Opleiding

Informatica en Economie

Data Preparation of Financial Accounts to Predict Bankruptcy of Companies

Vincent Sieraal

Supervisors:

Frank Takes & Ricardo Cachucho

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl 26/07/2017

Abstract

Since the beginning of the 20th century, researchers started to notify the importance of bankruptcy prediction because of the implications (e.g., economic and social) associated with this phenomenon for many stakeholders (e.g., investors, management, competitors and goverment). Much research has been done by using modern data mining techniques to make an accurate prediction model. However, the main criticism on previous works done is the lack of extensive data preparation. In this research different methods are used to ensure an effective dataset to train the model on. This will not only increase accuracy and performance but also gives us the possibilty to interpret the final model used. The possibility of interpretation is necessary for contribution to economics. Comparing the results of earlier work done, on the same data set, shows a huge increase in performance when performing extensive data preparation. For example, most of the work miss an efficient feature selection. Yet, because of the difficulty to collect quality data and a limited time scope the model is prone to overfitting. The main contributions of this work are:

- The extensive data preparation approach shows interesting hidden mistakes in the data and is therefore an underestimated crucial part this topic.
- By reducing the number of features used, we do not only see an increase in accuracy and performance but also the interpretability of the model.
- An interesting generic shared finding of the attributes chosen is the importance of the company's strategy of risk versus the total assets of the company.

Contents

1 **Introduction** 1

2 **Context** 3

2.1 Financial accounts 3

2.2 Data mining 4

3 **Related Work** 6

4 **Approach** 8

4.1 Prediction target 8

4.2 Data analysis 9

4.3 Attribute evaluators 9

4.4 Prediction models 9

4.4.1 Algorithms 9

4.4.2 Tuning of parameters 10

4.5 Validating models 10

4.5.1 Performance metric 10

4.5.2 Splitting dataset 10

5 **Data** 12

5.1 Dataset 12

5.2 Data preparation 14

5.2.1 Missing values 14

5.2.2 Outlier detection 14

5.2.3 Class imbalance 14

5.2.4 Curse of dimensionality 15

5.3 All features 15

6 **Results** 16

6.1 Analysing the data 16

6.1.1 General insights 16

6.1.2 Attribute evaluators 17

6.1.3 Correlations of features 19
6.1.4 Statistical characteristics features 21
6.1.5 Distributions of features 22
6.2 Performance of model 25
6.2.1 Final decisions for the model 25
6.2.2 Performance values 26
6.3 Interpretation of model 27

7 Conclusions **29**
7.1 Prediction 29
7.2 Future work 30

Bibliography **31**

Chapter 1

Introduction

Data mining gives new insights and predictions which were unattainable in the past. This is already proven within different industry sectors, together with many financial advantages. Yet, the question of bankruptcy prediction still keeps scientists busy for the last few decades. To be aware of a company that is on track to bankruptcy is very interesting to know for many beneficiaries. Financial institutes would, for example, like to know the financial position of a company before giving them a credit. Directors of the company itself would like to know the financial position to see how well they are doing. Stakeholders would like to know in which companies they should invest. Even more generalized, the government of the country would probably like to keep track of how their economy is doing. Because of this reason it is understandable that many researchers would like to predict bankruptcy.

As discussed later in Section 3 lots of research done on this topic is mainly focussed on finding the perfect algorithm to predict bankruptcy. However, the problem with any predictive model is that it heavily relies on the data input which the model should train on. Financial data, used to predict bankruptcy, acquired from balance sheets is error sensitive numerical data. The sensitivity of making errors is not only when reporting the data, but also while collecting and researching it.

That is why the focus of attention within this research is on data preparation for bankruptcy prediction. Hereby we make an effective feature selection to split nessecary and less relevant data. This feature selection is important for several reasons. First, we reduce the dimensionality of the model which leads to a higher accuracy. Second, a simple model is faster in making predictions. Finally, with less features it is easier to analyze and interpret the results. Modern data mining algorithms and domain knowledge are combined to surpass standing prediction models. Therefore, in this research there will be two main questions:

- To what extent can we find a pattern in financial data of failed companies?
- Given a company's financial data of the past 3 years, can we accurately predict bankruptcy in the future?

This research is based on a standardized data mining concept explained in [1]. The data mining process follows: Data understanding, data preparation, data dodeling, model evaluation and final analysis and eventual deployment of the model. This approach fits perfectly to this research because it focuses, compared to others, also on the explaining factor of the data and models. This is important for the discovery of hidden mistakes, patterns and the contribution to economics.

We will first look at the context of the research, by taking a look at the companies, the accounting sector and data mining in Section 2. After this, we take into account what related work has been done and could eventually be relevant to this topic in Section 3. Next, the approach and considerations for all choices made in the research will be explained in Section 4. After, we look at the preparation of the data before training in Section 5. This is followed by the results of the research in Section 6. Finally, contributions made to this topic are described in Section 7.

Chapter 2

Context

This chapter describes the most important concepts and context which are relevant to this research. First, we take a closer look at the companies which are analyzed in this research. Second, we will look at commonly used attributes in accountingr. Finally, important aspects of data mining of this research are explained in the last subsection.

2.1 Financial accounts

The challenge of our research question is that it combines two very different fields of expertise. Namely, the economics and computer science fields. In economics, the knowledge related to the analyzed companies is explained to have a better understanding. To limit the scope of this research and because of limited sources of data, the choice has been made to analyze only Polish companies.

We should also note that companies are divided into many industry sectors. The sector says something about in which market segment the company is operating. It should be taken into account that between these sectors there are differences in business management. In the manufacturing sector of Poland, many companies went bankrupt as noted in [2]. Because of the high occurrence of bankruptcy in this sector, we have a good amount of training data. This is important to reduce potential overfitting.

The financial data of companies we are looking into in this research consists out of 10500 companies. From all of these companies relative (ratios) and absolute financial numbers are collected for analysis. Within the financial world ratios (relative numbers) are very often used to evaluate businesses. These ratios are extracted from the yearly balance sheets and the profit and loss accounts of the companies. A balance sheet is a financial statement that summarizes the company's assets, liabilities and shareholders' equity at a specific point in time. It is important for the company itself and the stakeholders to see their current financial position. Ratios are interesting because they tell us the relation between two attributes. Such a relation could explain much more than a non related attribute. For example, if a company is seen with a huge debt we could conclude that bankruptcy will occur in the future. However, if this company has twenty times more profit in the same year, the huge debt does not say much anymore.

Because it is so complicated, it makes common sense that the demand for an accurate bankruptcy model is high for many different stakeholders.

2.2 Data mining

Data mining is the computing process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics and database systems [3]. In the beginning this was mostly done by using easy algorithms, but since datasets are becoming much larger, we use machine learning algorithms. Data mining could be done for many purposes, but in this research we are trying to find patterns so we can finally make a prediction on bankruptcy. The prediction will allocate a classification to a company with two possible values: the company stays healthy or the company goes bankrupt. There are many machine learning algorithms to make such a prediction, where no algorithm is rated best forehand.

Within data mining we make a difference between supervised or unsupervised learning problems. Supervised learning means that we train on a model with labeled data, so we already know the outcome when training. Unsupervised learning is about finding patterns within unlabeled data. There are no instances from where we know the final classification. In this research, we have a dataset available including labeled companies, so in this case we are looking at a supervised learning problem.

A very important aspect of data mining is choosing the right algorithm for the final model. In this paper, we look at the *C4.5* and *RandomForest* algorithms which are explained later in Section 4.4.1. The main difference between these algorithms is the complexity. Where as *C4.5* only uses a single prediction classifier, *RandomForest* uses many and is therefore more accurate. However, the increase of complexity also increases the time to build the model. Therefore the *C4.5* algorithm is first used to analyze the data before making a final prediction with the *RandomForest* algorithm.

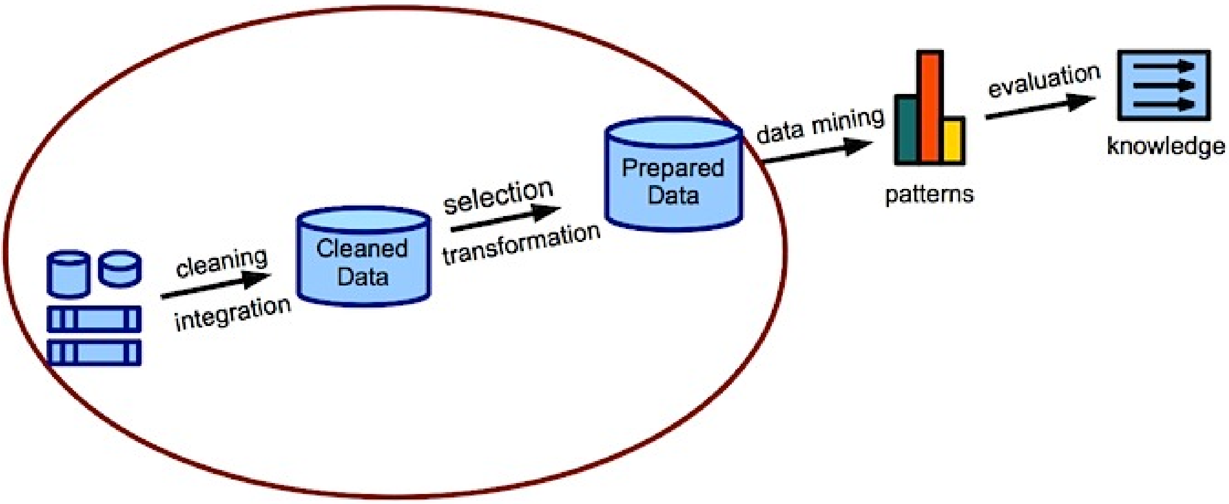

Figure 2.1: Global process of data mining

To build an efficient model the data has to be prepared before making the model as shown in Figure 2.1. The focus in this paper will be the optimization of this data, in scientific terms we call this data preparation. Data preparation itself could be divided in different parts [4]:

- Data Discovery: Process of searching for patterns of specific items in the data.
- Data Profiling: Statistical analysis and assessment of data for consistency, uniqueness and logic.
- Data Cleaning: Process of detecting and removing corrupt data.
- Data Validation: Process of ensuring that useful, correct (clean) data is used.
- Data Transformation: Application of a chosen mathematical function to each data point in data set.

For this research it is important to ensure the accuracy and correctness of insights. If the insights are neither accurate or correct the results are no added value to the topic, discussed in Section 4.5.

Chapter 3

Related Work

In this chapter related work on the topic will be shortly described and criticized.

Because of the high interest by different stakeholders lots of related work has been done on this topic. The first attempts of bankruptcy prediction go back to the start of the 20th century when first indicators were proposed to describe bankruptcy by Fitzpatrick [4] . In the sixties more interesting models were shown, because of the use of statistical models as seen in the work of Beaver [5]. Altman [6] followed this way of thinking and proposed to use multidimensional analysis to predict bankruptcy which was followed by many others. Next to this, there was huge interest in generalized linear models by Olhson [7] because of the explaining factor of the indicators used. Moreover, the manual models used were either too simple to be accurate or too hard to use, manually, in practice.

From 1990, artificial intelligence and machine learning have become a growing factor in the research direction of bankruptcy prediction. The increase of data led to insights suggesting that the simplest models used so far are too inaccurate. Also, the increase in data needed a new approach of processing all the data.

Since then, different methods were simultaneously researched in order to find the best prediction model. To start off, the use of first-order logic in combination with evolutionary programming [8]. The classification accuracy of these models was not sufficient enough because of the lack of predictive possibilities of the decision rules. Next to this, researchers tried support vector machines [9]. They were highly accurate so far, but could not do the prediction within the accounting context and also needed time consuming hand tuning. A whole different approach is the use of automatic feature extraction from data that avoids the hand tuning of support vector machines. Neural networks were used in that case [10], [11], [12], [13], [14], [15], [16], [17], [18], but failed at the relative large proportions of variance in the data.

Research has been done by ensembling classifiers as seen in [19]. The advantage of ensembling classifiers is that it combines the strenghts (and weaknesses) of different algorithms to improve one final model. It has been shown that an ensemble classifier can be successfully used in this field of research [20].

Prediction of bankruptcy with the same dataset as used here has been done by Zieba and Tomczak [2] using an ensembled boosted classifier which is known to be successful in many classification problems. Their research is using additional synthetic features which are developed by using arithmetic operations.

However, the paper by Zieba and Tomczak and most papers mentioned above put the main focus on the algorithm part. Whereby data preparation is neglected and could harm the final results. Next to this, arithmetic operations on the features are used to make a prediction model. The complexity of the different formalus used, results in good performance values, but this approach is prone to overfitting.

This research will focus on analyzing, simplifying and optimizing the data before training the final model. The data is criticised, evaluated by different algorithms and is statistically researched for different characteristics as correlations, distrubitions and others. With this approach the model should be less prone to overfitting. Data mining algorithms and financial domain knowledge are combined to surpass the existing models.

Chapter 4

Approach

This chapter describes the approach and considerations of this research.

The first step in this research is to decide the target variable. The target answers binary, with yes or no, to our bankruptcy prediction question. To answer this question different analyses are made on the data. As a third step, attribute evaluators are used to rank subsets of all features. Then an algorithm is chosen to predict the bankruptcy of companies. Afterwards, the model will be validated to see if there is an increase in performance by using extensive data preparation. Finally, the model with the final choices are compared to related work to see if there are interesting contributions to be made from the perspective of an extensive data preparation approach.

4.1 Prediction target

In Section 2.1 we observed that the definition of bankruptcy itself is interesting between countries as well as within countries many different classes of bankruptcy are stated. This makes it difficult to make a single model which is representative for worldwide companies. However, the approach for getting this model could be used in domains worldwide.

Next to this, the research has been mainly focussed on choosing the right features used to derive such a prediction. For this feature selection analysis, it is important to simplify the problem into manageable parts. This way the choices made can be easiliy visualized and analysed. On these grounds, the choice has been made for a more simple binary classification. The target will either classify whether a given company is bankrupt or not, within three years from now on. The choice of three years is mainly because of the amount of data available.

4.2 Data analysis

To get more insight in the data and to find the most useful features we made different analyses. The dissection has been made from both the economics and computer science point of view. Because of the large number of features we started with using different attribute evaluators to make subsets of interesting features. These subsets are evaluated by using domain knowledge, correlations, statistics, distributions and relevant work. Results, choices and explanations of these are found in Chapter 6.

4.3 Attribute evaluators

Attribute evaluators are used to rank attributes of the chosen data set. There is not one "best" evaluator for all cases, because for each case there is a different optimal trade-off. This is why many attribute evaluators were tested in the experiment to ensure the best results. In this particular case we are looking for a highly interpretable model, low complexity and a good accuracy. In this research, the *OneREvaluator* comes out the best tested on AUC performance values and interpretability of the model with different amount of features used and thus shall be explained in detail.

The *OneREvaluator* ranks attributes by the OneR, short for "One Rule", classifier. The algorithm generates one rule (or frequency table) for each attribute against the class. The error of this rule, compared to the class, says something about the contribution to the model. A lower error means a higher contribution to the predictability of the model and is therefore ranked higher. Because of its linear relation to the class, evaluation is simple and interpretable. However, since only linear relations are tested the accuracy could be criticised.

4.4 Prediction models

As noticed earlier, the main focus of this research will be on feature selection. However, to compare different subsets of features we have to validate those by making a final prediction model. One of the most important things in this model is to predict not only companies that will stay healthy, but also the companies that went bankrupt. In Section 4.5 we will describe how we validate the final model, while this section will describe which choices are made in choosing the final model and related parameters.

4.4.1 Algorithms

In this research the right choices in data preparation are made by validating the model built by the commonly used C4.5 algorithm. This recursive algorithm generates a decision tree by splitting the attributes which have the highest normalized information gain. This means that for each step in the decision tree, attributes with higher knowledge are preferred above lower ones. The model can be built within seconds and is fairly accurate for many data mining cases [21].

However, as stated in the paper about choosing the right algorithm for bankruptcy prediction [2], a better option for the final model would be to choose the RandomForest algorithm. The RandomForest algorithm works with a concept called bagging. This technique is focused on combining different learning models to increase the classification accuracy. This algorithm creates many decision trees and use these to make a final classification. More about this algorithm can be read in [22].

4.4.2 Tuning of parameters

The parameters of the models are tuned to optimize final results. To keep the scope of this researched limited the choice has been made to choose the parameters with a exhaustive algorithm. Namely, all possible values are taken into account and the one with the best results are finally chosen. In case results do not improve by changing the value, the initial (recursive) value will be selected. For example, the depth of the decision trees (15) and the number of neighbours (3) to look at by choosing the oversampling percentage of companies that went bankrupt as explained in Section 5.2.3.

4.5 Validating models

Validating models is an important step in data mining. Model validation is mainly done to see how accurate the model is and how it compares to other models.

4.5.1 Performance metric

The performance metric is describing how good the model performs by, mostly, using a single number. This makes it easier to compare and test with other models. There are many performance metrics which could be used in this case. However, most of the performance measures can not handle class imbalance which is a difficult problem in this case as discussed in Section 5.2.3.

A well known measure is the AUROC shown in Figure 4.1. This metric is a good choice for class imbalance problems, because it uses both the true positive rate and the false positive rate. The true positive rate tells us more about the probability of detection, where the false positive rate says more about false detections. The bigger the area under the curve, the better the model performs than a random choice [23]. To get a single numbered performance metric, we only look at area under the ROC curve. So, the higher this number is the better the model performs.

4.5.2 Splitting dataset

The dataset is split for training, validating and testing. The training set is used to build the model, the validation set for tuning the parameters and the test set for evaluating the final model. To avoid too much

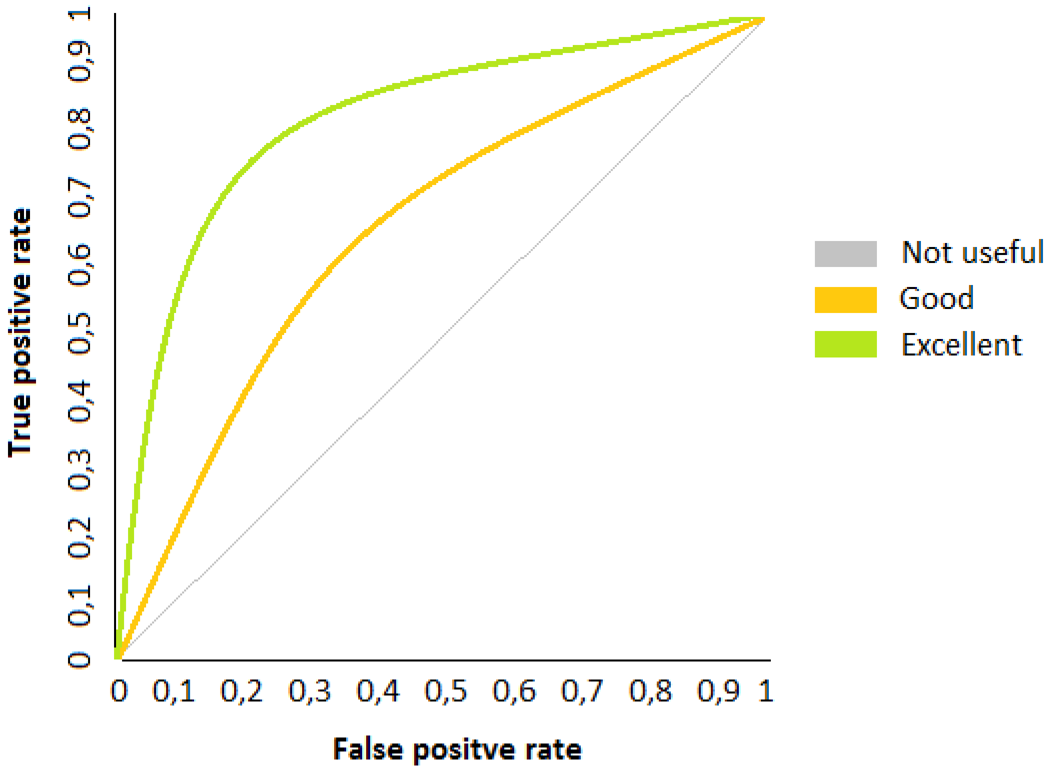

Figure 4.1: Example of Receiver Operating Characteristic curve (ROC)

data loss, the training and validation set are fine tuned with 10-cross-fold-validation. So, all instances in this data set are used for training and validation and the average out of 10 folds will give us a performance value. However, in this way the model will attempt to improve the output of 10-cross-fold-validation. That is why it is important to test the final, tuned model on new data to give a better impression on the accuracy of the model. Next to this, it should be taken into account that the data used is time specific. The financial data used is gathered in the beginning of the 21th century.

Chapter 5

Data

This chapter describes the used data and the implications on this research. It is split into three sections, in the first part we discuss the used data set and in the second part we will describe how data preparation is done. Finally, we will have a short look at the initial attributes.

5.1 Dataset

The main issue with data mining is that the accuracy of the model highly depends on the quality of the data used. Next to this, the main focus of this research was to analyze features used for the prediction. For those particular reasons, one of the main tasks was to search for a usable dataset with a various amount of features.

The dataset [24] finally used was allocated by the University of Science and Technology in Wroclaw, Poland. The data was originally collected by the Emerging Markets Informatic Service [25]. It consists of around 10.500 instances with approximately 500 companies that went bankrupt and 10.000 companies that stayed healthy within three years. The dataset contains 64 features with financial numbers and ratios derived from the yearly balance sheets and a final binary target attribute which indicates bankruptcy. The interesting choice that has been made for this particular dataset is the large amount of relative numbers instead of absolute numbers. As we noticed before in Chapter 3 we assume that relative numbers say more than absolute numbers. In Figure 5.1 we can see that most common financial ratios are taken into account.

The main reasons why this specific dataset has been chosen are the large number of features, the consistency of the data and the number of instances. Those aspects are needed to do data mining.

ID	Attribute name	ID	Attribute name
X1	net profit / total assets	X33	operating expenses / short-term liabilities
X2	total liabilities / total assets	X34	operating expenses / total liabilities
X3	working capital / total assets	X35	profit on sales / total assets
X4	current assets / short-term liabilities	X36	total sales / total assets
X5	(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)) * 365	X37	(current assets - inventories) / long-term liabilities
X6	retained earnings / total assets	X38	constant capital / total assets
X7	EBIT / total assets	X39	profit on sales / sales
X8	book value of equity / total liabilities	X40	(current assets - inventory - receivables) / short- term liabilities
X9	sales / total assets	X41	total liabilities / ((profit on operating activities + depreciation) * (12/365))
X10	equity / total assets	X42	profit on operating activities / sales
X11	(gross profit + extraordinary items + financial expenses) / total assets	X43	rotation receivables + inventory turnover
X12	gross profit / short-term liabilities	X44	(receivables * 365) / sales
X13	(gross profit + depreciation) / sales	X45	net profit / inventory
X14	(gross profit + interest) / total assets	X46	(current assets - inventory) / short-term liabilities
X15	(total liabilities * 365) / (gross profit + depreciation)	X47	(inventory * 365) / cost of products sold
X16	(gross profit + depreciation) / total liabilities	X48	EBITDA (profit on operating activities - depreciation) / total assets
X17	total assets / total liabilities	X49	EBITDA (profit on operating activities - depreciation) / sales
X18	gross profit / total assets	X50	current assets / total liabilities
X19	gross profit / sales	X51	short-term liabilities / total assets
X20	(inventory * 365) / sales	X52	(short-term liabilities * 365) / cost of products sold)
X21	sales (n) / sales (n-1)	X53	equity / fixed assets
X22	profit on operating activities / total assets	X54	constant capital / fixed assets
X23	net profit / sales	X55	working capital
X24	gross profit (in 3 years) / total assets	X56	(sales - cost of products sold) / sales
X25	(equity - share capital) / total assets	X57	(current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)
X26	(net profit + depreciation) / total liabilities	X58	total costs / total sales
X27	profit on operating activities / financial expenses	X59	long-term liabilities / equity
X28	working capital / fixed assets	X60	sales / inventory
X29	logarithm of total assets	X61	sales / receivables
X30	(total liabilities - cash) / sales	X62	(short-term liabilities * 365) / sales
X31	(gross profit + interest) / sales	X63	sales / short-term liabilities
X32	(current liabilities * 365) / cost of products sold	X64	sales / fixed assets

Table 5.1: List of all features.

5.2 Data preparation

Data preparation is the process of collecting, cleaning and consolidating data. The process of data preparation is important to maximize the quality prior to its use for prediction. Data with low quality can contribute to misleading outputs and thus should be avoided.

As noted earlier in Section 3 in the paper Zieba and Tomczak [2], the main focus was on the algorithmic part. The research started with the process of selecting data, choosing the sector, the research period and the number of financial attributes. However, they missed crucial aspects such as missing values, outlier detection, class imbalance and more in general, the curse of dimensionality. That is why, for this paper the focus of attention is on data preparation.

5.2.1 Missing values

It is important to train on a training set which does not have any missing values. This could lead to inaccuracy of the final model. To eliminate these there are different methods as explained in [26]:

- Remove every instance with one or more missing values
- Replace the missing values by the mean or median of the feature
- Use a prediction model to predict the missing values

The simplest method would be to remove all instances with any missing values. However, this would remove to approximately 70% of the training set and thus making it not valid for this data. To keep the scope of this research limited and keep all features available at start, the choice has been made to replace all missing values by the mean or median of the feature.

5.2.2 Outlier detection

Outliers are data points that are at an exceptionally large range from the common observations. Outlier detection within financial datasets are difficult because of the high dimensionalty of data [27]. This high dimensionality leads to a space with very spread data points. Because of the sparsity of most data points in this high dimensionality space outliers are hard to recognize. Thus, making it very hard to reveal the real outliers which could lead to inaccuracy of the final model. Real outliers within financial data could be for example, type mistakes while entered data.

5.2.3 Class imbalance

To predict bankruptcy of a company, we make use of a classification algorithm. However the input to this algorithm of bankrupt companies and healthy companies is in a ratio of 1:20. The problem with classification algorithms is that they tend to learn on the dominated part of the dataset [28], in this case the companies that are healthy. This means that the model will perform quite well if it classifies all companies, including the bankrupt ones, healthy. However, this does not solve the problem.

To avoid this problem, the training set is oversampled with companies that went bankrupt by using the popular SMOTE [29] technique to a 1:1 ratio. The idea of this algorithm is to generate artificial instances based on their k-nearest neighbours. For this particular dataset the best results were received with the Nearest Neighbour parameter chosen as three. This will create artifiicial instances based on the three closest neighbours. This was tested by looking at the AUC-performance values for different values of k. With this balanced training set the algorithm has a better chance to identify the bankrupted companies.

5.2.4 Curse of dimensionality

Data mining within a financial context is often paired with a large number of features. The first thought behind this, is that an increase in features will lead to more information and thus an increase of accuracy in the prediction model. Nonetheless, the increase of features is also affecting the number of dimensions in the space where the instances are. Because of the large number of dimensions, it is difficult for the model to perform well.

A solution to this would be to exponentially let the dataset grow, so the model can learn on more data. The problem with this is that there is no huge amount of high quality data available on this topic. A different approach to avoid the curse of dimensionality is to reduce the amount of features used in the model. Because of the small effective feature set, the data available will be even more interpretable and thus the model will perform better.

5.3 All features

All features used at start are shown in Table 5.1. As noted earlier, feature selection will be done over the listed features for reducing dimensionality, increasing accuracy and interpretability of the model. Noteable things are written down here:

- Attribute X21 uses a yearly growth factor of sales and should therefore be interesting for bankruptcy prediction. However, because it only takes the growth factor of one year it is probably prone to economic bubbles or other economic aspects.

- Attribute X55, working capital, is a single number and is prone to a large spread distribution for different companies worldwide. The working capital varies for different industry sectors. The assumption is that therefore the attribute is not useful for the prediction model on its own. However, if a prediction is needed for different sectors the attribute would be very interesting.

Chapter 6

Results

This chapter describes the most important findings obtained in this research. These findings are acquired by using the approach explained in Chapter 3. Most analysis is done by using Weka 3.8, which is a collection of data mining algorithms [30]. The figures in this section show the healthy company class colored in red and the bankrupted ones in blue.

6.1 Analysing the data

6.1.1 General insights

The number of companies that are healthy is way higher than the companies that went bankrupt. As noted in Section 5.2.3, the data is oversampled with the SMOTE algorithm to prevent generalization of healthy companies. The data we are looking at this section consists out of 16082 companies with an 1:1 ratio of true and false instances as shown in Figure 6.1.

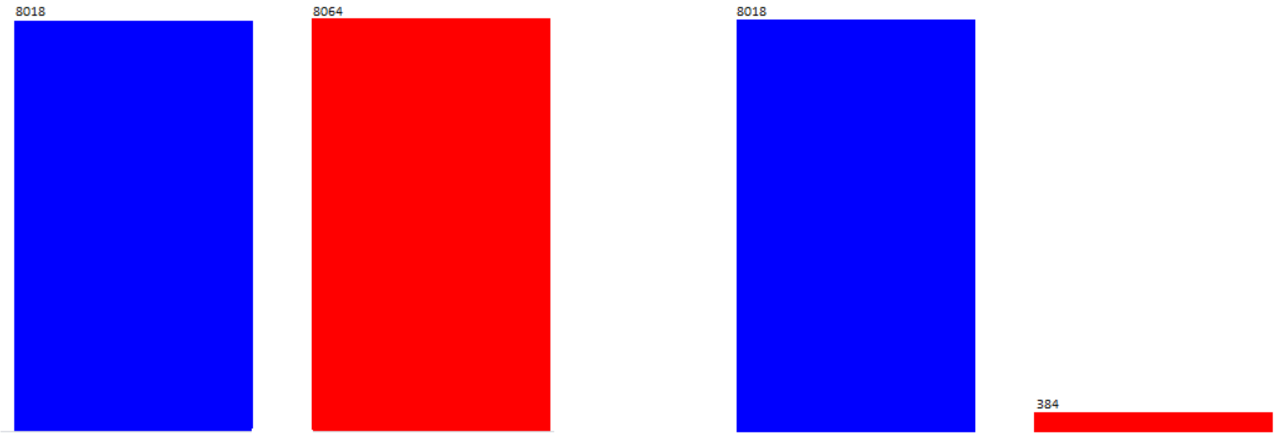

Figure 6.1: Solving the class imbalance problem on the left side we see a 1:1 ratio after oversampling compared to the 1:20 ratio on the right side.

Our first look at the total data set, consisting out of 65 features, shows us that for some attributes there is a relatively high number of missing values. Too many missing values could lead to an inaccurate model, and should be avoided. The features with more than 5% missing values in increasing order, peeking at 50% are shown in Table 6.1.

ID	Attribute name	Missing values (%)
X45	net profit / inventory	7%
X60	sales / inventory	7%
X21	sales (n) / sales (n-1)	17%
X27	profit on operating activities / financial expenses	27%
X37	(current assets - inventory) / long-term liabilities	50%

Table 6.1: Features with missing values.

As noticed in Section 4, missing values are replaced by the modes or means of the relevant features by using the appropriate *ReplaceMissingValues* filter in weka. This method came out best by looking at the AUC-performance values of different methods.

Next to this, it is noticeable that the data has high variance. And as such the standard deviation is fairly to extremely high for many features. A high standard deviation tells us more about the variance in data. The five features with the highest standard deviations are shown in Table 6.2. After these 5 features the standards deviations are declining at exponential rate.

ID	Attribute name	Standard deviation vs Mean
X47	(inventory * 365) / cost of products sold	25235 vs 371
X27	profit on operating activities / financial expenses	29879 vs 814
X55	working capital	44030 vs -736
X15	(total liabilities * 365) / (gross profit + depreciation)	87419 vs 2734
X5	((cash + short-term securities + receivables - short term liabilities) / (operating expenses - depreciation)) * 365	95766 vs -925

Table 6.2: Features with high standard deviations.

This could potentially be caused by outliers or extreme values. However, earlier we noticed that financial data is prone to high variance. The high dimensionality of the data it is too high to guarantee outliers, which is why we asume for now that there are no outliers or extreme values in this particular data.

6.1.2 Attribute evaluators

Before we take an in-depth look at the data, attribute evaluators are used to evaluate how relevant each feature is. We do not want to make any assumptions of which kind of relations should be good for the prediction model. Therefore, different attribute evaluators were used:

- *CfsSubsetEval*: Evaluates the quality of a subset by considering the individual predictive ability of each feature and the correlation between the attributes each other. Features that have high correlation with the target and low correlation between attributes are preferred.

- *ConsistencySubsetEval*: Evaluates the worth of a subset by considering the consistency in class values when the instances are projected on the subset. The consistency can never be lower than that of the full set of attributes, a random or exhaustive search looks for a smaller subset with consistency equal to that of the full set of attributes.

- *CorrelationAttributeEval*: Evaluates the worth of attributes by looking at the correlation between them and the target. For nominal attributes, each value acts as an indicator. Then an overall correlation for a nominal attribute is set via a weighted average.

- *GainRatioAttributeEval*: Evaluates the worth of attributes by measuring gain ratio with respect to class. The gain ratio is defined as information gain divided by the instinct value. It is mainly used to reduce bias towards multi-valued attributes by taking the number and size of branches into account when choosing an attribute.

- *InfoGainAttributeEval*: Evaluaties worth of attribute by measuring information gain with respect to class. Attributes with a higher information gain are preferred above lower ones.

- *OneRAttributeEvall*: Evaluates the worth of an attribute by using the OneR classifier. The classifier looks for those features which have the smallest error, discretizing numeric attributes [31]. As this is our main evaluator further explanation is in Section 4.3.

- *PrincipalComponents*: PCA tries to embed data from high dimensionality to a lower dimensional space while keeping all relevant structure intact, more about PCA can found in [32].

- *SymmetricalUncertAttributeEval*: Evaluates the worth of an attribute by measuring the uncertainty with respect to the class. It mainly answers the question: given attribute X, what fraction of the bits of Y can we predict? It is useful for measuring the validity of a statistical classification algorithm, while avoiding relative fractions of different classes.

AUC values	**21 Attributes**	**18 Attributes**	**15 Attributes**	**10 Attributes**	**7 Attributes**
CfsSubsetEval	0,897	n/a	n/a	n/a	n/a
ConsistencySubsetEval	0,88	n/a	n/a	n/a	n/a
GainRatioAttributeEval	0,897	0,895	0,892	0,818	0,786
GainRatioAttributeEval	0,925	0,917	0,905	0,873	0,838
InfoGainAttributeEval	0,918	0,901	0,881	0,847	0,838
OneRAttributeEval	0,938	0,914	0,922	0,912	0,855
PCA	0,886	0,899	0,893	0,735	0,788
SymmetricalUncert	0,926	0,911	0,908	0,88	0,841

Table 6.3: AUC performance values for different subsets of the top attributes suggested by evaluators.

The results of these attribute evaluators are shown in Table 6.3. For all evaluators explained above, we took a look at different subsets of attributes and see how well they performed. All AUC performance values are tested with the C4.5 algorithm as explained in Section 4.5. For most of the evaluators, there was a noticeable performance drop around fifteen features. The oneR attribute evaluator stays quite consisted and only started dropping around 10 attributes chosen. The subset selected by oneR are closer analyzed because of a higher performance and less performance decrease compared to others. The ten attributes chosen by the

OneRAttributeEval algorithm are shown in table 6.4.

Rank	ID	Attribute Name
1	27	profit on operating activities / financial expenses
2	13	(gross profit + depreciation) / sales
3	26	(net profit + depreciation) / total liabilities
4	23	net profit / sales
5	42	profit on operating activities / sales
6	16	(gross profit + depreciation) / total liabilities
7	19	gross profit / sales
8	6	retained earnings / total assets
9	2	total liabilities / total assets
-	65	class attribute

Table 6.4: Most important attributes evaluated by *OneRAttributeEval*.

6.1.3 Correlations of features

	X2	X6	X13	X16	X19	X23	X26	X27	X42
X2	1	-0.97	0.01	0	0.03	0.03	0	0	-0.01
X6	-0.97	1	-0.01	0.06	-0.03	-0.03	0.06	0	0.01
X13	0.01	-0.01	1	0	0.2	0.2	0	0	0.24
X16	0	0.06	0	1	0.01	0.01	1	0	0
X19	0.03	-0.03	0.2	0.01	1	1	0.01	0	0.98
X23	0.03	-0.03	0.2	0.01	1	1	0.01	0	0.99
X26	0	0.06	0	1	0.01	0.01	1	0	0
X27	0	0	0	0	0	0	0	1	0
X42	-0.01	0.01	0.24	0	0.98	0.99	0	0	1

Table 6.5: Correlation matrix from attributes evaluated by *OneRAttributeEval*.

To take a closer look at these ten features, we plotted a scatter plot as shown in Figure 6.2. The scatter plot is very useful to spot potential correlations. It shows us whether there might be any perfect correlations between features. Next to the scatter plot, exact correlations are shown in Table 6.5. The following remarks have been noticed:

- Because of the earlier used attribute evaluator *OneRAttributeEval*, we can guarantee a correlation between the class attribute and other attributes. This means that all attributes have a relationship with our prediction class.

- Most of the data points, except from the linear relations, are clustered. It tells us that the variance between the features for these attributes is quite low.

- An almost perfect correlation (-0,97) is seen between X2 (*total liabilities / total assets*) and x6 (*retained earnings / total assets*). Which could be explained, because both attributes are ratios divided by a variable realted to sales.

- A weak correlation (0.24) is seen between X13 ((*gross profit + depreciation) / sales*) and x42 (*profit on operating activities / sales*). Which could be explained, because both attributes are again ratios divided by the same value. However, the added value of *gross profit* and *depreciation* causes a relevant difference and therefore both attributes should be taken into account.

- A perfect correlation (1) is seen between X26 ((*net profit + depreciation) / total liabilities*) and x16 ((*gross profit + depreciation) / total liabilities*). This suggests that the differences between *gross and net profit* are so small that adding both attributes X26 and X16 to the model should be avoided. It could also be that for this particular industry sector the variables are equal. Next to this, it should be also noticed that both features use the attribute *total liabilities* and *depreciation* which causes a high correlation.
- An almost perfect correlation is seen between X19 (*gross profit / sales*) and x42 (*profit on operating activities / sales*). Which could be explained, because both attributes are ratios divided by the same value.
- The same relation is seen between X23 ((*net profit / sales*) and x42 ((*profit on operating activities / sales*). The contribution of X19 (*gross profit / sales*) and X23 ((*net profit / sales*) should therefore be taken into account.

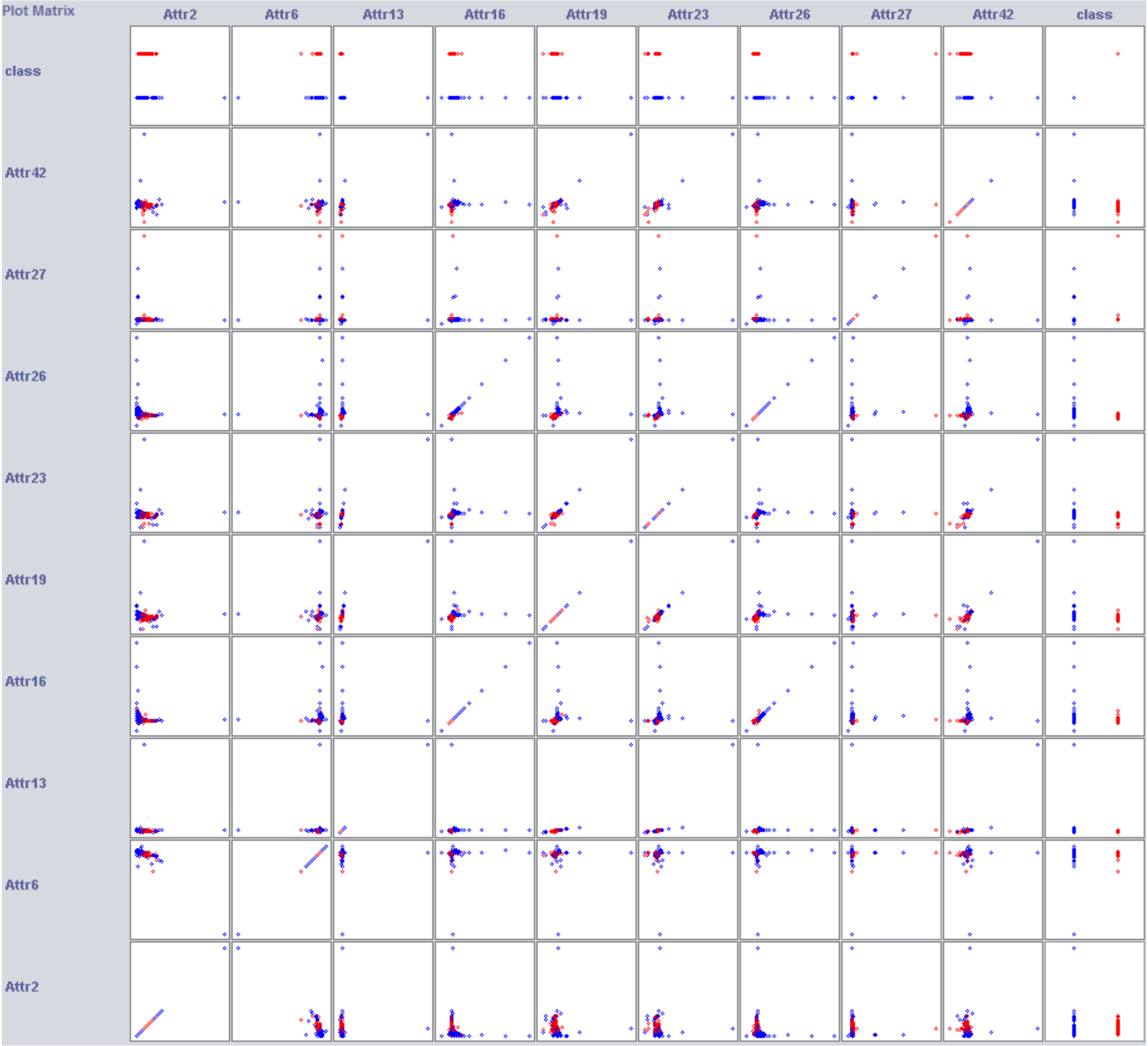

Figure 6.2: Scatterplot of the ten features suggested by *OneRAttributeEval*.

6.1.4 Statistical characteristics features

	X2	X6	X13	X16	X19	X23	X26	X27	X42
Minimum value	0	-508	-632	-204	-771	-771	-204	-190130	-765
Maximum value	409	46	4972	8260	124	124	8262	2723000	166
Mean	0,70	-0,11	0,22	1,54	-0,15	-0,16	1,48	814,71	-0,1
Standard deviation	3,83	4,13	40,05	86,76	8,98	8,96	86,46	25554,16	8,69
Unique (%)	33%	64%	91%	95%	92%	91%	95%	67%	87%

Table 6.6: Statistical values from attributes evaluated by *OneRAttributeEval*.

To investigate the features as proposed by the *OneRAttributeEval* even further, we have looked to the statistical values and distributions of them. The statistical values are shown in Table 6.6. Statistical values could give us new insights about the used features which could lead to an improvement on the accuracy of the predictive model. Important notices are:

- In general, most of the features chosen by the evaluator have a relatively small standard deviation compared to the original set of features. The standard deviation tells us something about the variance of the data. It should be taken into account that financial data always has been sensitive for a high variance, because of the large financial differences between companies. That means that a standard deviation between zero and hundred percent is not an uncommon phenomenon.

- Attribute X27 *profit on operating activities / financial expenses* shows us a very high standard deviation. Yet, it is one of the most important attributes rated by the *OneRAttributeEval*. This suggests that the final model trained should be always used carefully and be fine tuned to a set domain of companies. As it is expected that results will vary too much between different domains.

- As noted before in Section 6.1.3, there is a correlation between attributes X19 (*gross profit / sales*) and X23 ((*net profit / sales*). In the statistics we can now also see that both attributes almost have the same values for different metrics. This could identify a potential duplicate of an attribute and one of them should therefore be removed.

- Attributes X2 (*total liabilities / total assets*) and X6 (*retained earnings / total assets*) show a low percentage for the statistic unique numbers. Since they share the value of $totalassets$, we could assume that values within this attribute are rounded, estimated or misentered.

- For most of the attributes shown in the table, we notice that the minimum values of them are negative. Theoretically, most of them are not supposed to be negative.

6.1.5 Distributions of features

To explore the chosen attributes chosen by *OneRAttributeEval* even more, distributions are made. To make these distributions, outliers and all extreme values are removed (rougly six percent), by taking the statistical values of Table 6.6. This way, we can visualize the distrubition in such a way that results are interpretable. The distributions tells us something about how the data is spread within the statistical values named in Table 6.6. Interesting sightings are:

- The skewness of attribute X2 (*total liabilities / total assets*), as shown in figure 6.3, is tending to the right. However, the statistics in Table 6.6 tell us that there is a small standard deviation which implies there could be a normal shaped distribution. This effect could be caused by the low number of unique numbers. We can also notice a significant difference between the healthy companies and bankrupted ones. From this distribution we could imply that the proportion between *total liabilities* and *assets* is an important factor for bankruptcy prediction. Yet, because of the skewed distribution we should be careful with using this feature.

- As suspected, attributes 19 (*gross profit / sales*) and X23 ((*net profit / sales*) are showing an almost identical distribution. The features, which show identical values, were probably named wrong when the initial dataset was acquired. As noted earlier, this also could be a phenomenon for the industry sector only.

- Figure 6.10 with attribute X27 (*profit on operating activies / financial expenses*) shows an interesting peak around one specific value. A financial explanation could be that a significant proportion of the companies are working towards a treshold. This could be, for example, to avoid an increase in taxes above a certain value. It is probably that the data is biased, because of the high frequency of a certain odd value, and thus should be used carefully within a model.

- The attributes X6, X13, X16, X19, X23, X26 and X42 show a normal distribution, which says that the data is less varianced and therefore could be more reliable because of the normal occurence of frequencies.

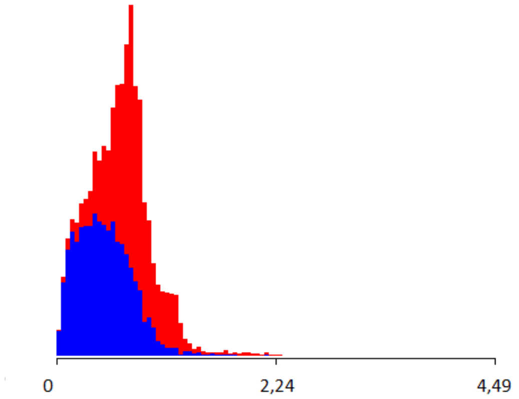

Figure 6.3: Distribution of attribute X2 (*total liabilities / total assets*).

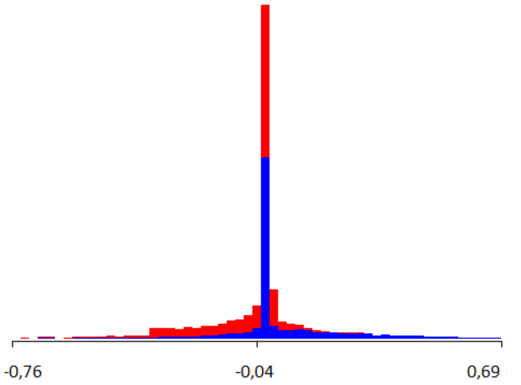

Figure 6.4: Distribution of attribute X6 (*retained earnings / total assets*).

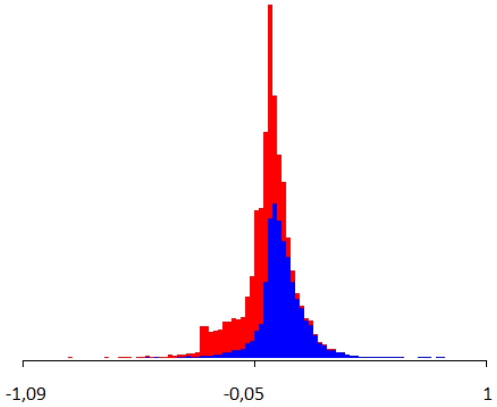

Figure 6.5: Distribution of attribute X13(*(gross profit + depreciation) / sales*).

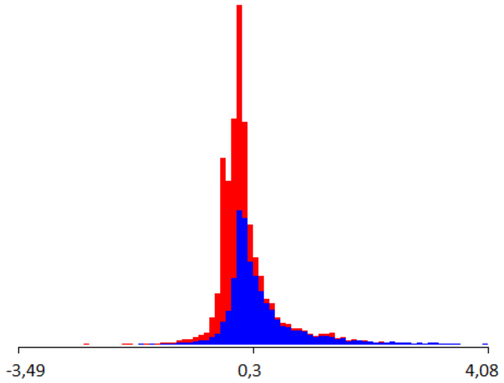

Figure 6.6: Distribution of attribute X16 (*(gross profit + depreciation) / total liabilities*).

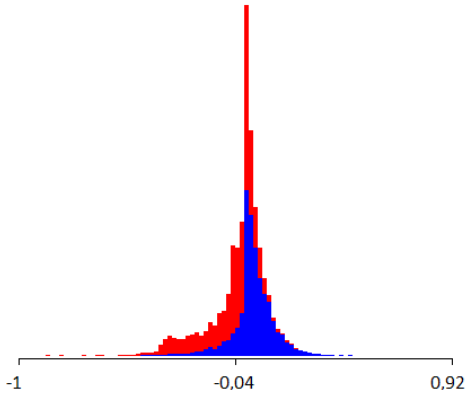

Figure 6.7: Distribution of attribute X19 (*gross profit / sales*).

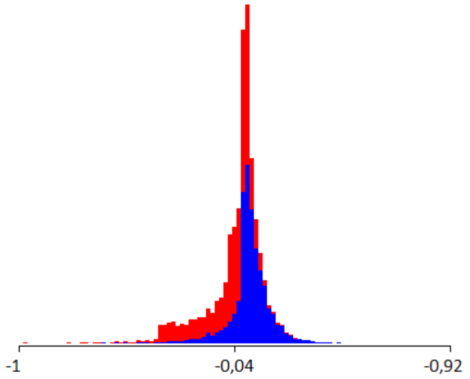

Figure 6.8: Distribution of attribute X23 (*net profit / sales*).

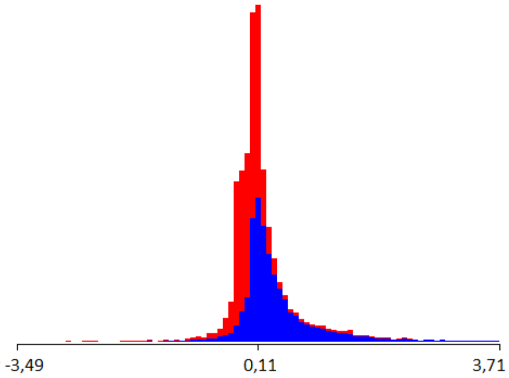

Figure 6.9: Distribution of attribute X26 ((*net profit + depreciation) / total liabilities*).

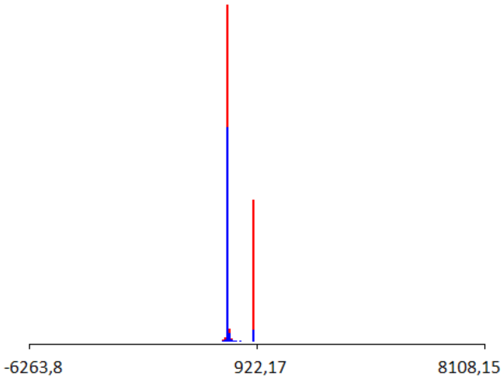

Figure 6.10: Distribution of attribute X27 (*profit on operating activies / financial expenses*).

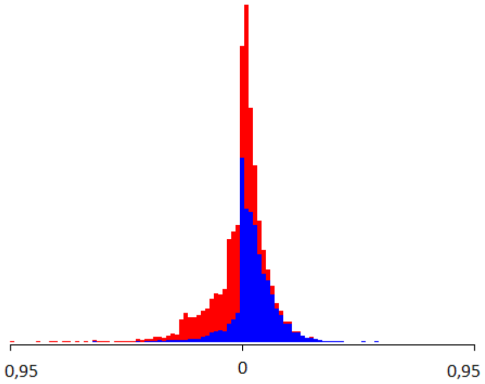

Figure 6.11: Distribution of attribute X42 (*profit on operating activities / sales*).

6.2 Performance of model

In this chapter, we will look at the AUC performance values of the different models following the approach as noted earlier in Chapter 4. This section is divided in two parts where we look at the *J48* and *RandomForest* algorithms. As stated earlier, the *J48* is used to make quick data preparation decisions and the *RandomForest* algorithm will be used to make a final prediction model.

6.2.1 Final decisions for the model

We have already discussed reducing the dimensionality of the model in Section 5.2.4. Looking at the AUC performance values now, we can see a huge improvement in performance when using an efficient feature selection as shown in Table 6.7. This table looks at all the features in the start, compared to the 10 features chosen by the oneRevaluator. The difficult part of feature selection is the trade off between accuracy, complexity and explainabilty of the model. In this research, the goal is to keep the model as simple, yet accurate as possible.

AUC values	**Initial feature set**	**10 features**
J48 — Train	0,715	0,913

Table 6.7: AUC performance values of the initial feature set compared to the ten features chosen by the oneR evaluator.

By taking different charactereristics into account, as noted in previous Section 5.1, we have come up with an even more efficient feature set. In Table 6.8 below AUC performance values are shown by removing single attributes of the feature set suggested by the oneR evaluator.

AUC values	**- X2**	**- X6**	**- X13**	**- X16**	**- X19**	**- X23**	**- X26**	**- X27**	**- X42**
J48 - Train	0,925	0,913	0,920	0,909	0,911	0,912	0,917	0,880	0,910

Table 6.8: AUC performance values of the OneR evaluator set minus one selected attribute.

We can see that by removing one of the following attributes (X2, X6, X13, X16, X26) there is an expected increase of the performance. However, removing them all will make the model to simple for an accurate prediction. We take earlier noticed aspects into account for the final feature selection:

- Attributes X19 (*gross profit / sales*) and X23 (*net profit / sales*) are showing an almost identical distribution, same statistical values and an almost perfect correlation. This indicates that the attributes actually have the same values. Attribute 19 is therefore removed.

- Attributes X26 (*(net profit + depreciation) / total liabilities*) and X16 (*(gross profit + depreciation) / total liabilities*) are showing almost identical values in Table 5.6, show a perfect correlation and have almost the same distribution. This also indicates that both attributes share the same values. Attribute 16 is therefore removed.

- Attributes X23 (*net profit / sales*) and X42 (*profit on operating activities / sales*) are also showing almost identical values, an equal distribution and an almost perfect correlation. The difference between net profit and profit on operating activities seems to be small in the manufacturing sector. Attribute 42 is therefore removed.

- Attribute X2 (*total liabilities / total assets*) shows a right skewed distribution and a potential increase in performance as shown in Table 6.8. The skewness of the model is an indication that it would perform less well on new data, because it indicates a high variance and therefore a difficulty for the model to perform.

This leads to our final feature selection consisting out of the attributes shown in Table 6.9. A more in-depth analysis to this feature set is made in the next section 5.3.

ID	Attribute name
6	retained earnings / total assets
13	(gross profit + depreciation) / sales
23	net profit / sales
26	(net profit + depreciation) / total liabilities
27	profit on operating activities / financial expenses
65	class attribute

Table 6.9: The final attributes chosen.

6.2.2 Performance values

With the obtained selection made of features, AUC performance values are calculated to see how well the model works. In Table 6.10 we see both the AUC values for the J48 and RandomForest algorithms based on the training and test set. This is important, because we want to know how well the model is working on unseen data.

Model / Dataset	AUC value
J48 - Train	0,911
J48 - Test	0,895
RandomForest - Train	0,965
RandomForest - Test	0,988

Table 6.10: The AUC performance values of the J48 and RandomForest algorithm based on the training and test set.

As suspected, the RandomForest algorithm performs much better than the J48 algorithm because of its higher complexity, higher accuracy but lower interpretability. We can also notice a difference between the training and test set as probably a result of overfitting to the training set. The data set for training and testing was relatively small, so this is an expected result.

The main aim of this research was to reduce the number of features to increase accuracy and interpretability. The AUC performance values are so high, that the model probably only performs well on very equal data as presented in this single data set. Therefore, new data is tested with the same prediction model as shown in Table 6.11. This data was set available from [2]. The table shows us the AUC performance values of datasets, including companies with a classification of bankruptcy within the number of years as shown in the table. The AUC values in this table are more likely, as noted in Section 7.1, and could be labeled more reliable.

AUC values	**5 years**	**4 years**	**3 years**	**2 years**	**1 year**
RandomForest - Test	0,880	0,775	0,988	0,780	0,827

Table 6.11: The AUC performance values tested on new data. The number of years indicates when the classification of bankruptcy is set.

6.3 Interpretation of model

This section describes how the model could be used in practice and describes why the chosen attributes are useful in bankruptcy prediction. This section will refer to the performance of the model as shown in Section 5.2. Most models used today are kept private likely, for (anti-) competitive reasons. This section could especially be interesting for accounting experts to understand the reasoning behind such a prediction.

As noted earlier, financial data are error prone. Using different data sets will make the prediction less sensitive to overfitting to a single domain of data. The data used, approximately 10,000 polish instances, is a relatively small number to guarantee a working prediction model in the manufacturing sector for worldwide companies. Therefore, more testing and maybe fine tuning of the model should be done on different datasets before using it in daily pursuits. As a contribution to previous research, the importance of data preparation at this certain topic is shown in this research. The same approach could be used to ensure maximization of accuracy and interpretability of the model.

The used attributes for the final prediction model, as shown in Table 6.9, are analyzed to get an even better understanding of bankruptcy and the model:

- Attribute X6 *retained earnings / total assets* shows us the ratio between the percentage of profit that is held within the company versus the amount of total assets. Low retained earnings could imply that the company has an inconsistent income, and thus management have not figured out to improve profitability. A high percentage of retained earnings could imply an aggressive growth strategy and therefore taking more risk to profit. So, the percentage of retained earnings tells us something about the company strategy to profit. It is interesting to have a ratio putting that to a firm's safe assets, which values a company.

- Attribute X13 *(gross profit + depreciation) / sales* is interesting because it shows a ratio between profit and sales. The ratio tells us a lot about the companys strategy, because it takes into account how much variabele costs are made to make profit. Adding depreciation to this gives not only insight in the variable costs, but also the fixed costs.

- Attribute X23 *net profit / sales*, as explained above, tells us a lot about the company's strategy because it takes into account how much variable costs are made to make profit.

- Attribute X26 *(net profit + depreciation) / total liabilities* indicates the ratio of profit versus the liabilities. The suggestion it makes, is that it is crucial to a company's decision making, to keep track of the total amount of liabilities versus the profit and the fixed costs.

- Attribute X27 *profit on operating activities / financial expenses* is the ratio between profit on operating activities (as noted earlier, in this sector, it is almost the same as net profit) and the expenses made on total assets of the company. This tells us more about the ratio of profit against the costs of total assets and therefore implies the importance of a solid cash flow for the company.

In current accounting, known attributes for bankruptcy prediction are for example: *Current Assets/Current Liabilities*, *Operating Cash Flow/ Sales*, *Debt/Equity ratio* and *Cash flow/Debt ratio*. It is interesting to see that some attributes of this research are new and others are excisting bankruptcy indicators. The assets to risk strategy can be clearly seen from both set of attributes. Yet, our research shows that costs and depreciation are importants factors to take into account to. It should be noted that as discussed earlier, our research is based on the industry sector only and based on a relative small dataset.

Chapter 7

Conclusions

This chapter describes the contributions made to this research topic. Also, the earlier discussed research questions are answered:

- To what extent can we find a pattern in financial data of failed companies?
- Given a company's financial data of the past 3 years, can we accurately predict bankruptcy?

7.1 Prediction

Previous work done has shown that data mining is important to get an accurate model for the prediction of bankruptcy. In this research, we saw that data preparation should not be neglected and is of great significance to data mining results. As suspected, there is a great increase in performance compared to results without sufficient data preparation. However, there is always a trade-off between the amount of information gained, accuracy and the interpretability of the model.

The related work of Zieba and Tomczak achieved AUC performance values of 0.701 for the J48 and 0.831 for the RandomForest algorithm, where this research performs with the respective values of 0.911 and 0.965. It should be noted that all of these 4 values are based on 10-cross-fold validation. Yet, performance values on different data shows values around 0.8 and thus shows signs of overfitting to the used data set. This will be discussed more in the next section on future work.

The interpretability of the model, and thus the pattern we could see in financial data, is difficult to understand from a Randomforest algorithm. This is because the RandomForest algorithms make use of different trees with huge depths and is therefore hard to understand. However, we can interpret the model out of the final features used. The generic finding out of the used attributes is that bankruptcy is related to the company's strategy of how much risk the company is taking versus the amount of fixed assets. We could see this from different attributes, for example, retained earnings were held against total assets, or costs against profits made.

7.2 Future work

As noted earlier, the approach of this research by focussing on data preparation instead of choosing the right algorithm is expected to show a great increase in performance. However, because of the size of the dataset, we have seen a model that was prone to overfitting. Using the same approach on a larger dataset and taking more time for data preparation, with the main focus on feature selection, should avoid this problem. The combination of extensive data preparation and choosing ensembled boosted classifiers would lead to a very accurate model.

However, to actually find quality data for such research is a challenge. Financial data, as noted earlier, is prone to errors. A suggestion for stakeholders (for example, financial institutions or governments) is to start creating gigantic quality dataset with proper data quality to accomplish such research in the future.

Bibliography

[1] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[2] M. Zieba, S. K. Tomczak, and J. M. Tomczak, "Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction," *Expert Systems with Applications*, vol. 58, pp. 93–101, 2016.

[3] Wikipedia. https://en.wikipedia.org/wiki/Datamining, accessed 27-07-2017.

[4] P. J. FitzPatrick, *A comparison of the ratios of successful industrial enterprises with those of failed companies*. 1932.

[5] W. H. Beaver, "Financial ratios as predictors of failure," *Journal of accounting research*, vol. 4, no. 1, pp. 71–111, 1966.

[6] E. I. Altman, "Multidimensional graphics and bankruptcy prediction: a comment," *Journal of Accounting Research*, vol. 21, no. 1, pp. 297–299, 1983.

[7] J. A. Ohlson, "Financial ratios and the probabilistic prediction of bankruptcy," *Journal of accounting research*, vol. 18, no. 1, pp. 109–131, 1980.

[8] Y. Zhang, S. Wang, and G. Ji, "A rule-based model for bankruptcy prediction based on an improved genetic ant colony algorithm," *Mathematical Problems in Engineering*, vol. 2013, no. 1, pp. 3–10, 2013.

[9] K.-S. Shin, T. S. Lee, and H.-j. Kim, "An application of support vector machines in bankruptcy prediction model," *Expert Systems with Applications*, vol. 28, no. 1, pp. 127–135, 2005.

[10] A. Koster, N. E. Sondak, and W. Bourbia, "A business application of artificial neural network systems," *Journal of Computer Information Systems*, vol. 31, no. 2, pp. 3–9, 1991.

[11] D. T. Cadden, "Neural networks and the mathematics of chaos-an investigation of these methodologies as accurate predictors of corporate bankruptcy," in *Artificial Intelligence Applications on Wall Street, 1991. Proceedings., First International Conference on*, pp. 52–57, IEEE, 1991.

[12] K. Y. Tam, "Neural network models and the prediction of bank bankruptcy," *Omega*, vol. 19, no. 5, pp. 429–445, 1991.

[13] R. C. Lacher, P. K. Coats, S. C. Sharma, and L. F. Fant, "A neural network for classifying the financial health of a firm," *European Journal of Operational Research*, vol. 85, no. 1, pp. 53–65, 1995.

[14] K. Y. Tam and M. Y. Kiang, "Managerial applications of neural networks: the case of bank failure predictions," *Management science*, vol. 38, no. 7, pp. 926–947, 1992.

[15] R. L. Wilson and R. Sharda, "Bankruptcy prediction using neural networks," *Decision support systems*, vol. 11, no. 5, pp. 545–557, 1994.

[16] C. Serrano-Cinca, "Self organizing neural networks for financial diagnosis," *Decision Support Systems*, vol. 17, no. 3, pp. 227–238, 1996.

[17] G. Zhang, M. Y. Hu, B. E. Patuwo, and D. C. Indro, "Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis," *European journal of operational research*, vol. 116, no. 1, pp. 16–32, 1999.

[18] R. Geng, I. Bose, and X. Chen, "Prediction of financial distress: An empirical study of listed chinese companies using data mining," *European Journal of operational research*, vol. 241, no. 1, pp. 236–247, 2015.

[19] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, "On combining classifiers," *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[20] L. Nanni and A. Lumini, "An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring," *Expert systems with applications*, vol. 36, no. 2, pp. 3028–3033, 2009.

[21] J. R. Quinlan, *C4. 5: Programs for machine learning*. Elsevier, 2014.

[22] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[23] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve.," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[24] M. Zikeba, S. K. Tomczak, and J. M. Tomczak. https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data, accessed 01-03-2017.

[25] EMIS. https://www.emis.com/, accessed 01-03-2017.

[26] J. Grzymala-Busse and M. Hu, "A comparison of several approaches to missing attribute values in data mining," in *Rough sets and current trends in computing*, pp. 378–385, Springer, 2001.

[27] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," in *ACM Sigmod Record*, vol. 30, pp. 37–46, ACM, 2001.

[28] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent data analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[29] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, no. 1, pp. 321–357, 2002.

[30] Waikato. http://www.cs.waikato.ac.nz/ml/weka/, accessed 01-01-2017.

[31] R. Holte, “Very simple classification rules perform well on most commonly used datasets,” *Machine Learning*, vol. 11, no. 1, pp. 63–91, 1993.

[32] C. Boutsidis, M. W. Mahoney, and P. Drineas, “Unsupervised feature selection for principal components analysis,” in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 61–69, 2008.