VU University Amsterdam

MSc. Business Analytics

Research Paper

Predicting the profitability level of companies regarding the five comparability factors

March 31, 2017

Manon Wintgens (2558262)

Manon Wintgens (2558262)

Predicting the profitability level of companies regarding the five comparability factors

University:
VU University Amsterdam
MSc. Business Analytics
Faculty of Science
De Boelelaan 1105, 1081 HV
Amsterdam

Company:
EY - Transfer Pricing
Cross Towers
Antonio Vivaldistraat 150
1083 HP
Amsterdam

Supervisor University:
Dr. E.N. Belitser

Supervisor EY:
K. Lukosz

Abstract

This research is executed for the EY Transfer Pricing department. Transfer pricing is the price set between companies in the same group when services or goods are exchanged. Transactions within two companies in the same group are legal, as long as the price set (the transfer price) is at arm's length. The law and regulations are set by the OECD (Organisation for Economic Co-operation and Development).

In order to set the profitability margins, transfer pricing analysis the identifaction for independent comparable companies. Modern mathematical methodologies and a mix of data- and statistical-analysis could provide more information and insights of the comparability analysis.

For this research three different methodolagies are implemented to predict the profitability level for individual companies. The three models make an indication of the influential variables that should be used in the comparability analysis. Two methodolagies originate from machine learning algorithms (Random Forest and Gradient Boosting) and one is more based on statistical analysis (Lasso).

The Lasso method makes the best prediction of the profitability level. The variables independence, industry, size and financial stability of an individual company are the most important features to predict the profitability level of an individual company.

Preface

The research paper is one of the compulsory courses within the Master Business Analytics. In the research paper the student should analyse a relevant problem which embraces aspects of business, mathematics and computer-science.

This research paper is written in coloboration with the EY department Transfer Pricing. The EY Transfer Pricing department has a license to access the database Amadeus. Amadeus is the database of the company Bureau van Dijk, which collects comprehensive data information of 21 million individual companies world wide. For this research the dataset was generated by the database Amadeus.

All analysis on the dataset are done with the program R (Studio). Several packages within R are used to implement the different models, which will be referred to in this paper. Since a license is required to access the database Amadeus, the dataset and R code of the analysis could be requested by Manon Wintgens.

Correspondance to: Manon Wintgens, *m.s.wintgens@student.vu.nl* or *manonwintgens@gmail.com*.

Contents

1. Introduction

The research paper will be dedicated to the application of mathematical methods in the area of transfer pricing, as part of the assignment to the transfer pricing department of EY. This department covers various areas and topics of the transfer pricing worldwide, which are related to, amongst others, to intercompany tangible goods, intangibles, services and financial transactions. One of the focus areas in the transfer pricing is to set the arm's length level of pricing/profitability (e.g. EBIT margin[1]) to be reported by group companies engaged in the wholesale of goods. Such group distributors purchase finished products from related companies with the purpose of selling such products to third parties. While the revenue reported by such distributors is by definition at arm's length (i.e., reflecting the market prices), as it reflects transactions between the independent parties, the level of costs incurred by such distributors (i.e., price of the finished products), and subsequently the profitability level of such group distributors, can be influenced by the internal stakeholders.

In order to set the profitability margins for such related group distributors, the objective of the transfer pricing analysis is to identify independent comparable companies, which are engaged in similar type of activities, under the similar facts and circumstances. Once such comparable companies to a respective group distributor are identified, the comparable companies profitability level (e.g. EBIT margin) can be used to set the profitability level for that group distributor.

However, it is inherently difficult to identify and capture in the comparability analysis all the factors that influence the profitability of comparable (independent) companies. It also has to be acknowledged that there are no two companies that are identical, and therefore even companies identified as comparable to the group distributors will show certain comparability differences. Effectively, this makes the comparability analysis a somewhat subjective exercise. However, the objectivity of the analysis can be increased by making certain comparability adjustments to the identified comparable companies (if and to the extent possible).

Modern mathematical methodologies and a mix of data- and statistical-analysis could provide more information and insights, as well as enabling the increase in the reliability of the comparability analysis. EY has access to a dataset (on Amadeus[2]) with all information regarding those companies (e.g. industry, location, financial data, size of company). Specifically, mathematical tools help to identify as to which factors are the most relevant in light of the profitability margins. These factors could then be used for the purpose of the comparability adjustments, if the difference exist between potential comparable companies and group distributors.

The main principles of the transfer pricing (based on the OECD Transfer Pricing Guidelines) prescribed the five comparability factors are:

1. characteristic of the property or service transferred
2. functions performed by the parties taking into account assets employed and risk assumed, in short, termed as functional analysis (FAR)
3. contractual terms
4. economic circumstances
5. business strategies pursued

The focus of this research will be on comparability factor 1, 2, 4 and 5. Since 3 are out of scope for a mathematical approach.

Therefore the problem statement is:

What are the main qualitative and quantitative variables in Amadeus that drive the profitability levels of companies and reflect the comparability factors prescribed by the OECD?

[1] EBIT stands for Earnings Before Interest and Taxes and will be described in Chapter 3 in more detail.
[2] Amadeus is the database of the company Bureau van Dijk.

2. Background

EY describes transfer pricing as the price set between companies in the same group when services or goods are exchanged, EY Tax (2016). The price between two companies within the same complex corporate group is still referred to as the transfer price. Transactions within two companies in the same group are legal, as long as the price set (the transfer price) is at arm's length.

Arm's length principle means that the transfer price should be set as if the two companies were independent companies. An independent company would compare the price to other available options and will eventually agree with the best transaction, which is called arm's length principle. The arm's length principle therefore sets the transfer price by market forces. The law and regulations are set by the OECD (2010) (Organisation for Economic Co-operation and Development).

2.1 Related work

In April 2007 Bastiaan Stucken did research for EY for his thesis 'Addressing Transfer Pricing Issues Using Quantitative Methods', Stucken (2007). In his research he analysed if there were effects that influences the Profit Level Indicator (PLI) and modeled a linear regression analysis. For the PLI he took the Operating Profit Margin (OPM) as well as the Return on Assets (ROA) for the period 2003-2005. He concluded that the accounts receivables, inventories, accounts payable and industry were significantly important for the linear regression model. This research will take his foundings into account. Having said, the dataset for this research contains companies with different industries and information from the period 2014.

In 2011 Deloitte has done research to the factors that drive profitability of distributors, Crespo and Clark (2011). The factors for which they measured their relation to the profitability were:

- net sales (often referred to as turnover)
- SG&A expenses (Selling General & Administrative expenses)
- sales per employee
- net working capital
- fixed assets
- other assets-primarily cash and cash equivalents
- intangible assets
- geographic location
- industry

The analysis was done over the reported financial accounts for fiscal year 2008 and 2009. The outcome concluded that the following factors had influence on the profitability level of the distributors:

- SG&A expenses relative to sales
- sales per employee
- asset intensity (net working capital, fixed assets, and other assets relative to sales)

Since this research is deprecated, new research could help in finding other factors that drive the profitability.

3. Data description

The mission of this research is to investigate which factors have influence on the profitability of a company. EY Transfer Pricing has access to Amadeus, which is the database of the company Bureau van Dijk. Bureau van Dijk collects data of over 21 million individual companies. Since EY Transfer Pricing is interested in a specific type of companies a search strategy was used to make a selection of companies for the dataset. The search strategy used to select the dataset in Amadeus is explained in Appendix I.

The dataset exist of 18867 rows and 48 columns. The rows contain individual companies while the columns contain information about these companies (e.g. location, financial information and product). The main goal of this research is to make a predictive model of the profitability of a company and find factors that influence the profitability.

3.1 Response variable

The profitability is measured by the EBIT Margin of 2014, which will be referred to as the response variable in this research. The EBIT Margin stands for Earnings Before Interest and Taxes, Berk and DeMarzo (2013) p. 30, and reflects the profit of a company before depreciation of the interest and taxes to be paid. The period 2014 is chosen, since it is the most recent EBIT Margin available for (most) companies in Amadeus.

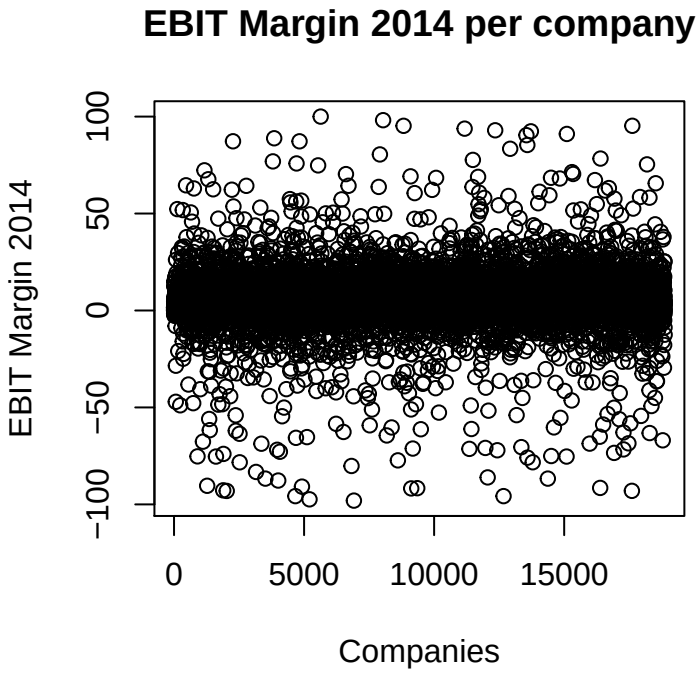

Figure 1: Plot of EBIT Margin 2014

Figure 1 shows the EBIT Margin of 2014 for each individual company in the dataset. In 201 cases the EBIT Margin of 2014 was not applicable. Since there are quite some missing values for the EBIT margin 2014 as well as missing datapoint in the 47 features, data crunching was required. Companies with missing values in the response variable, EBIT Margin 2014, or in one of the 47 features have been removed from the dataset.

Eventually this has lead to a dataset with 358 companies. *Figure* 2a,..,d illustrate the EBIT Margin of 2014. *Figure* 2a shows the EBIT Margin 2014 against each individual company. There are 2 extreme positive and 2 extreme negative points for the EBIT Margin of 2014 (outliers). These outliers are not due to errors in the measurement but due to the fact that some EBIT Margins of 2014 could fall out more extreme than other companies. *Figure* 2b shows the box plot for the EBIT Margin of 2014, one could observe that there are far more extreme outliers than the 4 observed in *Figure* 2a. The box between the first quantile (1.2795) and third quantile (5.21525) is extremely small. *Figure* 2c shows the histogram of the EBIT Margin 2014 with the line of the normal distribution $\mathcal{N}(\mu = 3.1721061, \sigma^2 = 5.0883759)$. The histogram is a bit skewed to the left (negative skewed), this means that the tail on the left side is longer and that there are more extreme negative observations than positive observations. The frequency of the companies with an EBIT Margin comparable to the mean is higher than the normal distribution indicates. *Figure* 2d gives the QQ-plot of the EBIT Margin 2014. If the observations of the EBIT Margin 2014 create a straight diagonal line, normality would be a right assumption. The middle part of the QQ-plot does not show a lot of deviance from the diagonal line, although the tails bend which gives the QQ-plot a S-shape. A S-shape indicates heavy tails, which where already observed in *Figure* 2c.

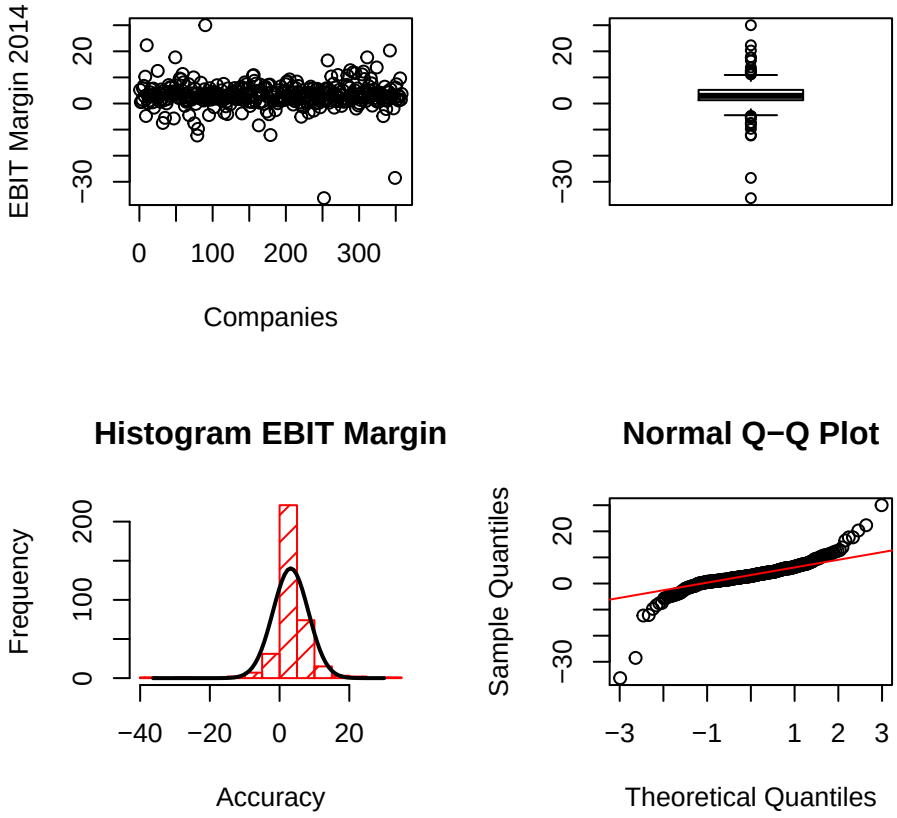

Figure 2: Plots of EBIT Margin 2014 after removing missing values

The dataset consist of companies which are active in different industries and with a variety of size levels. Therefore one expects the response variable, EBIT Margin 2014, not to be homogeneous. *Figure* 2a,..,d gave the presumption that normality is indeed doubtful. More comprehensive tests will give evidence if normality could be rejected.

Shapiro-Wilk test

The Shapiro-Wilk tests if observations (in this case the EBIT Margin 2014) are independent and come from a normal distribution with mean μ and variance σ^2, Bijma (2015), p. 29. The Shapiro-Wilk test statistic W $\in (0, 1]$ is rejected for p-value $\leq \alpha$ for $\alpha = 0.05$. The null hypothesis and alternative hypothesis are:

H_0 : The observations of the EBIT Margin 2014 are coming from a normal distribution.

H_1 : not H_0

Table 1: Shapiro-Wilk normality test: EBIT Margin 2014

Test statistic	P value
0.8106	3.597e-20 * * *

Table 1 shows the outcome of the Shapiro-Wilk Normality test. Since the p-value $\leq \alpha$ for $\alpha = 0.05$ the null-hypothesis is rejected, the observations of the EBIT Margin 2014 are probably not from a normal distribution.

The EBIT Margin 2014 of the individual companies is probably not following the normal distribution, the heavy outliers should be taken into account to make an efficient predictive model. Some statistical tools are based on normally distributed data, which will not be applicable in this case. For this particular dataset one should concentrate on non-parametric predictive models which does not assume the data to be drawn from a specific distribution.

3.2 Explanatory variables

Some of the data cannot be used for the predictive model (e.g. unique company names and unique identification numbers). The columns with variables which cannot be included in the model are dropped out of the dataset. The features that are left in the data exist of different types of data (categorical and numerical). The data has been listed in a dataframe in the program R. The variables with numerical data has been converted into numeric. The categorical features are left as characteristic variables in the dataset. Since the number of features is quite high and some features might not even be relevant for this research, an explanation of the most important features will be given after modeling and finding the influential factors.

Dummy variable

The categorical features (e.g. location) cannot be taken into a predictive model. Converting the categorical features into a dummy variable will make it possible to include categorical data into a predictive model. A categorical variable with I values can be transferred into a $I - 1$ dummy variable (binary), Breiman (2001) p. 15, where the value 1 indicates if a company is from a specific category and the value 0 if a comany is not from that specific category.

Included in the dataset is the categorical variable *Country*. This variable could be converted into a dummy variable. The I different countries could be converted into $I - 1$ dummy variables. The dummy variable for the country France will be 1 if the company is indeed located in France and 0 otherwise (when the company is located in another country). In this dataset, companies are only located in one particular country. So if the $I - 1$ dummy variables for the variable Country are all 0 for a company, than this company is located in the remaining country I.

Other categorical features that are converted into dummy variables are:

- The variable *NACE Reverence primary code*, which gives an indication of the type of wholesale (e.g. textiles, household goods).
- The variable *Category of company*, which gives the size category per company. The categories are: very large, large, medium and small. (E.g. “very large” stands for a company with an operating revenue

bigger than 100 MIL Euro, total assets above 200 MIL Euro and more than 1000 employees. All information about the different categories could be found on Amadeus).

- The variable *BvD independence indicator* is a characteristic given by Bureau van Dijk (BvD) to identify the degree of independence of each company. There are five possible degrees: A, B, C, D, U which are transformed into dummy variables.

3.3 Correlation

The dataset consist of multiple variables. It is possible that some of the variables show the same information in a slightly different way. Correlation measures the linear dependency among a pair of variables. By making a pairwise plot, one could compare variables among each other and get insight if there might be linear dependence. The number of features is too large to make pairwise plots for each feature. Therefore there is chosen to make a selection here of 6 variables. Since the variables are all related to the employees of a company, linear dependence is expected.

variable 1 = Profit per employee, EURO, 2014

variable 2 = Operating revenue per employee, EURO, 2014

variable 3 = Costs of employees divided by Operating revenue, 2014

variable 4 = Average cost of employee, EURO, 2014

variable 5 = Working capital per per employee, EURO, 2014

variable 6 = Number of employees, 2014

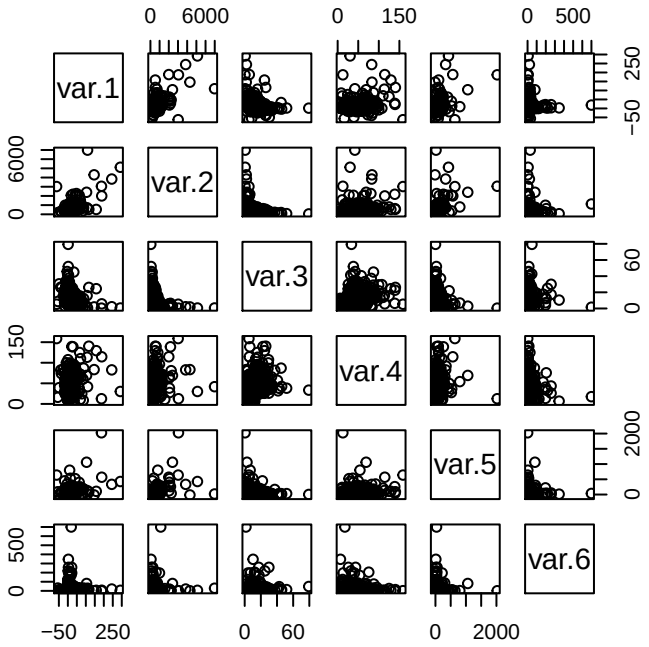

Figure 3: Pairwise plots between 6 variables

Figure 3 shows the pairwise plots for the 6 variables. The plots show linear dependence among a couple of

variables (e.g. variable 1 and variable 2). A correlation test will identify the degree of linear dependence and test which variable to neglect in this research.

To obtain the level of correlation among the variables one could carry out a correlation test. Most used correlation test is the Pearson correlation test. Since the Pearson correlation test assumes normality of the variables, this test is not useful in this particular research since outliers will get a higher weight than they should in the Pearson test.

The Spearman rank test takes the Pearson correlationcoefficient per ranked variable into account. When all variables are integers (which is the case for this research) the correlationcoefficient for the Spearman rank test could be simplified and be calculated by equation 1, Buijs (2008), p. 377. The Spearman Rank test measures the correlationcoefficient by ranking two explanatory variables, X_i and Y_i, from highest to lowest level. For each explanatory variable, the observation with the highest level will get rank number 1 and the observation with the lowest level will receive the rank which is equal to the number of observations (number of different companies). Per company the difference in ranking number between the two explanatory variables could be calculated and is denoted as d_i.

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n^3 - n} \tag{1}$$

Where:

n = the number of observations per pair (X_i, Y_i)

d = the difference between the ranking number of X_i and Y_i

The Spearman correlation coefficient will have a value between -1 and 1, which could be interpreted as a highly negative and highly positive linearity between the pair of variables.

The dataset consist of 48 variables. To get an investigation of the correlation between the pairs one should calculate a matrix of 48 by 48 correlations coefficient. The diagonal line could be filled with 1 since the correlation coefficient of X_1 against X_1 is 1 since it reflects the same. Only half of the rest of the correlation matrix should be calculated due to the fact that correlation coefficient for X_1 against X_2 or the other way around is the same. Still this will give a large matrix which is hard to analyse since there are so many factors in the dataset. Therefore there is chosen to show the visualization of the correlation matrix for the six variables from above. This gives a good insight in the correlation between each pair of variables.

The correlation matrix in *Figure* 4, made with the R package `corrplot`, shows the Spearman Rank correlation coefficients between six variables. A blue circle indicates that there exist positive correlation while a red circle shows negative correlation. The larger the circles are, the higher the correlationcoefficient. One could observe that there exist quite some positive as well as negative correlation between the variables. The pairwise plot gave the presumptions about strong correlation between the first and second variable, this correlation plot shows that the correlation between variable 1 and variable 2 seems to be quite high indeed.

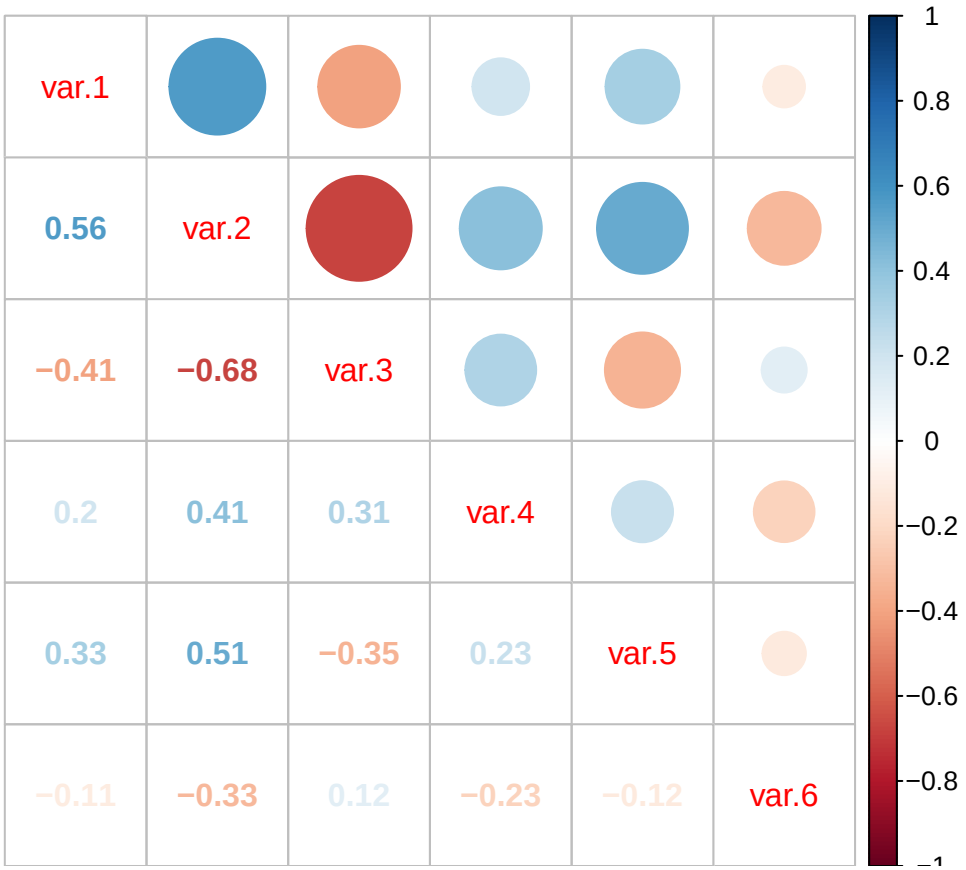

Figure 4: Correlation matrix plot between 6 variables

The correlation between the variables should be taken into account for the eventual predictive model. But since there are quite a lot of variables there is a change that there is not only correlation between the variables as pairs but also between more than two variables. Therefore a sufficient research in multicollinearity should indicate if there is indeed correlation between multiple variables.

3.4 Multicollinearity

Multicollinearity is the existence of correlation between multiple variables. Multicollinearity is a problem for making a prediction model since it causes inaccurate coefficients for the predictive model, fails in measuring which factors are significant and therefore it is not able to identify the importance of the variables which should be taken into the model.

Eigenvalues

To identify multicollinearity the eigenvalues of the matrix $X^T X$ could be calculated. Where the matrix X consist of all explanatory variables in the dataset. In general an eigenvalue is the scalar λ that identifies $M\bar{X} = \lambda\bar{X}$ where M is a $n \times n$ matrix (n = 48) and $\bar{X}$ is a nonzero vector.

Small eigenvalues imply that multicollinearity might be a problem. The condition number K_p indicates if there is indeed multicollinearity, Faraway (2002), p. 117.

$$K_p = \sqrt{(\frac{\lambda_1}{\lambda_p})}$$

for $p = 1, \ldots, n$ all explanatory variables.

The condition number K_p, is the square root of the first eigenvalue divided by all p eigenvalues. If the range between the maximum and minimum eigenvalues is wide and the condition number K $\geq$ 30 than there is more than one collinearity found.

Table 2 shows the maximum and minimum eigenvalues for the dataset, which are in a wide range.

Table 2: Maximum and minimum eigenvalues of the explanatory variables

Maximum Eigenvalue	Minimum Eigenvalue
8.64e+11	0.3675

The first row of the condition number K:

1, *4.457*, *7.427*, *20.03*, *37.65* and *64.39*

Since the maximum and minimum eigenvalues are in a wide range and some condition numbers are above 30 there is more than one collinearity problem. By calculating the variance inflation factors the variables that causes collinearity could be removed.

Variance Inflation Factor

To test the goodness of fit of a regression model the coefficient of determination, the R squared (R^2) could be calculated (see calculation of R^2 in chapter 5). To explore which variables do have an inaccurate influence on the predictive model due to overfitting one could calculate the variance inflation factor (VIF), Faraway (2002), p. 118.

$$VIF = \frac{1}{1-R_i^2}$$

for $i = 1, \ldots, n$ all explanatory variables.

A VIF outcome of 70, means that the standard error for this explanatory variable is $\sqrt{70}$ as high than when there would not have been multicollinearity.

One could remove multicollinearity with the aid of the R package `usdm`. The function `vifcor()` makes it possible to set a threshold. Here the threshold was 0.6, which was also used in the handbook of this package. By setting the threshold, the program will find the two explanatory variables with the highest correlation above the threshold and will delete the variable with the highest VIF. After removing this explanatory variable the program will continue till there are no variables with a correlation coefficient above the threshold.

After measuring the VIF and setting the threshold, 13 variables were excluded from the model and the eventual dataset to make the prediction on exist of 35 variables. The excluded variables could be found in Appendix II.

4. Approach

There are several methods to create a predictive model for the profitability of the EBIT Margin 2014. As has been mentioned above, the EBIT Margin 2014 can not be assumed to be normally distributed. Therefore the predictive model should not be based on normality assumptions. By trial and error multiple methods has been modeled and found influential factors for the profitability of companies.

4.1 Models

At the start of this research step wise regression was used to find the features that captured the profitability level of the companies best. Step wise regression is a helpful technique to find the features that are considered most appropriate to predict the profitability. But since the dataset exist of quite some explanatory variables, step wise regression is time consuming and eventually did not execute the best prediction. For this research one more statistical based model and two machine learning techniques are implemented.

4.1.1. Lasso

Lasso stands for Least Absolute Shrinkage and Selection Operator and is a method to select the best variables to predicts the response variable from a high dimensional size of variables. The lasso is applicable since it is a generalized linear model which does not assume normality.

The outcome of the lasso is similar as a regular linear regression and is of the form:

$$Y = X\beta + \epsilon$$

with $Y = (Y_1, .., Y_n)^T$ a $(n \times 1)$-vector with the response variable, X a $(n \times (p+1))$-matrix with all explanatory variables with the i-th row $x_i^T = (1, x_{i1}, .., x_{ip})$. The first column of X exist of 1, this is to obtain the intercept. The $(p \times 1)$-vector exist of $\beta = (\beta_0, .., \beta_p)^T$, which are the coefficients of the $i = 0, .., p$ explanatory variables and the $(n \times 1)$- vector with the errors $\epsilon = (\epsilon_1, .., \epsilon_n)^T$. The errors are assumed to be normally distributed with $\epsilon \sim N(0, \sigma^2 I)$ for I the $n \times n$ identity matrix. Furthermore we assume $\mathbb{E}\,\epsilon_i = 0$, $Var\epsilon_i = \sigma^2$ and all errors are independent, Gunst (2013), p. 2.

The lasso method was first introduced by Tibshirani (1996). Lasso is an extension of the Ordinary Least Squares (OLS) method and requires knowledge of vector norm calculation. Appendix III gives some more explanation of the vector norm calculations used and the OLS method.

The lasso belongs to the penalized regression family in which a regularization parameter λ governs the level of shrinkage, Glad and Richardson (2016). The goal in lasso is to find the optimal λ for the following minimization:

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{N} ||y - X\beta||_2^2 + \lambda ||\beta||_1 \right\}$$

Where y is the response variable, matrix X consist of the explanatory variables, β are the coefficients calculated by lasso and λ the shrinkage level and N the number of observations.

The first part of the lasso consist of the OLS method and uses the ℓ_2-norm (see Appendix III for more detail), in the second part of the lasso λ is introduced to control the behavior of the β's by the ℓ_1-norm.

Since there are some limitations of using only the ℓ_1-norm, Hastie and Qian (2014) introduced an elastic net parameter alpha (α), with $\alpha \in [0, 1]$. An α of 1 uses the ℓ_1-norm, α of 0.5 combines the ℓ_1 and ℓ_2-norm equally, and an α of 0 uses the ℓ_2-norm, also known as ridge regression. The combination of lasso and ridge regression solves the following problem:

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{N} ||y - X\beta||_2^2 + \lambda \Big[(1-\alpha)||\beta||_2^2/2 + \alpha ||\beta||_1 \Big] \right\}$$

Finding the optimal λ and α manually is time-consuming, therefore the R package `glmnet` is used to find the optimal parameters. The tuning of the Lasso method to identify the optimal parameters is explained in further detail in chapter 5.

On the Abel Symposium of 2014 (a symposium in mathematics) one of the topics was Statistical Analysis for High Dimensional Data. The winning papers are bundled and give a comprehensive discussion of the lasso method, Frigessi et al. (2014).

4.1.2. Random Forest

Random Forest is a machine learning algorithm that uses multiple decision trees to learn to make a prediction based on decision trees. It is excellent method for regression and classification problems, where the focus will be on regression here.

The random forest algorithm consist of multiple trees. Each tree learns on a subset of variables, which are randomly selected out of the set of explanatory variables. Than a decision tree is grown for each subset.

The regression splitting criteria in each tree are based on the greatest decrease of the Residual Sum of Squares (RSS), Breiman (1984), p. 263.

$$RSS = \sum_{Left}(Y_i - Y_L^*)^2 + \sum_{Right}(Y_i - Y_R^*)^2$$

for Y_i the response variable for the i^{th} company and Y_L^* the mean of the Y value on the left node and Y_R^* the mean of the Y value on the right node Cutler (2003), p. 30.

If there are n subsets with randomly selected explanatory variables the random forest algorithm will learn n decision trees. After training on all subsets, the average of each feature $\{T_n\}_1^N$ in the different decision trees could be taken as predictor variable and is calculated with the following formula:

$$\hat{f}_{RF}^N = \frac{1}{n}\sum_{n=1}^{N} T_n(x)$$

for N the number of trees, Kim (2015).

With the aid of recursive algorithms one could find the optimal number of decision trees (ntrees) and the number of variables within each subset (mtry).

Pruning (setting the depth of the decision tree) is not necessary for the random forest algorithm since it selects a subset of features for each decision tree that the algorithm learns. Therefore it automatically reduces the risk of overfitting.

Note: instead of using the ℓ_1-norm (as is used in the lasso method) the ℓ_2-norm is used to calculate the the leaf weights, Chen (2014), p. 4.

4.1.3. Gradient Boosting Machine

The third model to make a prediction of the profitability levels of the companies is the machine learning technique Gradient Boosting Machine (GBM). GBM combines the decision trees, which were introduced in the paragraph Random Forest, with weak learners (weak predictors).

The regression model $F(x)$ adds an additive model for the weak learner $h(x)$ to the regression model to improve $F(x) + h(x)$ and minimize the RMSE[3]. The additive model improves the eventual model for the weak learners, $h(x_i) = y_i - F(x_i)$ for y_i the actual outcome and $F(x_i)$ the predicted model, Li (2016), p. 20. By learning this for each individual tree, weak learners have the change to improve the model. By setting a learning rate $\lambda \in (0, 1]$ (shrinkage) the GBM learns in small steps, and errors made in an iterations could be corrected in the following step, Natekin and Knoll (2013). The interaction depth is the maximum depth of variable interaction within each tree and the n.minobsinnode is the minimum number of observations in each

[3]RMSE stands for Root Mean Squared Error and measures the goodness of fit of a model. Chapter 5 describes the RMSE in more detail.

tree. To tune the GBM model, the interaction depth and n.minobsinnode could be calculated iterative to find the best prediction.

The importance (influence) of each variable is calculated for each tree in the same way as the influences for the Random Forest were calculated.

Note that the regularization of the leaf weights are calculated by the ℓ_2-norm, Chen (2014), p. 4.

4.2 Training- and testset

To give an insight in the goodness of fit of the predictive model the data could be separated in a trainingset and a testset. The trainingset will be used to learn the predictive model which variables to use to make a prediction. The testset could be used to test the usability of the model. Better performance on the trainingset than on the testset might appear due to overfitting. Note that to measure the goodness of fit of the predictive model, the trainingset and testset must be independent, otherwise the predictive model predicts what is already known. For this research the trainingset and testset has been set in the following way:

- Randomly select 75% of the data to be the trainingset and use the other 25% as the testset.

To split the dataset in a training- and testset the R package `caret` was used. The function `createDataPartition()` gives the option to split the data by defining percentage p. The `set.seed()` function was used to reproduce the exact same training- and testset.

```
dataset #dataset after pre-processing
set.seed(998)
splitInTrainingSet <- createDataPartition(y = dataset$y, p = .75, list = FALSE)
trainingSet <- dataset[splitInTrainingSet,]
testingSet  <- dataset[-splitInTrainingSet,]
```

4.3 Crossvalidation

To improve the performance of the predictive model cross-validation is used on the trainingset. Cross-validation is a technique whereby the trainingset is randomly split in k parts which are also referred as folds. The model will train the model on $k-1$ folds and will test on the remaining fold. This will be repeated till all k folds has been learned and tested, Flach (2012), p. 349.

Each fold should at least consist of 30 observations. Oftenly k is set to 10 and is called 10-fold crossvalidation. Since the dataset is limited in this research $k = 10$ should not be extended to a higher level of k. The 10-fold crossvalidation is repeated 10 times to optimize the predictive model.

To implement the crossvalidation function the R package `caret` was used. The function `trainControl()` is an excellent method to define the number of folds (number = 10) and how many times the crossvalidation should be repeated.

```
crossValidation <- trainControl(method = "repeatedcv",
                                number = 10,
                                repeats = 10)
```

5. Results

For this research three different models use their own technique to identify the important variables to predict the profitability level EBIT Margin 2014. This chapter will show the results and the belonging influential factors.

5.1 Goodness of fit

To test which model performance best the models could be compared by the Root Mean Squared Error (RMSE) and the coefficient of determination also referred to as the R Squared (R^2).

Root Mean Squared Error

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where $\hat{y}$ are the predicted values by the model and y_i are the actual outcomes. It is the goal to minimize the RMSE, so that the mean of the residuals (the difference between the actual outcome and the predicted) would be minimized.

R Squared

The R Squared (R^2) is the coefficient of determination which measures the proportion of the response variable that could be predicted from the explanatory variables. The R^2 lies between 0 and 1, where 0 does not imply a good fit whereas 1 implies a perfect fit.

$$R^2 = 1 - \frac{SSE}{SST}$$
$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where SSE stands for Sum of Squared Error and measures the sum of the predicted values minus the actual outcomes squared.

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

where SST stands for Total Sum of Squares and measures the sum of the actual outcome minus the mean of the actual outcome.

Note: it is not always efficient to compare the R^2 if data is not from the normal distribution, since the R^2 takes the mean and variance from not normally distributed data into account. Ford (2015) does not even seem to find a reason to compare the R^2 between models and prefers the Mean Squared Error. Therefore there is choosen to put more power on the RMSE to decide which model performs best, although the R^2 is calculated as well.

5.2 Result per model

5.2.1. Lasso

The Lasso method tries to find the best fit by optimizing the λ. Lasso is a combination of the OLS and ℓ_1-norm. Since there are some limitations of using only the ℓ_1-norm, alpha (α) is introduced as an elastic net parameter, and combines the optimal combination between the ℓ_1- and ℓ_2-norm.

To fit the Lasso method and find the optimal λ and α the R packages `glmnet` was used, which gives the options to find the best tuning parameters. By using the function `train()` from the `caret` package one could fit a predictive model on the trainingset. The crossvalidation implemented in chapter 4 is invoked in this training process for the Lasso method.

```r
lasso <- train(x = trainingX, y = training$y, #for trainingX all explanatory variables
               method = "glmnet",
               trControl = crossValidation)
```

Figure 5 shows the tuning process for different λ and α. The optimal λ for the Lasso method is 0.2239772 and the optimal α is 1.

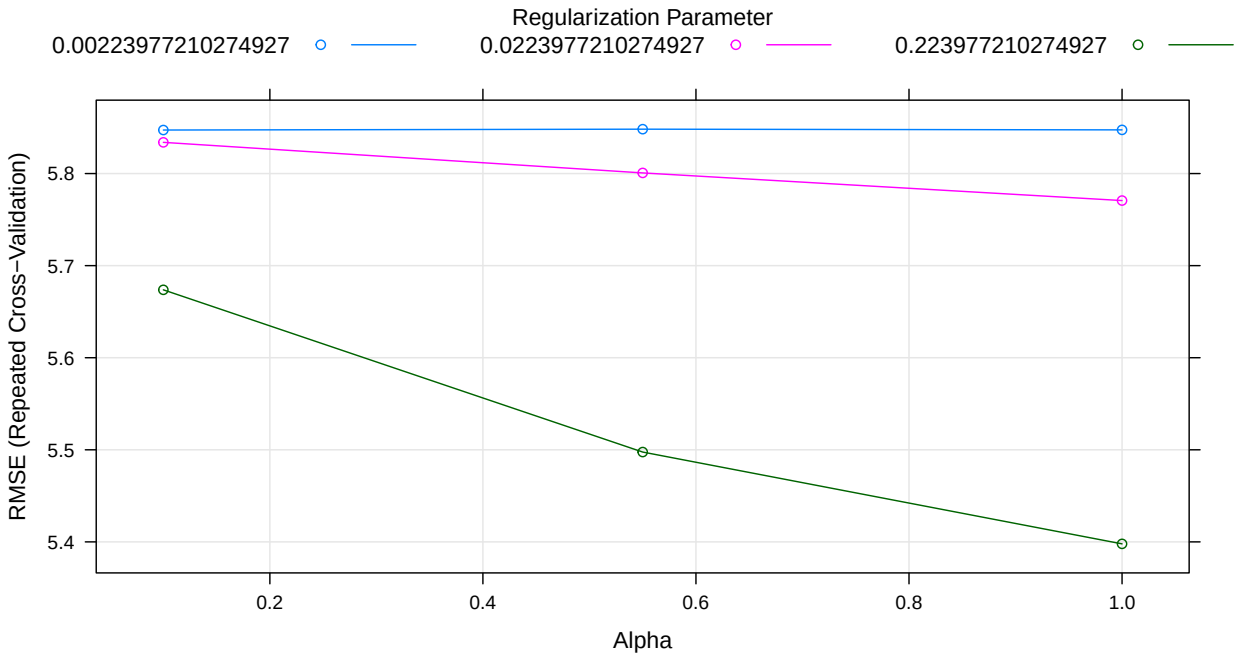

Figure 5: Plot of the tunning process for the Lasso method

This leads to the following performance of fit:

RMSE
3.473

The plot in *Figure* 6 shows which features were selected and the belonging importance of this features. The variable importance were calculated with the function `varImp()` from the R package `caret`. The `varImp()` function uses a ranking method which was used for the selection of features on the subsets. Eventually the average was taken over all subsets to compute the overall value for the final set of most important features. An importance number of 70, means that this variable was taken into account in 70% of the features selection.

In Appendix IV the intercept α and the coefficients β's are shown in a table.

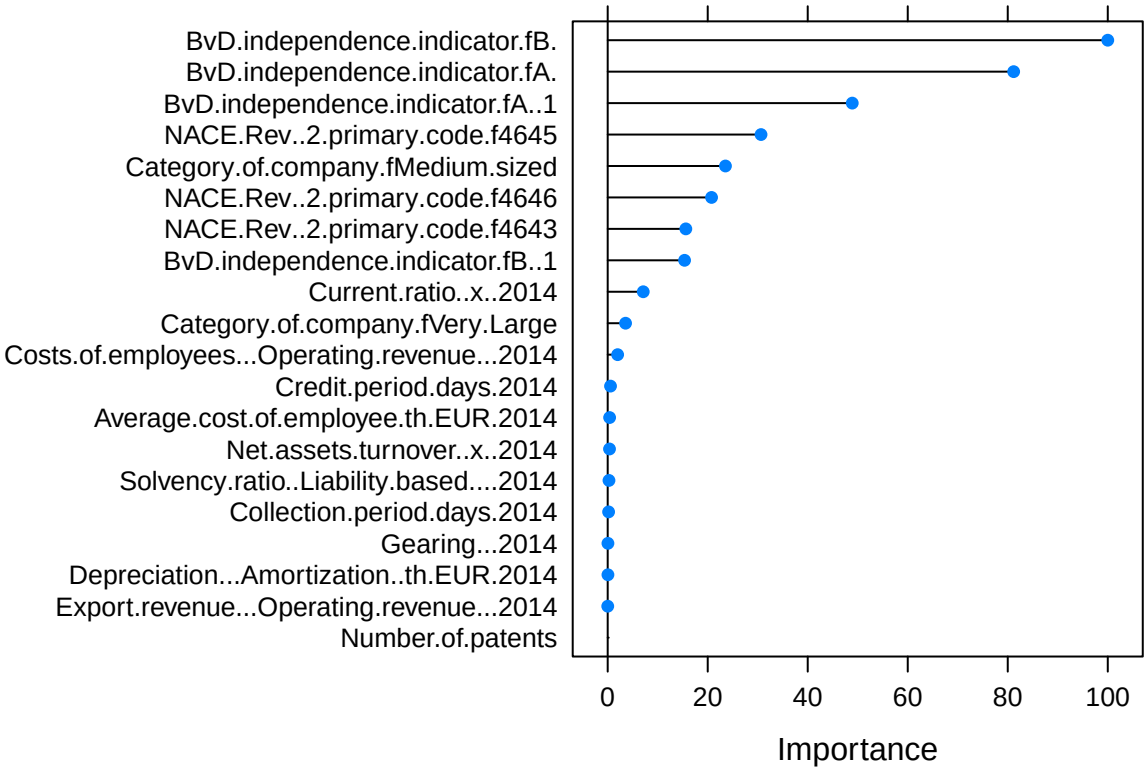

Figure 6: Importance of variables, Lasso method

5.2.2. Random Forest

To make the best predictive model using the random forest algorithm, the randomly selected variables for each individual tree and the number of trees n should be calculated. Due to lack of computer capacity it is unfortunately not possible to select the number of trees and number of randomly selected variables in one run. First an indication of the best number of trees is calculated. By recursively train the random forest model for $n = 1, 5, 10, 100, 500, 1000, 2000, 5000$ the RMSE decreased the most for $n = 2000$ trees, and is therefore chosen as the optimal number of trees.

The random forest model should re-run with the optimal number of trees to find the optimal number of randomly selected variables per tree. The random forest model is trained with the aid of the R package `rf`. The function `train()` from the R package `caret` was used to invoke the crossvalidation and to make sure that the importance of the variables could be reproduced.

```
rf <- train(y ~ ., data = training,
                  method = "rf",
                  ntree = 2000,
                  trControl = crossValidation,
                  importance = TRUE)
```

The optimal number of variables randomly selected per tree is 2. In *Figure* 7 the RMSE is shown for different numbers of randomly selected variables.

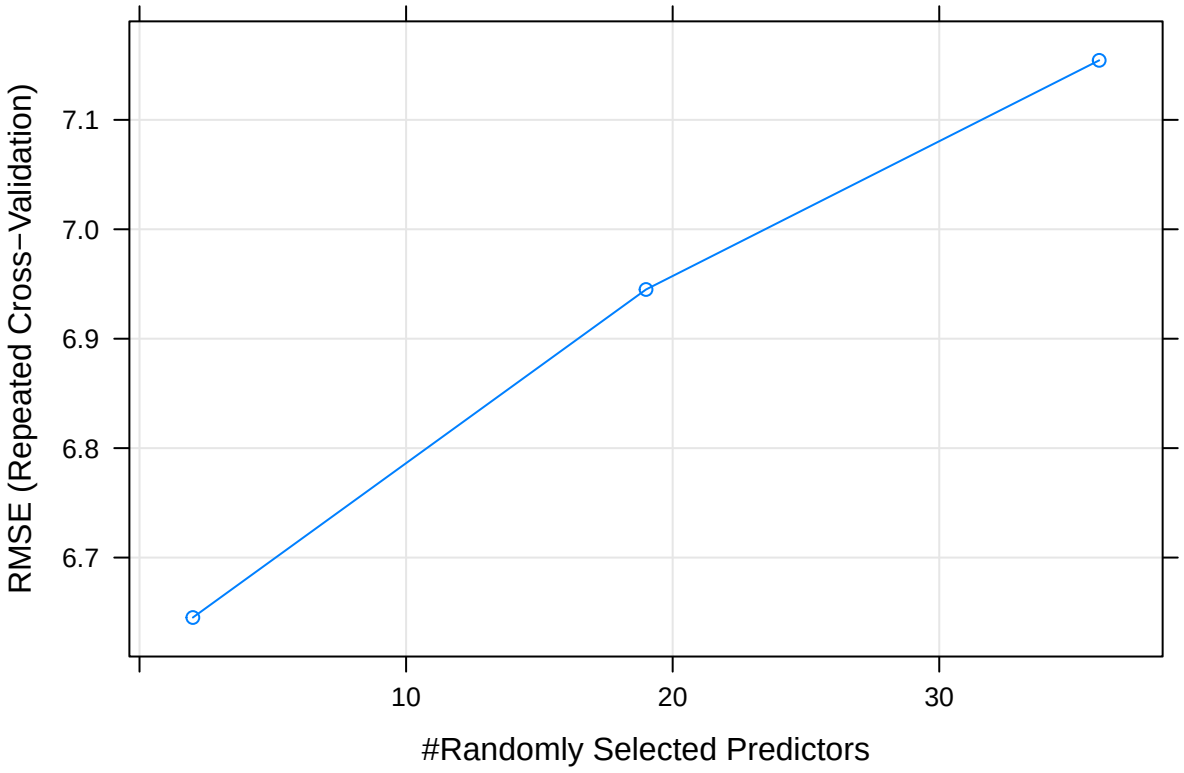

Figure 7: RMSE vs. number of randomly selected variables, random forest

The optimal number of decision trees $n = 2000$ and the optimal number of randomly selected variables per tree 2 leads to the following results:

RMSE
5.987

The selected features and their importance are shown in the *Figure* 8. The variable importance were calculated in the same way and with the same code as the Lasso method.

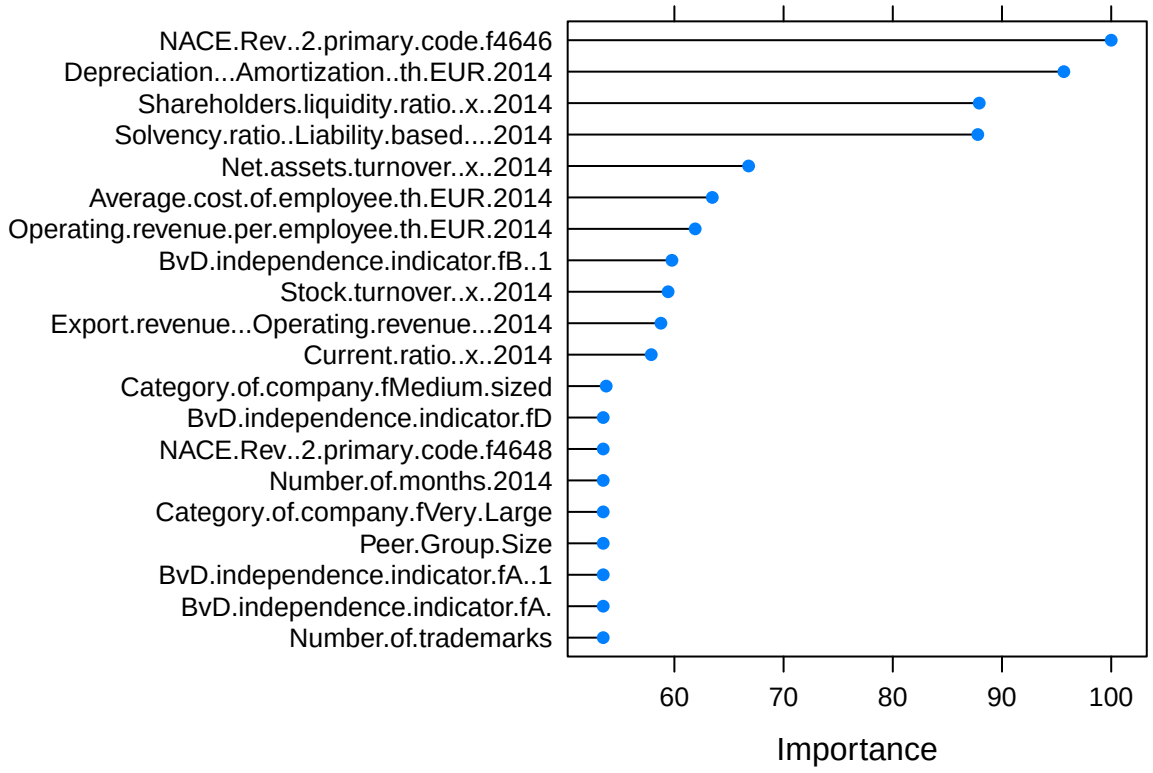

Figure 8: Importance of variables, Random Forest

5.2.3. Gradient Boosting

To implement the gradient boosting one made use of a tuning grid to select the optimal parameters. This grid contained the interaction depth, number of trees, learning rate (shrinkage) and the minimum observations per tree. The actual model was implemented with the aid of the R package `gbm`. By calling the grid in the trainingprocess the best parameters could be selected which decreases the RMSE the most.

```
tuningGridGBM <-  expand.grid(interaction.depth = c(1, 5, 9),
                        n.trees = (1:30)*50,
                        shrinkage = 0.1,
                        n.minobsinnode = c(5,10,20))

gbm <- train(y ~ ., data = training,
                        method = "gbm",
                        trControl = crossValidation,
                        verbose = FALSE,
                        tuneGrid = tuningGridGBM)
```

The best predictive GBM model was fitted with $n = 50$ number of trees, the maximum interaction depth was 5, the learning rate (shrinkage) used was $\lambda = 0.1$ and the best n.minobsinnode (the minimum observations per tree) was 20. One could obtain from the *Figure* 9 that the interaction depth of 5 is slightly better than interaction depth of 9.

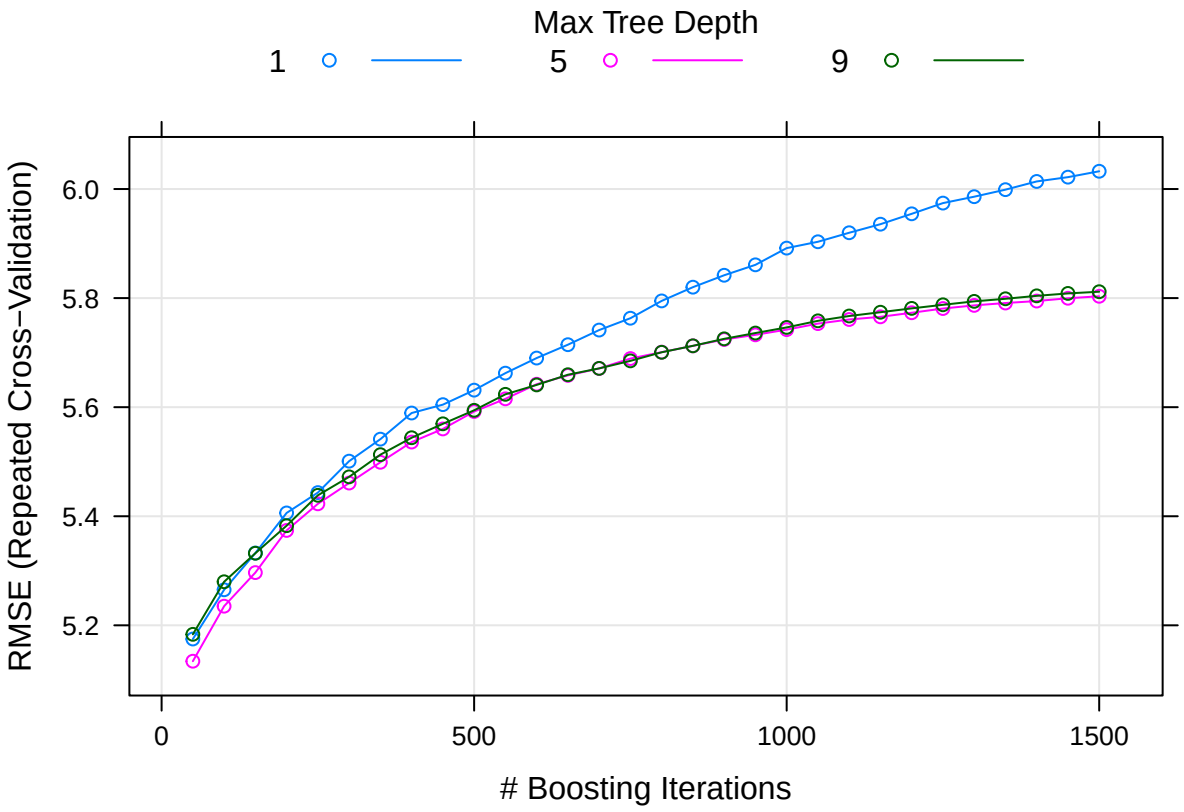

Figure 9: RMSE vs. maximum tree depth, gradient boosting

With these best tuned parameters the following goodness of fit was observed.

RMSE
3.604

The features and there belonging influence are shown in the *Figure* 10. The importance of variables are calculated in the same way as for the Lasso method.

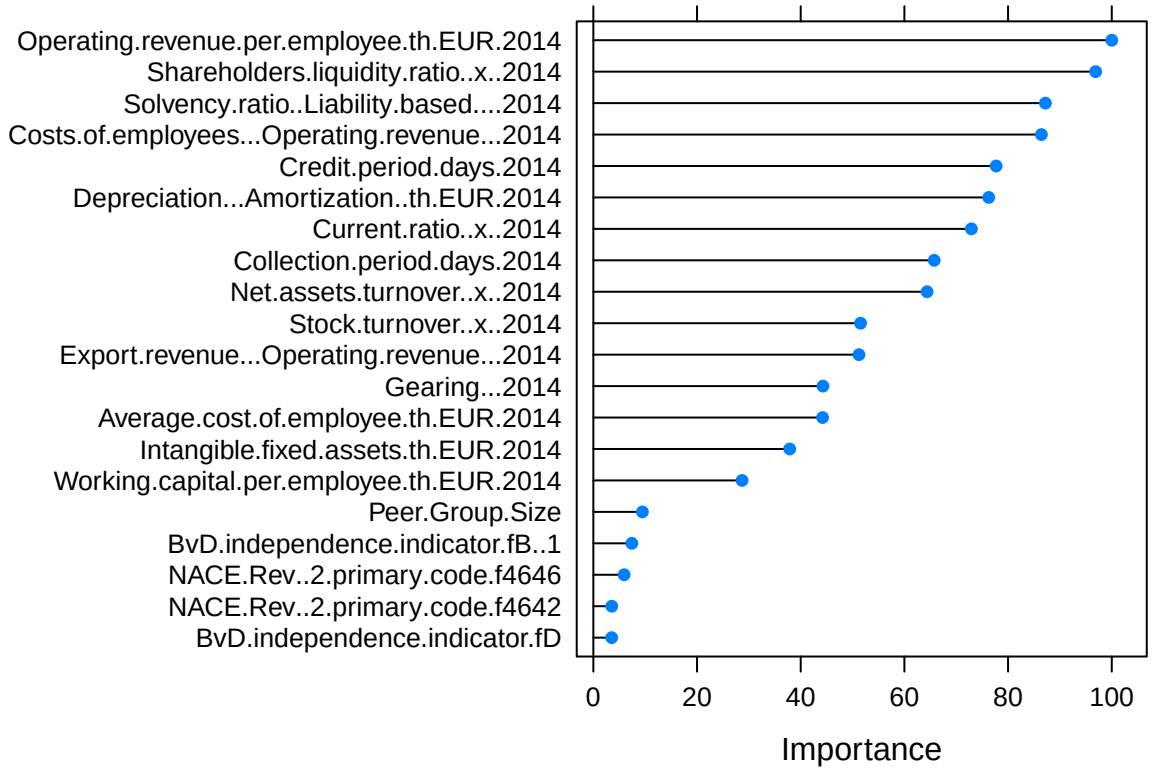

Figure 10: Importance of variables, Gradient Boosting

5.3. Residual Analysis

A residual is the difference between the actual observation and the prediction of a model. The assumptions for a good fitted model is that the residuals are normally distributed. The residuals are calculated by the following formula:

$$r_i = Y_i - \hat{Y}_i$$

with r_i the residual for each observation, Y_i the actual observation and $\hat{Y}_i$ the predicted.

An comprehensive analysis of different plots (see Appendix V) shows that normality for the residuals of the three fitted models seem adequate.

Predicted and Actual EBIT Margin vs. most important variable

A plot where the actual EBIT Margin and predicted values for the EBIT Margin could be plotted against the most important variable. *Figure* 11 shows the actual vs. predicted EBIT Margin against the most important variable: *Operating revenue per employee EUR 2014* for the Gradient Boosting Machine model.

A blue value will indicate that the actual EBIT Margin is less than the predicted EBIT Margin. A red value indicates the other way around, so the actual EBIT Margin is higher than the predicted EBIT Margin. The amount of red and blue dots seem to be equal. More red than blue dots would have indicated that the gradient boosting model predicts the EBIT Margin higher than would be appropriate and a investigation in overfitting would be necesarry. Another observation from *Figure* 11 is that the model founds it difficult to predict for companies with a high variable *Operating revenue per employee, EUR, 2014*. This is probably due to the lack of data for companies with a *Operating revenue per employee, EUR, 2014* above 2000.

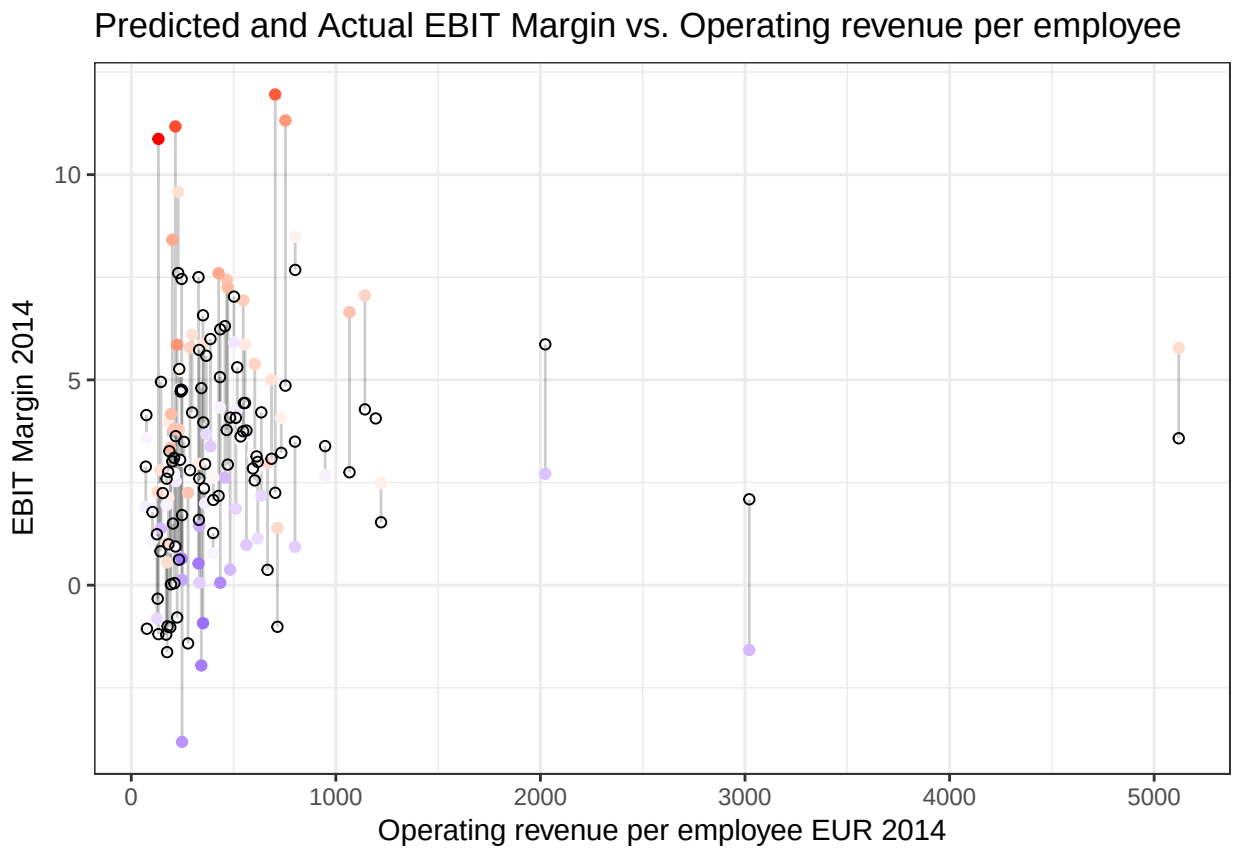

Figure 11: Predicted and Actual EBIT Margin vs. Operating revenue per employee EUR 2014 (Gradient Boosting)

5.4. R Squared

As mentioned in chapter 4, the R^2 might not be a good indicator to explore the goodness of fit of a model when the data does not originate from a normal distribution. The R^2 takes the mean and variance into account and when the data is not normally distributed this will give a misleading conclusion. Since the models that are fitted are working quite well, and they try to minimize the RMSE one could compare the R^2 to see if this gives a good indication of the goodness of fit. The R^2 for each model is calculated and shown in *Table* 6. The $R^2 \in (0, 1]$, for 0 not so good fitting model and 1 a perfect fitted model. The R^2 is negative for the random forest model, while the $R^2 \in (0, 1]$. This is not mathematically impossible due to the fact that the data is probably not normally distributed. One could obtain that the R^2 for all models are quite low and therefore we can emphasize that the the R^2 is not a good indicator for the goodness of fit for this particular dataset and belonging models.

Table 6: R Squared per fitted model

Lasso	Random Forest	Gradient Boosting
0.001045	1.267e-05	0.009687

6. Conclusion

The objective of this research was to find which variables are important for the profitability level of companies.

Nowadays a standard path of steps is followed to compare companies and obtain the transfer price at arm's length to be paid between companies within the same group. The EY Transfer Price department helps companies to set this transfer price at arm's length. Therefore knowing what influences the profitability level of companies would help them to set the most efficient transfer price which would still be at arm's length as required by the OECD.

A dataset of 18867 companies with 47 corresponding variables was deducted from the database Amadeus. Due to missing values, some companies were deleted from the dataset.

The variable EBIT Margin 2014 was chosen as response variable of the profitability level of a company. After plotting a histogram, box plot and QQ-plot assumptions of non normality strengthened. The Shapiro-Wilk Normality test strengthened the assumption and the null-hypothesis was rejected. To find variables that are of importance to predict the profitability level a non-parametric model was chosen.

Since the dataset consisted of 47 variables, the existence of correlation and multicollinearity was expected. Therefore a Spearman Rank correlation test as well as a comprehensive multicollinearitytest (calculating the eigenvalues and VIF[4]) were conducted. By setting a threshold, variables with the highest VIF were excluded till there was no VIF above the threshold.

Three different non-parametric models were trained using 10-fold-crossvalidation on 75% of the dataset.

- Lasso
- Random Forest
- Gradient Boosting Machine

The eventual models were tested on the remaining 25% of the data (the testset). The goodness of fit was tested by the RMSE, since the R^2 is not a good measure for non-normally distributed data. Based on the minimization of the RMSE, the Lasso method was chosen as the best model since the RMSE was the lowest as could be obtained in *Table* 7.

Table 7: RMSE per fitted model

Lasso	Random Forest	Gradient Boosting
3.473	5.987	3.604

The residuals of the three different models were analysed and seemed to be normally distributed. No clear patterns were observed from plots between the residuals and the predicted profitability levels.

The EY transfer pricing department was specifically interested which variables influences the profitability level of companies. The more machine learning based algorithms (random forest and gradient boosting machine) came up with variables which were expected to be influential. However, the Lasso method did find other important features.

From the Lasso method the conclusion could be made that the following variables are influential on the profitability level EBIT Margin 2014:

- The variable *BvD independence indicator* is a character given by Bureau van Dijk (BvD) to identify the degree of independence of each company. There are five possible degrees: A, B, C, D, U which are transformed into a dummy variable.
- The variable *NACE Reverence primary code*, which gives an indication of the type of wholesale (e.g. textiles, household goods)

[4]VIF stands for Variance Inflation Factor

- The variable *Category of company*, gives the size category per company. The categories are: very large, large, medium and small. (E.g. "very large" stands for a company with an operating revenue bigger than 100 MIL Euro, total assets above 200 MIL Euro and more than 1000 employees.)
- The variable *Current ratio*: is a ratio which measures the ability to pay it short term as well as it long term liabilities. The ratio is the current assets divided by the current liabilities.

The EY Transfer Pricing department uses the following five comparability factors (based on the OECD Guidelines) to compare independent companies:

1. characteristic of the property or service transferred
2. functions performed by the parties taking into account assets employed and risk assumed, in short, termed as functional analysis (FAR)
3. contractual terms
4. economic circumstances
5. business strategies pursued

In this research variables from the 1, 2, 4 and 5 comparibility factors were taken into account to obtain a predictive model.

The most important variables from the lasso method reflect the comparibility factors. The variable *NACE Reverence primary code* is part of the first comparibility factor. The variable *BvD independence indicator* could be seen as the second comparability factor. The variables *Category of company* and *Current ratio* reflect the economic circumstances (the fourth comparibility factor).

Summarized, the independence, industry, size and financial stability of the companies are important to predict the profitability level EBIT Margin and reflect the comparibility factors denoted by the OECD. This research could help the EY Transfer Pricing department by setting the most optimal transfer price.

7. Recommendations

In this research the dataset contained missing values. There was chosen to delete companies with missing values, since the lasso method and random forest could not handle missing values. The gradient boosting on the other hand could deal with missing values. Since EY Transfer Pricing uses the actual dataset from the database Amadeus and deals with missing values as well, the gradient boosting algorithm could handle this data.

The random forest algorithm was slow in training the model (double the time of the two other models together) and did not obtain the best results. Unless higher computer power is available and the results will be better, this algorithm is not recommended to use for this research.

A recommendation for the dataset is to change the variable location which is now denoted by country to the variable region (e.g. Western Europe). There was quite a large collection of countries but not so many observations per country. While the variable region will reduce the collection of regions but will enlarge the number of observations.

On the transfer price department they did not use Machine Learning algorithms or other high dimensional based models to predict the importance of variables for the profitability level. The EY transfer price department is curious in further implementations and offered an internship at their department.

Appendix I - Search Strategy

Amadeus provide the option to set your own search strategy to select data. The search strategy which is used to select the dataset in Amadeus is based on the following steps and in consultation with the Transfer Pricing department:

1. Only companies located in the 28 members of the European Union, Iceland, Norway and Switzerland are accepted.
2. The NACE Rev. code is an industry classifier. For this research only NACE Rev. codes 464 (wholesale of household goods) are taken into account. The wholesale of household goods (NACE Rev. code 464) could be divided in 9 sub-categories (e.g. 4641 = Wholesale of textiles, 4642 = Wholesale of clothing and footwear). Note: all NACE Rev. codes could be found on Amadeus.
3. Companies should be owned by at least one shareholder, of one of the following types: Banks and Financial companies, Insurance companies, Industrial companies, Private Equity firms, Hedge funds, Venture capital, Mutual & Pension Funds/Nominees/Trusts/Trustees, Foundations/Research Institutes, Public authorities, States, Governments, owning between 25% and 100%
4. Companies should own at least one subsidiary, owned between 25% and 100%, given by the Operating Revenue.
5. The variable Operating revenue (Turnover) in Euro's must have a minimum of 0 and be in the data for 2014.
6. The variable Operating Profit/Loss [=EBIT]: should be a known value for 2014.
7. The type of accounts for each company should be: U1 (companies with unconsolidated accounts only).
8. The status of each company selected should be active.

Appendix II - Excluded Variables

After calculating the eigenvalues and the Variance Inflation factors, multicollinearity could be removed by setting a threshold. For this research 13 variables were removed. *Table* 8 shows the variables that are excluded for this research due to multicollinearity.

Table 8: Excluded variables due to multicollinearity

k.excluded
Solvency.ratio..Asset.based. . . .2014
EBITDA.th.EUR.2014
Total.assets.th.EUR.2014
Liquidity.ratio..x..2014
Total.assets.per.employee.th.EUR.2014
Operating.revenue..Turnover..th.EUR.2014
Number.of.employees.2014
Costs.of.employees.th.EUR.2014
Tangible.fixed.assets.th.EUR.2014
Cash.flow.th.EUR.2014
NACE.Rev..2.primary.code.f4649
Working.capital.th.EUR.2014
BvD.independence.indicator.fU

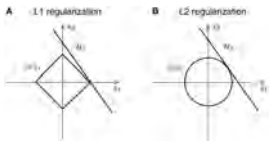

Figure 12: Jianing V. Shi (2013) p. 4

Appendix III - Vector Norm

There are two norms two measure the magnitude from a vector, Tibshirani (1996). Assume vector is:

$$V = (V_0, \ldots, V_n)^T$$

$$\ell_1\text{-norm} = ||\beta||_1 = \sum_{i=1}^{n} |\beta_i|$$

$$\ell_2\text{-norm} = ||\beta||_2 = \sqrt{\sum_{i=1}^{n} \beta_i}$$

for $i = 1, \ldots, n$.

Figure 12 visually shows the regularization for the ℓ_1-norm and the ℓ_2-norm by setting the vector equal to the line H_0. The ℓ_1-norm gives the shape of a diamond while the ℓ_2-norm has the shape of a circle. This means that setting the vector equal to H_0 line with the aid of the ℓ_1-norm the variable x_1 will be a number while the variable x_2 will be equal to zero. In this way it selects the most important features and will set less important features to zero. Note that Lasso uses the ℓ_1-norm and stands for Least Absolute Shrinkage and Selection Operator, so it selects the operators by the absolute value.

The ℓ_2-norm is used in Ordinary Least Squared and ridge regression and looks for the regression line which touches the circle. Therefore it will give some importance to both coefficients x_1 and x_2.

Ordinary Least Squares

Ordinary Least Squares (OLS) is a estimator to find the coefficients in linear regression. It sets the coefficients by minimizing the difference between the response variables and the predicted based on ℓ_2-norm. In OLS the following formula is used to find the coefficients:

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{N} ||y - X\beta||_2^2 \right\}$$

Appendix IV - Coefficients of Lasso

Table 9 shows the intercept (α) and the belonging coefficients (β's) of the lasso method.

Table 9: Intercept and coefficients of lasso

	X1
(Intercept)	5.852
Peer.Group.Size	0
Number.of.months.2014	0
Intangible.fixed.assets.th.EUR.2014	0
Depreciation. . . Amortization..th.EUR.2014	-0.001865
Net.assets.turnover..x..2014	-0.01495
Stock.turnover..x..2014	0
Collection.period.days.2014	0.0074
Credit.period.days.2014	-0.02344
Export.revenue. . . Operating.revenue. . . 2014	0.0002521
Current.ratio..x..2014	-0.2954
Shareholders.liquidity.ratio..x..2014	0
Solvency.ratio..Liability.based. . . .2014	0.01058
Gearing. . . 2014	-0.001957
Operating.revenue.per.employee.th.EUR.2014	0
Costs.of.employees. . . Operating.revenue. . . 2014	-0.08247
Average.cost.of.employee.th.EUR.2014	0.01582
Working.capital.per.employee.th.EUR.2014	0
Number.of.patents	0
Number.of.trademarks	0
Country.fGermany	0
Country.fHungary	0
Country.fUnited.Kingdom	0
NACE.Rev..2.primary.code.f4642	0
NACE.Rev..2.primary.code.f4643	-0.6492
NACE.Rev..2.primary.code.f4644	0
NACE.Rev..2.primary.code.f4645	1.274
NACE.Rev..2.primary.code.f4646	0.8632
NACE.Rev..2.primary.code.f4647	0
NACE.Rev..2.primary.code.f4648	0
Category.of.company.fMedium.sized	-0.9786
Category.of.company.fVery.Large	-0.1485
BvD.independence.indicator.fA.	3.374
BvD.independence.indicator.fA..1	-2.032
BvD.independence.indicator.fB.	-4.156
BvD.independence.indicator.fB..1	-0.6388
BvD.independence.indicator.fD	0

Appendix V - Residuals per model

- QQ Plot of residuals

A normal QQ plot of residuals should follow the diagonal line to indicate if the residuals are normally distributed.

- Residuals vs. Predicted

A good fitted model will not show a clear patterns between the the residuals against the predicted values.

Residuals Lasso

Figure 13 show the residuals of the Lasso method, which seem to be normally distributed on behalve of 1 outlier. The residuals are on average smaller than for the other models, except for the outlier. The residuals versus the predicted does not show a clear pattern.

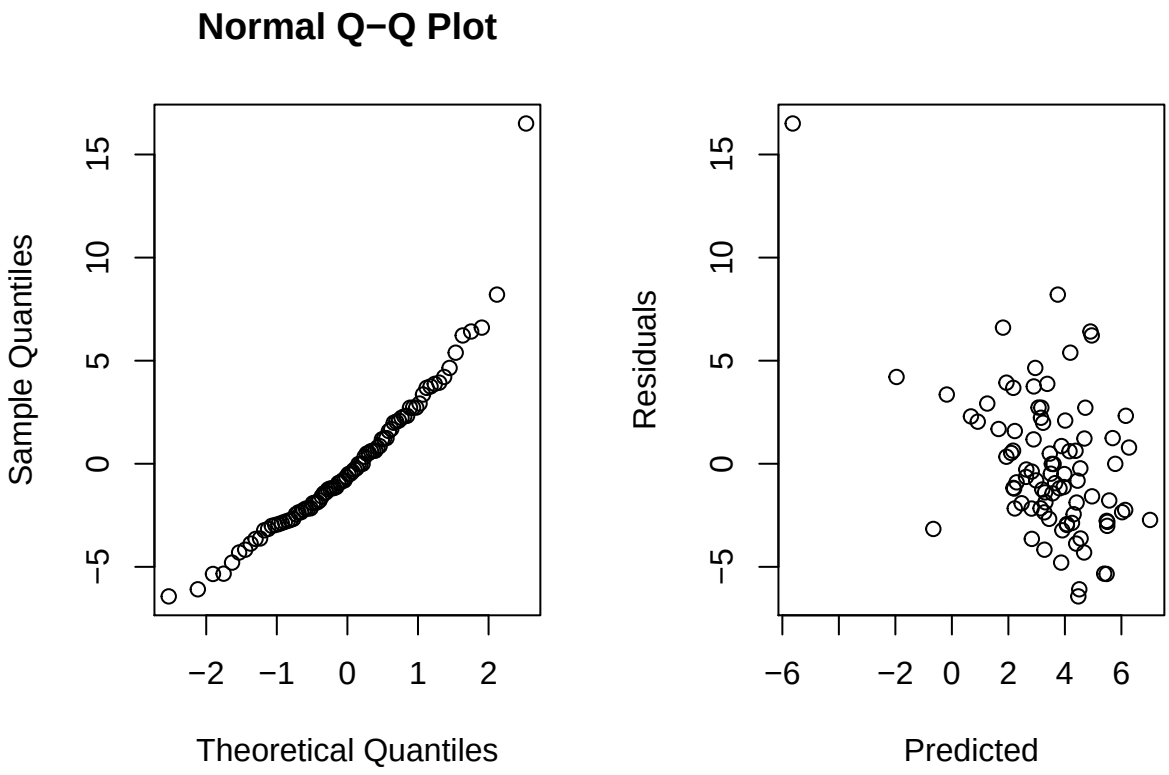

Figure 13: a. Normal QQ Plot for residuals of lasso, b. Residuals vs. fitted

Residuals Random Forest

The QQ-plot, *Figure* 14, seem to follow the diagonal line, so it is plausible that the residuals are normally distributed. The residuals vs. the predicted do not show a pattern.

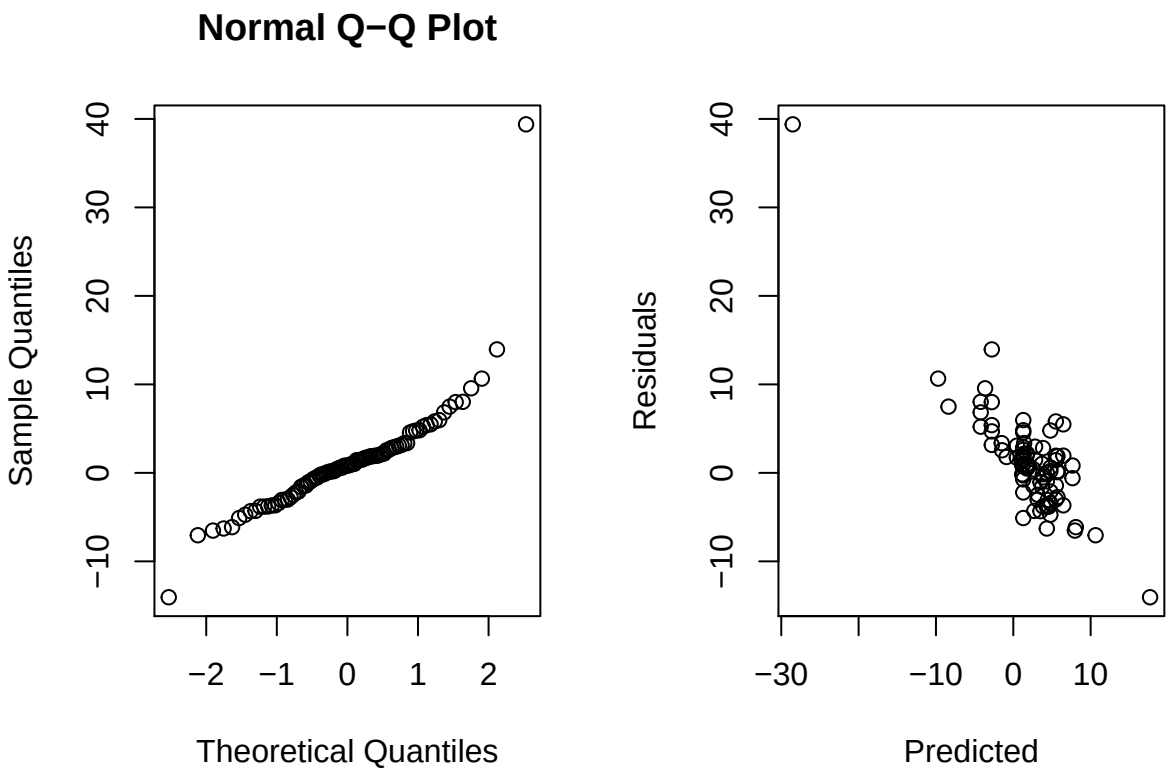

Figure 14: a. Normal QQ Plot for residuals of random forest, b. Residuals vs. fitted

Residuals Gradient Boosting

The residuals of the gradient boosting, *Figure* 15, seem to be normally distributed. The QQ plot shows a diagonal line and the residuals vs. the predicted does not show any pattern.

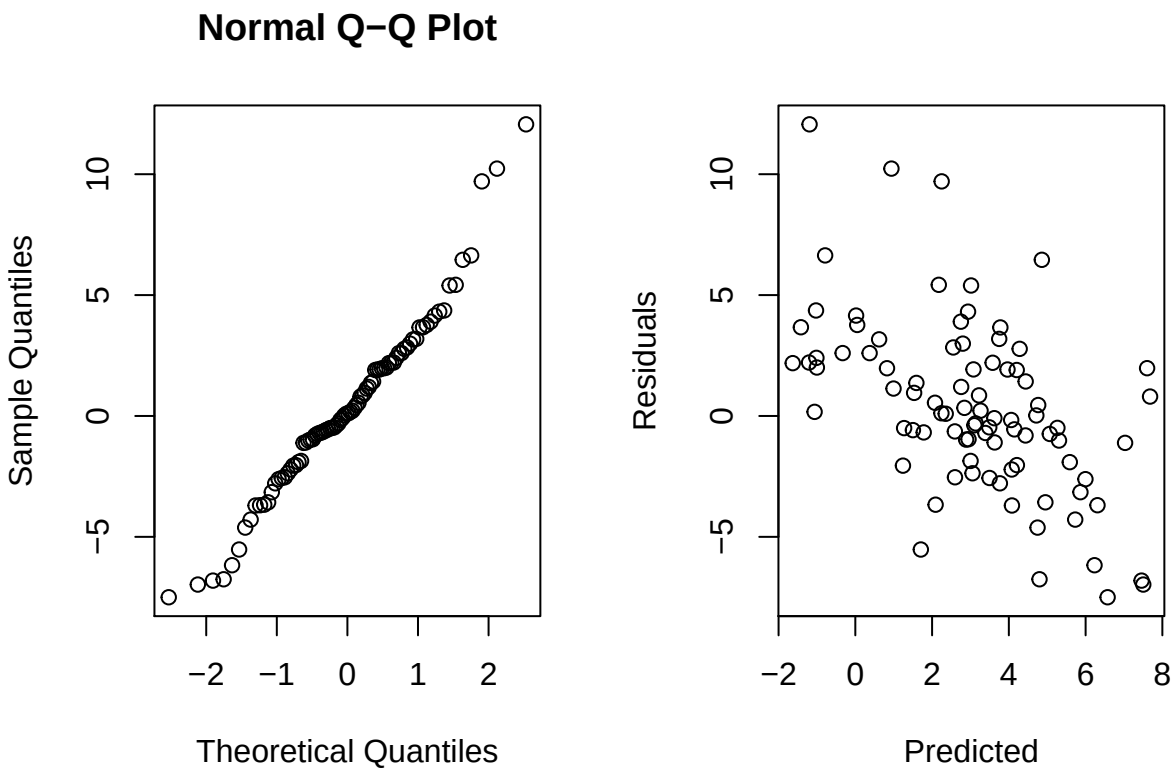

Figure 15: a. Normal QQ Plot for residuals of gradient boosting, b. Residuals vs. fitted

References

Berk, Jonathan, and Peter DeMarzo. 2013. "Corporate Finance." In, 30. Pearson.

Bijma, F. 2015. "Statistical Data Analysis." Department of Mathematics, Faculty of Sciences, VU Amsterdam.

Breiman, Leo. 1984. "Classification Algorithms and Regression Trees." In, 263. https://rafalab.github.io/pages/649/section-11.pdf.

———. 2001. "Random Forests." *Machine Learning* 45 (October). https://link.springer.com/article/10.1023%2FA%3A1010933404324: 15.

Buijs, Arie. 2008. "Statistiek Om Mee Te Werken." In, 377. Noordhoff Uitgevers.

Chen, Tianqi. 2014. "Introduction to Boosted Trees." October. https://homes.cs.washington.edu/~tqchen/pdf/BoostedTree.pdf.

Crespo, Gonc Alo, and Richard A. Clark. 2011. "Analyzing the Determinants of Profitability: Evidence from European Distributors." *Tax Management Inc.* 20 (March): 528–35.

Cutler, Adele. 2003. "Trees and Random Forests." http://www.math.usu.edu/adele/RandomForests/UofU2013.pdf.

EY Tax, Services. 2016. "Tax Services Transfer Pricing." http://www.ey.com/uk/en/services/tax/transfer-pricing-and-operating-model-effectiveness/tax---transfer-pricing-practice.

Faraway, Julian J. 2002. "Practical Regression and Anova Using R." In, 117.

Flach, Peter. 2012. "Machine Learning, the Art and Science of Algorithms That Make Sense of Data." In, 349–50. Cambridge.

Ford, Clay. 2015. "Is R-Squared Useless?" http://data.library.virginia.edu/is-r-squared-useless/.

Frigessi, A., P. Bahlmann, I.K. Glad, M. Langaas, S. Richardson, and M. Vannucci. 2014. *Statistical Analysis for Highdimensional Data*. Springer.

Glad, and Richardson. 2016. "Preselection in Lasso-Type Analysis for Ultra-High Dimensional Genomic Exploration." Edited by Springer.

Gunst, M.C.M. de. 2013. "Statistical Models." Department of Mathematics, Faculty of Sciences, VU University Amsterdam.

Hastie, Trevor, and Junyang Qian. 2014. "Glmnet Vignette." https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html.

Jianing V. Shi, R. Theodore Smith, Jim Wielaard. 2013. "Perceptual Decision Making 'Through the Eyes' of a Large-Scale Neural Model of V1." *Frontiers in Psychology*, 1–12.

Kim, Chul. 2015. "Deep Learning Vs. Random Forest." https://www.chulkim.net/single-post/2015/06/01/Deep-Learning-vs-Random-Forest.

Li, Cheng. 2016. "A Gentle Introduction to Gradient Boosting." http://www.chengli.io/tutorials/gradient_boosting.pdf.

Natekin, Alexey, and Alois Knoll. 2013. "Gradient Boosting Machines." In. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3885826/#B19.

OECD. 2010. *OECD Transfer Pricing Guidelines for Multinational Enterprises and Tax Administrations 2010*. OECD Publishing.

Stucken, Bastiaan. 2007. "Addressing Transfer Pricing Issues Using Quantitative Methods." Master's thesis, Erasmus University Rotterdam.

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society* 58: 267–88.