universität wien

MASTERARBEIT

Titel der Masterarbeit

„Effectivness of machine learning algorithms in predicting insolvencies"

Verfasser

Hubert Tchorzewski

angestrebter akademischer Grad

Master of Science (MSc.)

Wien, 2014

Studienkennzahl lt. Studienblatt:	A 066 920
Studienrichtung lt. Studienblatt:	Quantitative Economics, Management and Finance
Betreuer:	Univ.-Prof. Gyöngyi Lóránth

Acknowledgments

I would like to extend my sincere thanks to the people without whom the creation of the following work would not have been possible.

First of all, to univ.-prof. Gyöngyi Lóránth for the patience and support with the theory of insolvence.

To univ.-prof. Marcus Hudec Friends for ensuring that the code is doing what it is supposed to be doing.

To all of my friends for keeping my spirit up in the direst of times.

Last, but not least, to my family for the everlasting support.

Thank you!
Dziękuję!
Danke!

Abstract

The task of predicting insolvencies is important for both policy makers and companies alike. There exists substantial research in this area focused on creating models that could identify troubled firms ahead of time or analyze the changes in default probabilities. There also exists a growing interest in machine learning classifiers that could potential help future development in this area. This study uses the Compustat database to compare the effectiveness of both approaches by investigating their corresponding goodness of fit statistics, specificity and sensitivity rates. Discussed are the logistic regression, decision trees, random forests, naive Bayes, k nearest neighbors and support vector machines. The research has found that each method is highly effective and has different strengths.

Keywords

logistic regression, machine learning, decision tree, random forest, naive Bayes, k nearest neighbor, support vector machines, corporate insolvency

Zusammenfassung

Die Aufgabe, Insolvenzen vorherzusagen ist sehr wichtig sowohl für politische Entscheidungsträger als auch für Unternehmen selbst. Es existieren umfangreiche Studien auf diesem Gebiet, die sich auf auf das Erstellen von Modellen fokussieren, die solche Firmen voraus identifizieren können und die Wahrscheinlichkeit einer Insolvenz untersuchen. Es besteht auch ein wachsendes Interesse an Klassifikationsverfahren, die auf Machine Learning-Algorithmen aufgebaut sind, die als potenzielle Hilfsmittel für zukünftige Entwicklung in diesem Bereich dienen könnten. Diese Studie benutzt den Compustat-Datensatz, um die Effektivität von beiden Ansätzen auf Basis ihrer entsprechenden Anpassungsgüte, Sensitivität und Spezifität vergleicht. Die Methoden der logistischen Regression, Entscheidungsbäumen, Random Forests, Bayes-Klassifikatoren, Nächste-Nachbarn-Klassifikationen und Support Vector Machines werden angewendet. Diese Studie hat herausgestellt, dass jede Methode sehr effektiv ist und verschiedene Stärken aufweist.

Stichwörter

Logistische Regression, Machine Learning, Entscheidungsbaum, Random Forest, Bayes-Klassifikator, Nächste-Nachbarn-Klassifikation, Support Vector Machines, Insolvenz

Contents

Acknowledgments ii

Abstract iii

Chapter 1. Introduction 2

Chapter 2. Literature Review 4
2.1 On the difference between machine learning and regression analysis 4
2.2 Studies on insolvencies 5
2.2.1 Regression analysis 6
2.2.2 Machine Learning 8
2.2.3 Comparison of machine learning and regression analysis 10

Chapter 3. Data 13
3.1 Database Description 13
3.2 Financial Ratios and Descriptive Statistics 14
3.3 Balancing the Dataset 16

Chapter 4. Research Methods 20
4.1 Regression Analysis 21
4.1.1 Weight of Evidence transformation 24
4.2 Machine Learning Algorithms 24
4.2.1 Decision Trees 25
4.2.2 Random Forest 27
4.2.3 Naïve Bayes 29
4.2.4 k- Nearest Neighbor 32
4.2.5 Support Vector Machines 33
4.3 Comparison statistics 35

Chapter 5. Results 38
5.1 Logit 38
5.2 Decision Tree 41
5.3 Random Forest 43
5.4 k Nearest Neighbor 45
5.5 Naïve Bayes 46
5.6 Support Vector Machines 48
5.7 Discussion of predictive power of the methodologies 49

Chapter 6. Conclusion 52

Bibliography 55

Appendix 1. R Code 59

Appendix 2. Figures 65

Curriculum Vitae 81

List of Tables

3.1 Explanatory Variables, Definitions, Relevant Research Examples . . . 15
3.4 Prior Bankrupcy Proportions by Years 16
3.5 Total Insolvencies in Populations 17
3.2 Main Data Characteristics of the Variables 18
3.3 Main Data Characteristics of the Variables by the Dependent 19

4.1 Classification Table 36

5.1 Characteristics of the Logit Model 40
5.2 Classification Table of the Logit Model for the Test Set 40
5.3 Predictive Power of the Logit Model 41
5.4 Classification Table of the Decision Trees 42
5.5 Predictive Power of the Decision Trees 43
5.6 Variable Importance of the Random Forest 44
5.7 Classification Table of the Random Forest for the Test Set 45
5.8 Classification Table of the kNN for the Test Set 46
5.9 Predictive Power of kNN 46
5.10 Classification Table of the Parametric Bayes for the Test Set 47
5.11 Classification Table of the Non- Parametric Bayes for the Test Set . . 47
5.12 Predictive Power of the Naïve Bayes 48
5.13 Classification Table of the SVM for the Test Set 48
5.14 Predictive Power of the SVM 48
5.15 Summary of Predictive Power 51

List of Figures

6.1 Denisities of Variables 'x1' – 'x6' 65
6.2 Denisities of Variables 'x7' – 'x12' 66
6.3 Denisities of Variables 'x13' – 'x18' 67
6.4 Denisities of Variables 'x19' – 'x24' 68
6.5 Denisities of Variables 'x25' – 'x26' 69
6.6 Boxplot Distribution Plots for Variables 'x1' – 'x4' 69
6.7 Boxplot Distribution Plots for Variables 'x5' – 'x10' 70
6.8 Boxplot Distribution Plots for Variables 'x11' – 'x16' 71
6.9 Boxplot Distribution Plots for Variables 'x17' – 'x22' 72
6.10 Boxplot Distribution Plots for Variables 'x23' – 'x26' 73
6.11 Comparison of Gini and Information Index Values 73
6.12 Receiver Operator Curves 74
6.13 Decision Trees 75
6.14 Variable Importance Plot for Random Forest 76
6.15 Fitted Distributions of Variables 'x1' – 'x3' 76
6.16 Fitted Distributions of Variables 'x4' – 'x9' 77
6.17 Fitted Distributions of Variables 'x10' – 'x16' 78
6.18 Fitted Distributions of Variables 'x17' – 'x22' 79
6.19 Fitted Distributions of Variables 'x23' – 'x26' 80

Chapter 1

Introduction

The topic of predicting bankruptcies within industries is one of the key research areas in the current economic environment. While experts still argue whether the financial crisis of 2008 has already been averted in the US, the Federal Reserve is still unclear whether the accommodative policy in the form of the quantitative easing should be continued or not [45, 50]. At the same time in Europe there is no foreseeable end to the raging debt and austerity crisis[1]. There is no arguing that the economic downturn takes its toll on the operating enterprises and their ability to generate profit and remain solvent. There is persistent interest in the ability to forecast which companies might be experiencing troubles regarding their profitability in the foreseeable future. However, generating losses has been proved to be not sufficient for a company to decide to leave a market. In fact, bankrupt companies may persist on the market due to regulatory reasons or as going concerns [18] which emphasizes the need for research on the corresponding determinants.

Therefore, forecasting the probabilities of default is one of the most pressing issues in the financial markets. It is of high importance to stockholders such as shareholders, creditors, customers, employees, suppliers or auditors, just to name a few [23]. Considering the importance of financial data for relying information about the state of the company, numerous researchers investigated the connection between balance sheet data and insolvency probabilities. Numerous potential classifiers have been developed, one after another (See Chapter 2). However, only few of them try to comprehensively analyze the effectiveness of these models in comparison to different

[1] At the moment of writing, January 2014.

classifying techniques. Thus, the purpose of this study is to assess the effectiveness of two popular approaches used for classification purposes.

Methods used in this research stem mostly from the field of machine learning. These methods range from simple nearest neighbor classifiers, through parametric and non parametric Bayes predictors, decision trees, random forests and finally support vector machines. However, this study is not restricted to solely analyzing algorithms coming from the field of data mining. In order to provide comparison to regression analysis an additional logistic model is constructed. The exact methodology and examples of studies following these models are discussed in Chapter 2.

This research is directed at measuring the predictive power of the aforementioned statistical modeling techniques. That is, to compare the necessary assumptions regarding the data structure and theoretical background behind them, as well as to obtain necessary diagnostic statistics and compare the returned goodness of fit measures (where applicable). Due to the binary structure of the problem; rates of sensitivity, sensitivity and the corresponding receiver operating curves can be easily analyzed and directly compared, resulting in easily explainable conclusions.

The structure of this study is as follows. In Chapter 2 a review of existing research in the field of forecasting insolvencies is presented. The review focuses on more recent studies in the field. In addition, it also includes a brief discussion on the main differences regarding the different data analysis approaches. Chapter 3 then describes the data and variables that are used in this research. It provides necessary definitions and explains how the data has been transformed. Chapter 4 provides an in depth discussion of the used methodologies and presents the essential mathematical part of the research. Finally, in Chapter 5 the results of applying the methods are presented, followed by a discussion in which the differences in predictive power are highlighted and reviewed. Chapter 6 provides concluding remarks. In addition, the necessary R code to perform or to reproduce the analysis is included in Appendix 1. All figures are grouped together in Appendix 2.

Chapter 2

Literature Review

In this chapter we intend to provide a discussion on the existing literature in the field of predicting insolvencies and outline some essential writings that were relevant for the creation of this study. Elementary discussion of the necessary assumptions and formulated definitions is being presented. In addition, a brief review of some comparison studies is presented. Particular attention will be given to the methods which are later described in the Methodology chapter.

2.1 On the difference between machine learning and regression analysis

There are two approaches regarding data analysis. The dominant approach in the field of statistics comes from the presumption that the relation between the analyzed phenomena can be estimated following certain models [9]. The model is then verified following certain testing procedures and an analysis of fitted statistical properties. The other approach would be to abandon any assumptions on the nature of the relation between data and response and restrict the analysis only to the creation of accurate predictors that do not require any additional assumptions on data characteristics. This would be the way of machine learning. A field that finds its roots in computer science, where it was first used to obtain descriptive statistics. Descriptive statistics do not take any assumptions, they simply exist to convey information about the database. However, it has been noted that creating descriptions of increasing detail can lead to quite efficient estimations of the whole data, much alike

a map of an appropriate scale. In turn, this lead to an ever increasing interest in creating algorithms that would generate said statistics and be able to predict responses for future data. This approach is in strict contradiction with the general rule: ‚Correlation does not imply causation', but is nonetheless effective. One one side, no assumptions are required for predictions following this methodology. On the other, it is a complete black box as we have no insight in what rules over the relation between data and response.

However, modeling approaches also quite often lead to dilemmas where one has to compare numerous different models describing the same relation, but yielding complete opposite results. Deciding to value one model over the other often stands on fragile foundations [9]. Logistic regressions are widely accepted for classification task, yet no one actually believes in the assumption of multivariate normality [9]. In the end, with the modeling approach we just try to explain how the world behaves by fitting data to mathematical equations. It is well known that with enough variables every model will yield good results. Every econometrics textbook emphasized that the identification of correct predictors is the most important part of each regression analysis [20, 30], but in the end no other choice than semi-arbitrary variable selection is being offered. Of course, there exist numerous testing procedures for identification of significant relations, but these test impose even more assumptions on data. If these assumptions are violated we still have a predictor, maybe even an accurate one, but no information on the nature of the relation between data and response. This becomes even more evident when applying a model to data which should follow the same principle (e. g. company characteristics from one year $t_1 = 2001$ and another $t_2 = 2002$), but produce different coefficients than the original model. And even then, the variables that are to be considered at the start of the feature selection process is always restricted by the initial choice of potential predictors.

2.2 Studies on insolvencies

Early detection of financial distress and the study of reasons behind insolvencies is an important concern for many regulatory agencies [11]. Regression analysis methods are still very much popular and numerous researchers deem them sufficiently good

to base their insolvency analysis on them [15, 33, 47, 31, 24]. However, a wide array of machine learning techniques has also been found to be an effective tool for forecasting insolvencies. These method range from rule association [11, 28], SVM [21, 22], Bayesian interference [37, 28] or random forests [26, 47, 28, 41, 10]. Other than that, decision trees have been very popular in a growing area of research, ranging from credit scoring, through credit fraud detection to marketing research [14, 39]. The empirical research also points out the arbitrary structure of financial statements, as due to the fact that the fiscal year only rarely corresponds to a calendar year which means that possible shifts and lags are expected to occur in the dataset. Another issue is the problem of "creative accounting", but this should only be valid for outliers and therefore is expected by statistical methods. However, should it be a significant aspect then it would result in an increase of the error terms and lower the overall predictive power.

2.2.1 Regression analysis

When discussing insolvency predictions the work of Altman and Sabato (2007) [4] must be mentioned. Altman, who is the author of the most prominent classifier, the Z- Score [2, 3] has a well established place in the field of corporate financial predictions. In this newer study, the researchers focus on a sample of American companies and try to identify potential more suitable predictors than the ones used in the Z- Score method. The model uses the logistic regression approach with several financial ratios considered as explanatory variables. They conclude that the logistic regression is a classifier of high accuracy.

In the case of European countries, Dzidzeviciute (2010) [15] considers an analysis based on the logistic regression methodology to create an insolvency prediction model for Lithuanian companies. There are a few things to note related to this research. Firstly, in order to cope with the unbalanced data structure, as solvent companies drastically outnumbered insolvent ones, a simple random sampling without replacement process was used. Secondly, in order to obtain better predictive power a weight of evidence transformation was applied on quantiles of continuous variables followed by creating the according information values. Unfortunately, the study does not include any form of validation and reports only the model accuracy

based on the training sample. It would also seem that the researcher did not inspect any potential non linear correlation of the explanatory variables.

Another example for applying a logistic model to forecast company insolvencies was conducted by Nikolic *et al.* (2013) [33]. The researchers focused on data originating from the Serbian financial statements. Here the researchers decided to take an unusual way of dealing with the high amount of potential regressions. Instead of choosing an arbitrary selection of explanatory characteristics, 350 potential financial ratios were constructed. Their values have then been transformed using the weight of evidence transformation to limit the influence of outliers. Next a variance minimization algorithm was used to group these variables into clusters. From these clusters a brute force computation of all possible combinations was executed in order to identify the ones reporting the highest predictive power in a logistic regression. While there is a clear disconnection between theoretical framework for selecting potential explanatory variables, the brute force solution definitely alleviates numerous concerns regarding the feature selection process[1].

When providing examples of applying logistic regression methodologies for predicting insolvencies the work of Nehrebecka and Dzik (2013) [31] seems to be a worth mentioning. Here, companies from the Polish market serve as the data set and an arbitrary selection of financial ratios as the explanatory variables. In order to deal with the unbalanced data structure a random selection of solvent companies was applied, such that they outnumber insolvent ones in a 1:4 proportion with preservation of representativeness of the whole population of financially stable companies. The study is of rather empirical nature and focuses on identifying significant factors for the identification of distressed companies. Although the analysis is based on a logistic regression executed on variables transformed following the weight of evidence approach, no final model is presented. Instead, a ranking system is proposed followed by a discussion of the distribution characteristics of the potential grading methods.

In relation to the Austrian market a notable reference would be the work of Hofmarcher *et al.* (2011) [24] who tried to identify macroeconomic determinants of aggregated default probabilities. Even though the aggregated rate has little to do

[1] Again, while this is a very potent method it is not viable in the case of this study due to high requirements of computing power.

with the microeconomic case, the work of Hofmarcher *et al.* (2011) presents an interesting discussion on how to deal with high collinearity in logistic regressions. Concluding from these works it would seem that applying logistic regressions for business demographic is rational and widely accepted. It does not, however, mean that it has already reached its most refined form- several tuning algorithms are still being developed[2] and waiting for an implementation in business field.

2.2.2 Machine Learning

Next, we present several studies representing the machine learning approach for predicting insolvencies and provide a discussion on their results. The discussion starts with decision trees, through random forests and concludes on support vector machines. Other techniques are discussed in the next section, which presents studies using methods from both regression and data mining approaches.

Decision trees are commonly used in practice and tackle the concept of non-parametric dependencies in a simple yet elegant manner. They do not depend on assumptions regarding the relationship between the explanatory and dependent variable. In fact, the linking function does not even need to be of monotonic nature. Another interesting aspect of this approach is that it provides potential insight about possible variable clusters through a graphical representation of their hierarchical structure. Despite the simplicity of the method, decision trees prove to be a valuable tool for binary problems [27].

The next classification algorithm often implemented by researchers to predict insolvencies would be the random forest approach. Although being only published in 2001, it grows rapidly in popularity and is often used for business purposes [8, 47]. Random forests are a very popular data mining algorithm that supposedly delivers more accurate predictions than the usual logistic predictors [10]. Rather than judging the forecasting model as a whole it ranks the explanatory variables by their predictive power in the same way as a decision tree does. However, because the random forests runs the decision tree algorithm multiple times, it produces more stable results than single decision trees and due to the random sub sampling algo-

[2]For example Bayesian and random forest updated prior probabilities or distance based variable clustering.

rithm applied by the random forest, it is also more reliable when dealing with a large number of predictors which is another important characteristic of our data set. Therefore, the random forest approach should theoreticaly yield better results when applied on an unbalanced database [26]. Among other techniques, random forest have also been the subject of the research of Mukkamala *et al.* (2006) [27, 28], who focus on French companies to compare the effectiveness of such methods as decision trees, neural networks or random forests. The original sample size consisted of 2800 companies total, of which 311 were insolvent. In order to deal with the unbalanced data structure the whole analysis was performed with different variations of resampling proportions (without replacement). The research has found evidence suggesting that methods based on linear programming outperform the other considered models. However, the random forest and decision trees approaches did also achieve high predictive power and should not be dismissed.

Another example of applying machine learning techniques would be the work of Härdle *et al.* (2007) [21] who examine the effectiveness of smooth support vector machines (SSVM) for predicting defaults among German firms. SSVM differ slightly from the simpler version[3] by dropping the constrains in minimization problem. The researchers did not restrict themselves to just analyze the predictive power of SSMVs, but also looked into the influence of the variable selection process and how the structure of the training sample affects the trade off between error rates.

As in most cases, the researches do not fail to point out the shortcomings of existing models and strongly argue that the restrictions of regression based classifiers in regard to data structure are, more often than not, not satisfied. Therefore, non parametric estimation methods are expected to yield better prediction accuracy. One of the biggest strengths of the work of Härdle *et al.* (2007) [21] is their enormous sample size (21 000 observations total) which provides the necessary reliability and limits variance in estimation results. Again, a set of financial ratios describing each company is used as the explanatory space. Also, since the researchers dealt with a database overly saturated with solvent companies a bootstrapped resampling procedure was used to improve the statistical properties of the estimators. Not surprisingly the researchers found that SVM are an effective classifier that deals well

[3]Which is applied in this study.

with both a large number of predictors and a large sample size.

The concept of using SVM to model company insolvencies was also the topic of Härdle *et at.* (2010) [22] where the researchers once again focus on German companies. However, this time no resampling techniques are applied and the researchers try to develop instruments based on the SVM methodology which would be able to cope with the challenge of dealing with a highly unbalanced sample. They conclude that with careful parameter tunning good accuracy rates can still be obtained and reaffirm that they often give the best results, when compared to other methods.

2.2.3 Comparison of machine learning and regression analysis

In this section we try to provide insight into essential studies using both regression and machine learning techniques to predict insolvencies. We try to provide examples of studies using the methods used for data analysis in this study.

The topic of using machine learning techniques for predicting insolvencies has been the main point of Diaz *et al.* (2005) [11]. For this purpose the performance of non- parametric learning methods were compared in terms of their performance to popular techniques such as linear discriminant analysis and logistic regressions. The researchers decided that it is worthwhile to consider the influence of dropping distribution assumptions on data for predicting insolvencies of Spanish non-life insurance companies. While insurance companies are usually behaving in a different way and follow a different financial model a industry based companies, the techniques used to model their financial distress and general methodologies are very much in common with other operating sectors. Also, in this case financial ratios serve as the explanatory data vector for classifiers and as per usual they do not satisfy the general statistical assumptions. While noticing that non-parametric machine learning classifiers operate on the principle of a black box the researchers still consider them to be extremely viable, despite the obvious problems related to their interpretation. Although the results of this study should be approached with some caution as the analysis was performed on a extremely limited sample size consisting of only 72 companies in total, the researchers still managed to compare 15 different variations of data and classification techniques and in the end reaffirmed that in presence of violated data assumptions machine learning predictors fare significantly better

than standard data analysis techniques. Another identified limitation would be that having a high number of potential predictors negatively influences the accuracy of regression classifiers while increasing the effectiveness of machine learning algorithms this could be very much an indicator of potential overfitting problems. However, since most real life data are connected with a multitude of potential explanatory variables the researchers recommend techniques that can better cope with the abundance of data.

Of particular interest is the work of Kartasheva and Traskin (2011) [26] which again focuses on the comparison of insolvency classifiers, but this times compares their effectiveness to one based on the random forest approach. For this purpose a sample of American insurers data characteristics was used. As in most cases, a selection of financial ratios was calculated and served as a vector of explanatory variables. The researchers point out several advantages of the random forest approach, such as high resistance to unbalanced data sets, high number of potential predictors and non linear classification methodology. A logistic regression is used as a benchmark comparison model with some variations in the variable selection of data transformation process. The analysis results of this research lead to a conclusion that the random forest approach is superior in both prediction quality and the identification process of significant explanatory variables. It also reaffirms the claim that the correlation between insolvencies and financial ratios is of highly nonlinear nature.

The question of the effectiveness of random forest techniques has also been approached by Wrzosek and Ziemba (2009) [47]. The researchers focus on Polish market data and, as usual in these cases, a vector of financial ratios was used as the explanatory space with a sample size of 11505 companies total. However, this was then restricted to obtain a better proportion between solvent and insolvent companies. For this purpose a simple random sampling without replacement was used. The logistic regression is being used as the benchmark model to which other (i. e. Altman model, Random Forest) are compared to. Since the research was conducted by a private company the scope is more focused on the identification of significant predictors and accuracy maximization and not the discussion of methodological effectiveness. The results are of particular interest, as the overall effectiveness of potential models with different combinations of explanatory variables seems to per-

form reasonably well no matter what kind of methodology is applied. This would suggest that despite the shortcomings regarding statistical properties of data, both the logistic regression and Altman model still produce viable results.

Another comparative study comes from the German SME market by Fantazzini and Figini (2009) [16] who compare a random survival forest approach to logistic regressions. A set of financial ratios has been used for the explanatory variables. Interestingly enough the researchers find the random forest approach to be more accurate than the logic classifier in the training sample, but at the same time it would perform worse in the test sample. This suggest that we cannot expect the random forest approach to outperform the logit classifier at all times. Another question raised by the research is whether potential linking transformations between the results of random forests and logistic regressions can be obtained, as there exists empirical research that supports the claim combining the results from different methods yield better performance [10]. However, other studies prove the contrary and find that simple logistic regressions outperform random forests and other hybrid forms [17]. This would indicate that the potential comparison of these approaches cannot be simply predicted on the principle of “the more complex the better". These discrepancies are likely to be traced back to the fact that random forests are better at describing the in-sample data whereas logistic regression fare better at predicting the behavior of out-of-sample data.

With this we end our review of existing literature and believe to have a proper understanding of both regression and machine based data analysis approaches. In the next chapter essential information regarding the data set and choice of variables is provided.

Chapter 3

Data

In this Chapter all necessary information on the structure of the dataset is provided. We start with a short description of the dataset and provide some descriptive statistics. Next the variables are introduced. A definition for each one is listed and the performed data transformations are discussed.

3.1 Database Description

For the purposes of this research the COMPUSTAT North America Fundamental Annual is used. The database is provided by Wharton Research Data Services and offers 382 balance sheet items and 328 income statement items from a selection of more than 86k companies from between the years 2004 and 2013. This database has also already been used in the research of Altman (2007) [4] which gives us confidence that the data is of sufficient quality for this analysis.

For the scope of this research the data was restricted to the period between 2004 and 2011 in order to maximize full data instances and the initial number of observations in the database was set to 64902 companies. Compustat classifies companies depending whether they are present on the market or not. Therefore, a company can be 'active' or 'inactive'. Further, a company can be 'inactive' due to market exit, acquisition, merger, bankruptcy etc. For the purposes of this study we pool all companies that are 'inactive' and not bankrupt (or marked as 'no information') into the 'active' group. Thus, we have a restricted 'inactive' group consisting only of bankrupt companies. Next, we corrected that group to include only companies

that went bankrupt between the years 2004 and 2011.

3.2 Financial Ratios and Descriptive Statistics

What followed was the creation of the explanatory variables. According to literature a set of financial ratios was constructed that describes each company in terms of its profitability, leverage, liquidity, activities, size and growth characteristics (See Table 3.1). If somewhere a division by zero was required the observation was deleted from the data base. The descriptive statistics of the remaining subjects are presented in Table 3.4. We immediately notice that there exist multiple large outliers for nearly every variable and in all but a few cases (x11, x18, x21, x23) the median significantly differs from the mean value. Another important feature of the data set is the high number of missing values. For example, x11 reports 50389 missing cases, most of which were solvent companies. We decided that imputing values for such a large percentage of the database is pointless and discarded variable x11 from the set of explanatory variables. Another important observation is that the range between the 1st and 3rd quantile is almost universally small when compared to the maximums and minimums. In addition, in the case of 16 variables (x1, x2, x3, x4, x5, x6, x7, x8, x10, x14, x15, x16, x17, x19, x21, x22) the mean value resides outside of the interquantile range, which emphasizes the influence of the outliers in dataset. Next, we take a closer look at the descriptive statistics broken down into subset of solvent ($y = 0$) and insolvent ($y = 1$) companies presented in Table 3.5. When looking at minimal and maximal values we notice that the outliers are occurring in both categories across the whole database and in most instances they are fairly comparable. However, there are numerous differences in median values, which suggest that the classification task may yield interesting results.

In order to get a clear overview of the structure of the variables their values have been plotted over several graphs (See Appendix 2). What we notice is that in most cases there is a clear difference between the density of the explanatory variables. Variables x1, x2, x3, x8, x9, x12, 19, x20, x25 and x26 have a larger spread of values for insolvent companies than solvent ones. There is also a clear distinction in the

Table 3.1: Explanatory Variables, Definitions, Relevant Research Examples

Variable	Definition	Studies applying them
	Profitability Ratios:	
x1	Net Income / Assets Total	[4, 10, 16, 21, 22, 31, 33, 37, 47]
x2	Net Income / Sales Net	[4, 10, 21, 22, 31, 33, 47]
x3	Earnings Before Interest and Taxes / Assets Total	[21, 22, 33, 47]
x4	Earnings Before Interest and Taxes / Sales Net	[4, 10, 21, 22, 27, 28, 33]
x5	(Earnings Before Interest and Taxes+Depreciation) / Assets Total	[4, 21, 22, 31, 33]
	Leverage Ratios:	
x6	Equity / Assets Total	[10, 16, 21, 22, 27, 28, 31, 33, 47]
x7	(Equity-Intangible Assets Total)/(Assets Total-Intangible Assets -Cash- Property)	[21, 22, 33]
x8	Liabilities Current Total/Assets Total	[21, 22, 33]
x9	(Liabilities Current Total-Cash and Short-Term Investments)/Assets Total	[16, 21, 22, 33]
x10	Liabilities Total/Assets Total	[4, 21, 22, 33, 37, 47]
x11	Earnings Before Interest/Interest Expense Total	[4, 16, 21, 22, 31, 33]
x12	Debt Total/Assets Total	[4, 10, 16, 21, 22, 27, 28, 31, 33, 47]
	Liquidity Ratios	
x13	Cash and Short-Term Investments/Assets Total	[4, 21, 22, 33]
x14	Cash and Short-Term Investments/Liabilities Total	[21, 22, 33, 37]
x15	(Assets Current Total - Inventories Total)/Liabilities Current Total	[4, 21, 22, 31, 33, 47]
x16	Assets Current Total/Liabilities Current Total	[21, 22, 27, 28, 31, 33, 37, 47]
x17	(Assets Current Total-Liabilities Current Total)/Assets Total	[21, 22, 27, 28, 33]
x18	Liabilities Current Total/Liabilities Total	[16, 21, 22, 31, 33]
	Activity Ratios	
x19	Assets Total/Sales Net	[4, 21, 22, 33, 47]
x20	Inventories Total/Sales Net	[16, 21, 22, 33]
x21	Receivables Total/Sales Net	[4, 21, 22, 33]
x22	Accounts Payable/Sales Net	[4, 21, 22, 33]
	Size and Growth Ratios	
x23	log(Assets Total)	[10, 21, 22, 33, 47]
x24	Inventory Increase/Inventories Total	[21, 31]
x25	Accounts Payable Increase/Liabilities Total	[21, 31]
x26	Cash and Cash Equivalents Increase/Cash and Short-Term Investments	[21, 31]

distribution form of variables x10, x11, x18 and x23 which suggests that a kernel transformation might provide additional insight into the relationship. However, there are also several variables where no noticeable difference can be spotted, such as x6, x7, x14, x15, x16, x21 or x22.

The differences between medians across both categories are easily visualized on the boxplot graphs (See Appendix 2). These graphs have been rescaled to present 90% through 80% of all observations to limit the distortion caused by outliers. From these graphs we can notice that in the case of variables x5, x7, x13 and x24 it is difficult to notice any difference to the median level between classes. Other than that, in each group we can notice a significant amount of outliers which raises the question to which degree it can decrease the overall classification accuracy.

Table 3.4: Prior Bankrupcy Proportions by Years

Year	Active	Insolvent	Probability
2004	925	1	0.1%
2005	7932	10	0.1%
2006	7859	9	0.1%
2007	7811	37	0.5%
2008	7618	69	0.9%
2009	7556	51	0.7%
2010	7600	2	0.02%
2011	7233	8	0.1%

3.3 Balancing the Dataset

For the next step, the variable x11 has been removed from the dataset due to an enormous number of missing values. In order to cleanse the database from additional noise, observations that reported more than 20% missing values in the explanatory set have been removed. Thus, the total sample listed 54721 observations, of which 187 were insolvent, down from the initial 64902 with 430 insolvent companies. The breakdown by years (Table 3.4) unsurprisingly reveals an increase of insolvencies in the years 2007, 2008 and 2009, which can be easily attributed to macroeconomic

shocks. For the remaining observations the missing values have been imputed from the corresponding median values following the approach of Torgo (2011) [44]. Next, like in many previous studies discussed in Chapter 2, the data set went through a balancing procedure. The new sample consist in 30% of insolvent companies and 70% solvent ones which were drawn at random with replacement from the solvent sample with preservation of representativeness. Then, in order to ensure out of sample validation we split the data into a training test set with the proportions 1 to 4. The final number of solvent and insolvent companies in both test and training sets is reported in Table 3.5.

Having provided a presentation of the available database and discussing preliminary data characteristics as well as providing insight in the balancing process the next chapter is devoted to present the detailed methodological side of this research. Essential mathematic features are presented and discussed, while drawing connections with our database specification.

Table 3.5: Total Insolvencies in Populations

Data	Active	Insolvent	Probability
Total Sample	54534	187	0.34%
Study Sample	415	178	30%
Training Sample	332	142	30%
Test Sample	83	36	30%

Table 3.2: Main Data Characteristics of the Variables

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
x1	-25880.000	-0.038	0.017	-0.867	0.064	244.800	11144
x2	-29320.000	-0.064	0.039	-7.062	0.111	609.100	11134
x3	-581.200	-0.008	0.045	-0.172	0.099	49.750	11310
x4	-30180.000	-0.012	0.083	-5.717	0.195	193.100	11297
x5	-581.000	0.070	0.220	0.071	0.402	65.410	17245
x6	-25970.000	0.157	0.404	-0.692	0.636	2.856	10944
x7	-5.191e+18	0.000e+00	1.000e+00	-1.853e+14	2.000e+00	1.153e+18	13053
x8	0.000	0.131	0.217	1.554	0.352	25970.000	19462
x9	-0.998	-0.064	0.091	1.345	0.222	25970.000	19468
x10	0.000	0.358	0.584	1.687	0.833	25970.000	11055
x11	-5.30	0.52	0.90	1.17	1.51	20.28	50389
x12	0.000	0.001	0.137	0.305	0.351	120.900	17181
x13	-0.015	0.030	0.092	0.185	0.258	1.000	10948
x14	-0.013	0.041	0.145	1.830	0.604	24100.000	11060
x15	-0.019	0.809	1.318	3.458	2.338	24110.000	19903
x16	-0.019	1.087	1.758	3.898	2.935	24110.000	19596
x17	-25970.000	0.018	0.175	-1.067	0.390	0.999	19595
x18	0.002	0.303	0.562	0.571	0.860	2.233	19577
x19	-50780.00	0.82	1.54	13.66	4.25	60270.00	11144
x20	-17.400	0.000	0.053	0.163	0.143	93.420	11700
x21	-223.300	0.097	0.160	1.559	0.270	787.200	11387
x22	-290.100	0.044	0.081	2.038	0.176	3546.000	11518
x23	-6.908	4.328	6.234	6.064	7.860	15.140	10941
x24	-2943.000	-0.193	-0.048	-0.152	0.088	584.200	26244
x25	-28.560	-0.012	0.005	0.033	0.049	685.900	26383
x26	-12440.000	-0.288	0.049	-1.240	0.349	24660.000	11622

Table 3.3: Main Data Characteristics of the Variables by the Dependent

x	y	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
x1	0	-2.588e+04	-3.810e-02	1.672e-02	-8.677e-01	6.451e-02	2.448e+02
	1	-3.266e+00	-5.160e-01	-2.634e-01	-4.091e-01	-2.246e-02	4.135e-01
x2	0	-2.932e+04	-6.412e-02	3.958e-02	-7.040e+00	1.107e-01	6.091e+02
	1	-1.118e+03	-1.272e+00	-1.755e-01	-3.329e+01	-5.593e-02	6.495e-01
x3	0	-5.812e+02	-7.715e-03	4.547e-02	-1.714e-01	9.949e-02	4.975e+01
	1	-2.524e+00	-3.300e-01	-1.217e-01	-2.924e-01	1.243e-02	4.133e-01
x4	0	-3.018e+04	-1.150e-02	8.308e-02	-5.691e+00	1.949e-01	1.931e+02
	1	-1.261e+03	-6.493e-01	-8.370e-02	-3.685e+01	1.204e-02	6.492e-01
x5	0	-5.810e+02	6.960e-02	2.206e-01	7.128e-02	4.027e-01	6.541e+01
	1	-1.968e+00	-1.406e-01	1.259e-01	-5.751e-03	3.225e-01	1.169e+00
x6	0	-2.597e+04	1.568e-01	4.041e-01	-6.931e-01	6.359e-01	2.856e+00
	1	-1.415e+00	-5.702e-02	1.721e-01	1.690e-01	5.413e-01	9.744e-01
x7	0	-5.191e+18	7.428e-02	6.082e-01	-1.855e+14	1.518e+00	1.153e+18
	1	-1.078e+01	-4.699e-01	3.801e-01	2.209e+01	1.900e+00	3.520e+02
x8	0	0.000e+00	1.313e-01	2.170e-01	1.555e+00	3.518e-01	2.597e+04
	1	2.491e-02	1.885e-01	4.288e-01	5.257e-01	6.547e-01	2.225e+00
x9	0	-9.978e-01	-6.359e-02	9.130e-02	1.346e+00	2.217e-01	2.597e+04
	1	-9.691e-01	-3.627e-03	2.659e-01	2.752e-01	5.658e-01	2.190e+00
x10	0	0.000e+00	3.582e-01	5.841e-01	1.687e+00	8.326e-01	2.597e+04
	1	2.562e-02	4.587e-01	8.254e-01	8.329e-01	1.057e+00	2.415e+00
x11	0	-5.303000	0.524900	0.897700	1.174000	1.511000	20.280000
	1	-0.726200	-0.494900	-0.263500	-0.240400	0.002445	0.268400
x12	0	0.000e+00	1.177e-03	1.372e-01	3.048e-01	3.512e-01	1.209e+02
	1	0.000e+00	0.000e+00	4.497e-02	3.824e-01	5.947e-01	1.727e+00
x13	0	-0.015420	0.030350	0.092410	0.185200	0.258000	1.000000
	1	0.002974	0.027340	0.096130	0.249800	0.255600	0.999800
x14	0	-1.305e-02	4.082e-02	1.453e-01	1.829e+00	6.044e-01	2.410e+04
	1	4.054e-03	3.298e-02	9.985e-02	2.544e+00	7.728e-01	3.556e+01
x15	0	-1.902e-02	8.093e-01	1.318e+00	3.457e+00	2.338e+00	2.411e+04
	1	4.174e-02	4.072e-01	1.046e+00	3.813e+00	1.476e+00	3.563e+01
x16	0	-1.902e-02	1.087e+00	1.759e+00	3.898e+00	2.935e+00	2.411e+04
	1	7.217e-02	7.790e-01	1.392e+00	4.111e+00	2.124e+00	3.563e+01
x17	0	-2.597e+04	1.758e-02	1.750e-01	-1.068e+00	3.903e-01	9.994e-01
	1	-2.064e+00	-5.098e-02	1.722e-01	3.742e-02	3.102e-01	9.711e-01
x18	0	0.002335	0.302600	0.561400	0.570800	0.860000	2.233000
	1	0.057340	0.388500	0.867100	0.701700	0.989200	1.000000
x19	0	-5.078e+04	8.241e-01	1.536e+00	1.361e+01	4.248e+00	6.027e+04
	1	2.344e-01	4.603e-01	1.278e+00	6.728e+01	5.565e+00	2.270e+03
x20	0	-17.40000	0.00000	0.05345	0.16290	0.14310	93.42000
	1	0.00000	0.00000	0.06247	0.10710	0.14260	0.89240
x21	0	-223.30000	0.09695	0.15990	1.55800	0.27020	787.20000
	1	0.00000	0.06451	0.12800	2.92000	0.16870	65.22000
x22	0	-290.10000	0.04383	0.08115	2.03800	0.17640	3546.00000
	1	0.00000	0.04876	0.09933	2.73500	0.21910	49.67000
x23	0	-6.908	4.330	6.236	6.065	7.862	15.140
	1	1.510	3.447	4.918	4.619	6.084	7.412
x24	0	-2.943e+03	-1.932e-01	-4.790e-02	-1.521e-01	8.744e-02	5.842e+02
	1	-6.108e-01	-3.756e-02	1.189e-01	1.324e-01	2.288e-01	9.143e-01
x25	0	-28.560000	-0.012380	0.005347	0.033090	0.048640	685.900000
	1	-1.334000	-0.101300	-0.029010	-0.078650	0.003640	0.410600
x26	0	-1.244e+04	-2.880e-01	4.922e-02	-1.241e+00	3.492e-01	2.466e+04
	1	-6.199e+00	-8.453e-01	-1.338e-01	-4.993e-01	3.371e-01	8.156e-01

Chapter 4

Research Methods

Predicting insolvencies is a broad topic where numerous approaches are viable to use as potential predicting techniques. As seen in the previous chapter, this area of research has awoken a strong interest in both the academic world and the commercial market. Plentiful models have been developed to cope with the task of predicting liquidity difficulties in companies.

For this purpose regression analysis is often used. However, predicting methods based on machine learning algorithms are becoming more and more common as their explanatory power often matches and sometimes even exceeds regression model based ones. This and the fact that the necessary data characteristics for the application of machine learning algorithms are less strict than those necessary for a regression analysis, make machine learning methods a popular choice for classification tasks of this nature.

This chapter is therefore devoted to explain the logic and methods associated with some of the more popular techniques. In particular, in the field of regression analysis we focus on the logit regression family. Additionally, we describe a wide selection of machine learning techniques with the aim of applying them for evaluation purposes with the results being described in the next chapter. All methods are portrayed with definitions associated with the particular classification task.

4.1 Regression Analysis

The first method used to forecast insolvencies is based on the data regression approach. Because of the dichotomy character of insolvencies the classification task is a univariate one. A company can either become insolvent or not and there are no other options, therefore it is a binary dependent variable. We describe the modeling process of discrete dependent variables following mostly the logic of Greene (1997)[20] and Mycielski (2010)[30].

In regression analysis the prediction of solvency states of the company is modeled with probabilities of observing a company in a given state. In other words, the model returns the conditional probability of a company remaining solvent given its observable characteristics, which are the regression's explanatory variables. This relation can be expressed as:

$$Pr(y_i|x_i) = \begin{cases} 1 - p(x_i) & \text{for } y_i = 0 \\ p(x_i) & \text{for } y_i = 1 \end{cases} = [p(x_i)]^{y_i}[1 - p(x_i)]^{1-y_i}, \tag{4.1}$$

where:

y_i – is the observable state of the company i (whether it is solvent or not),

x_i – is a vector containing characteristics of the company i,

$p(\cdot)$ – is the probability function expressing the influence of the independent variables onto the dependent variable.

Because the probability function $p(x)$ has to attain values between 0 and 1 it is approximated with a distribution function of a continuous random variable which is given by a linear combination of independent variables [30]. Thus,

$$p(x_i) = F(\beta' x_i), \tag{4.2}$$

where:

$F(\cdot)$ – is a cumulative distribution function of a random variable.

Using this cumulative distribution function we can express the probabilities as fol-

lows:

$$\begin{aligned} Pr(y_i = 1) &= F(\beta' x_i) \\ Pr(y_i = 0) &= 1 - F(\beta' x_i) \end{aligned} \tag{4.3}$$

In the case of the logit regression this cumulative distribution function uses a logistic distribution:

$$\begin{aligned} Pr(y_i = 1) &= F(\beta' x_i) \\ &= \frac{e^{\beta' x_i}}{1 + e^{\beta' x_i}} \\ &= \Lambda(\beta' x_i), \end{aligned} \tag{4.4}$$

where:

$\Lambda(\cdot)$ – is the logistic cumulative distribution function.

Having specified the cumulative distribution function we can now model the relation between the dependent outcome and the explanatory characteristics with the use of the maximum likelihood method. On the assumption that each observation is independent of one another we can write the joint likelihood function as:

$$L(\beta) = \prod_{i=1}^{n} (\Lambda(\beta' x_i))^{y_i} (1 - \Lambda(\beta' x_i))^{1-y_i}. \tag{4.5}$$

In order to maximize the likelihood in respect to β we logarithm this equation,

$$\ln L(\beta) = \sum_{i=1}^{n} (y_i \ln(\Lambda(\beta' x_i)) + (1 - y_i) \ln(1 - \Lambda(\beta' x_i))). \tag{4.6}$$

Therefore, the first order conditions are as follows,

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n} (\frac{y_i \lambda_i}{\Lambda_i} + (1 - y_i) \frac{-\lambda_i}{1 - \Lambda_i}) x_i = 0. \tag{4.7}$$

Before we are able to solve this equation we need to state a convenient characteristic

of the logistic distribution,

$$\begin{aligned}\frac{\partial E(y|x)}{\partial x} &= \frac{d\Lambda(\beta' x)}{d(\beta' x)}\beta \\ &= \lambda(\beta' x)\beta \\ &= \frac{e^{\beta' x}}{(1+e^{\beta' x})^2}\beta \\ &= \Lambda(\beta' x)(1-\Lambda(\beta' x))\beta. \end{aligned} \tag{4.8}$$

Now we can observe that the necessary conditions can be expressed as,

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n}(y_i - \Lambda_i)x_i = 0. \tag{4.9}$$

Following the reasoning of Greene (1997)[20] the parameters can be computationally easily obtainable using the Newton's method. Also, the second order derivative of the likelihood function,

$$H = \frac{\partial^2 \ln L}{\partial \beta \partial \beta'} = -\sum_{i=1}^{n}\Lambda_i(1-\Lambda_i)x_i x_i{}' \tag{4.10}$$

is always negative definite, so the log-likelihood function actually achieves the global maximum. A popular measure for interpreting the influence of the explanatory variables is the *Odds ratio*, which measures the relation between the probability of observing a "1" and observing a "0",

$$Odds(x_i) = \frac{Pr(y_i = 1|x_i}{Pr(y_i = 0|x_i)} = \frac{\frac{e^{\beta' x_i}}{1+e^{\beta' x_i}}}{\frac{1}{1+e^{\beta' x_i}}} = e^{\beta' x_i}. \tag{4.11}$$

With this ratio we can easily compare the relative change in probability for different observations,

$$\frac{Odds(x_i)}{Odds(x_j)} = e^{\beta'(x_i - x_j)} = e^{\sum_{l=1}^{k}\beta' \Delta x_l} = \prod_{l=1}^{k} e^{\beta' \Delta x_l}. \tag{4.12}$$

This simple measure allows to approximate the change in probability in one variable, without computing the result for the whole vector x_i. Thus, $exp(\beta_k)$ is the relative change in probability for variable k when $\Delta x_k = 1$ *ceteris paribus*.

4.1.1 Weight of Evidence transformation

Another form of dealing with high collinearity is the popular[1] *Weight of Evidence* approach as it is very efficient at maximizing the use of available information [5]. This approach can be applied on categorical variables and improves the logit regressions' accuracy, especially if there are no interaction terms between the explanatory variables. The WoE transformation preserves the linear dependences in the linking logistic function and therefore does not influence the significance of results. The WoE value for categories of variables is expressed as $log(\frac{\%A_i}{\%I_i})$, where A and I indicates Active or Inactive companies and i the category of the variable. As the WoE approach can only be applied on categorical variables continuous ones need to be binned and transformed on some criteria. Numerous credit scoring models decide to go for the *Information Value* maximization [31] which measures the Kullback-Leibler divergence [34] between 'active' and 'inactive' observations, others decide to simply put them in quantiles, which is an approach that is less than perfect [6]. The credit scoring algorithm splits the continuous variables into categories that maximize the divergence between the binary dependent results. The next step is to assign the calculated WoE values as weights of the variables in the final model. From here a usual model building procedure can be applied for the variable selection process.

4.2 Machine Learning Algorithms

Machine learning is the field of science often associated with data mining which focuses on finding and describing patterns in data. It is a somewhat curious name, as machines are unable to learn as people do. Instead we define learning following the definition of Witten *et al.* (2011) [46]:

- To get knowledge of something by study, experience, or being taught
- To become aware by infromation or from observation
- To commit to memory
- To be informed of or to ascertain

[1] In credit scoring methods.

- To receive instruction

In order to refrain from philosophical questions we associate learning with performance instead of knowldege. The classifiers based on machine learning algorithms get better, i. e. more accurate, with the inclusion of more data. They are unable to explain why data performs as given, they are strictly restricted to classifying new information.

4.2.1 Decision Trees

The first machine learning technique applied in this research is the simple decision tree approach. Similarly to the previous section the methodology is explained based on existing literature. For this purpose we based the description mostly on the works of Witten *et al.* (2011) [46], Togo (2011) [44], Zhao (2012) [48] and the documentation provided by Therneau (2013) [43]. It is a non-parametric method and therefore we do not need to assume any prior hypothesis about the nature of relationship between the explanatory vector and the dependent variable [38].
The main idea behind the decision tree concept can be expressed with the phrase 'divide and conquer' as decision trees are constructed following a greedy algorithm scheme. The output of this algorithm is a hierarchy of logical tests on some of the potential explanatory variables. Because the algorithm automatically selects relevant explanatory variables not all are present in the final result. This selection follows a predefined selection criteria.
In the first step a single variable is selected which splits the database in the most informative way. After the separation, the process is applied independently again for both subgroups. This process is repeated recursively until a predefined stopping criteria. This depends on whether a minimal subgroup size has been achieved or whether no additional information can be obtained, whichever happens first. Because the process is extremely prone to overfitting and the resulting tree can be too complex for effective use, the next stage consists of algorithms which prune-back the tree.
To put things into perspective, we provide some mathematical background for the algorithm. Suppose that the sample of insolvent and solvent companies consists of n observations. These observations have only 2 classes. Namely, solvent and insolvent.

The decision tree will breakdown these observations into k terminal groups with a uniformly assigned predicted class. That is, all observations falling into a group share the predicted class value.

Firstly, we need to define the probability of correctly classifying a set of observations,

$$P(A) = \pi_1 P\{x \in A|\tau(x) = 1\} + \pi_2 P\{x \in A|\tau(x) = 2\} \tag{4.13}$$

$$\approx \pi_1 \frac{n_{1A}}{n_1} + \pi_2 \frac{n_{2A}}{n_2}, \tag{4.14}$$

where:

A – is a set of observations,

π_i – is the prior probability of class $i = 1, 2$,

$\tau(\cdot)$ – is the true class of an observation,

n_i – number of observations of class $i = 1, 2$,

n_A – number of observations in group A,

n_{iA} – number of observations of class $i = 1, 2$ in group A.

Then, we also need to define the risk associated with incorrectly classifying a set of observations,

$$R(A) = p(i = 1|A)L(i = 1, \tau(A)) + p(i = 2|A)L(i = 2, \tau(A)) \tag{4.15}$$

where:

$p(i|A)$ $= \pi_i P\{x \in A|\tau(x) = i\}$,

$L(i, j)$ – is the loss function for misclassification ($i, j = 1, 2$ and $L(i, i) \equiv 0$).

Therefore, a group A is split into subgroups B and C, if and only if,

$$P(B)R(B) + P(C)R(C) \leq P(A)R(A) \tag{4.16}$$

Obviously, the tree is built in such a way that maximizes the decrease in value of the risk function. Since the choice is basically linear, additional requirements are posed on the slit criteria. These criteria are based on some impurity functions assessing

the overall usefulness of a group,

$$I(A) = f(p_{1A}) + f(p_{2A}), \tag{4.17}$$

where:

$f(\cdot)$ is an impurity function,

p_{iA} – is the proportion of observations of class $i = 1, 2$ in group A.

Since the impurity has to be null for both a perfect group and a total misclassification, we notice that $f(\cdot)$ has to be concave and that,

$$f(0) = f(1) = 0. \tag{4.18}$$

The two most commonly used functions with these characteristics are the information index:

$$f(p) = -p\log(p) \tag{4.19}$$

and the Gini index:

$$f(p) = p(1 - p). \tag{4.20}$$

Plotting both functions we can easily observe that both obtain maximum in the same split (see Figure 6.11). The theoretical comparison of both measures also lead to the conclusion that there is no significant difference between them [36]. Thus, the maximization task for for splitting data can be expressed as,

$$\max_{B,C} \Delta I = p(A)I(A) - p(B)I(B) - p(C)I(C). \tag{4.21}$$

Which guarantees the optimal split.

4.2.2 Random Forest

Next, we take a closer look at the random forest approach. In order to convey the methodology in a proper and error free way this explanation is based on the original work of Breiman (2001) [8].

Random forest is build upon the concept of decision trees and as the name suggest it basically incorporates the idea of planting multiple trees and then using the

Strong Law of Large Numbers to generalize the obtained results. In other words, the random forest approach can be summarized in a set of step-by-step procedures in the following way. First, n observations are randomly selected with replacement and preservation of the distribution from the whole population N. Thus, creating a bootstrapped sample. The remaining $N - n$ observations are later used to estimate the out of sample error rate of the tree. Then, at each splitting moment randomly select m variables from the set of explanatory variables, which determine the explanatory variables for this split. Each tree is fully grown and unprunned. A random forest is more flexible regarding the data structure than logistic regressions and more accurate than decision trees [26].
The exact definition written by Breiman goes as follows:

A random forest is a classifier consisting of a collection of tree- structured classifiers $\{h(x, \Theta_k), k = 1, ...\}$ *where the* Θ_k *are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x.*

Again, a more formalized algorithm specification would not be amiss. We start by defining a decision tree k:

$$h_k(x) = h(x|\Theta_k) \tag{4.22}$$

where:

- $h(\cdot)$ – is a decision tree classyfier,
- Θ_k – is a set of parameters associated with tree k, such as sequence of splits, used variables and values,
- x – is the feature space.

Then, the follwoing margin function can be specified,

$$\hat{m}(x, y) = \hat{P}_k(h_k(x) = y) - \max_{j \neq y} \hat{P}_k(h_k(x) = j) \tag{4.23}$$

where:

- $\hat{P}_k(\cdot)$ – is the proportion of classifiers h_k for which a particular event occures (empirical probability),
- y – is the vector of the dependent variable.

Notice, that in the area of a binary classification tasks $j \neq y$ is just an incorrect classification. Therefore, the margin function measures the confidence in the classification approximating it with the extent to which the average number of classifiers (trees) votes for the correct class in comparison to the number of classifiers (trees) voting for the incorrect class.
Next, the generalization error is defined as follows,

$$\epsilon = P_{x,y}(\hat{m}(x, y) < 0). \tag{4.24}$$

As Breiman (2011) [8] proved, this generalization error converges almost surely to zero as the number of trees increases,

$$\epsilon \xrightarrow{K\to\infty} P_{x,y}(P_\Theta(h(x, \Theta) = y) - \max_{j\neq y} P_\Theta(h(x, \Theta) = j) < 0). \tag{4.25}$$

Therefore, the addition of more trees to the procedure does not produce an overfitting issue. However, since we always have limited number of trees for the analysis we need to obtain specific bounds on the generalization error. First, a margin function of a random forest is defined,

$$m(x, y) = P_\Theta(h(x|\Theta) = y) - \max_{y\neq y} P_\Theta(h(x|\Theta) = j). \tag{4.26}$$

As well as the strength of a forest,

$$s = \mathrm{E}_{x,y} m(x, y). \tag{4.27}$$

Then the generalization error can be obtained following the Chebyshev inequality,

$$P_{x,y}(m(x, y) < 0) \leq P_{x,y}(|m(x, y) - s| \geq s) \leq \frac{V(s)}{s^2}. \tag{4.28}$$

Which gives us a weak bound on the error term.

4.2.3 Naïve Bayes

The next method is an application of the well known Bayes Theorem. Here the main assumption is that there is a direct connection between the explanatory variables

and the dependent variable that can be approximated with conditional probabilities. We can write,

$$P(Y = y_i|X = x_k) = \frac{P(X = x_k|Y = y_i)P(Y = y_i)}{\sum_{j=1}^{M} P(X = x_k|Y = y_j)P(Y = y_j)}, \tag{4.29}$$

where:

y_n – is the dependent value of observation n,

x_n – is the value of the explanatory vector of observation n.

Notice, for a binary classification task the sum under the fraction is just the sum of two elements, namely when $y = 1$ and $y = 2$. The main idea of a Naïve Bayes classifier is to learn $P(Y|X)$ based on the estimates from $P(X|Y)$ and $P(Y)$ and then apply the results for new instances x_k in the form of $P(Y|X = x_k)$. This however is NP complex problem [19] and in order to reduce the requirements an additional assumptions is introduced. Namely, that the explanatory variables are conditionally independent of one another. An assumption that is only rarely satisfied, even less so when analyzing data from balance sheets. Still, in most cases the Naïve Bayes approach provides highly accurate predictions [19].

Using the conditional independence characteristic we can express the conditional probability as,

$$P(X_1, ..., X_n|Y) = \prod_{i=1}^{n} P(X_i|Y). \tag{4.30}$$

In the case of a binnary choice classification the Bayes rule looks as follows,

$$P(Y = y_k|X_1, ..., X_n) = \frac{P(Y = y_k)P(X_1, ..., X_n|Y = y_k)}{\sum_{j=1} P(X_1, ..., X_n|Y = y_j)P(Y = y_j)}. \tag{4.31}$$

Now, under the assumption of conditional independence this can be rewritten as,

$$P(Y = y_k|X_1, ..., X_n) = \frac{P(Y = y_k)\prod_{i=1}^{n} P(X_i|Y = y_k)}{\sum_{j=1} P(Y = y_j)\prod_{i=1}^{n} P(X_i|Y = y_j)}. \tag{4.32}$$

Since we are interested only in the classifier that maximizes the number of correct predictions the Naïve Bayes classifier can be formulated as,

$$Y \leftarrow \arg\max_{y_k} \frac{P(Y = y_k)\prod_{i=1}^{n} P(X_i|Y = y_k)}{\sum_{j=1} P(Y = y_j)\prod_{i=1}^{n} P(X_i|Y = y_j)}. \tag{4.33}$$

Because the part in the denominator does not depend on y_k it can be safely omitted, which brings us to the final version of the classifier,

$$Y \leftarrow \arg\max_{y_k} P(Y = y_k) \prod_{i=1}^{N} P(X_i|Y = y_k). \tag{4.34}$$

When dealing with discreet explanatory variables the necessary probabilities can be quite straightforwardly approximated with empirical probabilities. However, since in our case all of the variables are expressed in continuous terms we need to find a solution to represent the conditional probabilities. In the most basic form this is done by assuming that all of the variables follow a Gaussian distribution. The characteristics of each normally distributed variables, μ and σ^2, have to be independently estimated for each variable X_i and each value of the dependent y_k,

$$\mu_{ik} = E[X_i|Y = y_k], \tag{4.35}$$

$$\sigma_{ik}^2 = E[(X_i - y_{ik})^2|Y = y_k]. \tag{4.36}$$

In addition, the prior probabilities of Y are simply taken as the proportion of each value of the dependent variable,

$$\hat{\pi_k} = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}. \tag{4.37}$$

The Gaussian parameters are then obtained through a maximum likelihood estimation, and therefore

$$\hat{\mu}_{ik} = \frac{\sum_{j=1} X_i^{(j)} \mathbb{1}_{(Y^{(j)}=y_k)}}{\sum_{j=1} \mathbb{1}_{(Y^{(j)}=y_k)}}, \tag{4.38}$$

$$\hat{\sigma}_{ik}^2 = \frac{\sum_{j=1} (X_i^{(j)} - \hat{\mu}_{ik}) \mathbb{1}_{(Y^{(j)}=y_k)}}{\sum_{j=1} \mathbb{1}_{(Y^{(j)}=y_k)} - 1}, \tag{4.39}$$

where the superscript j refers to the jth training example. With these parameters the conditional probability can be expressed as,

$$P(X = x_k|Y = y_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp(-\frac{(x_k - \mu_i)^2}{2\sigma_i^2}). \tag{4.40}$$

An alternative is to drop the assumption that the variables follow a Gaussian dis-

tribution and instead approximate it with a kernel function. This would allow the distribution functions to be more flexible and in turn allow to obtain a better accuracy. The conditional probability would then change it for to

$$P(X|Y = y_i) = f_i(X), \tag{4.41}$$

where:
$f_i(\cdot)$ – is the kernel weighting function,

$$f_i(X) = \frac{\sum_{t=1}^{N} K_i(X, x_t)}{N}, \tag{4.42}$$

where:
$K(\cdot, \cdot)$ – is a kernel function.

Dropping the assumption of normality the weighing function and becomes

$$f_i(X) = \prod_{k=1}^{M} f_i^{(k)}(x_i k) = \prod_{k=1}^{M} (\sum_{i=1}^{N} K_i^{(k)}(x_k, x_{tk})). \tag{4.43}$$

With the one dimensional Gaussian kernel function

$$K_i^{(k)}(x_k, x_{tk}) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp(-\frac{(x_k - x_{tk})^2}{2\sigma_k^2}). \tag{4.44}$$

The drawback of this approximation is still quite influential as it needs to store more information and is computationally taxing [7].

4.2.4 k- Nearest Neighbor

In comparison to the previously presented methods is the K- Nearest Neighbor (kNN) a relatively simple one. This classifier does not really produce any models and only fits new observations to the pre-existing memorized instances from the training set. The value of y_{test} is classified using a majority vote among k nearest neighbors, hence the name. The nearness of a neighbor is based on the euclidean distance measure between the set of explanatory training data and the new test data,

$$d_i = \|x_i - x_{test}\|, \tag{4.45}$$

where

x_i – is the vector of explanatory variables for the ith observation,

x_{test} – is the new vector of explanatory variables for the new test observation.

Which concludes the construction of this classifier. More sophisticated versions and modifications are not needed for the purpose of this study.

4.2.5 Support Vector Machines

The last method used in this study would be the support vector machines (SVM) classifier, which is also the most computationally intensive one. The main idea of this classifier is to identify a hyperplane separating all observations with maximal accuracy. While this task seems relatively close to the decision tree problem, there is one major difference. Classification trees simply split the data maximizing some criterion values while SVM try to identify a data transformation function that would allow the split to be nearly perfect.

If we mask all insolvent companies with $y_i = 1$ and solvent ones with $y_i = -1$ the separating hyperplane can be expressed as,

$$x_i^\top \omega + b \geq 1 - \xi_i \text{ for } y_i = 1, \tag{4.46}$$

$$x_i^\top \omega + b \leq -1 + \xi_i \text{ for } y_i = -1, \tag{4.47}$$

where

x_i – is the explanatory vector associated with the ith observation,

ξ_i – is the classification error for the ith observation,

b – is the constant term,

ω is the vector of explanatory perimeters.

The task of the classifier is then to maximize the distance between these two supporting hyperplanes and to minimize the error rate. This can be phrased as a minimization problem (with a trade off constant Θ),

$$\min_{\omega,b,\xi_i} \frac{1}{2}\|\omega\| + \Theta \sum_{i=1}^{n} \frac{\xi_i}{\|\omega\|}, \tag{4.48}$$

with the following constrains,

$$\forall_i \; y_i(x_i^\top \omega + b) \geq 1 - \xi_i \tag{4.49}$$

$$\forall_i \; \xi_i \geq 0. \tag{4.50}$$

In order to solve this problem we need to formulate its primal form with the Lagrangian multipliers α_i and μ_i,

$$\min_{\omega,b,\xi_i} \max_{\alpha_i,\mu_i} L_p = \frac{1}{2}\|\omega\|^2 + \Theta \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i(y_i(x_i^\top \omega + b) - 1 + \xi_i) - \sum_{i=1}^{n} \xi_i \mu_i. \tag{4.51}$$

The optimality of this problem can be derived with the Kuhn- Tucker Conditions,

$$\frac{\partial L_p}{\partial \omega} = 0, \tag{4.52}$$

$$\frac{\partial L_p}{\partial b} = 0, \tag{4.53}$$

$$\frac{\partial L_p}{\partial \xi_i} = 0, \tag{4.54}$$

$$\alpha_i(y_i(x_i^\top \omega + b) - 1 + \xi_i) = 0, \tag{4.55}$$

$$\mu_i \xi_i = 0, \tag{4.56}$$

$$y_i(x_i^\top \omega + b) - 1 + \xi_i \geq 0, \tag{4.57}$$

$$\alpha_i \geq 0, \tag{4.58}$$

$$\mu_i \geq 0, \tag{4.59}$$

$$\xi_i \geq 0. \tag{4.60}$$

Therefore,

$$\sum_{i=1}^{n} \alpha_i y_i x_i = \omega \tag{4.61}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{4.62}$$

$$\Theta - \alpha_i - \mu_i = 0. \tag{4.63}$$

With these results it is possible to write the dual Lagrangian problem by substituting

elements with these equalities and reformulating the original problem,

$$\max_{\alpha_i} \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j, \tag{4.64}$$

with the following constrains,

$$0 \leq \alpha_i \leq \Theta, \tag{4.65}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0. \tag{4.66}$$

Which concludes the discussion of the linear application of this method. There still is though the possibility of finding non linear classifiers by applying kernel techniques. This happens by substituting the $x_i^\top x_j$ with a kernel function $K(x_i, x_j)$. Again, for the purposes of this study a Gaussian kernel was deemed sufficient. Since the Gaussian transformation preserves all of the necessary conditions of the inner product transformation the maximization problem can be read as,

$$\max_{\alpha_i} \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j), \tag{4.67}$$

under the same restrictions as before. Thus, it is also possible to obtain a nonlinear SVM classifier.

4.3 Comparison statistics

The last part of each research associated with data analysis and regressions in particular is related with assessing how well the model (or hypothesis) fits the reality. In the case of linear regressions this is usually expressed in the form of the adjusted R^2 value. However, due to the nonlinear nature of logistic regressions this popular statistic is unavailable and new measures are constantly introduced but cannot gain enough strength to become generally accepted [29]. With this many classifiers we need to specify also a variable to compare the effectiveness of their predictions. Since dealing with solvent and insolvent companies is a binary classification task the most obvious and simple method would derive from the values of the according

Table 4.1: Classification Table

Fitted Values		True Values		
		0	1	Total
	0	n_{00}	n_{01}	$n_{00} + n_{01}$
	1	n_{10}	n_{11}	$n_{10} + n_{11}$
	Total	$n_{00} + n_{10}$	$n_{01} + n_{11}$	n

classification tables (see Table 4.1).

The first most commonly encountered measure is the count R^2 ratio which is basically just the ratio of correct predictions,

$$R^2_{count} = \frac{n_{00} + n_{11}}{n}. \tag{4.68}$$

Also known as the accuracy rate. This ratio, however, can be misleading as with binary dependent variables one can always obtain an accuracy of at least 50% when predicting every observation to be of the more dominant class. Therefore this accuracy rate requires an adjustment,

$$\bar{R}^2_{count} = \frac{n_{00} + n_{11} - n_{max}}{n - n_{max}}, \tag{4.69}$$

where,

n_{max} – is the number of observations reporting the more dominating class.

This new adjusted ratio is only positive when the share of correct predictions exceeds the share of dominant responses in the whole population. It therefore takes into account the empirical probability and reports the adjusted percent of correct predictions.

The next method used to compare the methodologies relies on the sensitivity and specificity of the classifiers. Where with sensitivity we understand the empirical probability of observing a correct prediction for an observation with the true class $y_i = 1$

$$Pr(\hat{y}_i = 1 | y_i = 1) = \frac{n_{11}}{n_{01} + n_{11}}, \tag{4.70}$$

and specificity as the corresponding rate of correct predictions of class $y_i = 0$

$$Pr(\hat{y}_i = 0|y_i = 0) = \frac{n_{00}}{n_{00} + n_{10}}. \tag{4.71}$$

The second criterion variable measures the relation between the true positives rate (sensitivity) and the false positives rate, that is the probability of incorrectly classifying an observation of class $y_i = 1$ with class $y_i = 0$. There exists a direct connection between the false positives rate and specificity,

$$Pr(\hat{y}_i = 0|y_i = 1) = 1 - Pr(\hat{y}_i = 0|y_i = 0) = 1 - specificity. \tag{4.72}$$

Plotting the relation between the two we can obtain a very graphic illustration of the informativeness of the classifiers and directly compare their predictive power by comparing the area under the curve. In this case, we want to obtain a classifier for which the area under curve is as close to 1 as possible. This method however is by no means perfect [1, 49]. Other popular methods are the McFadden R^2, McKelvey & Zavoina R^2 and the adjusted count R^2[2]. All of them are accepted measures of model performance, but the simple nature of ROC analysis allows direct comparison between models based on different principles.

[2]The necessary R code for calculating each of them is widely accessible.

Chapter 5

Results

In this chapter we present and describe the results of our analysis. Each method is discussed separately. The last section offers a comprehensive comparison of the obtained results and provides an additional discussion of their overall effectiveness.

5.1 Logit

We start off with the logit model. The exact coefficient estimation results are reproduced in Table 5.1. As expected the model produces a good fit with many significant coefficients, where we assume variables to be significant if the hypothesis of the coefficient being equal to 0 being rejected with a 90% confidence level. A likelihood ratio test has reaffirmed the joint significance of the variables.

The first significant variable is the constant term of the regression, which would suggest that there exists some inherent risk of becoming insolvent that is not related to any of the provided financial ratios. This result seems quite logical and expected as many studies have reaffirmed hypotheses that macroeconomic shocks shape the insolvency probabilities of a market. This could also can be an indicator of missed variables. The model, however, reports high predictive power nonetheless.

Three profitability ratios report significance (x1, x2 and x4), but their direction is quite interesting. Whereas an increase in the value of ratios x1 and x4 (Net Income over Total Assets and EBITA over Net Sales) contributes to lower default probabilities, higher values of x2 (Net Income over Net Sales) are associated with higher insolvency probabilities. Variables x3 and x5 were removed from the model during

the backward feature selection process.

Next, we consider the leverage ratios and as it would seem all of them are included in the final model. Variable x11 was removed due to an abnormally high number of missing values. Of these variables all but one are significant. Apparently we cannot say much about the influence of x12 which reports the total debt to total assets ratio on the insolvency probability. Among the significant variables, an increase in x9 ((Liabilities Current Total - Cash and Short Term Investments)/Assets Total) corresponds with a lower default probability, an increase in all of the other corresponds with a higher insolvency probability.

Among liquidity ratios variables, x16 (Assets Current Total / Liabilities Current Total) was removed from the model during the feature selection process. Variables x13 and x15 (Cash and Short Term Investments / Assets Total) and (Assets Current Total - Inventories Total)/Liabilities Current Total) are related with lower insolvency probabilities, whereas x17 and x18 ((Assets Current Total -Liabilities Current Total)/Assets Total and Liabilities Current Total / Liabilities Total) lead in the opposite direction. The influence of x14 (Cash and Short Term Investments / Liabilities Total) is undetermined.

Out of the activity ratios, two ratios made it through the feature selection process, of which one is significant. A higher value in variable x22 (Accounts Payable / Sales Net) corresponds to a higher probability of default. The other, x20 (Inventories Total / Sales Net) is insignificant and we cannot interpret its influence.

In a similar fashion, two variables describing the size and growth aspects of a company are included in the final regression. Here variable x24 (Inventories Increase / Inventories Total) is of significance and an increase in its value is related to higher probabilities of insolvency.

Next we discuss the predictive power of the model in relation to how well it identifies insolvent and solvent companies in the test set. The corresponding confusion matrix is presented in Table 5.2. Out of 36 insolvent companies 29 were correctly classified. In the class of solvent companies 76 were correctly classified.

The resulting goodness of fit ratios are reported in Table 5.3. We notice that the model correctly identified 80.56% of all insolvencies and 91.57% of solvent companies, which would mean that an overall 88.24% of all observations were correctly

Table 5.1: Characteristics of the Logit Model

Variable	Coefficients	Std. Error	z value	Pr(> |z|)
Constant	-74.96062	23.54310	-3.184	0.001453 **
Profitabilty Ratios				
x1	-1.42254	0.53259	-2.671	0.007563 **
x2	0.33622	0.08413	3.997	6.43e-05 ***
x4	-0.34831	0.08250	-4.222	2.42e-05 ***
Leverage Ratios				
x6	70.66844	23.35322	3.026	0.002478 **
x7	0.43099	0.06393	6.742	1.57e-11 ***
x8	52.35627	9.20004	5.691	1.26e-08 ***
x9	-49.96624	8.15983	-6.123	9.16e-10 ***
x10	70.78910	22.83398	3.100	0.001934 **
x12	0.57018	0.99848	0.571	0.567970
Liquidity Ratios				
x13	-54.05801	8.30855	-6.506	7.70e-11 ***
x14	0.39159	0.53162	0.737	0.461368
x15	-2.68332	0.61734	-4.347	1.38e-05 ***
x17	4.37481	1.59338	2.746	0.006040 **
x18	4.90143	1.43168	3.424	0.000618 ***
Activity Ratios				
x20	-0.96650	2.86708	-0.337	0.736039
x22	0.34099	0.09366	3.641	0.000272 ***
Size and Growth Ratios				
x24	1.49899	0.78109	1.919	0.054972 .
x25	-3.17130	2.14296	-1.480	0.138909

. statistically significant at the 10% level.
* statistically significant at the 5% level.
** statistically significant at the 1% level.
*** statistically significant at the 0.1% level.

Table 5.2: Classification Table of the Logit Model for the Test Set

		True Values		
		Solvent	Insolvent	Total
Fitted Values	Solvent	76	7	83
	Insolvent	7	29	36
	Total	83	36	119

Table 5.3: Predictive Power of the Logit Model

Sensitivity	80.56%
Specificity	91.57%
ROC Area	89.06%
R^2_{count}	88.24%
$Adj.R^2_{count}$	61.11%

classified. However, if we take into consideration the correction for the most common class, the overall goodness of fit decreases to 61.11%. We can thus observe that the numerical overrepresentation of solvent companies heavily influences the overall goodness of fit statistics. Unfortunately, this influence is only hardly noticeable in ROC curve graph (see Figure 6.12).

5.2 Decision Tree

In the next section we take a closer look at the decision tree classifier methods. In particular we focus on two trees, one which makes decisions on the basis of the Gini coefficient and the other which maximizes information gain.

We start with the Gini coefficient based tree (see Figure 6.13a). As we see the decision tree method selected only two variables to be of importance for future classifications, that is variable x10 (Liabilities Total/ Assets Total) and x7 ((Equity - Intangible Assets Total)/(Assets Total - Intangible Assets- Cash and Short Term Investments - Property, Plant and Equipment)). After the first split two sub groups were created, one containing 5% insolvent companies and the other 77%. The second was then split based on the x7 variable into two new subgroups which subsequently report 17% and 97% insolvencies. The first group consist of 65% of all companies, the second of 9% and the third 26%. As seen in Figure 6.13c, the tree minimizes the relative error[1] and the tree is of maximal efficiency at just two splits.

Not surprisingly, using the information gain coefficients as the splitting criterion produced almost identical results (See Figure 6.13b). The same variable was identified as important for making the first splitting decision. The only difference is in the second splitting variable (here x6, which measures the relation between Equity

[1]Which compares the improvement in proportions to prevent overfitting.

Table 5.4: Classification Table of the Decision Trees

Fitted Values	Gini Decision Tree	True Values		
		Solvent	Insolvent	Total
	Solvent	79	9	88
	Insolvent	4	27	31
	Total	83	36	119
Fitted Values	Information Decision Tree	True Values		
		Solvent	Insolvent	Total
	Solvent	83	9	92
	Insolvent	0	27	27
	Total	83	36	119

and Assets Total), after which the second group has 28% insolvent companies and the third consists entirely of only insolvent companies. Ultimately, the second category groups 11% of all observations and the third 24%. Again, the tree obtains its minimal relative error at 2 splits (See Figure 6.13d).

Now we take a closer look at the confusion matrix produced after applying the classifier to the test set (Table 5.4). We immediately notice that even though both methods were astonishingly similar, the difference in validation results is staggering. The Gini based tree classified 31 companies as insolvent, 4 of which were incorrect. The information gain tree classified only 27 companies as insolvent, but made no false positive error.

Next, we focus on the corresponding goodness of fit ratios (Table 5.5). The Gini Decision Tree correctly classified 75% of all insolvencies and 95.18% of solvent companies. This results in an averaged accuracy of 89.08%. Again, after adjusting for predefined prior probabilities we obtain a classifier that explains 63.89% of total data variability. The Information Gain Decision Tree on the other hand correctly classified 75% of all insolvencies and all instances of solvent companies, therefore the overall accuracy rate is equal to 92.44% and remains a high value (75%) even after adjusting for priors. The corresponding ROC graphs for the training sets are reproduced in Figure 6.12.

Table 5.5: Predictive Power of the Decision Trees

	Gini Tree	Info Tree
Sensitivity	75.00%	75.00%
Specificity	95.18%	100.00%
ROC Area	88.42%	90.50%
R^2_{count}	89.08%	92.44%
$Adj.R^2_{count}$	63.89%	75.00%

5.3 Random Forest

Our next method relies on the random forest approach. As it is based upon a unsupervised machine learning algorithm no final model is produced. Therefore, the interpretation of this classifier is particularly difficult. One popular approach is to analyze which variables were the most important in the classifier construction process. We can do so by analyzing how much the average tree loses efficiency by omitting one particular variable from the feature selection process. Thus, two variables were created. One reporting the mean decrease of overall accuracy (true positive and negative rate combined) and the other, the mean decrease of the Gini coefficient. The results are presented in Table 5.6 and their corresponding graphical representation is available in Figure 6.14.

We notice that even though several variables have been selected as important by both methods there exists a certain variation in the results. Among the top five predictors with the highest importance only three are present in both results, that is x1, x24 and x25. In addition, variable x8 has been identified as important by only one of the algorithms.

If we would try to maximize the true positive rate, variables x1, x3, x23, x24 and x25 would be the most important for making decisions. On the other hand, for maximizing the True Negative rate the variables x1, x3, x6, x8, x9, x10, x23 and x24 would be the most important ones. As we see, these lists overlap in x1 and x24 which are among the most influential ones.

This is quite interesting as one of the variables, important for a single decision tree (x7) is not present in neither category and the other (x6, x10) show up as important only for maximizing the True Negative rate. Which suggests that if we consider the whole set of variables, these three have the highest explanatory power, but when we

Table 5.6: Variable Importance of the Random Forest

Variable	For Solvent	For Insolvent	For Accuracy	For Gini
Profitability Ratios				
x1	12.809844	9.476650	13.214271	16.775790
x2	8.046301	5.952271	8.355781	8.612045
x3	9.389691	7.845650	9.866903	11.131576
x4	7.976665	5.824483	8.374950	8.834374
x5	6.113404	6.381756	7.273305	5.296759
Leverage Ratios				
x6	9.839370	6.220716	9.695941	17.896324
x7	6.596589	4.185736	6.419547	5.335238
x8	9.650151	5.668727	10.240813	10.431442
x9	9.052869	4.386658	8.384251	8.949806
x10	9.671783	6.845017	9.891322	22.412084
x12	1.442062	2.543255	3.019935	2.032647
Liquidity Ratios				
x13	3.344476	5.099089	5.787288	2.813659
x14	5.049286	2.271365	4.805399	2.093007
x16	6.875580	4.441759	6.529269	3.635111
x17	7.427441	3.967435	6.664816	2.113517
x18	5.762230	5.325419	6.826140	6.810165
Activity Ratios				
x19	4.501386	4.983133	5.905863	3.021310
x20	1.737825	4.545820	4.666355	2.073684
x21	2.564005	4.293299	4.505368	2.538867
x22	5.216994	4.323180	5.456277	4.005832
Size and Growth Ratios				
x23	11.604544	7.281965	11.272150	5.334215
x24	9.403458	12.839703	13.618651	24.256939
x25	8.617498	12.392610	13.131122	20.491360
x26	0.693539	3.950166	3.915550	2.016915

Table 5.7: Classification Table of the Random Forest for the Test Set

Fitted Values		True Values		
		Solvent	Insolvent	Total
	Solvent	83	9	92
	Insolvent	0	27	27
	Total	83	36	119

aggregate the trees over multiple permutations of possible splitting variables their importance can be easily substituted by other variables.

Next we apply the forest to our test sample to measure how the classifier fares when confronted with real out-of-sample observations. The corresponding classification matrix is reproduced in Table 5.7

The confusion matrix of the random forest is much alike the confusion matrix produced by the information index based decision tree. As a matter of fact, it is identical. Even though the random forest fares better in explaining the variability of the training sample (see Figure 6.12d) when confronted with a real testing sample no improvement could be achieved. The resulting goodness of fit ratios are therefore identical to the values of the information based decision tree and can be accessed in Table 5.5 under the corresponding category.

5.4 k Nearest Neighbor

It is difficult to talk about a particular model in this case, as the k nearest neighbor is just a simple interpretation of the feature space and it does neither identify important variables nor does it interpret their influence. As a result we can only produce the outcome of categorizing new observations from the test set based on their k nearest neighbors in the training dataset (maximum efficiency was found to be with $k = 5$) with a confusion matrix (Table 5.8) and calculating the corresponding goodness of fit statistics (Table 5.9).

Apparently the method correctly identified 81 solvencies and 28 insolvencies, however 8 insolvencies and 2 solvencies were fitted to the wrong class. This supervised technique results with an overall accuracy of 92.44% and an adjusted accuracy of 72.22%. The corresponding sensitivity is equal to 77.78% and specificity to 97.59%.

Table 5.8: Classification Table of the kNN for the Test Set

Fitted Values		True Values		
		Solvent	Insolvent	Total
	Solvent	81	2	83
	Insolvent	8	28	36
	Total	89	30	119

Table 5.9: Predictive Power of kNN

Sensitivity	77.78%
Specificity	97.59%
R^2_{count}	92.44%
$Adj.R^2_{count}$	72.22%

Even though this is a simplistic fitting method based on euclidean distance it is highly effective nonetheless.

5.5 Naïve Bayes

Next we take a closer look at the predictions made by assuming that the variables are conditionally independent and follow a normal distribution (or any distribution at all in the non parametric case). Judging from the ROC graphs the parametric Bayes approach can correctly predict 80.42% of the total variability in the training sample and after dropping the assumption of normality this increases to 91.36% (See Figure 6.12e,f).

When we take a closer look at the fitted distributions (Figure 6.15, 6.16, 6.17, 6.18 and 6.19) we can observe the influence of imposing the normality assumptions on the variables. In all cases the non parametric distribution functions are characterized by the presence of sharp spikes and local extrema and only in a few cases the resemblance between the normal and kernel distributions is more or less preserved. In the case of variables x1 to x5 (belonging to the profitability ratios) it is hard to argue that the fitted non parametric distributions resemble anything different than simple spikes for particular values. The same situation occurs for all variables representing the leverage ratios. The picture changes however, when we observe the fitted distributions of the liquidity ratios. Here the variables x13, x15, x16 and x18 show

Table 5.10: Classification Table of the Parametric Bayes for the Test Set

Fitted Values		True Values		
		Solvent	Insolvent	Total
	Solvent	71	9	80
	Insolvent	12	27	39
	Total	83	36	119

Table 5.11: Classification Table of the Non- Parametric Bayes for the Test Set

Fitted Values		True Values		
		Solvent	Insolvent	Total
	Solvent	80	36	116
	Insolvent	3	0	3
	Total	83	36	119

similar distributions in both the parametric and non- parametric case. In the next category (activity ratios) this is only the case for variable x20. Variables of the size and growth ratios, with the exception of x26, show particular strong resemblance. As a next step we take a closer look at the corresponding confusion matrices (Tables 5.10 and Table 5.11) and we immediately notice that the non parametric Bayes fails to correctly identify even a single insolvency. Not only that, the classifier even incorrectly classified 3 solvent companies as insolvent. This would be a clear argument against overfitting models to the training database. In comparison, the non-parametric classifier was able to correctly identify 91.36% of all observations in the training sample, but when confronted with new data it proved to be of abysmal performance.

If we then take a look at the corresponding goodness of fit ratios we notice that the parametric Bayes correctly identified 85.54% of all solvent companies and 75% of insolvent ones in the test data which corresponds with a 82.35% overall accuracy rate (with the adjusted value of 41.67%). Another interesting observation would be that the non-parametric Bayes actually reports negative explanatory power. This classifier performs worse than randomly assigning 30% of observations to the insolvency class.

Table 5.12: Predictive Power of the Naïve Bayes

	Non Parametric	Parametric
Sensitivity	0.00%	75.00%
Specificity	96.39%	85.54%
ROC Area	91.36%	80.42%
R^2_{count}	67.23%	82.35%
$Adj.R^2_{count}$	-8.33%	41.67%

Table 5.13: Classification Table of the SVM for the Test Set

Fitted Values	True Values		
	Solvent	Insolvent	Total
Solvent	79	8	87
Insolvent	4	28	32
Total	83	36	119

5.6 Support Vector Machines

Last but not least, we look at the performance of the support vector machines. Again, since this approach only classifies new information based on existing supporting hyperplanes it is difficult to talk about interpreting the variables. Therefore, we immediately begin with discussing the corresponding confusion matrix (Table 5.12) and goodness of fit statistics (Table 5.13).

As we see, the support vector machines approach correctly classified 79 solvencies and 28 insolvencies, which lead to the overall accuracy of 89.91% for the test set in comparison to the 93.24% for the training set. After adjusting for the prior probabilities in the test set the accuracy drops to 66.67%. In total, the sensitivity estimate of 77.78% and specificity of 95.18%.

Table 5.14: Predictive Power of the SVM

Sensitivity	77.78%
Specificity	95.18%
ROC Area	93.24%
R^2_{count}	89.91%
$Adj.R^2_{count}$	66.67%

5.7 Discussion of predictive power of the methodologies

With this we finish presenting the results used to asses the effectiveness of individual methods and in the next part we compare their overall precision to one another. As an additional criterion the informativeness of particular classifiers is discussed. The important classifier characteristics are repeated in Table 5.15.

As mentioned before, the main goal of the techniques is to identify insolvencies as efficient as possible, thus the main goal is to maximize the sensitivity rate (true positives). However, this is not a particularly efficient approach as the best classifier was able to correctly identify only two more insolvencies than the fifth best classifier and only one more than the second and third best classifier. As we see, in this regard all of the classifiers were particularly efficient with the exception of the non parametric Bayes approach which was as bad as it possibly could be. Which again strongly emphasizes the danger of overfitting models to training data. The non-parametric Bayes was perfectly fitted for the training data, but once new observations had to be classified it became evident that it is completely biased towards the training data set.

We then take a closer look at the other classification rate, namely specificity, which measures the true negative rate. Here the picture is less obfuscated, as both the information based decision tree and random forest have achieved perfect scores. Other than that, all other methods were also able to break the 90% rate as well. Is seems that identifying solvent companies is easier than looking for insolvent ones.

If we consider the overall accuracy and the adjusted accuracy rate in the form of the R^2 and $\bar{R}^2$ we notice that three classifiers accomplished a score that surprasses the 90% threshold in the general case. These classifiers were based on the information based decision tree approach, random forest and the k nearest neighbor. However, if we adjust the values for the prior insolvency probabilities these percentages drop by a significant amount. Still, both the decision tree and random forest were 75% more accurate than randomly assigning insolvencies based on the prior probabilities.

Since some of the methods consider individual variables as important for the decision making process we should also compare the overall informativeness of the

classifiers. We immediately must consider the kNN and SVM approaches as the least informative ones as they do not offer any kind of information about what is important for making the decision or which variables are the most influential ones. Both are based on relatively simple mathematical fitting algorithms and as such do not offer any insight into the 'black box'. The naive Bayes approaches are also quite uninformative as both just jointly fit data to the approximated data distributions. However, possible insight can be reached based on these distributions. If there is a clear distinction between the distributions for solvent and insolvent companies we can argue that these variables could be indicative for potential risk. Unfortunately, we cannot speculate about their influence and the magnitude of particular associated risks. Thus, although the naive Bayes offers additional information over the aforementioned methods, the overall classification process is very much a mystery.

The next three methods have clearly specified decision making criteria. The most simple one, namely the decision trees, were both able to identify two variables based on which the classifications were made. These variables originated both from the group of variables representing various leverage ratios (x10 and x7 or x6). However, there is only a minor difference between x7 and x6 as both basically express the equity to assets ratio in ever so slightly different forms. The x10 variable compares the general balance sheet structure and by how much the liabilities outweigh the assets.

More information can be gathered from the random forest approach, which ranked all of the explanatory variables by their usefulness for the classification task. This method is therefore more informative regarding which ratios are important or need particular attention. In addition to the variables identified by the decision trees, the random forest also hints for potential significance of the general profitability ratio (x1) and two of the growth ratios (x24, x25). As such, it can be argued that it identified individual influential variables from every variable category, except for the liquidity ratios. However, the general equity to assets ratio is also listed as the forth most influential variable when considering the Gini coefficient, thus it should not be disregarded.

By all measures we have to admit that the logistic regression offers the most insight into the classification process as for each possible value of each variable we

Table 5.15: Summary of Predictive Power

Classifier	Sensitivity	Specificity	R^2	adj.-R^2
Logit	80.56%	91.57%	88.24%	61.11%
Decision Tree (Info)	75.00%	100%	92.44%	75.00%
Decision Tree (Gini)	75.00%	95.18%	89.08%	64.89%
Random Forest	75.00%	100%	92.44%	75.00%
kNN	77.78%	97.59%	92.44%	72.22%
parametric Bayes	75.00%	85.54%	82.35%	41.67%
non-parametric Bayes	0.00%	96.39%	67.23%	-8.33%
SVM	77.78%	95.18%	89.91%	66.67%

can determine a particular, individual risk of a single company. It also provides us with additional information regarding significant variables and the direction of their influence. As it turns out, numerous variables are listed as significant for the logistic regression. With the highest degree of confidence the variables measuring the profitability, leverage and liquidity characteristics of a company are especially important. Also, the dynamics of accounts payable is an important aspect as well. However, when it comes to creating individual decisions or interpreting particular coefficient values, the logistics regression is less understandable than decision trees. With the logistic approach the marginal values have to be computed on an individual basis and there is no straightforward interpretation for particular values, as the overall business risk is correlated with numerous variables and the changes in probabilities depending on one variable can only be formulated while keeping all other variables constant. Which is impossible in the business field.

To sum up, the logistic regression boasts the highest sensitivity rate, that means that it is the most efficient insolvency predictor. However, it is also burdened with a relatively high false positives rate, as it classifies a high number of solvent companies as insolvent as well. In this regard, the most efficient ones were the decision trees and the random forest approach. Overall, the latter are also characterized with the highest overall accuracy rate, both in the general and adjusted case. The easiest classifier to use in practice is the decision tree, as it depends on a simple two questions procedure. On the other hand, it is also the least informative of the three. The most informative would be the logistic regression.

Chapter 6

Conclusion

The purpose of this thesis was to assess the effectiveness of using machine learning algorithms in predicting insolvencies and to compare their efficiency to the popular logistic regression. The examined methods list such techniques as decision trees, random forests, naive Bayes, support vector machines and the k Nearest Neighbor. The investigation was founded on existing research and the methodologies are backed by mathematical evidences. The potential explanatory variables, as well as the dependent variables, are all consistent with previous studies in this field. The quantitative analysis was performed on a database provided by the Wharton Research Data Services to which access was gained through the subscription of the University of Vienna. Despite data limitations, such as high number of missing values as well as a strongly unbalanced sample, it was still possible to perform an analysis of performance of the aforementioned techniques. The efficiency of those techniques was verified on a set of out of sample observations and measured through several goodness of fit statistics.

As shown through the analysis, the machine learning classifiers are competitive alternatives to the logistic regression. Despite dropping the strong assumptions necessary for the regression approach, they still achieve high accuracy rates. Non surprisingly, the naive Bayes approaches are remarkably weak when confronted with out of sample observations which strengthens the claim that assuming normality or conditional independence of the variables is an especially poor choice of assumptions. All other methods achieved similar overall accuracy.

The logistic regression, decision trees and random forest were able to identify sig-

nificant variables for the decision making process. Despite the simple procedure the decision tree approach was able to achieve a higher adjusted accuracy than the logistic regression. Due to their construction the kNN and SVM classifiers do not offer any additional insight.

The practical application resulted in similar results for all methods, making it impossible to declare one to be superior to all other. All of the methods can however be useful for particular reasons. The Bayesian approach is able to verify assumptions on distributions, random forest to identify important variables and their interactions, decision trees for easy implementations and the logistic regression for a study of marginal effects.

In the end, all methods, with the exception of the non-parametric Bayes, were able to classify new data with a high degree of accuracy based on the predefined financial ratios. As mentioned before, several of them were also able to identify potential significant predictors among them. The classifiers were able to explain up to 75% of total data variance.

As a conclusion, we must state that selecting one classifier as the most efficient is impossible with current data. Each of them offers different potential uses and is viable for different goals.

The decision trees offer easy to understand rules which allow to identify troubled companies ahead of time and investigate. Due to their structure they are both inexpensive to implement and quick to provide results following simple measures, providing a great tool for initial pre-selection of potential defaults for banks and insurance companies alike.

On the other hand, machine learning techniques suffer from certain limitation. The strongest being that they only provide limited insight into the 'black box' of the decision making process. They do not model the reality, they merely follow association rules. Therefore, logistic regression, even though itself subject to powerful restrictions, is still an important tool for understanding the mechanisms and for investigating the relation between financial ratios and risk of default.

However, it would seem that machine learning techniques offer potential new insight in the model building process as well. The random forest approach works great for the feature selection process through identifying important variables, while Bayesian

approximations are able to verify assumptions on distributions. While these techniques cannot replace scientific modeling they serve as a powerful tool for research and practical application nonetheless.

Bibliography

[1] Ash A. & Shwartz M., (1999), R^2: a useful measure of model performance when predicting a dichotomous outcome, *Statistics in Medicine*, Vol. 18, pp.375 - 384

[2] Altman, E. I., (1968), Fiancial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy, *Journal of Finance*, Vol. 23, Issue 4, pp. 589 - 611

[3] Altman, E. I., Haldeman, R. G., Narayanam, P., (1977), Zeta-analysis. A New Model to Identify Bankruptcy Risk of Corporations, *Journal of Banking and Finance*, Vol.1 , Issue 1., pp. 29 - 54

[4] Altman, E. I., Sabato, G., (2007), Modelling Credit Risk for SMEs: Evidence from the U.S. Market, *Abacus*, Vol. 43, Issue 3, pp. 332 - 357

[5] Balls M., Amcoff P., Bremer S., Casati S., Coecke S., Clothier R., Combes R., Corvi R., Curren R., Eskes C., Fentem J., Gribaldo L., Halder M., Hartung T., Hoffmann S., Schectman L., Scott L., Spielmann H., Stokes W., Tice R., Wagner D. & Zuang V., (2006), The principles of weight of evidence validation of test methods and testing strategies, *Altern Lab Anim*, Vol. 34 (6), pp. 603 - 620

[6] Bennette C. & Vickers A., (2012), Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents, *BMC Medical Research Methodology*, Vol. 12 (21)

[7] Bouckaert, R. R., (2005), Naive Bayes Classifiers That Perform Well with Continuous Variables, *Lecture Notes in Computer Science*, Vol. 3329, pp. 1089 - 1094

[8] Breiman L., (2001), Random Forests, *Machine Learning*, Vol 45 (1), pp. 5 - 32

[9] Breiman L., (2001), Statistical Modeling: The Two Cultures, *Statistical Science*, Vol. 16, Issue 3, pp. 199-231

[10] Creamer, G., (2009), Using Random Forests and Logistic Regression for Performance Prediction of Latin American ADRS and Banks, *Journal of CENTRUM Cathedra*, Vol. 2 (1), pp.24-36

[11] Diaz, Z., Segovia, M. J., Fernandez, J., del Pozo, E. M., (2005), Machine Learning and Statistical Techniques. An Application to the Prediction of Insolvency in Spanish Non- Life Companies, *The International Journal of Digital Accounting Research*, Vol. 5, Issue 9, pp. 1 - 45

[12] Ding, Y., Entwistle, M. & Stolowy, H., (2007), Identifying and Coping with Balance Sheet Differences: A Comparative Analysis of U. S., Chinese, and French Oil and Gas Firms Using the "Statement of Financial Structure", *Issues in Accounting Education*, Vol. 22 (4), pp. 591-606

[13] Dixon, M., Chong, J., (2012), A Bayesian Approach to Ranking Private Companies based on Predictive Indicators, Available at SSRN: http://ssrn.com/abstract=2096425

[14] Duchessi P., Lauría E. J. M., (2013), Decision tree models for profiling ski resprts' promotional and advertising strategies and the impact on sales, *Expert Systems with Applications*, Vol. 40, pp. 5822 - 5829

[15] Dzidzeviciute, L., (2010), Statistical Scoring Model of Lithuanian Companies, *Ekonomika*, Vol. 89, Issue 4, pp. 96 - 115

[16] Fantazzini, D., Figini, S. (2009), Random Survival Forest Models for SME Credit Risk Measurement, *Methodology and Computing in Applied Probability*, Vol. 11 Issue 1, pp. 29-45

[17] Figini, S., Fantazzini, (2009), Random Survival Forests Models for SME Credit Risk Measurement, *Methodology and Computing in Applied Probability*, Vol. 11 (1)

[18] Franks, J., Lóránth, G., (2005), A Study of Inefficient Going Concerns in Bankruptcy, *CEPR Discussion Paper*, №5035, London, Centre for Economic Policy Research

[19] Gelman, A., Carlin, J. B., Stern, H. S., (2013), Bayesian Data Analysis, 3rd Edition, *CRC Press*

[20] Greene, W. H. (1997), Econometric Analysis, 3rd Edition, *Prentice Hall*, pp. 871 - 896

[21] Härdle, W., Lee, Y., Schäfer, D., Yeh, Y., (2007), The Default Risk of Firms Examined with Smooth Support Vector Machines, *German Institute for Economic Research Discussion Papers*, Nr 757

[22] Härdle, W., Moro, R., Hoffmann, L., (2010), Learning Machines Supporting Bankruptcy Prediction, *SFB 649 Discussion Paper 2010-032*

[23] Hensher, D. A., Jones, S., (2007), Forescasting Corporate Bankruptcy: Optimizing the Performance of the Mixed Logit Model, *Abacus*, Vol. 43, Issue 3, pp. 241 - 264

[24] Hofmarcher, P., Kerbl, S., Grün, B., Sigmund, M & Hornik, K., (2011), Model Uncertainty and Aggregated Default Probabilities: New Evidence from Austria, *Research Report Series*, Report 116, Vienna University of Economics and Business

[25] Kaski S., Sinkkonen J., Nikkilä J., (2001), Clustering Gene Expression Data by Mutual Information with Gene Function, *Lecture Notes in Computer Science*, Vol. 2130. pp. 81 - 87

[26] Kartasheva, A. V. & Traskin, M., (2011), Insurers' Insolvency Prediction using Random Forest Classification, *Journal of Risk and Insurance*, *revised and resubmited*

[27] Mukkamala, S., Sung, A. H., Ribeiro B., Vieira A. S. & Neves, J. C., (2006) Model Selection and Feature Ranking for Financial Distress Classification, *Proceedings of 8th International Conference on Enterprise Information Systems (ICEIS 2006)*, INSTICC - Institute for Systems and Technologies of Information, Control and Communication. (in press)

[28] Mukkamala, S., Tilve, G. D., Sung, A. H., Ribeiro, B. & Vieira, A. S., (2006), Computational Intelligent Techniques for Financial Distress Detection, *International Journal of Computational Intelligence Research*, Vol. 2 (1), pp. 60-65

[29] Murrell, B., Murell, D. & Murrell, H. (2013), Discovering general multidimensional associations, *arXiv*, eprint 1303.1828

[30] Mycielski, J. (2010), Ekonometria, 3rd Edition, *University of Warsaw*, pp. 261 - 279

[31] Nehrebecka, N. & Dzik, A. M., (2013), Business Demography in Poland: Microeconomic and Macroeconomic Determinants of Firm Survival, *Working Paper*, №8 (93), University of Warsaw

[32] Newson, R., (2002), Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences, *The Stata Journal*, Vol. 2 (1), pp. 45 - 64

[33] Nikolic, N., Zarkic- Joksimovic, N., Stojanovski, D., Joksimovic, I., (2013), The application of brute force logistic regression to corporate credit scoring models: Evidence from Serbian financial statementsm *Expert System with Applications 40*, Vol. 40, pp. 5932 - 5944

[34] Oruc, Ö. E., Kuruoglu E. & Gündüz, (2011), Entropy Applications for Customer Satisfaction Survey in Information Theory, *Frontiers in Science*, Vol. 1 (1), pp. 1 - 4

[35] Pranab, K. S., (2005), Gini diversity index, Hamming distance, and the curse of dimensionality, *International Journal of Statistics*, Vol. 63 (3), pp. 329 - 349

[36] Raileanu, L., E. & Stoffel, K., (2004), Theoretical comparison between the Gini Index and Information Gain criteria, *Annals of Mathematics and Artificial Intelligence*, Vol. 41, pp. 77 - 93

[37] Rudorfer, G., (1995), Early Bankruptcy Detection Using Neural Networks, *APL '95 Proceedings of the international conference on Applied programming languages*, pp. 171-178

[38] Rokach L. & Maimon O., (2005), Decision Trees, *The Data Mining and Knowledge Discovery Handbook*, pp. 165 - 192.

[39] Sahin Y., Bulkan S. & Duman E., (2013), A cost-sensitive decision tree approach for fraud detection, *Expert Systems with Applications*, Vol. 40, pp. 5916 - 5923

[40] Shi T. & Horvath S., (2006), Unsupervised Learning with Random Forest Predictiors, *Journal of Computational andGraphical Statistics*, Vol. 15 (1), pp. 118 - 138

[41] Strobl, C., Boulesteix, A., Kneib, T., Augustin T., Zeileis, A., (2008), Conditional variable importance for random forests, *BMC Bioinformatics*, Vol 9, Issue 307, pp. 1 - 11

[42] Sugiyama M., Yamada M., Kimura M. & Hachiya H., (2011), On information-maximization clustering: Tuning parameter selection and analytic solution, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 65-72

[43] Therneau, T. M., Atkinson, E. J., Mayo Foundation, (2013), An Introduction to Recursive Partitioning Using the RPART Routines

[44] Torgo, L., (2011), Data Mining with R; Learning with Case Studies, *Chapman & Hall/CRC*

[45] Torres, C. & Kearns, J., (2013, June 20), *Bernanke Sees Beginning of End for Fed's Record Easing* retrived from http://www.bloomberg.com/news/2013-06-20/bernanke-sees-beginning-of-end-this-year-for-fed-s-record-easing.html on July 11$^{\text{th}}$ 2013

[46] Witten, I. H., Frank, E., Hall, M. A., (2011), Practical Machine Learning Tools and Techniques, 3$^{\text{rd}}$ Edition, *Elsevier*

[47] Wrzosak, M.& Ziemba, A., (2009), Constructionof a Rating Based on a Bankruptcy Prediction Model, *Credit Research Centre Conference Achive*, Stat-Consulting

[48] Zhao, Y., (2012), R and Data Mining: Examples and Case Studies, *Elsevier*

[49] Zheng B. & Agresti A., (2000), Summarizing the predicitve power of a generalized linear model, *Statistics in Medicine*, Vol. 19, pp. 1771 - 1781

[50] Zumbrun, J., Torres, C. & Matthews, S., (2013, July 11), *Bernanke Supports Continuing Stimulus Amid Debate Over QE* retrived from http://www.bloomberg.com/news/2013-07-10/bernanke-backs-stimulus-for-foreseeable-future-amid-qe-debate.html on July 11$^{\text{th}}$ 2013

Appendix 1. R Code

```
# ==============================
# Author:   Hubert Tchorzewski
# Date:     November 2013 / Jannuary 2014
# Purpose: MSC Thesis
# ==============================

#Data Preparation
setwd("D:/Dropbox/MSC/R_Code_chunks") #Set up working Directory
data <- read.csv("collected2.csv", head=TRUE) #Loading initial Database

#Creating Dependable
summary(as.factor(data$dlrsn)) #1.686133% of Liquidated/Bankrupt companies

attach(data)
# We assume every company is active
y<- rep("A", length(gvkey))
n<-length(gvkey)-1

# If the company is inactive in the last year it is marked as "I"
for (i in 2:n) {
  if (gvkey[i]==gvkey[i-1] & gvkey[i]!=gvkey[i+1] & costat[i]=="I") {
    y[i]<-"I"
  }
}

# Correction for companies that are listed only once in the database
for (i in 2:n) {
  if (gvkey[i]!=gvkey[i-1] & gvkey[i]!=gvkey[i+1] & costat[i]=="I") {
    y[i]<-"I"
  }
}

# A company is insolvent if the deletion reason is '2' or '3'
for (i in 2:n) {
  if (y[i]=="I"){
    if (dlrsn[i]=="2" | dlrsn[i]=="3")
    { y[i]<-"I"  }
    else {y[i]<-"A"}
}
}

# Now we mask insolvency with 'y=1' and solvency with 'y=0'
for (i in 1:(n+1)) {
  if (y[i]=="I")
      {y[i]<-"1"}
  else {y[i]<-"0"}

}

#Remove unnecessary variables
rm(i)
rm(n)

aggregate(y, by=list(dlrsn), FUN="summary")
data$y<-as.factor(y)
data$dlrsn<-as.factor(data$dlrsn)
#Adding a new level to handle NA later on
levels(data$dlrsn) <- c(levels(data$dlrsn), "8")
detach(data)
rm(y)

# Remove uncecessary variables
data$indfmt     <- NULL
data$consol     <- NULL
data$popsrc     <- NULL
data$datafmt    <- NULL
data$curcd      <- NULL
data$artfs      <- NULL
data$xagt       <- NULL
data$opiti      <- NULL
```

```
data$datadate   <- NULL
data$row.names  <- NULL
data$gvkey      <- NULL
data$ch         <- NULL
data$cogs       <- NULL
data$dd1        <- NULL
data$dd2        <- NULL
data$dd3        <- NULL
data$dd4        <- NULL
data$dd5        <- NULL
data$dltt       <- NULL
data$gp         <- NULL
data$intpn      <- NULL
data$invwip     <- NULL
data$ppegt      <- NULL
data$rectr      <- NULL
data$teq        <- NULL
data$ugi        <- NULL
data$wcap       <- NULL
data$costat     <- NULL

#Creatin the financial ratios
attach(data)
data$x1         <-ni/at
data$x2         <-ni/sale
data$x3         <-ebit/at
data$x4         <-ebit/sale
data$x5         <-(ebit+dpact)/at
data$x6         <-seq/at
data$x7         <-(seq-intan)/(at-intan-che-ppent)
data$x8         <-lct/at
data$x9         <-(lct-che)/at
data$x10        <-lt/at
data$x11        <-ebitda/tie
data$x12        <-dt/at
data$x13        <- che/at
data$x14        <-che/lt
data$x15        <-(act-invt)/lct
data$x16        <-act/lct
data$x17        <-(act-lct)/at
data$x18        <-lct/lt
data$x19        <-at/sale
data$x20        <-invt/sale
data$x21        <-rect/sale
data$x22        <-ap/sale
data$x23        <-log(at)
data$x24        <-invch/invt
data$x25        <-apalch/lt
data$x26        <-chech/che
detach(data)

# Getting rid of infinite values
data<-data[!is.infinite(data$x1),]
data<-data[!is.infinite(data$x2),]
data<-data[!is.infinite(data$x3),]
data<-data[!is.infinite(data$x4),]
data<-data[!is.infinite(data$x5),]
data<-data[!is.infinite(data$x6),]
data<-data[!is.infinite(data$x7),]
data<-data[!is.infinite(data$x8),]
data<-data[!is.infinite(data$x9),]
data<-data[!is.infinite(data$x10),]
data<-data[!is.infinite(data$x11),]
data<-data[!is.infinite(data$x12),]
data<-data[!is.infinite(data$x13),]
data<-data[!is.infinite(data$x14),]
data<-data[!is.infinite(data$x15),]
data<-data[!is.infinite(data$x16),]
data<-data[!is.infinite(data$x17),]
data<-data[!is.infinite(data$x18),]
data<-data[!is.infinite(data$x19),]
data<-data[!is.infinite(data$x20),]
data<-data[!is.infinite(data$x21),]
```

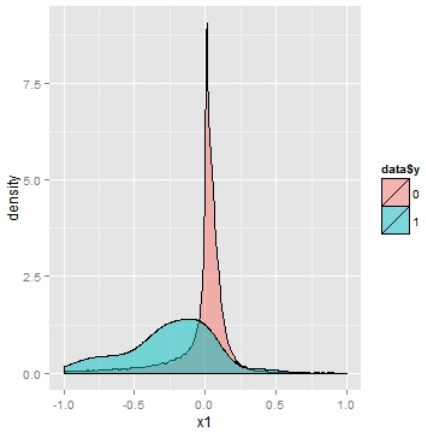
```
data<-data[!is.infinite(data$x22),]
data<-data[!is.infinite(data$x23),]
data<-data[!is.infinite(data$x24),]
data<-data[!is.infinite(data$x25),]
data<-data[!is.infinite(data$x26),]

# Subsetting for Insolvencies that happened not later than 2011
data2<-data
# With the subset command NA would disappear thus we replace them with a different value
data2$dldte[is.na(data2$dldte)] <- 20000001
data2<-subset(data2, data2$dldte<=20110000)
data<-data2
data2<-NULL

summary(data$y)
aggregate(data$y, by=list(data$fyear), FUN="summary")

# Removing now unnecessary variables
data$fyear      <-NULL
data$act        <-NULL
data$ap         <-NULL
data$apalch     <-NULL
data$at         <-NULL
data$che        <-NULL
data$chech      <-NULL
data$dpact      <-NULL
data$dt         <-NULL
data$ebit       <-NULL
data$ebitda     <-NULL
data$intan      <-NULL
data$invch      <-NULL
data$invt       <-NULL
data$lct        <-NULL
data$lt         <-NULL
data$ni         <-NULL
data$ppent      <-NULL
data$rect       <-NULL
data$sale       <-NULL
data$seq        <-NULL
data$tie        <-NULL


y<-as.factor(data$y)

# For the descriptive statistics
library(doBy)
# Repeat for x1 to x26
summary(data$x1)
# Repeat for x1 to x26
aggregate(x1 ~ y, data=data , FUN=summary)
library(ggplot2)
# Repeat for x1 to x26
qplot(x1, data=data, xlim=c(-1,1) ,geom='density', fill=data$y,alpha=I(0.5))
# Repeat for x1 to x26
qplot(y,x1, data=data, ylim=c(-2,2),geom='boxplot')

# Removing x11 because of high number of missing values
data$x11        <-NULL

# Sampling
# Subsample of only insolvent companies
positives<-subset(data, y=="1") # Bankrupts and Liquidated
# Only insolvencies after 2000
positives<-positives[!positives$dldte<=20000000,]
# Subsample of solvent companies
negatives<-subset(data, y=="0")
#Replacing missing values in deletion reason for solvent companies with code "8"
negatives$dlrsn[is.na(negatives$dlrsn)]<-"8"
# We remove all instances of "other" from Data
negatives<-subset(negatives, negatives$dlrsn!=7)
negatives<-subset(negatives, negatives$dlrsn!=10)
negatives<-subset(negatives, negatives$dldte>=20000000)
library(DMwR)
```

```
# Removing observations reporting 20% or more missing values
negatives<-negatives[-manyNAs(negatives),]
# For the rest missing values are imputed with the median
negatives<-centralImputation(negatives)
positives2<-centralImputation(positives)

# Random Sample with 70% solvent and 30% insolvent companies
set.seed(9876)
healthysample <- negatives[sample(1:nrow(negatives), 7*nrow(positives2)/3,replace=TRUE),]
totalsample<-rbind(positives2,healthysample)

#Training and Test Dataset (1:4 proportion)
d <- sort(sample(nrow(totalsample), nrow(totalsample)*.8))
train<-totalsample[d,]
test<-totalsample[-d,]
rm(d)

library(DMwR)
testBN          <-test
testBN$dlrsn    <-NULL
testBN$dldte    <-NULL

trainBN         <-train
trainBN$dlrsn   <-NULL
trainBN$dldte   <-NULL

# Logit
myFormula <- y ~ x1 + x2 + x3 + x4 +x5 + x6+x7 +x8+x9+x10+x12+x13
+x14+x15+x16+x17+x18+x19+x20+x21+x22+x23+x24+x25+x26
BenchLog<- glm(y ~ x1 + x2 + x3 + x4 +x5 +x6 + x7 +x8+x9+x10+x12+x13
+x14+x15+x16+x17+x18+x19+x20+x21+x22+x23+x24+x25+x26, data=trainBN, family=binomial)
step(BenchLog,direction="backward",trace=FALSE)
BenchLog<- glm(y ~ x1+x2+x4+x6+x7+x8+x9+x10+x12+x13
+x14+x15+x17+x18+x20+x22+x24+x25, data=trainBN, family=binomial)
library(ROCR)
summary(BenchLog)
score<-predict(BenchLog,type='response',testBN)
pred<-prediction(score,testBN$y)
perf<-performance(pred,"tpr","fpr")
performance(pred,"auc")
plot(perf) #Roc Curve
text(x=0.5,y=0.1,'Area ander Curve = 88.87%')
#KS Statistic
max(attr(perf,'y.values')[[1]]-attr(perf,'x.values')[[1]])
probabilities <- predict.glm(BenchLog, testBN, type="response")
predictions <- cut(probabilities, c(-Inf,0.5,Inf), labels=c("a","b"))
table(predictions, test$y)
plot(residuals(BenchLog, type="deviance"))
anova(BenchLog, test="Chisq")
library(lmtest)
lrtest(BenchLog)

#Decision Tree
library(rpart)
library(rattle)                                  # Fancy tree plot
library(rpart.plot)                              # Enhanced tree plots
library(RColorBrewer)                            # Color selection for fancy tree plot
library(party)                                   # Alternative decision tree algorithm
library(partykit)
treeGini <- rpart(myFormula, data=trainBN, cp=.04)
fancyRpartPlot(treeGini, xpd = NA,uniform=TRUE)
score1<-predict(treeGini,type='prob',testBN)
pred1<-prediction(score1[,2],testBN$y)
perf1<-performance(pred1,"tpr","fpr")
performance(pred1,'auc')
plot(perf1)
text(x=0.5,y=0.1,'Area ander Curve = 88.42%')
treeInfo<-rpart(myFormula,data=trainBN,parms=list(split='information'),cp=.072)
fancyRpartPlot(treeInfo, uniform=TRUE)
score2<-predict(treeInfo,type='prob',testBN)
pred2<-prediction(score2[,2],testBN$y)
perf2<-performance(pred2,"tpr","fpr")
performance(pred2,'auc')
```

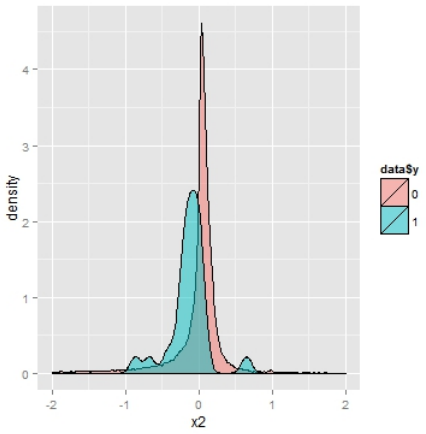
```
plot(perf2)
text(x=0.5,y=0.1,'Area ander Curve = 90.50%')
plotcp(treeGini,upper='splits')
plotcp(treeInfo,upper='splits')
predDT1<-predict(treeGini, newdata=testBN, type="class")
table(testBN$y, predDT1)
predDT2<-predict(treeInfo,newdata=testBN,type="class")
table(testBN$y,predDT2)
library(party)

plot(perf1,col='red',lty=1,main='Tree vs Tree with Prior Prob')
plot(perf2,col='green',add=TRUE,lty=2)
plot(perf, col='blue', add=TRUE, lty=3)
legend(0.6,0.6,c('simple tree','tree with 70/30 prior','Logit'),
col=c('red','green', 'blue'), lwd=3)

# RF
library(randomForest)
arf<- randomForest(y ~ x1 + x2 + x3 + x4+x5 +x6 +x7+ x8 +x9+x10+x12+x13
+x14+x16+x17+x18+x19+x20+x21+x22+x23+x24+x25+x26,
 data=trainBN, importance=TRUE, proximity=TRUE, ntree=500,
keep.forest=TRUE, na.action=na.roughfix,confusion=TRUE)
varImpPlot(arf)
importance(arf)

testRF<-predict(arf,testBN, type='prob')[,2]
predRF<-prediction(testRF,testBN$y)
perfRF<- performance(predRF,"tpr","fpr")
performance(predRF, 'auc')
plot(perfRF)
text(x=0.5,y=0.1,'Area ander Curve = 98.56%')
print(arf)

yRFTest<-testBN$y
testRFt<-predict(arf,testBN,type='class')
table(observed=testBN[,'y'], predict=testRFt,dnn=c('actual','predicted'))

# KNN
library(class)
trainKNN<-trainBN[,-1]
testKNN<-testBN[,-1]
cl<-trainBN[,1]
NN<-knn(trainKNN, testKNN, cl, k =5, prob=TRUE) #look at the prob arg. #then 2 5 4 3
table(NN,testBN[,1])

library(hmeasure)
scores.knn <- attr(NN,"prob")
# this is necessary because k-NN by default outputs
# the posterior probability of the winning class
scores.knn[NN=="0"] <- 1-scores.knn[NN=="0"]
results <- HMeasure(testBN$y,scores.knn)
plotROC(results,which=1)

# Naive Bayes Parametric
library(klaR)
library(caret)
yTrain <- trainBN$y
xTrain <- trainBN[,c("x1","x2","x3","x4","x5","x6","x7","x8","x9","x10","x12","x13"
,"x14","x15","x16","x17","x18","x19","x20","x21","x22","x23","x24","x25","x26")]
yTest  <- testBN$y
xTest  <- testBN[,c("x1","x2","x3","x4","x5","x6","x7","x8","x9","x10", "x12","x13",
"x14","x15","x16","x17","x18","x19","x20","x21","x22","x23","x24","x25","x26")]

# Naive Bayes Parametric
res.naive.p <- NaiveBayes(y ~ x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x12+x13
+x14+x15+x16+x17+x18+x19+x20+x21+x22+x23+x24+x25+x26, data=trainBN)
res.naive.p
plot(res.naive.p)
testNBp <- predict(res.naive.p, xTest, type='prob')
predNBp <- prediction(testNBp$posterior[,2],testBN$y)
perfNBp <- performance(predNBp,"tpr","fpr")
performance(predNBp, 'auc')
plot(perfNBp)
```

```
text(x=0.5,y=0.1,'Area under Curve = 80.42%')
table(predict(res.naive.p,xTest)$class,yTest)

# Naive Bayes Non-Parametric
res.naive.n <- NaiveBayes(y ~ x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x12+x13
+x14+x15+x16+x17+x18+x19+x20+x21+x22+x23+x24+x25+x26,
                          usekernel = TRUE, data=trainBN)
res.naive.n
plot(res.naive.n)
testNBn <- predict(res.naive.n, xTest, type='prob')
predNBn <- prediction(testNBn$posterior[,2],testBN$y)
perfNBn <- performance(predNBn,"tpr","fpr")
performance(predNBn,'auc')
plot(perfNBn)
text(x=0.5,y=0.1,'Area under Curve = 91.36%')

table(predict(res.naive.n,xTest)$class,yTest)

# SVM
library(e1071)
svm.model<- svm(y ~ x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x12+x13
+x14+x15+x16+x17+x18+x19+x20+x21+x22+x23+x24+x25+x26,data=trainBN,probability=TRUE)
xSVTest  <- testBN[,c("x1","x2","x3","x4","x5","x6","x7","x8","x9","x10","x12","x13",
"x14","x15","x16","x17","x18","x19","x20","x21","x22","x23","x24","x25","x26")]
ySVTest  <- testBN$y
testSV <- predict(svm.model,xTest,type='prob',probability=TRUE)
table(testSV,ySVTest)
predSV <- prediction(attr(testSV,"probabilities")[,1],ySVTest)
perfSV <- performance(predSV,"tpr","fpr")
performance(predSV,"auc")
plot(perfSV)
text(x=0.5,y=0.1,'Area under Curve = 99.3%')
```

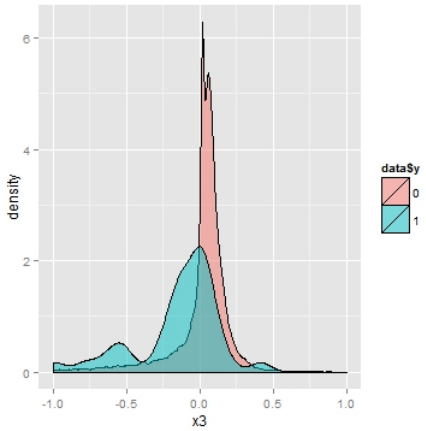
Appendix 2. Figures



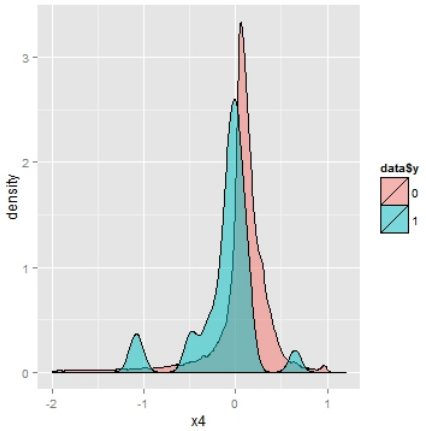
(a) Density of Variable 'x1'



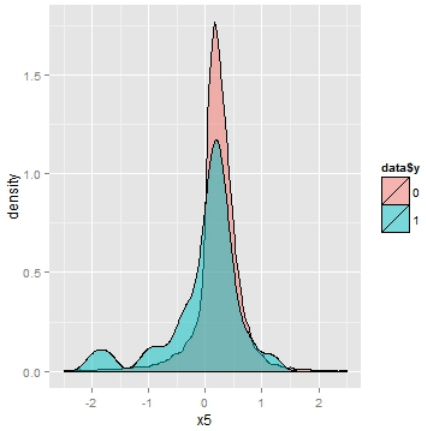
(b) Density of Variable 'x2'



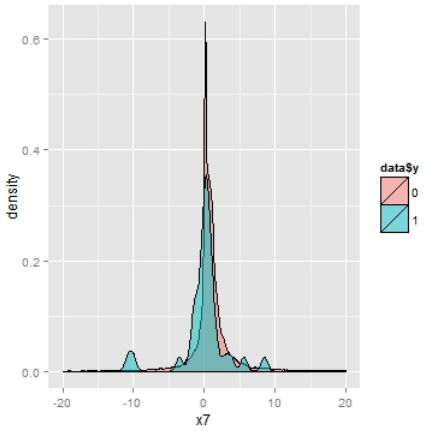
(c) Density of Variable 'x3'



(d) Density of Variable 'x4'

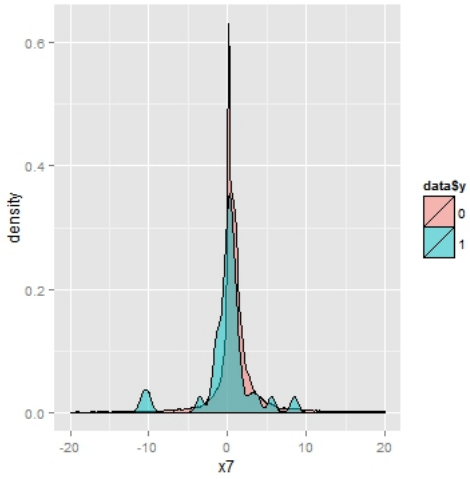


(e) Density of Variable 'x5'

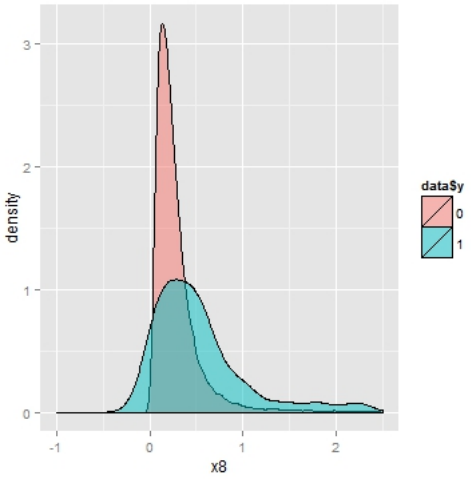


(f) Density of Variable 'x6'

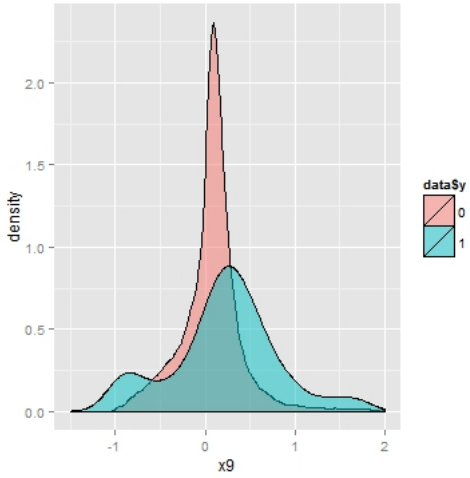
Figure 6.1: Denisities of Variables 'x1' – 'x6'



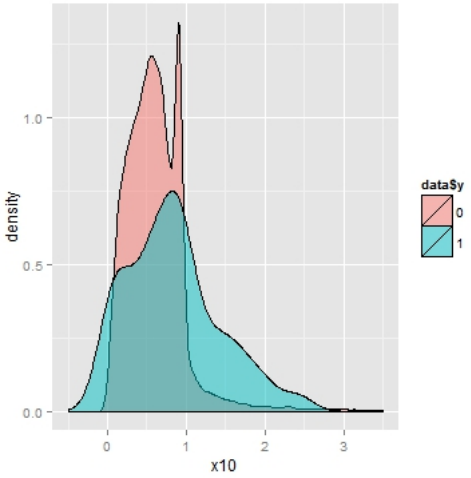
(a) Density of Variable 'x7'



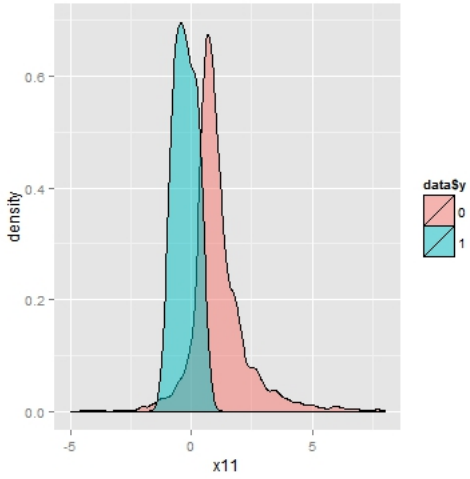
(b) Density of Variable 'x8'



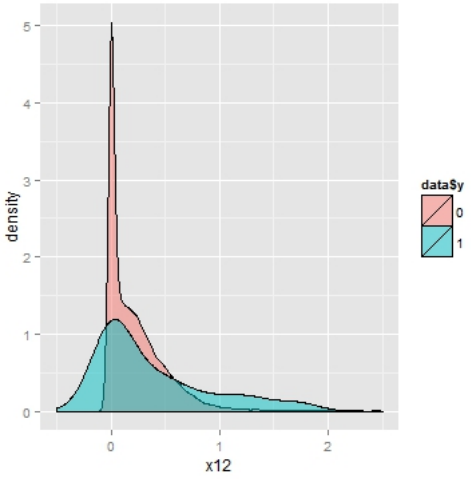
(c) Density of Variable 'x9'



(d) Density of Variable 'x10'

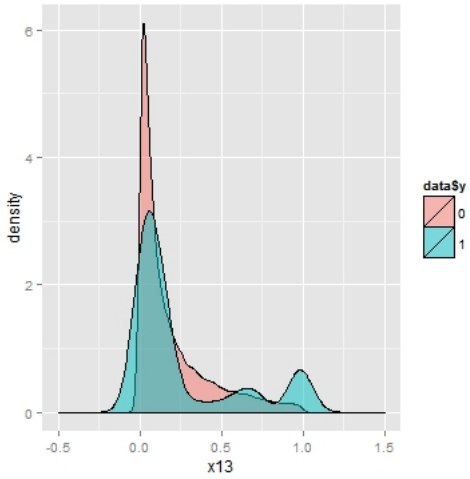


(e) Density of Variable 'x11'

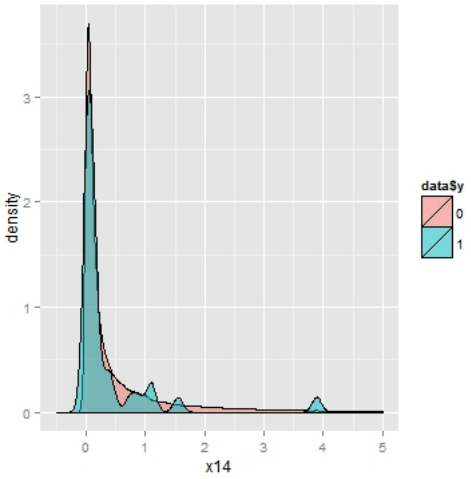


(f) Density of Variable 'x12'

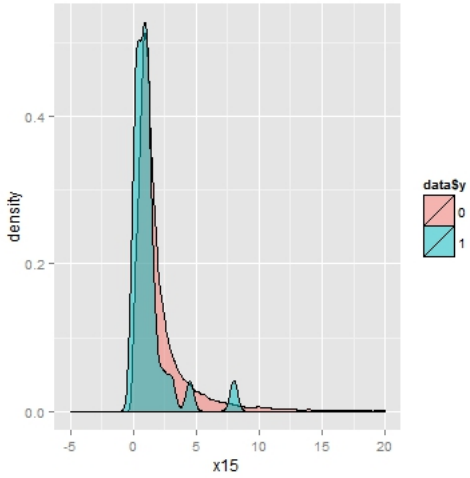
Figure 6.2: Denisities of Variables 'x7' – 'x12'



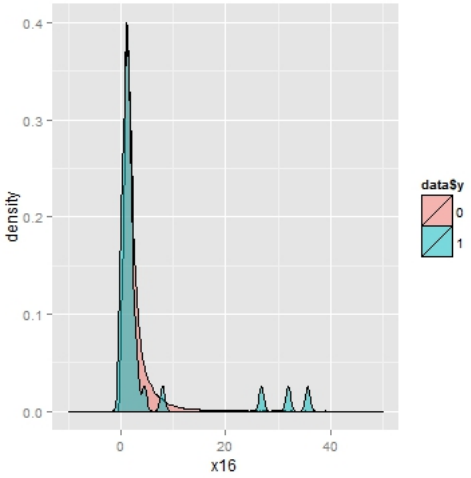
(a) Density of Variable 'x13'



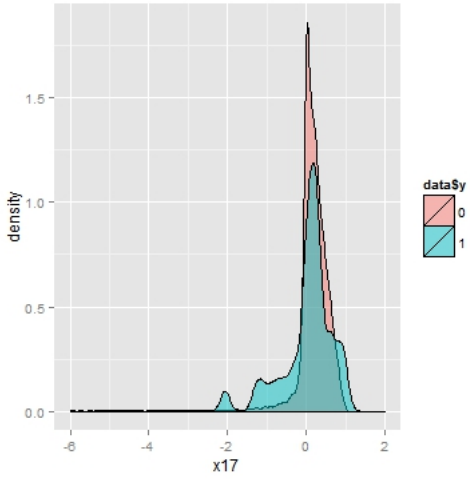
(b) Density of Variable 'x14'



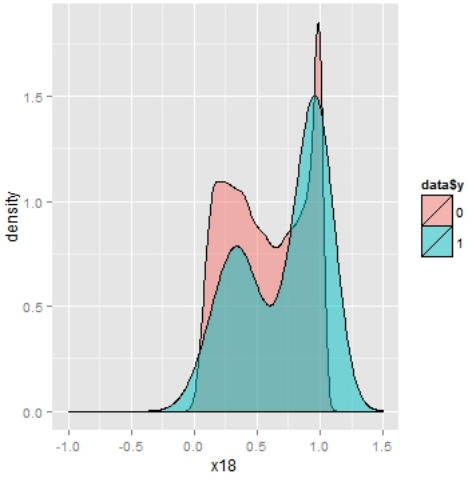
(c) Density of Variable 'x15'



(d) Density of Variable 'x16'

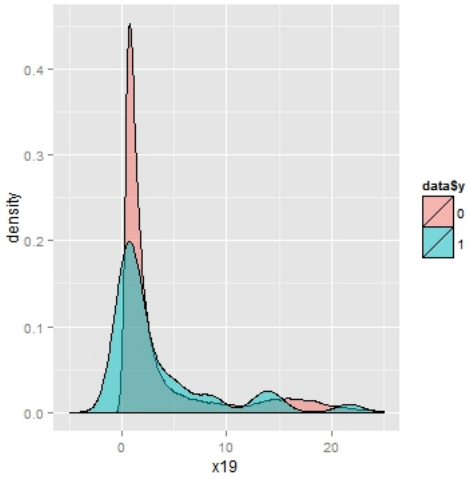


(e) Density of Variable 'x17'

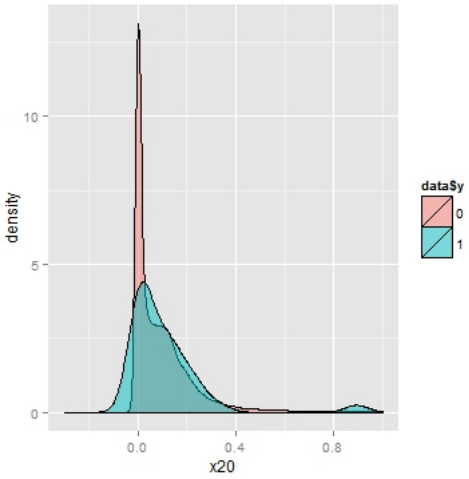


(f) Density of Variable 'x18'

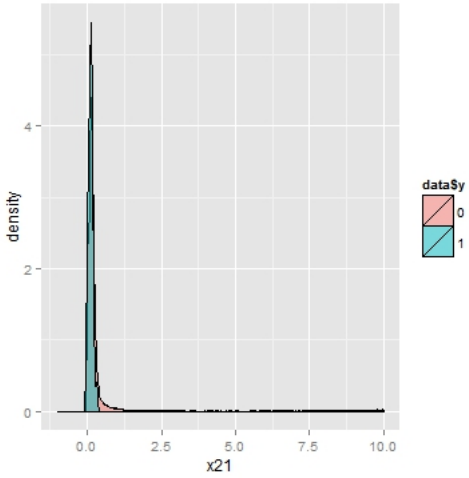
Figure 6.3: Denisities of Variables 'x13' – 'x18'



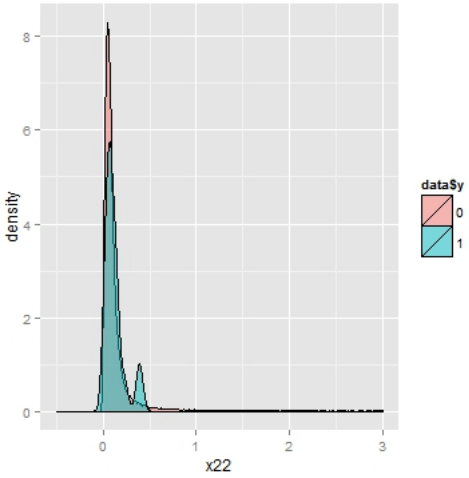
(a) Density of Variable 'x19'



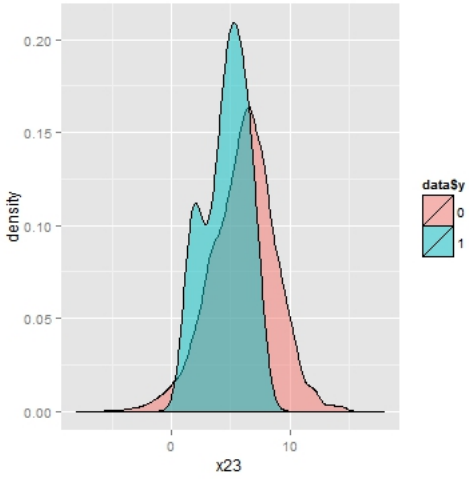
(b) Density of Variable 'x20'



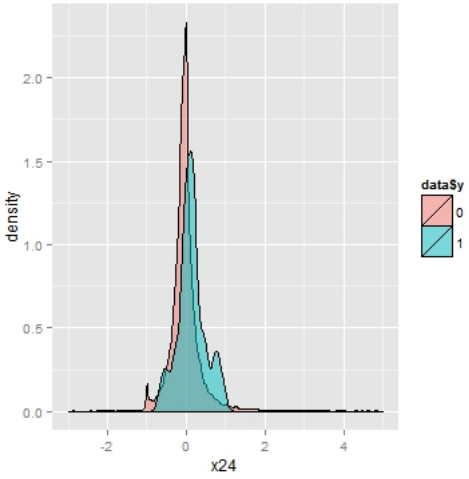
(c) Density of Variable 'x21'



(d) Density of Variable 'x22'

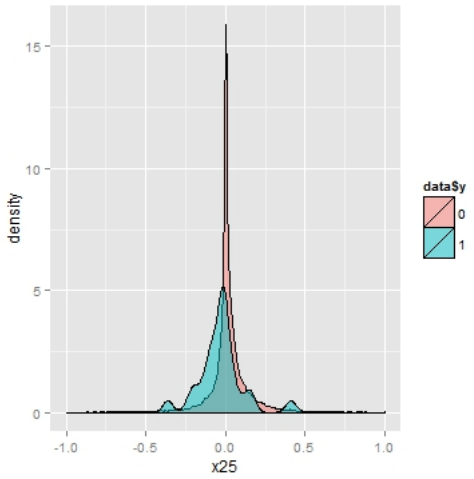


(e) Density of Variable 'x23'

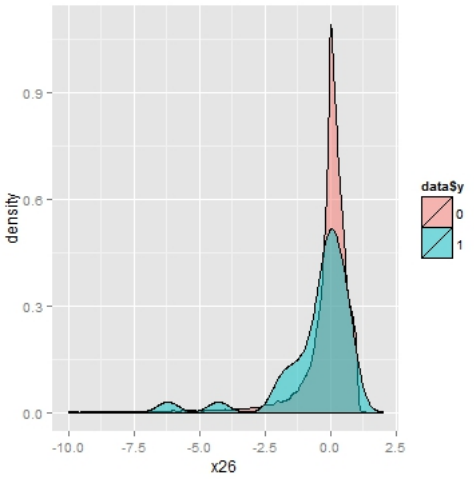


(f) Density of Variable 'x24'

Figure 6.4: Denisities of Variables 'x19' – 'x24'

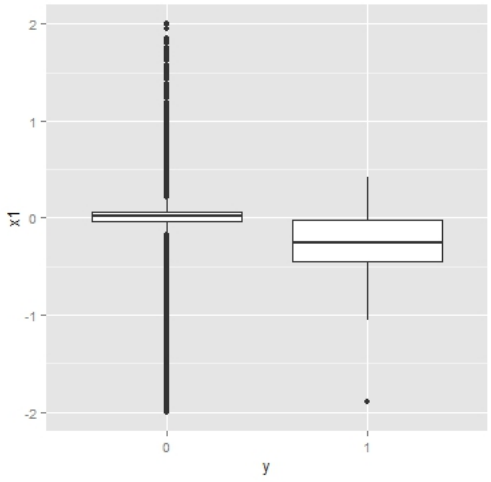


(a) Density of Variable 'x25'

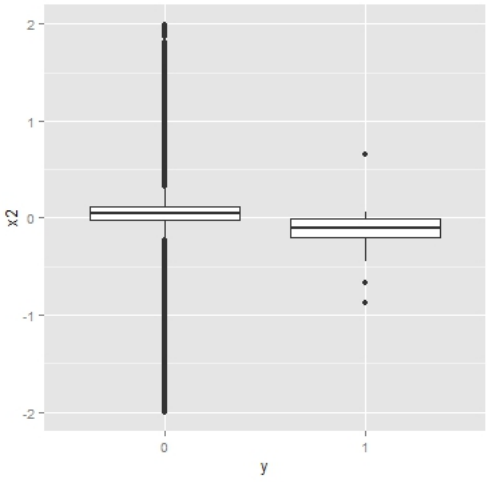


(b) Density of Variable 'x26'

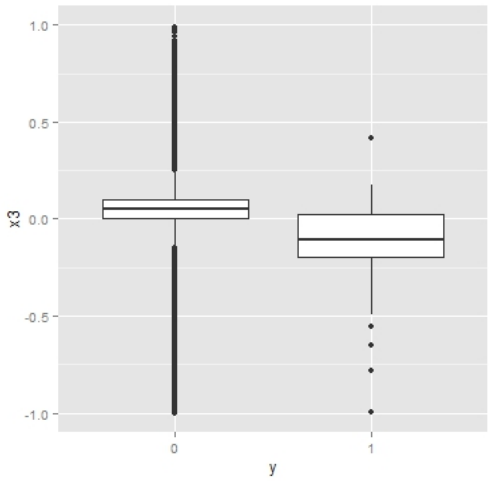
Figure 6.5: Denisities of Variables 'x25' – 'x26'



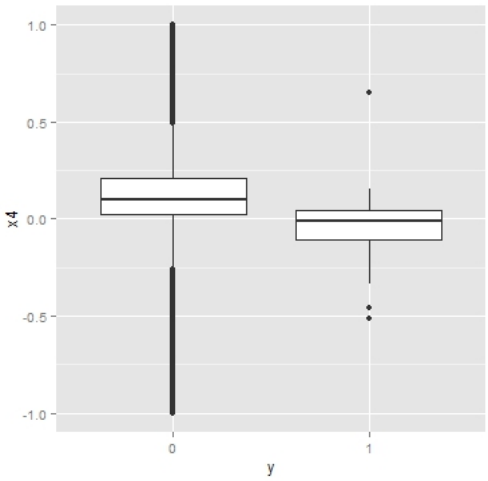
(a) Boxplot for Variable 'x1'



(b) Boxplot for Variable 'x2'

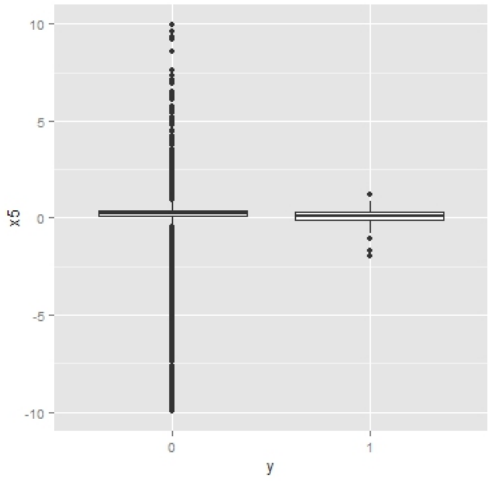


(c) Boxplot for Variable 'x3'

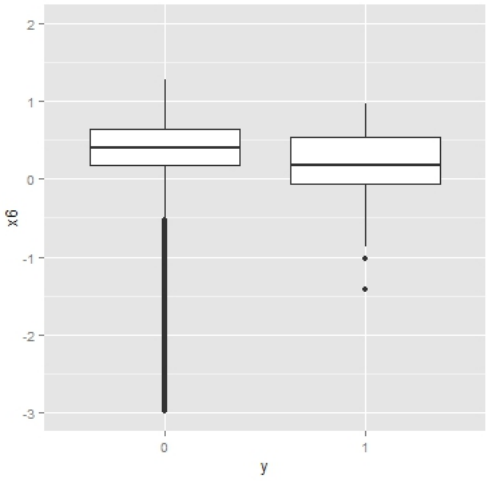


(d) Boxplot for Variable 'x4'

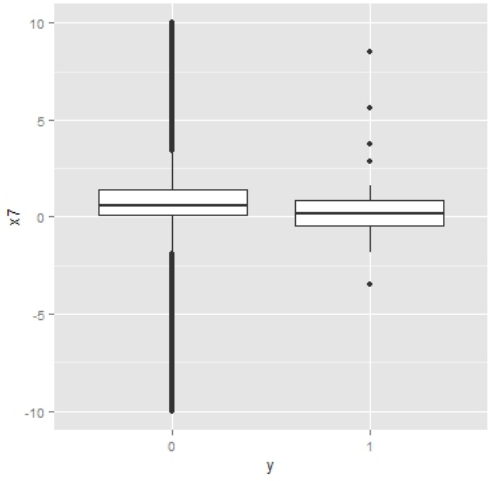
Figure 6.6: Boxplot Distribution Plots for Variables 'x1' – 'x4'



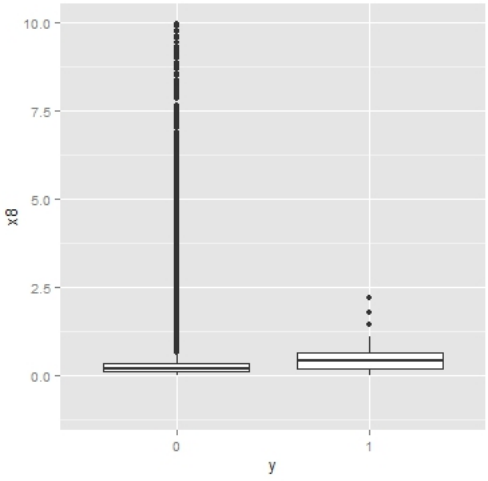
(a) Boxplot for Variable 'x5'



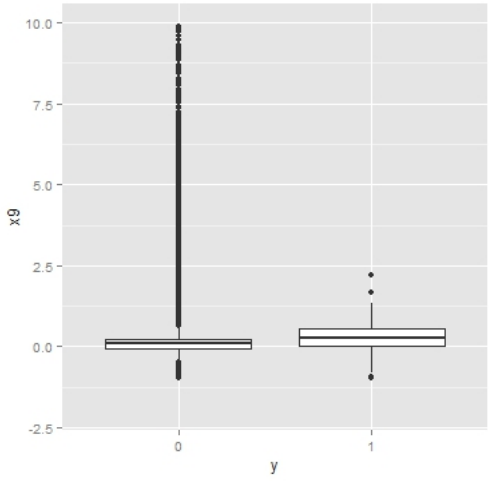
(b) Boxplot for Variable 'x6'



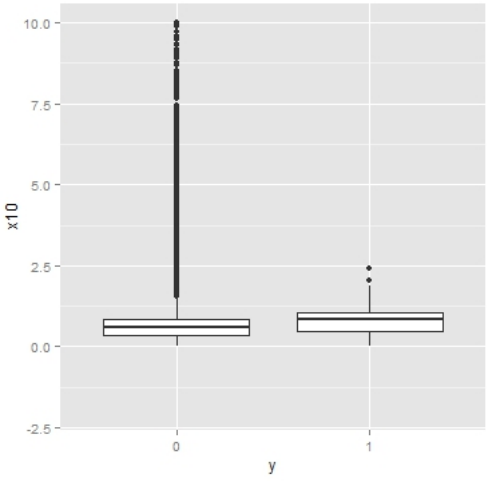
(c) Boxplot for Variable 'x7'



(d) Boxplot for Variable 'x8'

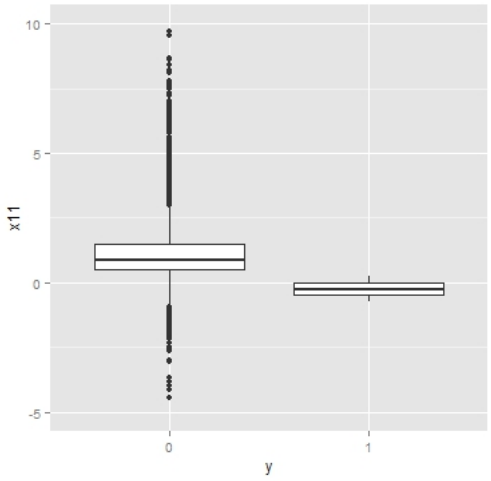


(e) Boxplot for Variable 'x9'

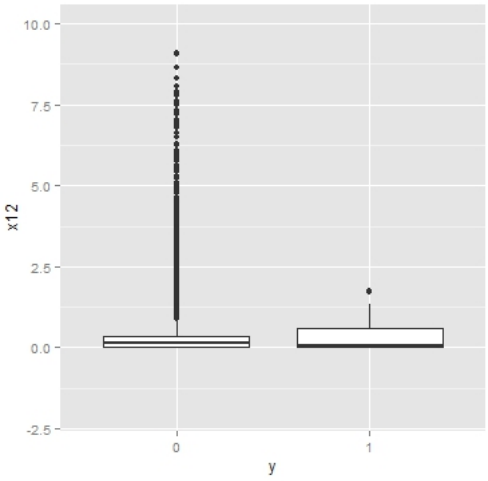


(f) Boxplot for Variable 'x10'

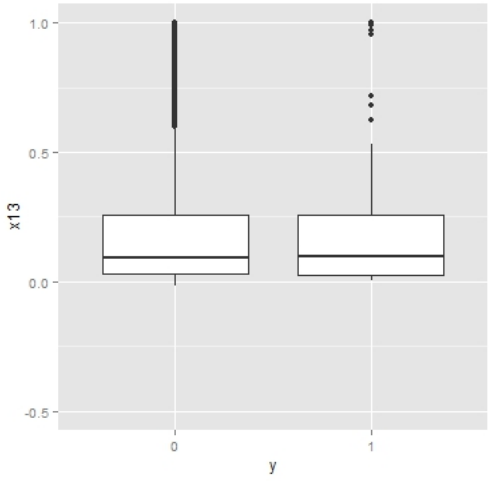
Figure 6.7: Boxplot Distribution Plots for Variables 'x5' – 'x10'



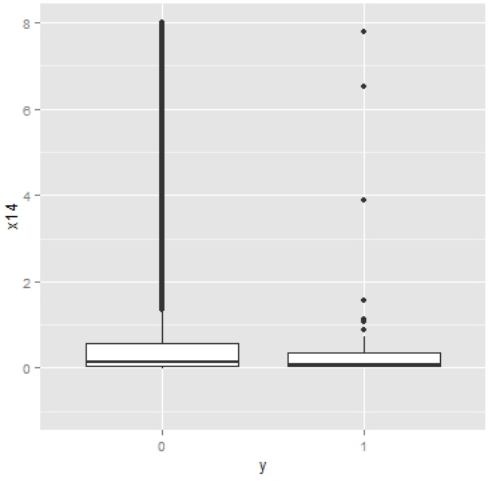
(a) Boxplot for Variable 'x11'



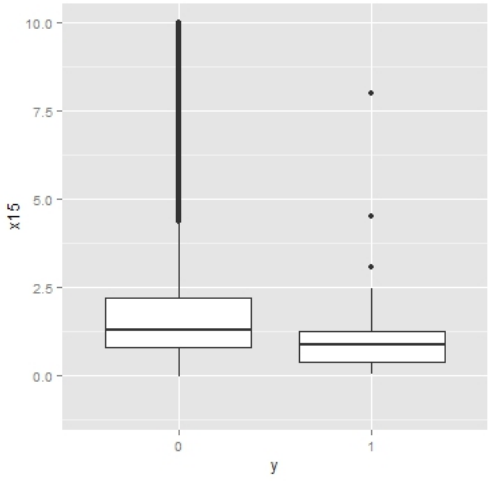
(b) Boxplot for Variable 'x12'



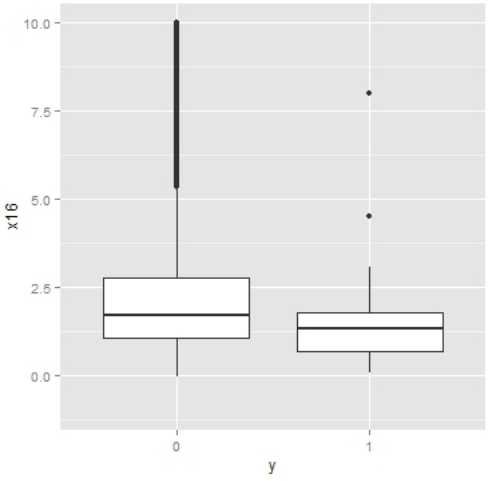
(c) Boxplot for Variable 'x13'



(d) Boxplot for Variable 'x14'

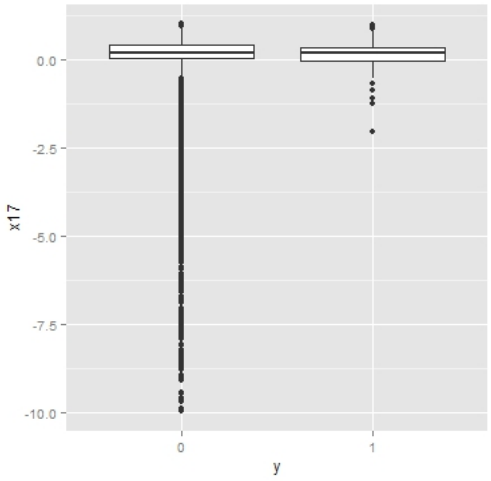


(e) Boxplot for Variable 'x15'

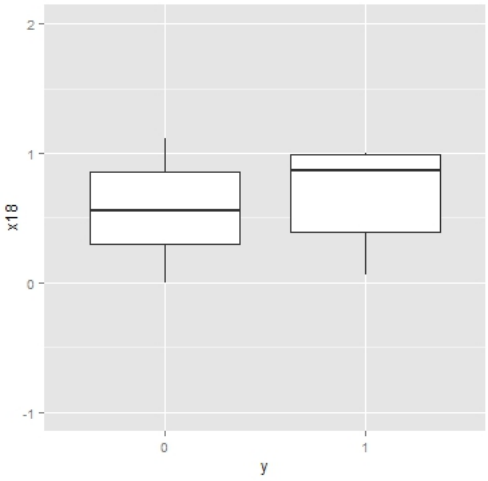


(f) Boxplot for Variable 'x16'

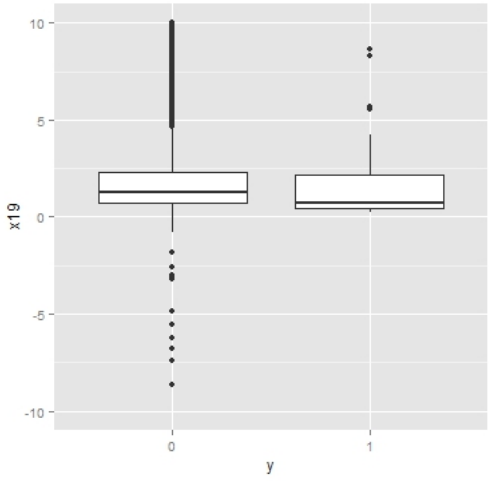
Figure 6.8: Boxplot Distribution Plots for Variables 'x11' – 'x16'



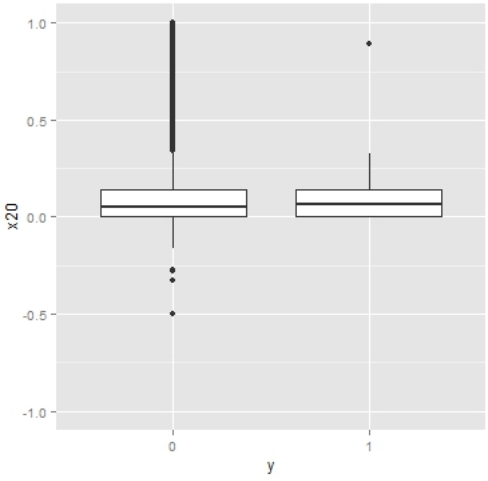
(a) Boxplot for Variable 'x17'



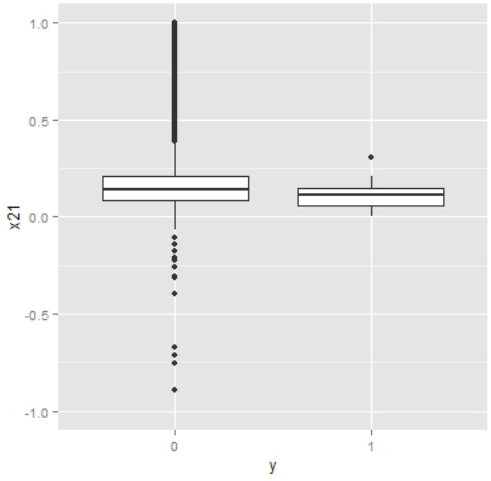
(b) Boxplot for Variable 'x18'



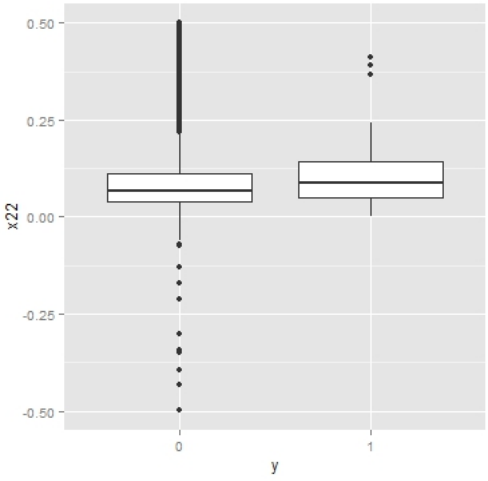
(c) Boxplot for Variable 'x19'



(d) Boxplot for Variable 'x20'

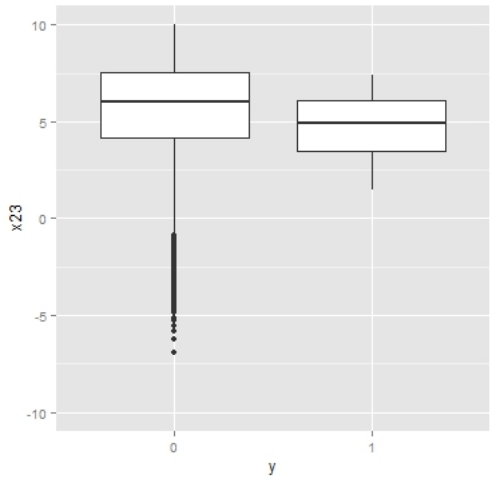


(e) Boxplot for Variable 'x21'

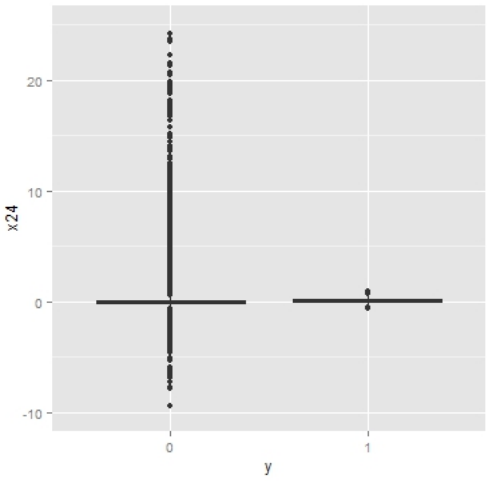


(f) Boxplot for Variable 'x22'

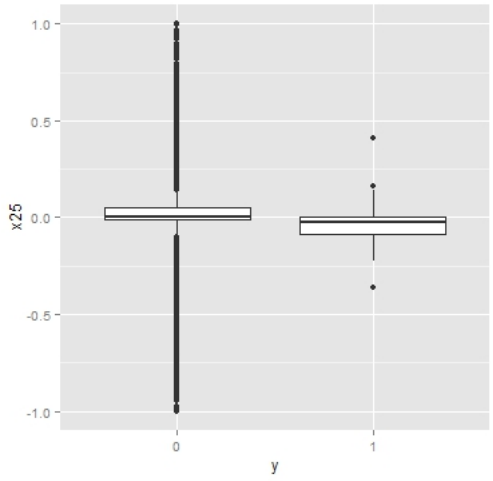
Figure 6.9: Boxplot Distribution Plots for Variables 'x17' – 'x22'



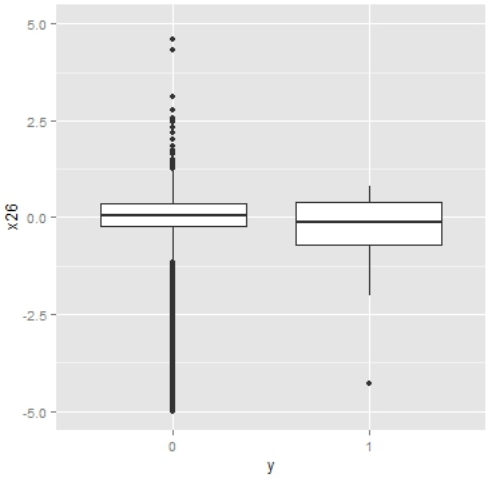
(a) Boxplot for Variable 'x23'



(b) Boxplot for Variable 'x24'

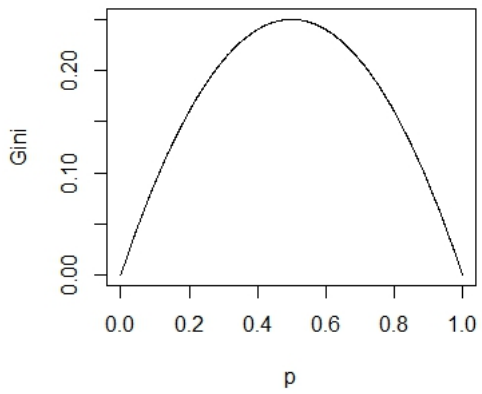


(c) Boxplot for Variable 'x25'

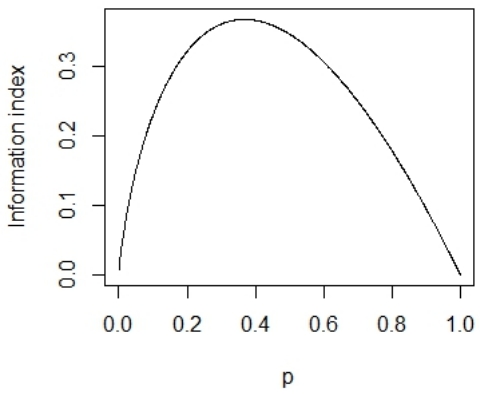


(d) Boxplot for Variable 'x26'

Figure 6.10: Boxplot Distribution Plots for Variables 'x23' – 'x26'

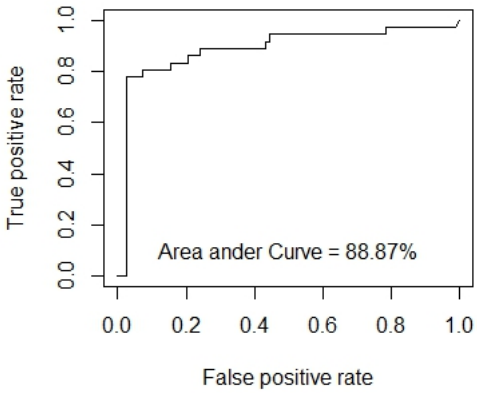


(a) Gini Values for Probabilities

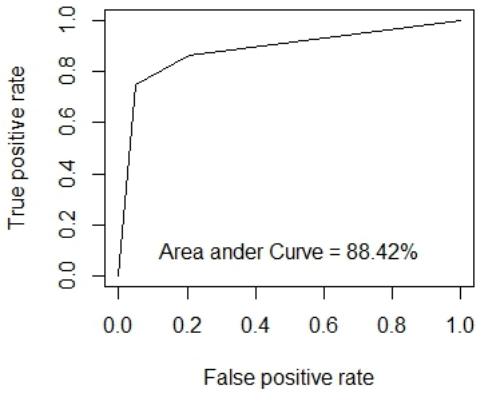


(b) Information Index Values for Probabilities

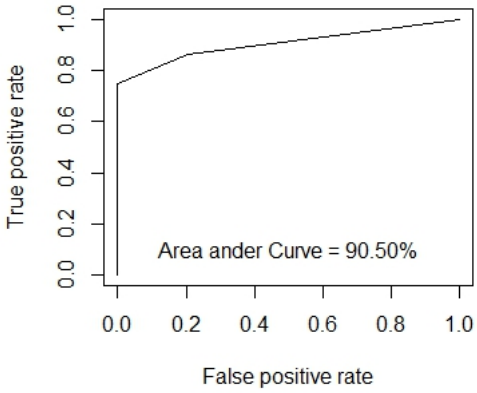
Figure 6.11: Comparison of Gini and Information Index Values



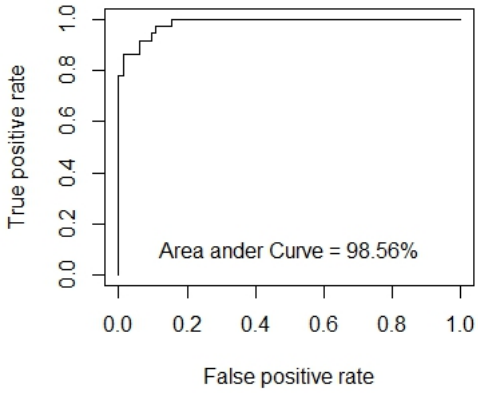
(a) ROC Curve of the Logistic Regression



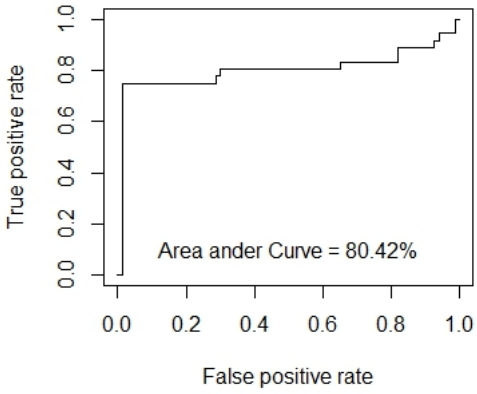
(b) ROC Curve of the Gini Decision Tree



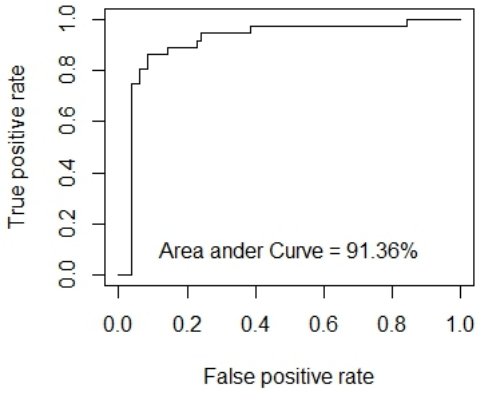
(c) ROC Curve of the IG Decision Tree



(d) ROC Curve of the Random Forest

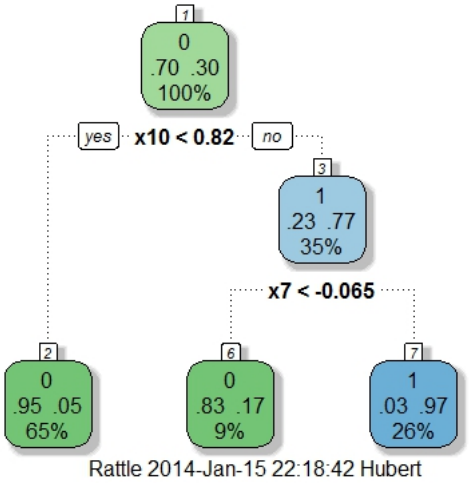


(e) ROC Curve of the Parametric Bayes

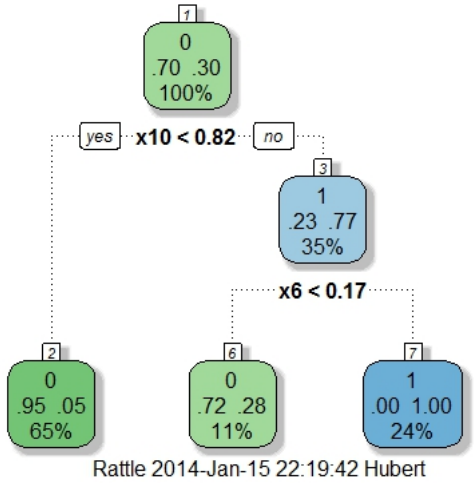


(f) ROC Curve of the Non Parametric Bayes

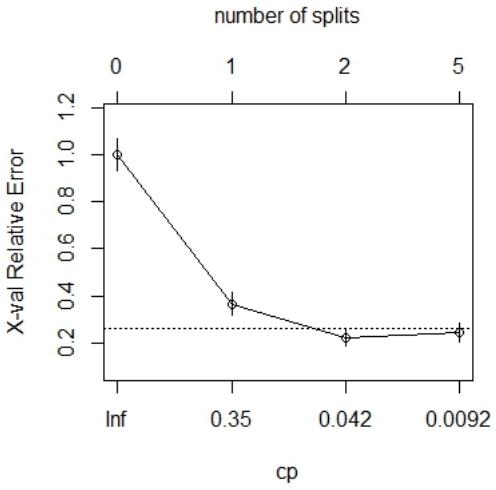
Figure 6.12: Receiver Operator Curves



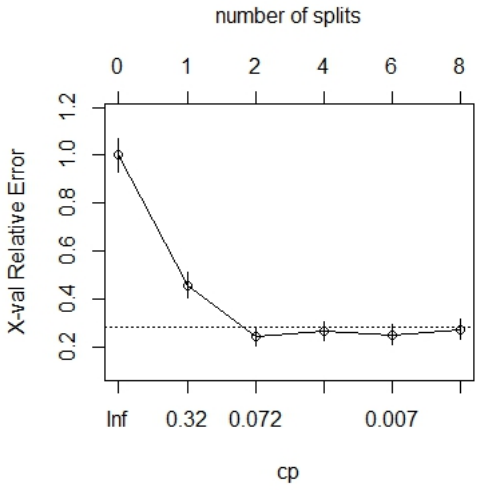
(a) Gini Based Decision Tree



(b) Information Gain Based Decision Tree



(c) Relative Error of the Gini Tree



(d) Relative Error of the Info Tree

Figure 6.13: Decision Trees

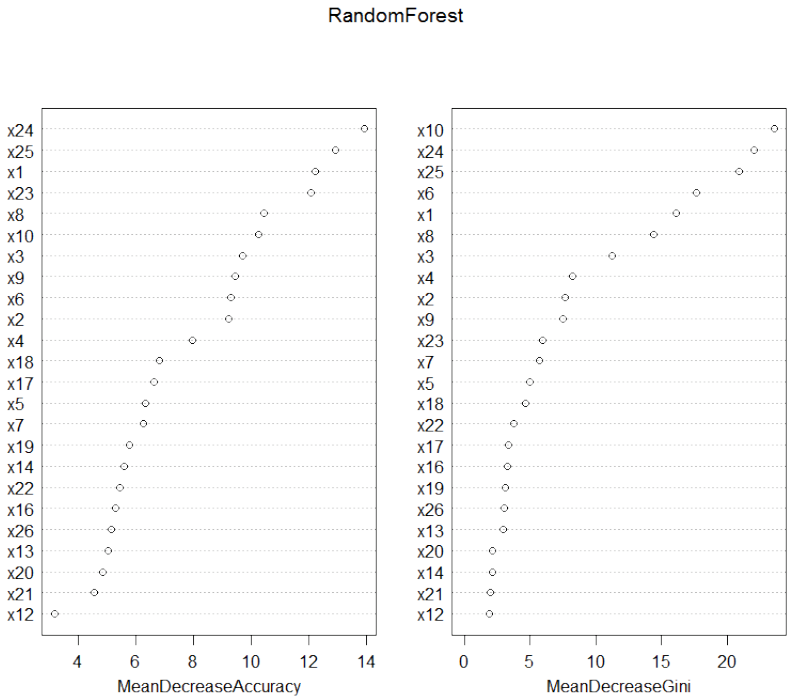
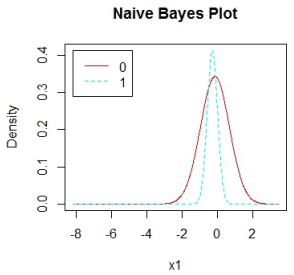
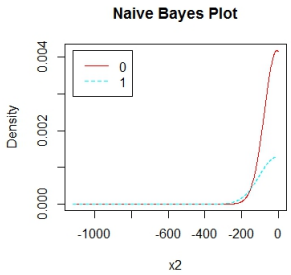


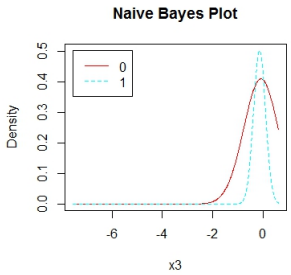
Figure 6.14: Variable Importance Plot for Random Forest



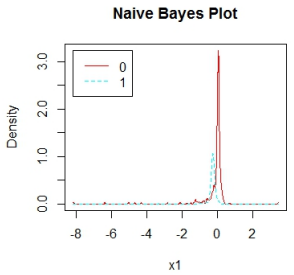
(a) Fitted Parametric Distribution of x1



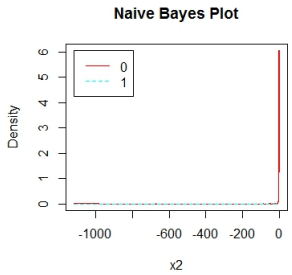
(b) Fitted Parametric Distribution of x2



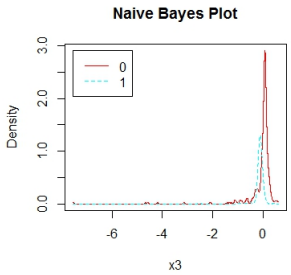
(c) Fitted Parametric Distribution of x3



(d) Fitted Non- Parametric Distribution of x1

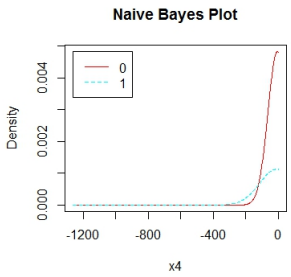


(e) Fitted Non- Parametric Distribution of x2

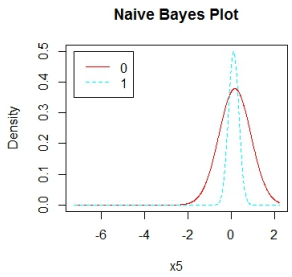


(f) Fitted Non- Parametric Distribution of x3

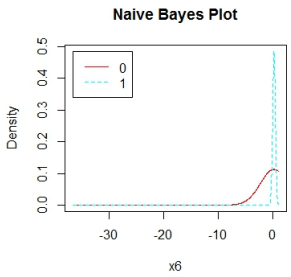
Figure 6.15: Fitted Distributions of Variables 'x1' – 'x3'



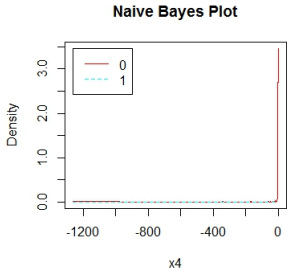
(a) Fitted Parametric Distribution of x4



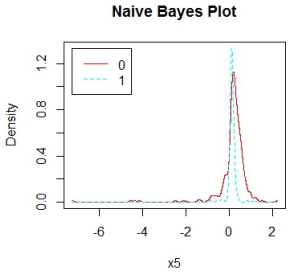
(b) Fitted Parametric Distribution of x5



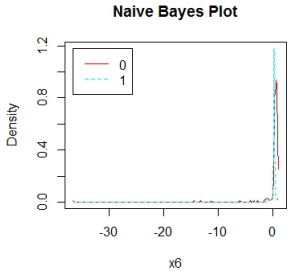
(c) Fitted Parametric Distribution of x6



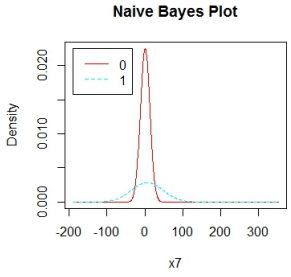
(d) Fitted Non- Parametric Distribution of x4



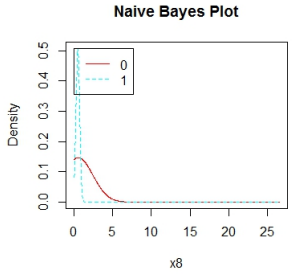
(e) Fitted Non- Parametric Distribution of x5



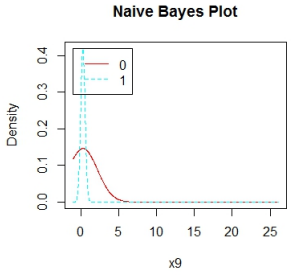
(f) Fitted Non- Parametric Distribution of x6



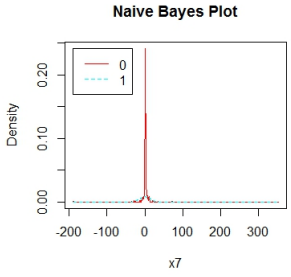
(g) Fitted Parametric Distribution of x7



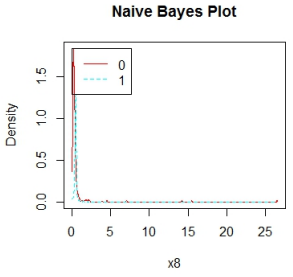
(h) Fitted Parametric Distribution of x8



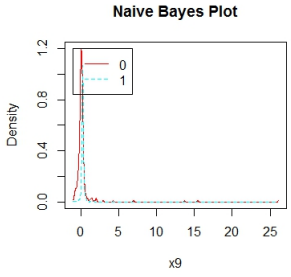
(i) Fitted Parametric Distribution of x9



(j) Fitted Non- Parametric Distribution of x7

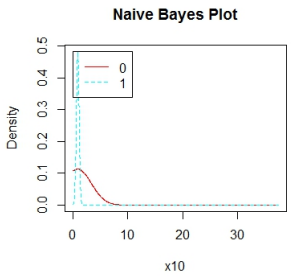


(k) Fitted Non- Parametric Distribution of x8

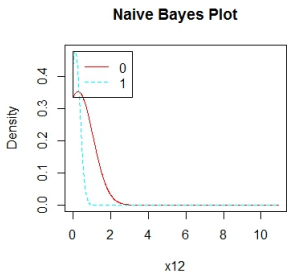


(l) Fitted Non- Parametric Distribution of x9

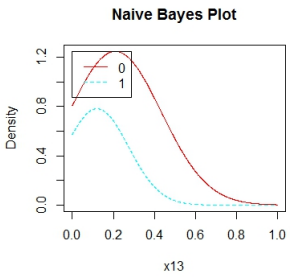
Figure 6.16: Fitted Distributions of Variables 'x4' – 'x9'



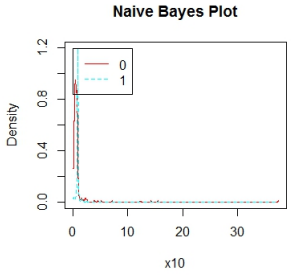
(a) Fitted Parametric Distribution of x10



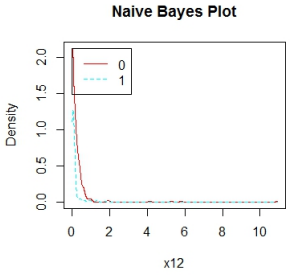
(b) Fitted Parametric Distribution of x12



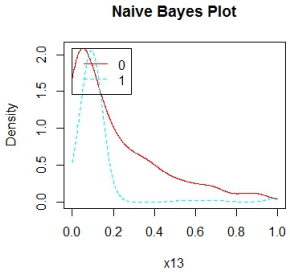
(c) Fitted Parametric Distribution of x13



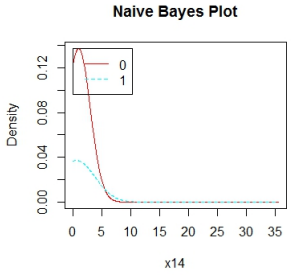
(d) Fitted Non- Parametric Distribution of x10



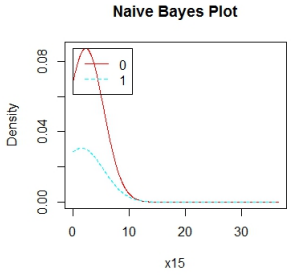
(e) Fitted Non- Parametric Distribution of x12



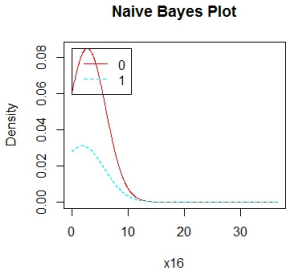
(f) Fitted Non- Parametric Distribution of x13



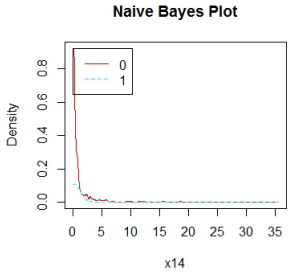
(g) Fitted Parametric Distribution of x14



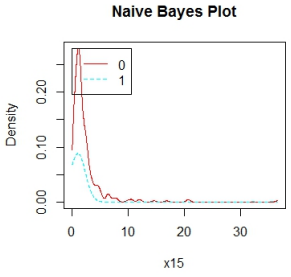
(h) Fitted Parametric Distribution of x15



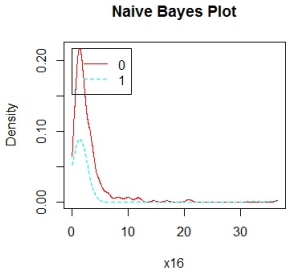
(i) Fitted Parametric Distribution of x16



(j) Fitted Non- Parametric Distribution of x14

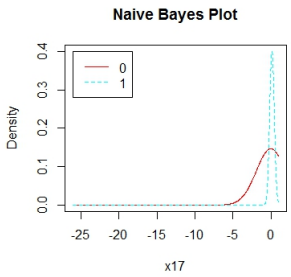


(k) Fitted Non- Parametric Distribution of x15

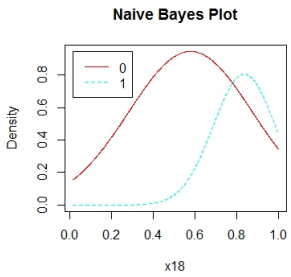


(l) Fitted Non- Parametric Distribution of x16

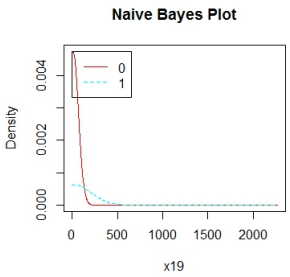
Figure 6.17: Fitted Distributions of Variables 'x10' – 'x16'



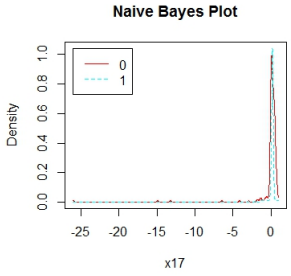
(a) Fitted Parametric Distribution of x17



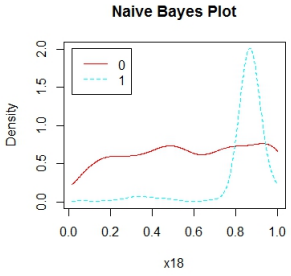
(b) Fitted Parametric Distribution of x18



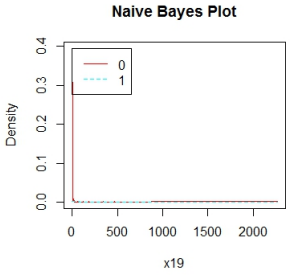
(c) Fitted Parametric Distribution of x19



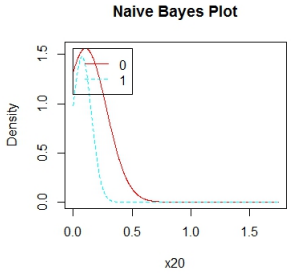
(d) Fitted Non- Parametric Distribution of x17



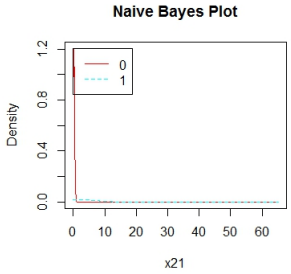
(e) Fitted Non- Parametric Distribution of x18



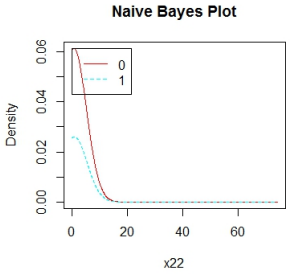
(f) Fitted Non- Parametric Distribution of x19



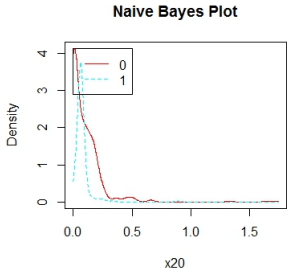
(g) Fitted Parametric Distribution of x20



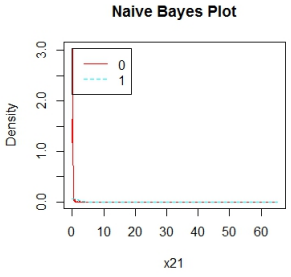
(h) Fitted Parametric Distribution of x21



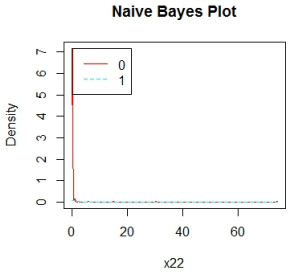
(i) Fitted Parametric Distribution of x22



(j) Fitted Non- Parametric Distribution of x20

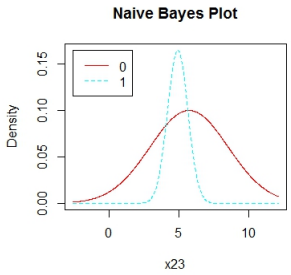


(k) Fitted Non- Parametric Distribution of x21

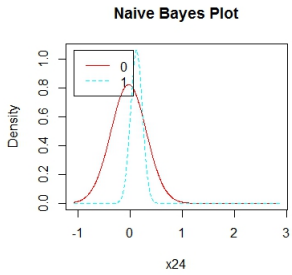


(l) Fitted Non- Parametric Distribution of x22

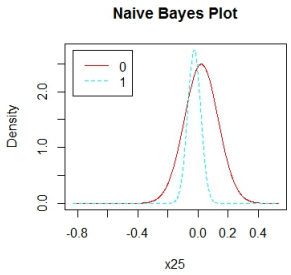
Figure 6.18: Fitted Distributions of Variables 'x17' – 'x22'



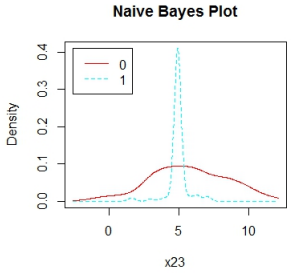
(a) Fitted Parametric Distribution of x23



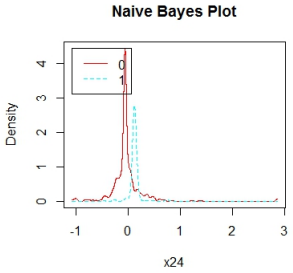
(b) Fitted Parametric Distribution of x24



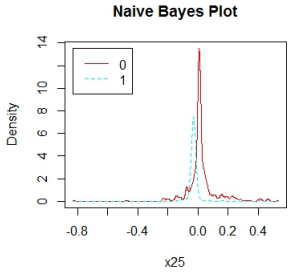
(c) Fitted Parametric Distribution of x25



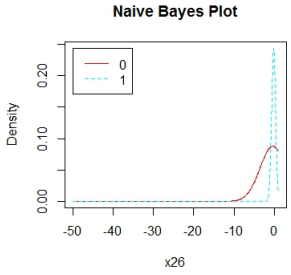
(d) Fitted Non- Parametric Distribution of x23



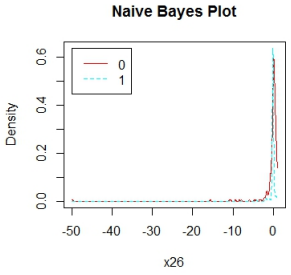
(e) Fitted Non- Parametric Distribution of x24



(f) Fitted Non- Parametric Distribution of x25



(g) Fitted Parametric Distribution of x26



(h) Fitted Non- Parametric Distribution of x26

Figure 6.19: Fitted Distributions of Variables 'x23' – 'x26'

CV

Hubert Tchorzewski • h.tchorzewski@gmail.com • Mobile: + 43 681 201 358 18

Curriculum Vitae

Hubert Tchorzewski

Personal Data:

Name and surname:	**Hubert Cyryl Tchorzewski**
E- Mail:	**h.tchorzewski@gmail.com**
Mobile number:	**+43 681 201 358 18**
Postal address:	AT 1010 Wien, Ebendorferstraße 8/1/308
Date and place of birth:	June 29th 1989, Vienna (AT)
Nationality:	Polish, Austrian

Education:

X 2011 –	University of Vienna, Quantitative Economics, Management and Finance, M. Sc. degree Topic of master thesis: **Effectiveness of machine learning algorithms for predicting insolvencies**
X 2008 – VII 2011	University of Warsaw, Interdisciplinary Economics- Management Studies, Bachelor degree, Topic of bachelor thesis: **Determinants of innovation. A univariate and bivariate analysis of microeconomic Polish data**
VI 2008	High School Nr 1 in Siedlce (PL), finished with a matura exam

Language skills:

English	Proficient
German	Bilingual
Polish	Native

Additional Skills:

Software affinity:	MS Office, STATA, Matlab, SAP, RStudio, TeXworks, MySQL
Programming languages:	Skilled: LaTeX2ε, R
Driving license:	B

Hubert Tchorzewski • h.tchorzewski@gmail.com • Mobile: + 43 681 201 358 18

Work experience:

VII 2011	Internship at the University of Warsaw (Managing databases for research purposes, programming cross database validation methods)
VII 2010	Internship at the Polish Security Printing Works in the Accounting Department (Managing client data in SAP and accounting data in Excel)
VI 2008 – X 2008	Non contracted work at a translating agency (Translating vehicle license branding, licenses, contracts etc.)

References:

Available on request

Vienna, June 1st 2013

Hubert Tchorzewski