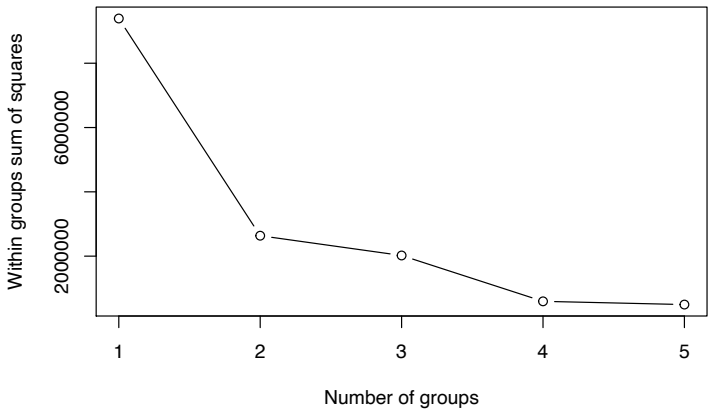

Figure 6: Optimal number of clusters

As can be seen in Figure 7 and in Table 6 the number of "Company" variables in two clusters differ significantly. In percentage it is 2% versus 98%, what is very far from the original data company groupings, where it was 49% versus 50%. The further investigation of the research question nr 2 with Clyster Analysis is irrelevant.

Table 6: Cluster assignment outcome Table

Var1	Freq
1	162
2	3

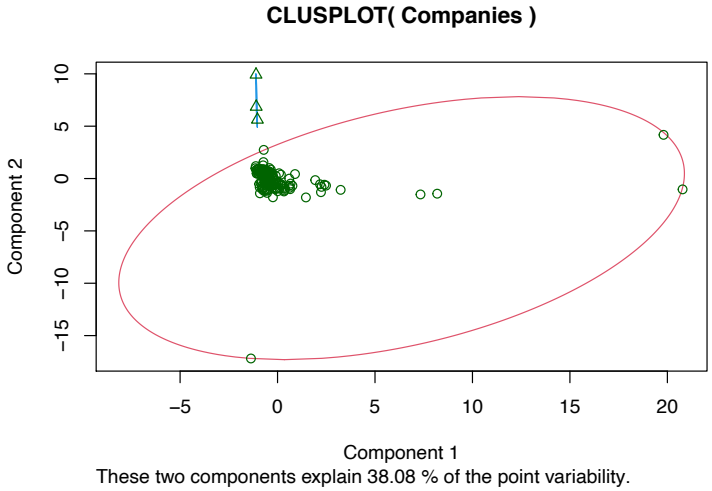

Figure 7: Clustering plot

4.4 Multidimensional Scaling

Trying to find the answer on the research question nr1 with MDA: "whether it is possible to apply the Altman Z-Score model on Estonian companies, and to predict the financial efficiency/Bankruptcy up to 2-3 years in advance only relying on the data possessed in Estonian e-Business Register" first of all we need to standardize the variable under consideration according to our problem statement.

By the next step, we will compute the distance matrix of our standardized variables and paste the specimen of first 5 observations from our distance matrix.

Now we will perform the multidimensional scaling of our distance matrix and store the results in our variable named as MDS. In order to make prettier plots, we have used the ggplot2. In the following plot (please see Figure 8) of two dimensional solution of multidimensional scaling, we have used numbers to specify the observation.

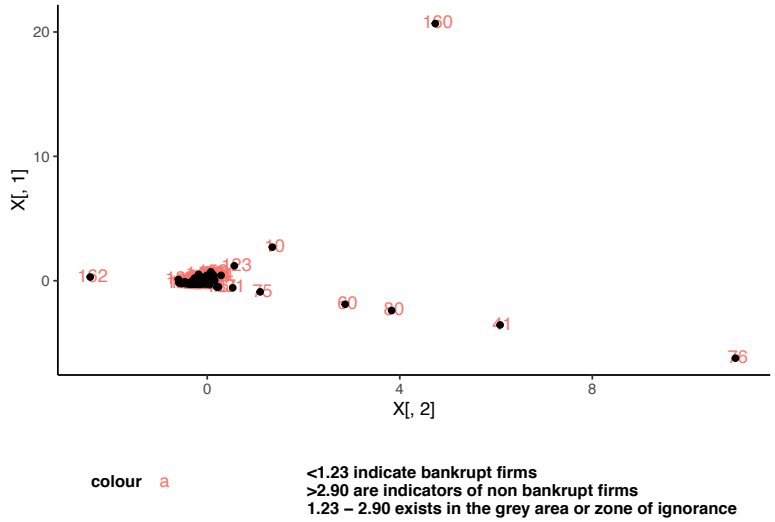

Figure 8: Multi-dimensional Scaling on predictor of bankruptcy Data

Then we should find eigen values and compute the goodness of fit of the solution. Evidence from the goodness of fit, it was shown that the fit is strong positive correlation at 0.7394865

In the Figure 9 we have plotted the fitted distances against the observed distances. The red line corresponds to the regression line. We can see that the regression line goes almost linearly with respect to the fitted and observed distances. This proves that there is a strong correlation between the observed and fitted distances which was our initial objective. Since the line is almost at 45 degree, we can say that correlation is strong and hence, visually, it gives a good goodness of fit.

By the next step was done the non-metric MDS with the isoMDS program.

In order to graphically examine the stress in the case of non-metric multidimensional scaling, we present Observed Dissimilarity graph in Figure 10. We can observe from the plot that the fitted distances are in almost perfectly linearly related with observed distances.

In order to obtain a good fit, we must have a stress of approximately 5% or equivalently, 0.05. However, so far up to three dimensions, the stress existing is equal to 7% approximately. Hence, it would be in our favor to obtain a good fit if we chose four dimensions where our stress reduces to 3% approximately (see Figure 11).

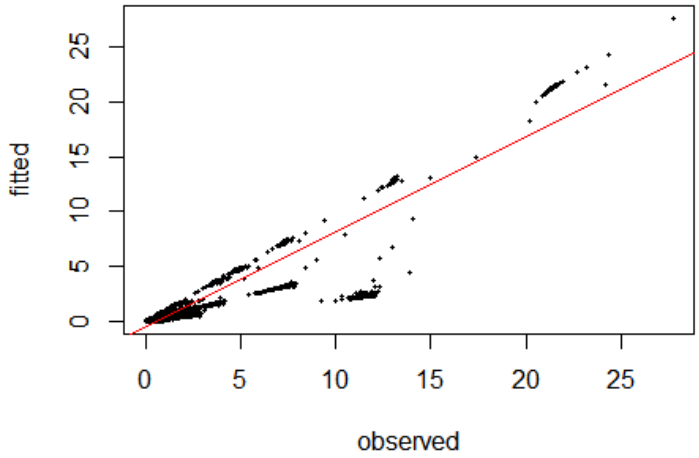

Figure 9: Fitted Distances Plot

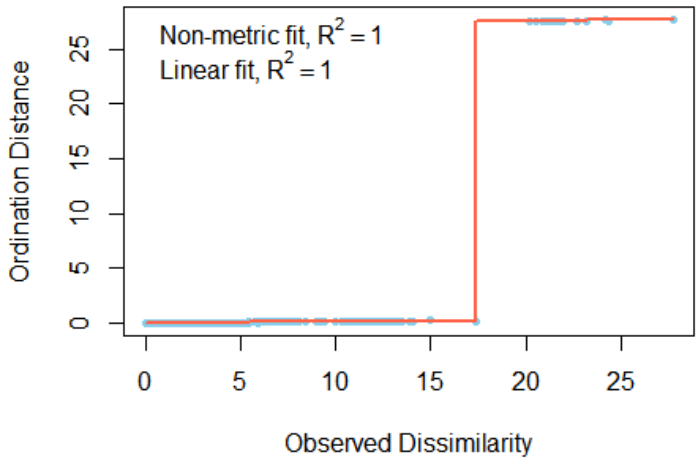

Figure 10: Stress plot

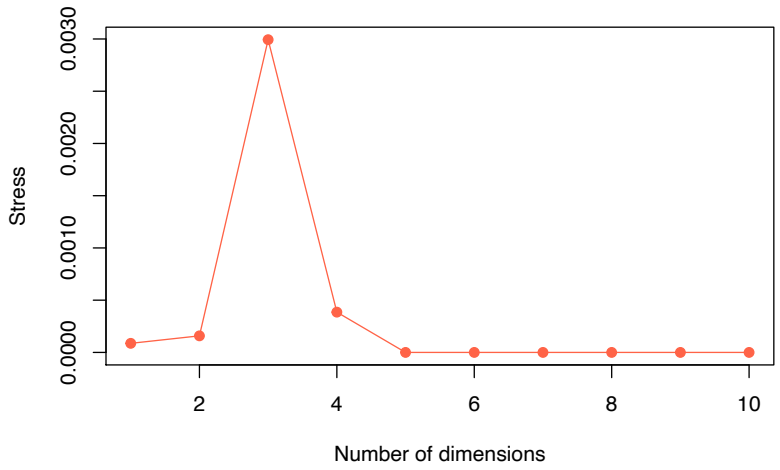

Figure 11: Stress level's dependence from the number of dimensions

5 Conclusion

The problem of the current research work was to investigate whether is it possible to apply the Altman Z-Score model on Estonian companies, and to predict the financial efficiency/Bankruptcy up to 2-3 years in advance only relying on the data possessed in Estonian e-Business Register (research question nr 1), or would instead a grouping technique be more effective for bankruptcy identification (research question nr 2)?

According to studies Altman's Z-score model produces results of 90.9% accuracy in bankruptcy forecasting at least one year prior to actual failure (Anjum 2012). This work analysis showed only 72.73% accuracy in bankruptcy prediction. The lower percentage is due to the fact that bankruptcy classification that Estonian e-Business Register is assigning to companies, is not a classical understanding of company financial insolvency.

The vivid example to that would be a company called MyJar IT OÜ. Z-score of which was among the highest 2 years before the bankruptcy declaration. Companany employeed 146 employees and was over 10 years on market. But came the owners decision for liquidating the company, and for some reasons the company was reclassified from "under liquidation" into "bankruptcy" group (Ruuda 2020).

If investigated further and to distinguish separately the phenomena referred below, the percent of accuracy received in this research work would be higher.

Finally answering to the research work question: "is it possible to apply the Altman Z-Score model on Estonian companies, and to predict the financial efficiency/Bankruptcy up to 2-3 years in advance only relying on the data possessed in Estonian e-Business Register", the answer would be "yes", but only after performing the additional reclassification of "Bankruptcy" type and based on the updated results to recalculate the Linear discriminant function Formula (2).

As was seen in Cluster Analysis part the grouping technique (research question nr 2) is not effective for bankruptcy identification. As in current research work the Clyster Analysis was not able, as a first "step", to replicate the similar percentage distribution of observations into groups. Still the technique managed to identify the number of bankruptcy types.

According to the two-dimensional solution of the multidimensional scaling problem, the result is similar with z-score coefficient of distinct as confirmed by distance matrix. At the same time using the non-standardized variables, makes distance matrix not accurate, as different variables will be on different scales and it will result in an absurd distance matrix. Therefore, it is important to use the standardized variables for computing the distance matrix in such cases, because different variables are based on different scales and have different units.

Lastly, needs mentioning that there exist three other significant types of Bankruptcy prediction models, after Discriminant Analysis. They are Logit and Probit Analysis, Recursive Partitioning Algorithm, and Neural

Networks etc. These have not been discussed in this work, but the results of the application of which on the current research work would be interesting to compare.

6 LIST OF APPENDIXES

6.1 Appendix 1:

Company names - Company names

ZscoreIIndModel - Z score value calculated based on the Z-score model Formula 1.

Type - Type of the company calculated based on the Z-score model Formula 1. “1” - bankrupt firms, “2” - companies in grey zone, “3” - non bankrupt companies.

Bankrupt - whether the company is declared a bankrcypt (1) or not (0).

WC_TA - Working Capital/Total Assets (X1)

RE_TA - Retained Earnings/Total Assets (X2)

EBIT_TA - Earnings Before Interest and Tax/Total Assets (X3)

BVE_TL - Book Value of Equity/Total Assets (X4)

TS_TA - Total Sales/Total Assets (X5)

CA - Current Assets

TA - Total Assets

CL - Current Liabilities

TL - Total Liabilities

RE - Retained Earnings

EBIT - Earnings Before Interest and Tax

BE - Book Value of Equity

TS - Total Sales

Employees - Number of employees in the company (X6)

Size - Size of the company. “1”-Micro company, “2”-middle size company, “3”- large company (X7).

Age - age of the company (X8).

Tallinn - whether the company is based in Tallinn or not (X9).

Dissolved_year - the year of the bankcypcy declaration.

Last_Report - the year of the report based on what the company data was collected.

6.2 Appendix 2: All R code for this group project

```
library(knitr)
library(readxl)
statistics <- read_excel("DataStatistics.xlsx")
kable(statistics, format = "pipe", digits = c(1, 1, 1, 0, 0, 1,1,1,1),
      format.args = list(big.mark = "'",scientific = FALSE), longtable = TRUE,
      caption = "A Summary statistics")
library(png)
library(knitr)
library(dplyr)
library(ggplot2)
setwd("/Users/nataljaneshumajeva/Documents/Isiklik/TTU/MSA/Project work")
img1_path <- "correlation_findata.png"
```

```r
include_graphics(img1_path)
img1_path <- "correlation_ratios.png"
include_graphics(img1_path)
rm(list=ls())
getwd()
library("readxl")
library("MASS")
library("DiscriMiner")
library(tint)
setwd("/Users/nataljaneshumajeva/Documents/Isiklik/TTU/MSA/Project work")
ratios <- read_excel("DataRMarkdown1012.xlsx")
summary(ratios)
fit <- lda(Bankrupt ~ WC_TA + RE_TA + EBIT_TA + BVE_TL + TS_TA + Employees + Size +
             Age + Tallinn, data=ratios)
fit
fit$scaling[,1]
fit.values <- predict(fit, ratios)
z<-fit.values$x
kable(fit$scaling[,1], format = "pipe", digits = 4,
      format.args = list(big.mark = "'",scientific = FALSE), longtable = TRUE,
      caption = "LDA scores for the first LDA function")
ct <- table(Actual=ratios$Bankrupt, Predicted=fit.values$class)
diag(prop.table(ct, 1))
sum(diag(prop.table(ct)))
average <- round(sum(diag(prop.table(ct)))*100,2)
ct <- table(Actual=ratios$Bankrupt, Predicted=fit.values$class)
ct
diag(prop.table(ct, 1))
sum(diag(prop.table(ct)))
kable(ct, format = "pipe", digits = 2,
      format.args = list(big.mark = "'",scientific = FALSE), longtable = TRUE,
      caption = "Assessment of the prediction accuracy")
kable(diag(prop.table(ct, 1)), format = "pipe", digits = 2,
      format.args = list(big.mark = "'",scientific = FALSE), longtable = TRUE,
      caption = "Percent correct for each category of Bankrupt")
fit$svd
ldahist(data = fit.values$x[,1], g=ratios$Bankrupt)
plot(ratios$Bankrupt, fit.values$x[,1], xlab="Bankrupt type", ylab="LDA scores")
text(ratios$Bankrupt, fit.values$x[,1],fit.values$class,cex=0.7,pos=4,col="blue")
dp = discPower(ratios[, c("WC_TA","RE_TA","EBIT_TA","BVE_TL","TS_TA","Employees",
                          "Size","Age","Tallinn")], ratios$Bankrupt)
correl_ratio <- round(dp$correl_ratio,digits = 6)
wilks_lambda <- round(dp$wilks_lambda,digits = 6)
p_value <- round(dp$p_value,digits = 2)
F_statistic <- round(dp$F_statistic,digits = 6)
matrix <-cbind(correl_ratio, wilks_lambda, F_statistic,  p_value)
rownames(matrix) =c("WC_TA","RE_TA","EBIT_TA","BVE_TL","TS_TA","Employees","Size",
                    "Age","Tallinn")
matrix
kable(matrix, format = "pipe", digits = 2,
      format.args = list(big.mark = "'",scientific = FALSE), longtable = TRUE,
      caption = "Significance test and measure of Discriminant Power of explanatory variables")
getwd()
```

```
library(ggplot2)
library(caret)
library(readxl)
library(readxl)
library(car)
library(MASS)
library(psych)
ratios <- read_excel("DataRMarkdown1012.xlsx")
ratios$Type <- factor(ratios$Type)
ratios_lda <- lda(Type ~ WC_TA + RE_TA + EBIT_TA + BVE_TL + TS_TA, data = ratios)
ratios_lda
ratios_lda.values <- predict(ratios_lda)
ldahist(ratios_lda.values$x[,1], g = ratios$Type)
ratios_lda$scaling[,1]
ratios_lda.values <- predict(ratios_lda, ratios)
z2<-ratios_lda.values$x
ct2 <- table(Actual=ratios$Type, Predicted=ratios_lda.values$class)
ct2
diag(prop.table(ct2, 1))
sum(diag(prop.table(ct2)))
kable(diag(prop.table(ct2, 1)), format = "pipe", digits = 2,
      format.args = list(big.mark = "'",scientific = FALSE), longtable = TRUE,
      caption = "Assessment of the prediction accuracy")
matrix2 = discPower(ratios[, c("WC_TA","RE_TA","EBIT_TA","BVE_TL","TS_TA")], ratios$Type)

kable(matrix2, format = "pipe", digits = 2,
      format.args = list(big.mark = "'",scientific = FALSE), longtable = TRUE,
      caption = "Significance test and measure of Discriminant Power of explanatory variables")
newdata <- data.frame(type = ratios[,1], lda = ratios_lda.values$x)
library(ggplot2)
newdataplot <- ggplot(newdata) + geom_point(aes(lda.LD1, lda.LD2, colour = ratios$Type),
                                            size = 2.5)
print(newdataplot + labs(y="LDA2", x="LDA1", colour = "Type based on Z-score model"))
library(readxl)
options(scipen = 999)
getwd()
setwd("/Users/nataljaneshumajeva/Documents/Isiklik/TTU/MSA/Project work")
ratios <- read_excel("DataRMarkdowncluster.xlsx")
head(ratios)
no.partitions<-function(n,g) {
  round((g^n)/factorial(g),0)
}
no.partitions(165,2)
ratios[,5:13]
PCA <- princomp(ratios[,5:13], cor = FALSE, scores = TRUE)
plot(PCA$scores[1:165,1],PCA$scores[1:165,2])
cl_kmeans<-kmeans(ratios[,5:13], centers=2)
cl_kmeans$centers
cl<-cl_kmeans$cluster
cl
xtabs(~Company_names+cl, data=ratios)
ratios_Scored <- data.frame(ratios, cl)
head(ratios_Scored)
```

```r
cl_kmeans$size
cl_kmeans$withinss
cl_kmeans$betweenss
cl_kmeans$totss
table(cl_kmeans$cluster)
library(cluster)
clusplot(ratios_Scored,cl_kmeans$cluster,color=TRUE,shade=FALSE,labels=0,lines=0)
wss<-rep(0,5)
wss[1]<-(nrow(ratios)-1)*sum(sapply(ratios[,5:13],var))
wss[1]<-sum(scale(ratios[,5:13], scale=FALSE)^2)
for (i in 1:5){
  wss[i]<- sum(kmeans(ratios[,5:13], centers=i)$withinss)
}
plot(1:5, wss, type="b", xlab="Number of groups", ylab="Within groups sum of squares")
table(cl_kmeans$cluster)
clustertable <- table(cl_kmeans$cluster)
kable(clustertable, format = "pipe", digits = 2,
      format.args = list(big.mark = "'",scientific = FALSE), longtable = TRUE,
      caption = "Cluster assignment outcome Table")
clusplot(ratios_Scored,cl_kmeans$cluster,color=TRUE,shade=FALSE,labels=0,lines=0, main = paste("CLUSPLO
# importing the data set
library(readxl)
Data1112 <- read_excel("Data1112.xlsx")
# removing the first row in the data set that has a subheading
mydata <- Data1112[-1, ]

#variable names in the dataset

names(mydata)
#first standardized the variable under consideration
mydata_standardized  =scale(mydata[,c(3,5,6,7,8,9)], center = TRUE, scale = TRUE)
head(mydata_standardized)

distance <- dist(mydata_standardized, method = "euclidean", upper = TRUE, diag = TRUE)
distance <- as.matrix(distance)
head(distance,5)
#perform the multidimensional scaling of our distance matrix using the cmdscale command and store the r
n = nrow(distance)
MDS <- cmdscale(distance, k= 2, eig = TRUE, x.ret = TRUE)
X <- MDS$points[,1:2]

#plot of two dimensional solution of multidimensional scaling
library(ggplot2)
g <- ggplot(data.frame(X, mydata),
            aes(X[,2], X[,1], label = rownames(mydata), group="
                                                              <1.23 indicate bankrupt firms
                                                              >2.90 are indicators of non bankrupt firms
                                                              1.23 - 2.90 exists in the grey area or zo
g + geom_text(aes(color = "
                          <1.23 indicate bankrupt firms
                          >2.90 are indicators of non bankrupt firms
                          1.23 - 2.90 exists in the grey area or zone of ignorance") ,
              angle=0, show.legend = T, nudge_x = 0.05, nudge_y = 0.15) +
```

```
  geom_point() + theme_classic() +
  theme(legend.position = "bottom", legend.title =
          element_text(size=10, face="bold"),
        legend.text = element_text(size=10, face="bold"))

ev <- MDS$eig
gof <- MDS$GOF
print(round(ev,digits=4))
print(gof)

# Goodness of fit
print(cmdscale(distance, 2, eig=TRUE)$GOF)

#Goodness of fit manually
print( (ev[1]+ev[2])/sum(abs(ev)) )

print( (ev[1]+ev[2])/sum(ev[ev>0]))

#fitted distances
fitted <- as.matrix(dist(X, method = "euclidean"))
fitted <- as.vector(fitted)
observed <- as.vector(distance)
reg <- lm(fitted~observed)
plot(observed, fitted,pch=19, cex=0.4)
abline(lm(fitted~observed), col="red")
#Coefficient of determination
print(paste("Coefficient of determination:", summary(reg)$r.squared))
setwd("/Users/nataljaneshumajeva/Documents/Isiklik/TTU/MSA/Project work")
library(png)
library(knitr)
img1_path <- "FittedDistancesPlot.png"
include_graphics(img1_path)
setwd("/Users/nataljaneshumajeva/Documents/Isiklik/TTU/MSA/Project work")
library(png)
library(knitr)
img1_path <- "ObservedDissimilarity.png"
include_graphics(img1_path)
#MDS with the isoMDS
library(MASS)
n <- nrow(distance)
init <- scale(matrix(runif(n*2),ncol=2),scale=FALSE)
nmmds.out <- isoMDS(distance, init, k=2, maxit = 100)

library(vegan)
nmmds <- metaMDS(comm = distance, distance = "euclidean", k=2)
gof <- goodness(nmmds)

#examine the stress in case of non-metric multidimensional scaling
stressplot(nmmds, pch = 19, cex=0.75, l.col = "tomato", p.col = "skyblue")

observed <- as.vector(as.matrix(distance))
reg <- lm(fitted~observed)
plot(observed, fitted,pch=19, cex=0.4)
```

```r
abline(lm(fitted~observed), col="red")


#Coefficient of determination
print(paste("Coefficient of determination:", summary(reg)$r.squared))

#computing the stress for the dimensions
stress_vec <- numeric(10)
for(i in seq(10)){
  stress_vec[i] <- metaMDS(distance, distance = "euclidean", k=i)$stress
}

plot(seq(10),stress_vec, type = 'o', ylab = "Stress", xlab = "Number of dimensions",
     col="tomato", pch=19)
abline(h=0.2, lty=2)
abline(h=0.05, lty=3)
```

References

Anjum, Sanobar. 2012. "Business Bankruptcy Prediction Models: A Significant Study of the Altman's Z-Score Model." *Available at SSRN 2128475*.

Ruuda, Lennart. 2020. *Enam Kui 100 Töötajaga Edukas It-Firma Pakkis Ootamatult Pillid Kotti. Postimees.* https://leht.postimees.ee/7103384/enam-kui-100-tootajaga-edukas-it-firma-pakkis-ootamatult-pillid-kotti?fbclid=IwAR1tr_Y9WcpFG4mIn8nxhiPMPtzRs-F5WVNjxLmI4vgEdD4hTjvN_AX5cPU/.