1. Data description

For the purpose of the default prediction analysis we used the Creditreform database obtained from the Lab for Empirical and Quantitative Research (LEQR) of the Humboldt University of Berlin (leqr.wiwi.hu-berlin.de). This dataset contains a random sample of 20,000 solvent and 1,000 insolvent German firms for the period 1996-2007 (except 2004). Due to incomplete data from 2003 onwards and missing data for insolvent firms in 1996 we will focus our analysis on the data of the period 1997 to 2002. About the half of these data refer to the years 2001 and 2002. The majority of firms appear several times in different years in the dataset, while the data for defaulted firms were collected two years before the default (Chen et al. (2011)). Each firm is described by several financial statement variables as those in balance sheets and income statements. A complete list of all variables of the Creditreform database as well as their descriptions is provided in the Appendix.

The firms of the database are classified into economic activity sectors according to the German Classification of Economic Activities, Edition 1993 (WZ93) issued by the German Federal Statistical Office (destatis.de). This classification method uses a five-digit code (VAR26 in the dataset), where the first two digits correspond to 17 broad sectors of the economy and the other three are used for a more precise classification into subsectors.

The industry composition for solvent and insolvent firms as described by Chen et al. (2011) is as follows. The insolvent firms are divided into the sectors of construction (39.7%), manufacturing (25.7%), wholesale and retail trade (20.1%), real estate (9.4%) and others (5.1%) (including agriculture, mining, electricity, gas and water supply, hotels and restaurants, transport and communication, financial intermediation and social service activities. The respective composition of solvent firms is manufacturing (27.4%), wholesale and retail trade(24.8%), real estate (16.9%), construction (13.9%) and others (17.1%) (including publishing, administration and defence, education and health). In our project we focus on the four largest industry sectors as in Chen et al. (2011), Zhang & Härdle (2010) and Härdle et al. (2012).

Following the methodology of the above mentioned articles, we prepare the dataset and construct 28 financial ratios to be used for classification and predictions about the solvency status of German firms. This task is completed by our first quantlet ("data.preparation"), which is described in the following two sections.

2. Data preparation

In the current project the free programming language R was used for the purpose of the task mentioned in the previous section. R is a free programming language and software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing (for more information on R see `www.r-project.org`).

We first install the dplyr package by using the command install.packages("dplyr"). The install.packages() command is used for the installation of additional packages in R when necessary with adding the name of the package in quotation marks inside the parenthesis like install.packages("name-of-the-package"). The dplyr package will allow as to manipulate data easier than with base R with less code typing. We insert the Creditreform database into R with the read.csv() command and store it as "data" before starting with the data cleaning. Using the filter() command of dplyr we choose only those observations with a value of the variable "JAHR" in the range 1997-2002 for the reasons mentioned in the previous section.

Then we want to choose only those observations belonging to the four industry sectors with the higher percentages in the industry composition for both solvent and insolvent firms, i.e manufacturing, wholesale and retail trade, real estate and construction. We extract the industry class of firms by using the substring() command and save it in a new column of data1 as follows:

data1$Ind.Klasse = substring(data1$VAR26, 1, 2)

These two digits contained in the variable "Ind.Klasse" are then used to identify in which of the 17 broad industry sectors each firm belongs. If the value of the variable is in the range 15-37 then the firm belongs to the manufacturing sector. Accordingly the ranges 50-52 and 70-74 correspond to "Wholesale and Retail Trade" and "Real Estate" respectively, while the value 45 corresponds to "Construction". We create four subsets of data1 for each of the above mentioned sectors by using the filter() command again and remove data and data1 with the rm() command as we do not need them anymore. The 4 subsets are then bound with rbind() into a new dataset with the name "data".

Then we turn our interest on the size of the companies and specifically the distribution of total assets which can be considered to be representative of the distribution of the companies' size (Chen et al., 2011). Following the methodology of Zhang & Härdle (2010), we keep in our study only firms with total assets (VAR6 in the dataset) in the range of $10^5 - 10^8$ Euros, because

the credit quality of small firms often depends as much on the finances of a key individual (e.g. the owner) as on the firm itself and largest firms rarely default in Germany.

Finally, we eliminate observations with zero values in variables used as denominators in the calculation of the financial ratios that will be used for the classification of the companies (lines $63 - 70$ in the code) and save the result as "data_clean". We end up with 9591 solvent and 783 insolvent firms, which is similar result to Chen et al.(2010). The number of solvent and insolvent firms per year is shown in table 1.

Table 1. Number of solvent and insolvent companies per year in the dataset

Year	Solvent	Insolvent
1997	1084	126
1998	1175	114
1999	1277	147
2000	1592	135
2001	1920	132
2002	2543	129

3. Financial Ratios

The Creditreform database contains many financial statement variables for each company. We follow the methodology of Chen et al. (2011) and we use 23 of them to create 28 financial ratios to be used in classification. The variables used in the creation of the financial ratios are summarized in table 2. These financial ratios can be divided into six main groups (risk factors): profitability, leverage, liquidity, activity, firm size and percentage change for some variables. In the next paragraphs we give a brief description for each of the six groups followed by some examples.

Profitability ratios are a class of financial metrics that are used to assess a firm's ability to generate earnings compared to its expenses and other relevant costs incurred during a specific period of time (investopedia.com). Profitability ratios have appeared in many studies to be strong predictors for bankruptcy. We calculate 7 ratios belonging to this group (ratios x1-x7). For example the return on assets ratio (x1) can inform investors on how effective is a firm in making use of its assets to generate income. A higher ratio means that a firm is able to earn more money on less investment.

Table 2. Variables used for the calculation of financial ratios and their description

Variable	Description	Variable	Description
VAR1	Cash and cash equivalents	VAR14	Bank debt
VAR2	Inventories	VAR15	Accounts payable
VAR3	Current assets	VAR16	Sales
VAR5	Intangible assets	VAR18	Amortization and depreciation
VAR6	Total assets	VAR19	Interest expenses
VAR3 - VAR2	Quick assets	VAR20	EBIT
VAR7	Accounts receivable	VAR21	Operating income
VAR8	Lands and buildings	VAR22	Net income
VAR9	Equity (own funds)	VAR23	Increase (decrease) inventories
VAR12	Total current liabilities	VAR24	Increase (decrease) liabilities
VAR12 + VAR13	Total liabilities	VAR25	Increase (decrease) cash
VAR3 - VAR12	Working capital		

Another ratio belonging to the profitability ratios is the net profit margin ratio (x2). It shows the percentage of sales that the firm actually keeps in earnings. A high ratio corresponds to a firm with more profitability and better control over its costs compared to the other firms.

Another key factor of risk measurement is leverage. Leverage ratios look at how much capital of a firm comes in the form of debt (loans), or assess the ability of a firm to meet its financial obligations. These ratios are important because as firms combine equity and debt to finance their operations, knowing the amount of debt held by a firm is useful in evaluating its ability to pay off its debts as they come due (investopedia.com) We calculate 7 ratios belonging to this group (ratios x8-x14). An example of a leverage ratio is the net indebtedness (x11) which measures the level of short term liabilities not covered by the firm's most liquid assets as a proportion to the firm's total assets. Except from measuring the short term leverage of a firm, this ratio provides a measure of liquidity as well. Another popular leverage ratio is the debt ratio (x13), which is defined as the debt of a company divided by its total assets. While this ratio performs well for public firms, it performs considerably worse for private firms compared to the total liabilities to total assets ratio (x12). The reason for that is that liabilities is a more inclusive term which includes debt, deferred taxes, minority interest, accounts payable and other liabilities (Chen et al.,2011).

The next six financial ratios we calculate belong to the family of liquidity ratios (ratios x15-x20). Liquidity is a common variable in many credit decisions and represents a firm's ability to convert an asset into cash quickly

(Chen et al.,2011). Ratios of this type are often used by bankruptcy analysts and mortgage originators to evaluate going concern issues, as they indicate cash flow positioning (investopedia.com). Chen et al. (2011) note that the cash to total assets ratio (x15) is the most important single variable relative to default in the private dataset. The quick ratio (x17) is an indicator of a company's short-term liquidity and measures the company's ability to meet its short-term obligations with its most liquid assets.

Another type of ratios which deliver important information on insolvency are the activity ratios (x21-x24). They measure a firm's ability to convert different accounts within its balance sheets into cash or sales. Activity ratios measure the relative efficiency of a firm based on its use of its assets, leverage or other such balance sheet items and show if the firm's management can efficiently generate revenues and cash from its resources (investopedia.com). An example of activity ratio is the accounts receivable turnover ratio (x23). It determines an entity's ability to collect cash from its customers and thus, a higher ratio signals a more efficient collection process.

As in Chen et al. (2011), we also compute an indicator of size risk which also in our case is the logarithm of total assets (x25) in order to study the insolvency risk of small, medium and large firms. Finally, we calculate the ratios of the percentage change of incremental inventories, liabilities and cash flow (x26-x28). As the increased cash flow is the additional operating cash flow that an entity receives from taking on a new project, a positive incremental cash flow means that the firm's cash flow will increase with the acceptance of a project, the ratio of which indicates that the firm should invest time and money in the project (Chen et al., 2011).

Table 3 presents all the financial ratios used in the current study, the formulas used for their calculation and their category. In our code, the calculation procedure of the financial ratios can be found in lines 95-124 of the quantlet "Data Preparation", where the ratios for each firm are computed and added as columns to the dataset by using the mutate() command of the dplyr package. Then we create a dataset (test_data_rel) where only relevant variables are kept, i.e. ID of the firm, solvency status, year and the 28 financial ratios.

Table 3. Definitions of financial ratios

Ratio No.	Formula	Ratio	Category
x1	VAR22/VAR6	Return on assets (ROA)	Profitability
x2	VAR22/VAR16	Net profit margin	Profitability
x3	VAR21/VAR6		Profitability
x4	VAR21/VAR16	Operating profit margin	Profitability
x5	VAR20/VAR6		Profitability
x6	(VAR20+VAR18)/VAR6	EBITDA	Profitability
x7	VAR20/VAR16		Profitability
x8	VAR9/VAR6	Own funds ratio (simple)	Leverage
x9	(VAR9-VAR5)/(VAR6-VAR5-VAR1-VAR8)	Own funds ratio (adjusted)	Leverage
x10	VAR12/VAR6		Leverage
x11	(VAR12-VAR1)/VAR6	Net indebtedness	Leverage
x12	(VAR12+VAR13)/VAR6		Leverage
x13	VAR14/VAR6	Debt ratio	Leverage
x14	VAR20/VAR19	Interest coverage ratio	Leverage
x15	VAR1/VAR6		Liquidity
x16	VAR1/VAR12	Cash ratio	Liquidity
x17	(VAR3-VAR2)/VAR12	Quick ratio	Liquidity
x18	VAR3/VAR12	Current ratio	Liquidity
x19	(VAR3-VAR12)/VAR6		Liquidity
x20	VAR12/(VAR12+VAR13)		Liquidity
x21	VAR6/VAR16	Asset turnover	Activity
x22	VAR2/VAR16	Inventory turnover	Activity
x23	VAR7/VAR16	Account receivable turnover	Activity
x24	VAR15/VAR16	Account payable turnover	Activity
x25	log(VAR6)		Size
x26	VAR23/VAR2	Percentage of incremental inventories	Percentage
x27	VAR24/(VAR12+VAR13)	Percentage of incremental liabilities	Percentage
x28	VAR25/VAR1	Percentage of incremental cash flow	Percentage

In order to avoid sensitivity to outliers in applying the SVM and the logit model, we follow the methodogy of Chen et al. (2011) and we replace extreme ratio values according to the following rule: For i = 1,...,28, if $x_i < q_{0.05}(x_i)$, then $x_i = q_{0.05}(x_i)$, and if $x_i > q_{0.95}(x_i)$, then $x_i = q_{0.95}(x_i)$, where $q_{0.05}(x_i)$ and $q_{0.95}(x_i)$ refer to the 0.05 and 0.95 quantiles of the ratio x_i respectively. This will make our results robust and insensitive to outliers. For that purpose we create the function "replace_extreme_values()" (lines 83-91 of the "data.preparation" quantlet), which is then separately applied to the subsets of solvent and insolvent companies (lines 134-145). The lower and upper quantile as well as the median of the financial ratios for solvent and insolvent firms are presented in table 4. Our results coincide almost entirely with those of Chen et al. (2011). The final clean dataset to be used in further analysis is then created by binding the two subsets of solvent and insolvent firms and is saved as "data_clean".

Table 4. Three number summary (lower quantile, median, upper quantile) of the financial ratios for solvent and insolvent firms.

	Insolvent			Solvent		
Ratio	$q_{0.05}$	Median	$q_{0.95}$	$q_{0.05}$	Median	$q_{0.95}$
x1	-0.19	0.00	0.09	-0.09	0.02	0.19
x2	-0.15	0.00	0.06	-0.07	0.01	0.09
x3	-0.22	0.00	0.10	-0.11	0.03	0.27
x4	-0.16	0.00	0.06	-0.08	0.02	0.13
x5	-0.09	0.02	0.13	-0.09	0.05	0.27
x6	-0.13	0.07	0.21	-0.04	0.11	0.35
x7	-0.14	0.01	0.10	-0.07	0.02	0.14
x8	0.00	0.05	0.40	0.00	0.14	0.60
x9	-0.01	0.05	0.56	0.00	0.16	0.96
x10	0.18	0.52	0.91	0.09	0.42	0.88
x11	0.12	0.49	0.89	-0.05	0.36	0.83
x12	0.29	0.76	0.98	0.16	0.65	0.96
x13	0.00	0.21	0.61	0.00	0.15	0.59
x14	-7.75	1.05	7.19	-6.76	2.16	74.37
x15	0.00	0.02	0.16	0.00	0.03	0.32
x16	0.00	0.03	0.43	0.00	0.08	1.41
x17	0.18	0.68	1.88	0.24	0.94	4.55
x18	0.57	1.26	3.72	0.64	1.58	7.15
x19	-0.32	0.15	0.63	-0.22	0.25	0.73
x20	0.34	0.84	1.00	0.22	0.86	1.00
x21	0.24	0.61	2.31	0.16	0.48	2.01
x22	0.02	0.16	0.88	0.01	0.11	0.56
x23	0.02	0.12	0.33	0.00	0.09	0.25
x24	0.03	0.14	0.36	0.01	0.07	0.23
x25	13.01	14.87	17.16	12.82	15.41	17.95
x26	-1.20	0.00	0.74	-0.81	0.00	0.57
x27	-0.44	0.00	0.47	-0.53	0.00	0.94
x28	-12.17	0.00	0.94	-7.03	0.00	0.91

4. Evaluation of predictions

In this section we explain the steps followed in the "evaluate_predictions" quantlet, which can be found in the Appendix. Purpose of this quantlet is to create a function that will take labels (actual solvency status) and predictions about the solvency status of firms and will return the confusion matrix, the ROC curve and the AUC as well as some evaluation metrics like sensitivity, sensitivity, precision and accuracy.

The confusion matrix, also known as error matrix, is a matrix that allows visualization of the performance of an algorithm. Each row of the matrix

represents the instances in a predicted class while each column represents the instances in an actual class. As its name states, this matrix makes it easy to see if the system is confusing two classes, i.e. if it is commonly mislabelling one as another. It is a special kind of contingency table, with two dimensions ("actual" and "predicted"), and identical sets of "classes" (solvent and insolvent in our case) in both dimensions (wikipedia.org). The cells of our confusion matrix will present the number of:

- True Positives (TP), i.e. cases that firms predicted to be insolvent were indeed insolvent (hits)
- False Positives (FP), i.e. solvent firms were wrongly predicted to be insolvent (false alarm or Type I error)
- False Negatives (FN), i.e. insolvent firms were wrongly predicted to be solvent (mass or Type II error)
- True Negatives (TN), i.e. firms were correctly predicted to be solvent (correct rejection)

These values will then be used to get some important metrics which are described next.

Sensitivity and specificity are statistical measures of the performance of a binary classification test. Sensitivity (also known as the true positive rate) measures the proportion of positives (defaulted firms in our case) that are correctly identified as such. It is defined as $TPR = TP/(TP + FN)$. Specificity (or true negative rate) measures the proportion of negatives (solvent firms in our case) that are correctly identified as such. It is defined as $TNR = TN/(TN + FP)$. Precision is analogous to the positive predictive value (PPV) and is a measure of exactness. It is defined as $PPV = TP/(TP + FP)$. Finally accuracy measures the fraction of correct predictions and is defined as $(TP + TN)/(TP + TN + FP + FN)$.

The values of specificity and sensitivity will allow us to plot the ROC curve. The ROC curve (Receiver Operating Characteristic curve) is a graphical plot presenting the diagnostic ability of a binary classifier system as its discrimination threshold (in our case, probability threshold above which a firm is predicted to be insolvent) is varied. It is created by plotting the sensitivity values(y axis) against their corresponding $1 - specifity$ values (x axis) as the threshold varies. The area under the ROC curve (AUC) can be interpreted as the average power of the test on default or non-default corresponding to all discrimination thresholds (Härdle et al.(2012)). A larger AUC corresponds to a better classification result. A model with perfect

discriminative power will have an AUC value of 1, while a random model without discriminative power will have an AUC value of 0.5. Thus, any reasonable rating model will have an AUC value between 0.5 and 1.

The first function we created in this quantlet is the get_prediction() function (lines $19-22$). This function takes as inputs the fitted probabilities for insolvency and the threshold above of which a firm would be predicted to be defaulted. Using the ifelse() function in the body of this function, get_prediction() will return predictions for the solvency status of each firm. We construct then the evaluate_predictions() function (lines $24-47$). Inputs of this function are the actual solvency status of each company (labels), the corresponding predictions about the status and the "verbose" parameter which is equal to FALSE by default. In the body of the function we create first the previously mentioned confusion matrix which we expect to be a 2x2 matrix. If it is not the case, the test in line 30 will show us a warning message. Then we get the values of TP, TN, FP and FN and we compute the values of sensitivity, specificity, precision and accuracy, which are then saved as a list in "reports". These reports are then printed in a data frame if we change the logical value of "verbose" to TRUE. The two previously mentioned functions are used in the body of the final function that we describe in the next paragraph.

The final function created in this quantlet is the evaluate_model() function (lines $49-101$), which takes the fitted probabilities for bankruptcy and the actual status (labels) as inputs. At first we create a threshold list in the body of the function with threshold values varying from 0 to 1. Then we apply the get_prediction() function to this list in order to get a list of predictions for each threshold value which we store as "pred_list". Then we apply the evaluate_predictions() function to pred_list and we get a list of reports. We use then the values of sensitivity and specificity from reports in order to compute the ROC curve and the area under it (AUC) (lines $58-70$). For calculating the AUC we create the function get_auc() which takes $1-specificities$ and sensitivities as inputs. This function uses the trapezoidal method to approximate the AUC. For each difference in $1-specificities$ there is one rectangle which underestimates the area under curve (corresponding to the "left" value of sensitivities) and one rectangle that overestimates it (corresponding to the "right" value of sensitivities). Therefore the function approximates the AUC by taking the average of the two rectangles.

We are next interested in finding the values of sensitivities and specificities for which the distance between the ROC curve and the 45-degree line (i.e. the ROC curve of a model with no discriminative power) is maximized.

We use the index of these values in order to find the optimal values of sensitivity, specificity and the threshold, which is then used to compute the optimal predictions. These predictions are then used to create the optimal confusion matrix (lines $73 - 82$). In the final part of the evaluate_model() function the ROC curve is plotted and the function returns a list with the optimal measures calculated above.

Appendix

1. Variables of the Creditreform database

Column	Value
ID	ID of each company
T2	indicator of solvency status (solvent=0, insolvent=1)
Jahr	Year
VAR1	Cash and cash equivalents
VAR2	Inventories
VAR3	Current assets
VAR4	Tangible assets
VAR5	Intangible assets
VAR6	Total assets
VAR7	Accounts receivable
VAR29	Accounts receivable against affiliated companies
VAR8	Lands and buildings
VAR9	Equity (own funds)
VAR10	Shareholder loan
VAR11	Accrual for pension liabilities
VAR12	Total current liabilities
VAR13	Total longterm liabilities
VAR14	Bank debt
VAR15	Accounts payable
VAR30	Accounts payable against affiliated companies
VAR16	Sales
VAR17	Administrative expenses
VAR18	Amortization and depreciation
VAR19	Interest expenses
VAR20	Earnings before interest and taxes (EBIT)
VAR21	Operating income
VAR22	Net income
VAR23	Increase (decrease) in inventories
VAR24	Increase (decrease) in liabilities
VAR25	Increase (decrease) in cash
VAR26	Industry classification code
VAR27	Legal form
VAR28	Number of employees
Rechtskreis	Accounting principle
Abschlussart	Type of account