Towards ML-based Platforms in Finance Industry – An ML Approach to Generate Corporate Bankruptcy Probabilities based on Annual Financial Statements

Mustafa Pamuk
University of Goettingen, mustafa.pamuk@uni-goettingen.de

René Oliver Grendel
University of Goettingen, reneoliver.grendel@stud.uni-goettingen.de

Matthias Schumann
University of Goettingen, mschuma1@uni-goettingen.de

Towards ML-based Platforms in the Finance Industry – An ML Approach to Generate Corporate Bankruptcy Probabilities based on Annual Financial Statements

Full research paper

Mustafa Pamuk
Chair of Application Systems and E-Business
University of Goettingen
Goettingen, Germany
Email: mustafa.pamuk@uni-goettingen.de

René Oliver Grendel
Chair of Application Systems and E-Business
University of Goettingen
Goettingen, Germany
Email: reneoliver.grendel@stud.uni-goettingen.de

Matthias Schumann
Chair of Application Systems and E-Business
University of Goettingen
Goettingen, Germany
Email: mschuma1@uni-goettingen.de

Abstract

The increasing interest in Machine Learning (ML)-based services and the need for more intelligent and automated processes in the finance industry brings new challenges and requires practitioners and academics to design, develop, and maintain new ML approaches for financial services companies. The objective of this paper is to provide a standardized procedure to deal with cases that suffer from imbalanced datasets. We thus put forward design recommendations on how to test and combine multiple oversampling techniques such as SMOTE, SMOTE-ENN, and SMOTE-TOMEK on imbalanced datasets with multiple ML models and attribute-based structures to reach higher accuracies. Moreover, this paper considers ways of finding an appropriate structure while maintaining systems that work with periodically changing datasets, so that the incoming datasets can be analyzed regularly.

Keywords: Bankruptcy Forecasting, Machine Learning, Oversampling Techniques, Finance Industry, Annual Financial Statements

1 Introduction

In Germany, 16,300 bankruptcy records were registered in 2019 (Creditreform 2020). The risk of insolvency and the creditworthiness of companies are therefore of great interest to stakeholders such as banks, suppliers, and employees (Bauer and Agarwal 2014). As a result of the 2007–2009 financial crisis, the Basel Committee on Banking Regulation issued stricter capital and liquidity requirements for banks as part of the Basel III reform (BaFin 2017). Banks were also encouraged to develop internal models for evaluating loans to businesses and their associated risks (Ala'raj and Abbod 2016). These models calculate corporate bankruptcy probabilities and are used as part of the risk management process to quantify the risk of bankruptcy and the creditworthiness of a company (Huang 2009; Kwon et al. 2013).

Annual financial statements provide relevant information to calculate these risk measures (Obermann and Waack 2016). Financial ratios are subsequently derived for the forecasting models (Andrés et al. 2012). However, these data possess different distribution and quality depending on the size of the company (Andreeva et al. 2016; Ciampi 2015). Statistical methods such as Logistic Regression (LR) and Machine Learning (ML) ones such as Decision Trees (DT) and Neural Networks (NN) are used to automate the prediction process of bankruptcy probabilities (Pai et al. 2015). Newer approaches, such as ensemble methods, combine several homogeneous and heterogeneous classification algorithms.

No approach has been developed to compare multiple models and oversampling techniques on imbalanced datasets (Lessmann et al. 2015). The need for standardization of ML approaches increases when the features that are used in periodic datasets to train ML models correlate with each other. These kinds of datasets must be analyzed regularly with an appropriate structure. In this study, we use an imbalanced dataset that makes the development as well as the training and maintenance of ML-based platforms even more challenging. We thus propose an attribute-based ML approach that can aid the elaboration of further design, development, and maintenance concepts regarding imbalanced datasets with oversampling techniques in the finance industry.

The rest of the paper is organized as follows. Section 2 provides an overview of the relevant literature and defines the key metrics for forecasting bankruptcy. The approach presented in Section 3 outlines each step of the construction and development of an attribute-based ML approach. In Section 4, we discuss the results of the prototype and how the development and maintenance of such prototypes can be automated. Other related works and possible future directions are presented in Section 5.

2 Basics

In this section, we first analyze the existing literature on ML in corporate bankruptcy forecasting. We then define the key metrics for financial balance sheet analysis that are used to forecast bankruptcy probabilities.

2.1 Machine Learning in Corporate Bankruptcy Forecasting

We conducted a systematic literature review to identify and analyze the best ML techniques for the three main steps (data preprocessing, modeling, and evaluation) in corporate bankruptcy forecasting based on the procedure presented in the articles by Fettke (2006), Hobert (2018), J. Brocke et al. (2009), and Webster and Watson (2002). A total of 60 publications dealing with corporate bankruptcy forecasting from 2008 to 2020 were identified. We selected the preprocessing, modeling, and evaluation techniques that have achieved the best results in previous studies to design a standardized process model.

Regarding data preprocessing, 63% of the publications did not mention a method for data cleaning. To construct their datasets for forecasting corporate bankruptcies, one-third of the publications used the paired matching method. In previous publications, the relevant and most widely used variables were selected through a literature review (Bai and Tian 2020; Lin et al. 2019; Tian et al. 2015). For example, in the first step, Liang et al. (2020) identified relevant variables based on a literature review, then used a stepwise discriminant analysis in the second step to select the variables. Only normalization or standardization was performed as variable transformation techniques.

In the literature, NN, LR models, Support Vector Machine (SVM), and DT are the most commonly used modeling techniques for insolvency forecasting. One possible reason is that these four ML algorithms provide reference methods for evaluation. These algorithms are also different in terms of their interpretability (e.g., white-box and black-box models) and their complexity. Furthermore, to predict bankruptcies, recent publications have predominantly applied ensemble methods and frameworks such as XGBoost. For instance, Zięba et al. (2016) compared a set of learning algorithms such as SVM, NN, DT, and LR with ensemble methods. These methods (e.g., XGBoost) achieve significantly better accuracy (almost 95%) than decision trees with the J48 algorithm (71.7%) and classical statistical methods such as LR (62%) (Zięba et al. 2016).

In summary, a large number of algorithms are used in the area of bankruptcy forecasting. The four most commonly used algorithms are NN, SVM, LR, and DT. Additionally, ensemble methods have been recently used more frequently and, according to Zięba et al. (2016), can achieve in some cases better results than the standard algorithms. However, ensemble methods must be viewed critically in terms of their interpretability. Against this background, metrics such as the Brier Score and ROC-AUC provide more accurate insights into the quality of the forecast. The Brier Score is used to quantify the deviation of the predicted insolvency probability from the true insolvency probability based on the squared deviation (Brier 1950). Likewise, ROC-AUC helps to assess the performance of classifications and represents the degree or measure of separability. As a result, we identified five widely used ML techniques (NN, DT, LR, XGBoost, and SVM) as having achieved the best results in corporate bankruptcy forecasting.

2.2 The Key Metrics of Financial Balance Sheet Analysis

In this section, we offer an overview of the key metrics used in the literature to forecast bankruptcy with ML that provide an insight into the financial structure and development of a company (Lachnit and Müller 2017). The literature review allowed us to identify the key metrics of financial balance sheet analysis to determine the relevant features used for bankruptcy forecasting. These financial metrics permit an understanding of the basic structure needed to periodically evaluate the company's business performance (Coenenberg 2016).

Based on historical data, the probability of bankruptcy quantifies the possibility of a company being declared bankrupt as a probability between zero and one. Financial balance sheet analysis distinguishes between ratios relating to capital structure, asset structure, and ratios relating to liquidity analysis, thus recognizing the relationship between capital utilization and capital raising (Coenenberg 2016). The metrics describing the asset structure provide information about the flexibility and liquidity of a company's assets (Coenenberg et al. 2021). The ratios used in the analysis of capital structure provide insight into the structure of financing in terms of maturity and composition (Coenenberg 2016). The key indicator of capital structure is the equity ratio (Lachnit and Müller 2017). This ratio measures the amount of leverage used by a company and determines how well a company manages its debts and funds its asset requirements (Coenenberg 2016).

$$(1)\ Equity\ Ratio\ (ER) = \ Total\ Equity\ /\ Total\ Assets$$

$$(2)\ Short\text{-}term\ Debt\ Ratio\ (STDR) = Short\text{-}term\ Liabilities\ /\ Total\ Assets$$

$$(3)\ Working\ Capital\ Ratio\ (WCR) = Working\ Capital\ /\ Total\ Assets$$

The next important ratio is the *Short-term Debt Ratio* (STDR), which indicates the likelihood that a company will be able to deliver payments on its outstanding short-term liabilities. Short-term debt includes bonds as well as liabilities with a residual maturity of less than a year (Peemöller 2013). The *Working Capital Ratio* (WCR) is the ratio of working capital to total assets. The *Return on Total Assets* (RTA) considers the success of a company with no regard to the origin of the capital employed (Coenenberg 2016). The *Return on Equity* (ROE) describes the return on the capital invested by the owners.

$$(4)\ Return\ on\ Total\ Assets\ (RTA) = (Net\ Income + Borrowing\ Costs)\ /\ Total\ Assets$$

$$(5)\ Return\ on\ Equity\ (ROE) = \ Net\ Income\ /\ Total\ Assets$$

$$(6)\ Asset\ Coverage\ Ratio\ (ACR) = \ Total\ Equity\ /\ Fixed\ Assets$$

$$(7)\ Second\text{-}Degree\ Liquidity\ (L2) = \ Monetary\ working\ capital\ /\ Short\text{-}term\ Debt$$

The liquidity analysis based on the asset and liability items, which can be taken from the balance sheet, attempts to assess the timely adherence to payment obligations (Brösel 2014). There is a distinction here between short-term liquidity ratios and medium- to long-term coverage ratios (Coenenberg 2016). Furthermore, *Asset Coverage Ratio* (ACR) is a long-term ratio that considers the degree of asset coverage by fixed assets. *Second-degree Liquidity* (L2) indicates whether current liabilities can be covered by cash and cash equivalents and current receivables.

3 Research Approach

The Cross-Industry Standard Process for Data Mining (CRISP-DM) provides the basis for this analysis. The CRISP-DM reference model consists of six phases (Shearer 2000). In this section, we consider those of Data Understanding (3.1), Data Preprocessing (3.2 and 3.3), Modeling (3.4), and Evaluation (3.5).

3.1 Context and Data

The dataset for this paper contains financial reports of German companies from 2000 to 2012. It consists of 3,309,007 entries with 74 measured variables; it also contains 2,040 insolvent firms. According to El Kalak and Hudson (2016) and Ciampi (2015), small- and medium-sized companies provide less and mostly unprecise information as part of their disclosures. Thus, the prediction of bankruptcy probabilities should be modeled divided by business size. Table 2 gives an overview of the division of (insolvent) companies in the dataset according to business size. The four categories based on assets (micro, small, medium, and large) follow the German Commercial Code (§267 and §267a).

Before data preprocessing, the analysis revealed that 95% of observations in the entire dataset, and almost 97% among insolvent firms, belonged to micro- and small-sized enterprises. Based on the findings of Ciampi (2015) and El Kalak and Hudson (2016), the dataset was then split according to micro, small, and medium sizes. Large-sized businesses were ignored because the dataset contained only 13 cases of corporate bankruptcies involving such businesses.

3.2 Data Preprocessing

To provide ML approaches with the necessary structure and increase data quality, preprocessing consists of four main phases: data cleaning, sampling, feature selection, and feature transformation. First, all the features with more than 30% of missing values were deleted, because methods such as imputation with the median or the arithmetic mean would have distorted the data. The feature *"insolvency date"* is an exception here, because it is necessary for determining solvency. As a result, 50 features were removed from the dataset. Second, features that contained irrelevant information (e.g., IDs and postal codes) were deleted. The remaining 13 variables were analyzed to identify the relevant features for generating corporate bankruptcy probabilities. Six features (ER, STDR, WCR, ACR, Liquidity Ratio, and L2; see Table 1) exhibited a clear positive or negative trend in terms of their value as the credit rating level got worse. We considered the correlation among these features since a high correlation can lead to prediction problems (Kim and Kang 2012). This resulted in a Pearson correlation between *Liquidity Ratio* and *Second-Degree Liquidity* (L2). The variable L2 was considered as a feature in the dataset. Furthermore, since the calculation method for the variables ACR and WCR (see Table 1) could not be reconstructed from the data, these variables were recalculated.

Metric Types	**Names**	
Capital Structure	**ER**	(1) Equity Ratio
	STDR	(2) Short-term Debt Ratio
	WCR	(3) Working Capital Ratio
Profitability	***RTA***	(4) Return on Total Assets
	ROE	(5) Return on Equity
Liquidity	***ACR***	(6) Asset Coverage Ratio
	L2	(7) Second Degree Liquidity

Table 1. Key Metrics to Generate Corporate Bankruptcy Probabilities

In the second part of the data cleaning process, all the companies with missing values in key metrics (Table 1) were deleted from the dataset. Moreover, companies that had negative values in the variables *Assets*, *Asset and Working Capital Intensity*, and *STDR* were deleted from the dataset. Firms with invalid values or manual entries for rating levels (e.g., "don't know") were likewise removed. The recalculation of four metrics of liquidity, capital structure, and profit ratios (ACR, ROE, RTA, and WCR) was conducted based on the work of Brodag (2010) and Du Jardin (2016). Moreover, the target variable *"bankrupt"* was created for all the observations based on bankruptcy dates to provide a relevant structure for the ML models. Thus, a value of one indicates an insolvent company and zero a solvent company. In the final cleaning step, the dataset was cleaned from the values that arose from these calculations but whose result was undefined or could not be represented accurately.

	Raw Data				**After Data Cleaning**			
Business Size (asset size)	**All**	**%**	**Number of Bankruptcies**	**%**	**All**	**%**	**Number of Bankruptcies**	**%**
Micro *< 350k €*	2.087.867	63,10%	1144	56,08%	1.443.739	56,28%	875	51,68%
Small *350k € - 6 Mio. €*	1.077.832	32,57%	834	40,88%	1.028.480	40,09%	769	45,42%
Medium *6 Mio. € - 20 Mio. €*	97.253	2,94%	49	2,40%	93.044	3,63%	49	2,89%
Large *> 20 Mio. €*	46.055	1,39%	13	0,64%	-	-	-	-
Σ	**3.309.007**		**2.040**		**2.565.263**		**1.693**	

Table 2. Raw and Preprocessed Dataset

After cleaning, the dataset contained a total of 2,565,263 data points, including 1,693 bankrupt companies. The analysis of data preprocessing revealed that observations of micro-sized companies were almost the only ones to be deleted (see Table 2). This supports the findings of (Ciampi 2015) that micro-sized companies have lower data quality. Nevertheless, the percentage of the three categories in the dataset remained almost constant. Furthermore, a comparison of the quantiles of the features used (divided by company size) showed that these features had different values.

3.3 Oversampling Structure

Based on the arguments of Ciampi (2015) and El Kalak and Hudson (2016), the training dataset was split into three training sets (“micro”, “small”, and “medium”) using the attribute “business size”. The purpose of this split was to improve the prediction quality of the ML models. After performing this step, there was still a strong imbalance between insolvent and solvent companies. This was due to the fact that the bankrupt companies in the dataset were significantly fewer than the solvent ones. This problem is also known as the class imbalance problem, which causes major challenges when using classification algorithms (Batista et al. 2004).

To correct the class imbalance in each of the three training datasets, the oversampling method Synthetic Minority Oversampling Technique (SMOTE) and its combination with the cleaning algorithms Tomek Links (SMOTE-TOMEK) and Edited Nearest Neighbor (SMOTE-ENN) were used (Batista et al. 2004; Chawla et al. 2002). With SMOTE, the minority class (bankruptcies) was oversampled by synthetically generated observations (Chawla et al. 2002). The k next bankrupt companies were thus determined and some or all of them were connected with a line (Chawla et al. 2002). The synthetic bankrupt companies were then generated along this line. To remove the overlap created by the SMOTE algorithm, the data cleaning techniques Tomek Links and ENN were used in combination with SMOTE. Tomek Links are based on the calculation of the distance between two observations of different classes (He and Garcia 2009). A Tomek Link exists if no observation of the same class with a shorter distance is present in the dataset for these two observations. The identified Tomek Links were deleted from the dataset.

Data cleaning with ENN removed all the observations whose class did not match the class of k nearest (Batista et al. 2004). In this paper, both SMOTE-TOMEK and SMOTE-ENN only cleared observations of the majority class. The default value of k = 5 nearest neighbors was considered for the SMOTE algorithm. After oversampling, the ratio between the minority and majority class was two-to-one. In SMOTE-ENN, the k = 5 nearest neighbors were considered while deleting observations of the majority class. As a result, these three oversampling techniques were used to generate synthetic observations for each of the three datasets (“micro”, “small”, and “medium”).

3.4 Model Selection

In section 2.1 we identified the following ML techniques as having achieved the best results for forecasting bankruptcy: LR, NN, DT, XGBoost, and SVM. However, SVM was not used in this study because it has a runtime complexity between n^2 (best case) and n^3 (worst case). Thus, SVMs are not appropriate for datasets as large as the one being considered here (Bottou 2007). Ten-fold stratified cross-validation was used to train and test the models of LR, DT, and XGBoost (Olson et al. 2012). Each of the three datasets was split into ten disjoint test and training datasets (Liang et al. 2015). The stratified ten-fold cross-validation ensured that the class frequencies in these datasets were more balanced. For more complex methods using NN, we chose to split the data into training, testing, and validation datasets. In this case, 60% of the observations were in the training dataset, 20% in the test dataset, and 20% in the validation dataset.

Following this division, a standardization process was applied to all the features to obtain a mean of zero and a standard deviation of one. This ensured that the features did not dominate the learning process due to their scale. Doing so allows the weights in algorithms such as NN to be updated faster (Raschka 2014). Metrics based on the confusion matrix (Accuracy, Recall, ROC-AUC, F1, and Brier Score) were then used to evaluate the prediction quality of the ML models.

Neural Network		XGBoost		Logistic Regression		Decision Tree	
Parameter	**Values**	**Parameter**	**Values**	**Parameter**	**Values**	**Parameter**	**Values**
1st Layer Number of Neurons	50	Base Score	0.5	Inverse of Regularization Strength	2	CCP Alpha	0.0
1st and 2nd Layer Activation Function	Rectified Linear Unit	Learning Rate	0.300012	Maximum number of iterations	100	Criterion	Gini
2nd Layer Number of Neurons	20	Max Depth	6	Penalty	l2	Max Depth	5
3rd Layer Number of Neurons	1	Number of Gradient Boosted Trees	100	Solver	lbfgs	Max Features	7
3rd Layer Activation Function	Sigmoid			Tolerance	0.0001	Min Samples Leaf	1
Optimizer	RMSProp	Reg_alpha	0			Splitter	best
Learning Rate	0.001						

Table 3. Hyperparameters Considered for NN, XGBoost, LR, and DT

To achieve a better prediction, a grid search was performed for the models LR and DT to determine the optimal hyperparameters (Bergstra et al. 2011). Furthermore, the hyperband algorithm was applied to optimize the number of neurons in the first layer of the NN and the learning rate (Li et al. 2018). In the validation dataset, configurations with a lower accuracy were eliminated through an iterative process to determine the best configuration for the hyperparameters (Li et al. 2018). To avoid overfitting, early stopping was applied to the NN based on the validation error (Montavon et al. 2012). The training of the NN was stopped if the validation error did not improve after k = 3 epochs. An overview of the hyperparameters used for NN, XGBoost, LR, and DT is shown in Table 3.

3.5 Evaluation of the Attribute-based ML Approach

Focusing on a standardized bankruptcy forecasting method, a design-oriented research approach was followed. A software prototype was thus constructed to derive findings for research and business purposes (Österle et al. 2010). The core idea of this prototype was the splitting of the underlying dataset based on the business size of each company. Figure 1 provides an overview of the approach that integrates the steps outlined above.

The models were trained and tested for each dataset (i.e., the three oversampling techniques SMOTE, SMOTE-TOMEK, and SMOTE-ENN were independently applied to each dataset). Furthermore, the learning algorithms were applied to each of the resulting datasets. Following a process of iteration, SMOTE was implemented first for oversampling. SMOTE-TOMEK and SMOTE-ENN were then used to evaluate the influence of algorithms that combine oversampling and data cleaning. To establish comparability between the different oversampling techniques, the same number of neurons was used in the first layer for all NN models. However, depending on the configuration and the size of the company, tests with other configurations showed a partially better prediction performance for SMOTE (see Table 4). These results indicate that this approach requires hyperparameter tuning for each firm size and model.

XGBoost provided the best forecasting quality. The forecast quality also improved with fewer observations in the oversampled dataset. The results of the LR model, though, did not show this trend. Based on recall (zero), the LR model was unable to correctly forecast bankrupt firms. Because too high a correlation between the features can lead to problems in the regression, we examined the correlation before and after oversampling. An especially strong negative correlation (-0.97) was found in the "micro" dataset between ER and STDR. However, no significant difference in the correlation could be found before and after oversampling. Following this correlation, different combinations of features were tested for the LR model, especially to improve its recall. Removing the variables STDR and L2, as well as WCR and L2, produced no improvement. Removing only the L2 variable also did not improve the model's ability to identify bankrupt companies. Moreover, the forecasts were evaluated to investigate the effect of the oversampling techniques SMOTE-TOMEK and SMOTE-ENN on the prediction of the ML models. The data cleaning process based on Tomek Links did not have a strong positive effect on the results. The LR model especially continued to suffer from poor recall.

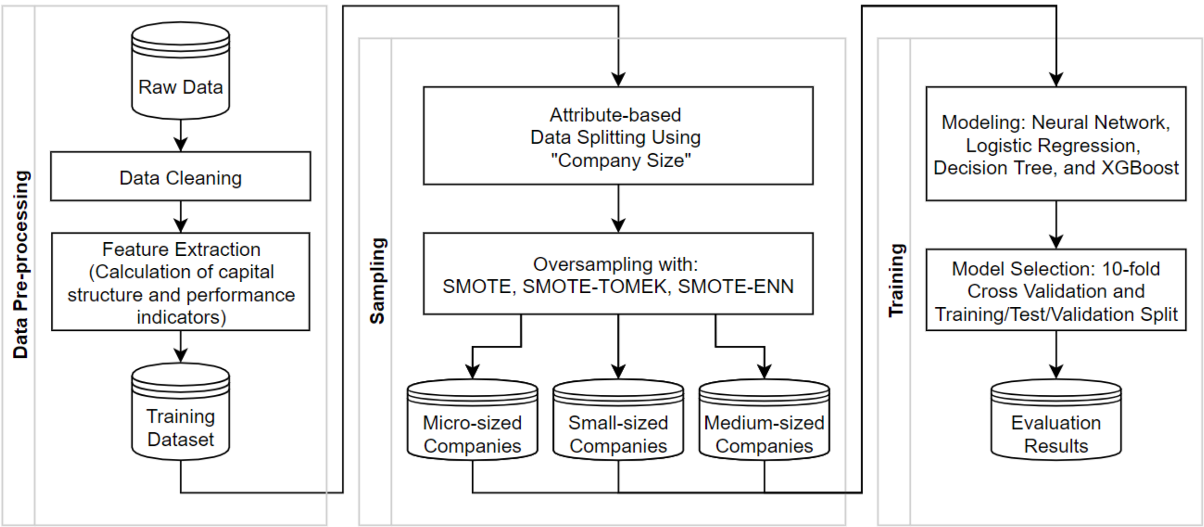
Data Pre-processing
Raw Data
Data Cleaning
Feature Extraction (Calculation of capital structure and performance indicators)
Training Dataset
Sampling
Attribute-based Data Splitting Using "Company Size"
Oversampling with: SMOTE, SMOTE-TOMEK, SMOTE-ENN
Micro-sized Companies
Small-sized Companies
Medium-sized Companies
Training
Modeling: Neural Network, Logistic Regression, Decision Tree, and XGBoost
Model Selection: 10-fold Cross Validation and Training/Test/Validation Split
Evaluation Results

Figure 1. Attribute-based Machine Learning Approach

Except for the LR, all the models benefited from data cleaning with SMOTE-ENN instead of simply oversampling with SMOTE. Furthermore, as Table 4 shows, SMOTE-ENN provided the best prediction quality for the DT. To improve the recall of the LR model, another feature, the *Provisioning Rate*, was calculated and combined with the best performing oversampling method (SMOTE-ENN). Thus, XGBoost (highlighted in bold in Table 4) achieved an accuracy of almost 90% or more in all three datasets; the predicted probabilities also hardly deviated from the true values, especially in the "medium" dataset (see Brier Score). The stratified cross-validation ensured that each entry was used for validation just once and helped examine the importance of each feature in the original dataset within the model (see Table 6). Hence, the attribute-based ML approach easily aided training and evaluation. Moreover, the NN produced significantly better results in the case of small- and medium-sized firms, whereas the DT and the LR model did not show improvements in prediction quality.

SMOTE												
	NN			**XGBoost**			**LR**			**DT**		
Company Sizes	**Mic**	**Sm**	**Med**	**Mic**	**Sm**	**Med**	**Mic**	**Sm**	**Med**	**Mic**	**Sm**	**Med**
Accuracy	0,718	0,788	0,828	0,876	0,895	0,982	0,667	0,709	0,651	0,716	0,760	0,818
Brier Score	0,189	0,144	0,115	0,093	0,080	0,016	0,215	0,190	0,213	0,189	0,160	0,129
ROC-AUC	0,734	0,856	0,911	0,947	0,959	0,998	0,647	0,751	0,651	0,729	0,816	0,875
F1 Score	0,469	0,653	0,743	0,811	0,841	0,972	0,000	0,325	0,036	0,423	0,638	0,741
Recall	0,374	0,600	0,744	0,797	0,833	0,986	0,000	0,210	0,020	0,313	0,634	0,781
SMOTE-TOMEK												
Accuracy	0,712	0,787	0,816	0,877	0,893	0,982	0,666	0,71	0,651	0,716	0,76	0,823
Brier Score	0,19	0,146	0,125	0,092	0,081	0,015	0,215	0,189	0,213	0,189	0,16	0,129
ROC-AUC	0,745	0,853	0,894	0,948	0,958	0,998	0,655	0,751	0,651	0,726	0,816	0,876
F1 Score	0,387	0,655	0,712	0,813	0,838	0,974	0	0,327	0,036	0,456	0,638	0,748
Recall	0,273	0,608	0,685	0,798	0,829	0,987	0	0,212	0,02	0,356	0,634	0,789

SMOTE-ENN												
Company Sizes	**Mic**	**Sm**	**Med**	**Mic**	**Sm**	**Med**	**Mic**	**Sm**	**Med**	**Mic**	**Sm**	**Med**
Accuracy	0,718	0,799	0,862	0,887	0,905	0,990	0,652	0,710	0,637	0,719	0,765	0,829
Brier Score	0,188	0,138	0,105	0,086	0,073	0,009	0,218	0,188	0,215	0,189	0,158	0,127
ROC-AUC	0,752	0,873	0,924	0,956	0,967	0,999	0,655	0,762	0,652	0,734	0,824	0,881
F1 Score	0,469	0,681	0,795	0,836	0,862	0,986	0,000	0,374	0,037	0,459	0,652	0,756
Recall	0,358	0,620	0,778	0,826	0,856	0,993	0,000	0,251	0,020	0,344	0,637	0,774
SMOTE-ENN (with PR)												
Accuracy	0,716	0,806	0,898	**0,896**	**0,915**	**0,994**	0,653	0,712	0,638	0,718	0,765	0,818
Brier Score	0,189	0,13	0,073	**0,081**	**0,066**	**0,005**	0,218	0,185	0,215	0,189	0,158	0,127
ROC-AUC	0,742	0,894	0,965	**0,962**	**0,973**	**0,999**	0,654	0,765	0,651	0,731	0,822	0,881
F1 Score	0,491	0,673	0,853	**0,850**	**0,876**	**0,991**	0	0,393	0,037	0,449	0,612	0,716
Recall	0,394	0,58	0,857	**0,848**	**0,877**	**0,999**	0	0,271	0,02	0,332	0,538	0,671

Table 4. Evaluation Results of each ML Model and Oversampling Technique (Mic: Micro-sized; Sm: Small-sized; Med: Medium-sized)

Finally, a weighting approach was applied to the LR model that assigned a stronger weight to individual observations while training. Therefore, each observation of the minority class was adjusted in terms of its weight so that the algorithm perceived an equal ratio of bankrupt to solvent companies. Hence, the prediction of the LR model achieved a very high recall of 0.90 in the "micro" dataset. However, this had a negative impact on the other metrics, such as accuracy or the Brier Score. This raises the question of which of the metrics should be considered as the most important one in the context of forecasting bankruptcy probabilities.

Weighted LR			
Company Sizes	**Mic**	**Sm**	**Med**
Accuracy	0,522	0,684	0,584
Brier Score	0,241	0,202	0,235
ROC-AUC	0,653	0,764	0,654
F1 Score	0,567	0,610	0,520
Recall	0,900	0,717	0,657

Table 5. Evaluation Results of Weighted LR

XGBoost & SMOTE-ENN (with PR)			
Company Sizes	**Mic**	**Sm**	**Med**
Accuracy	0,887	0,905	0,990
Brier Score	0,086	0,073	0,009
ROC-AUC	0,956	0,967	0,999
F1 Score	0,836	0,862	0,986
Recall	0,826	0,855	0,993

Table 6: Validation Results for the best ML Model

4 Discussion

The performance of the developed prototype indicates that the attribute-based approaches and selective oversampling on imbalanced datasets provide a solid infrastructure for ML-based platforms in the finance industry. Beginning with the preprocessing, the attribute "business size" was the key indicator for this study. Based on this approach, it is possible to balance the datasets iteratively. Furthermore, this approach enables to undertake additional operations (e.g., for micro-sized companies) to increase accuracy in highly imbalanced datasets. Within the stratified cross-validation, it ensures that the original training dataset is used both for training and validation. Moreover, the results of the forecasting show that the AI methods used, especially the ensemble one XGBoost, achieve very good outcomes for all the sampling methods. XGBoost achieves a forecast quality of 87% and increased accuracy for each of the three datasets. This satisfying prediction quality is also in line with the results of the benchmarking study conducted in Lessmann et al. (2015), where ensemble methods also achieved better forecasting quality.

Splitting the dataset based on the explanatory variable "firm size", and training and testing the AI models for each of the three datasets, provide several findings. First of all, the argument made by (Ciampi 2015) that smaller firms have poor data quality is confirmed based on the results of the forecasting and the exploratory data analysis. For each of the four classification algorithms used, the prediction performance for the "micro" dataset is the worst. Furthermore, data cleaning with ENN and Tomek Links removed a much higher percentage of observations from the "micro" dataset than from the "small" and "medium" ones. Consequently, further cleaning of the dataset and selection of explanatory variables are necessary for these companies to provide accurate forecasts, especially for more complex problems such as rating class forecasting.

The attribute-based ML approach (see Figure 1) and the prototype were developed based on a constant dataset. However, in a productive environment, existing datasets change annually. The items in the financial statements and the key metrics derived from them are formed based on financial accounting processes. Legal reforms, such as the German Accounting Law Modernization Act (i.e., Bilanzmodernisierungsgesetz), lead to a decreased comparability among different years due to new selection rights for certain items on the balance sheet. These two issues generate the need for frequently optimizing the models and their hyperparameters and selectively determining the training dataset. As a result, it is necessary to investigate the development of a standardized process for maintaining these attribute-based ML systems because the numerous models they contain need to be iteratively trained, evaluated, and supported.

5 Conclusions and Future Work

This paper focuses on techniques for data preprocessing, modeling, and evaluation based on the CRISP-DM process. The key aspect here is to split and adjust the imbalanced dataset based on the attribute “business size”. This helps to increase the model’s accuracy based on imbalanced parts of the dataset. Likewise, it provides an easy comparison between multiple ML models and datasets. This approach can be customized or extended for similar cases that contain imbalanced datasets, such as credit scoring. Following this approach, it is easy to identify inconsistent structures and work on them by dividing the dataset properly. The detected patterns on the datasets need to be monitored and maintained regularly due to periodic changes. However, this maintenance can be undertaken using attributes for specific groups (e.g., micro-sized companies) regardless of the other ones in the dataset.

DT, LR, and NN are the most commonly used ML models. In recent research, these are combined with ensemble methods. This paper uses different oversampling algorithms to balance the distribution of the number of bankrupt and solvent firms. The results indicate that ensemble methods such as XGBoost, and more complex methods such as NN, provide the best outcomes in combination with SMOTE-ENN. Both algorithms achieve an accuracy of 72%–99% in the prediction of bankruptcy probabilities. These findings show that the approach presented here is appropriate. Thus, the concept offers a basis for information systems in practice. However, the higher number of parameters and models entailed by this approach, which increases training time, must be considered critically. The constantly changing nature of the dataset, due to legal regulations, means there is an increased maintenance effort.

The division based on company size must also be viewed critically, especially in terms of the complexity of the results and the maintenance of the developed artifact. Therefore, it is sensible to establish a dashboard that can extract the text-based results and automatically visualize them in the form of various graphics, such as bar charts and boxplots. Previous publications have primarily focused on the development of ML models or simple procedures. For this reason, our future research direction includes the problem of maintenance in ML-based platforms for the finance industry.

6 References

Ala'raj, M., and Abbod, M. F. 2016. “A new hybrid ensemble credit scoring model based on classifiers consensus system approach,” *Expert Systems with Applications* (64), pp. 36-55 (doi: 10.1016/j.eswa.2016.07.017).

Andreeva, G., Calabrese, R., and Osmetti, S. A. 2016. “A comparative analysis of the UK and Italian small businesses using Generalised Extreme Value models,” *European Journal of Operational Research* (249:2), pp. 506-516 (doi: 10.1016/j.ejor.2015.07.062).

Andrés, J. de, Landajo, M., and Lorca, P. 2012. “Bankruptcy prediction models based on multinorm analysis: An alternative to accounting ratios,” *Knowledge-Based Systems* (30), pp. 67-77 (doi: 10.1016/j.knosys.2011.11.005).

BaFin 2017. *Jahresbericht der Bundesanstalt für Finanzdienstleistungsaufsicht*.

Bai, Q., and Tian, S. 2020. “Innovate or die: Corporate innovation and bankruptcy forecasts,” *Journal of Empirical Finance* (59), pp. 88-108 (doi: 10.1016/j.jempfin.2020.09.002).

Batista, G. E. A. P. A., Prati, R. C., and Monard, M. C. 2004. "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter* (6:1), pp. 20-29 (doi: 10.1145/1007730.1007735).

Bauer, J., and Agarwal, V. 2014. "Are hazard models superior to traditional bankruptcy prediction approaches? A comprehensive test," *Journal of Banking & Finance* (40), pp. 432-442 (doi: 10.1016/j.jbankfin.2013.12.013).

Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. 2011. "Algorithms for Hyper-Parameter Optimization," in *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira and K. Q. Weinberger (eds.), Curran Associates, Inc.

Bottou, L. 2007. *Large-scale kernel machines*, Cambridge, Mass., London: MIT Press.

Brier, G. W. 1950. "Verification of forecasts expressed in terms of probability," *Monthly Weather Review* (78:1), pp. 1-3 (doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2).

Brodag, T. 2010. *PAC-Lernen zur Insolvenzerkennung und Hotspot-Identifikation*: *Anwendung statistischer Modelle des algorithmischen Lernens auf betriebswirtschaftliche und bioinformatische Probleme der Praxis*, Saarbrücken: Suedwestdeutscher Verlag fuer Hochschulschriften.

Brösel, G. 2014. *Bilanzanalyse*: *Unternehmensbeurteilung auf der Basis von HGB- und IFRS-Abschlüssen*, Berlin: Erich Schmidt.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. 2002. "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research* (16), pp. 321-357 (doi: 10.1613/jair.953).

Ciampi, F. 2015. "Corporate governance characteristics and default prediction modeling for small enterprises. An empirical analysis of Italian firms," *Journal of Business Research* (68:5), pp. 1012-1025 (doi: 10.1016/j.jbusres.2014.10.003).

Coenenberg, A. G. 2016. *Jahresabschluss und Jahresabschlussanalyse*: *Betriebswirtschaftliche, handelsrechtliche, steuerrechtliche und internationale Grundlagen - HGB, IAS/IFRS, US-GAAP, DRS*, Stuttgart: Schäffer-Poeschel Verlag für Wirtschaft . Steuern . Recht Gm.

Coenenberg, A. G., Haller, A., Mattner, G., and Schultze, W. 2021. *Einführung in das Rechnungswesen*: *Grundlagen der Buchführung und Bilanzierung*, Stuttgart: Schäffer-Poeschel.

Creditreform 2020. *Insolvenzen in Deutschland*. https://www.creditreform.de/aktuelles-wissen/pressemeldungen-fachbeitraege/news-details/show/insolvenzen-in-deutschland-jahr-2020. Accessed 26 April 2021.

Du Jardin, P. 2016. "A two-stage classification technique for bankruptcy prediction," *European Journal of Operational Research* (254:1), pp. 236-252 (doi: 10.1016/j.ejor.2016.03.008).

El Kalak, I., and Hudson, R. 2016. "The effect of size on the failure probabilities of SMEs: An empirical study on the US market using discrete hazard model," *International Review of Financial Analysis* (43), pp. 135-145 (doi: 10.1016/j.irfa.2015.11.009).

Fettke, P. 2006. "State-of-the-Art des State-of-the-Art," *WIRTSCHAFTSINFORMATIK* (48:4) (doi: 10.1007/s11576-006-0057-3).

He, H., and Garcia, E. A. 2009. "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering* (21:9), pp. 1263-1284 (doi: 10.1109/TKDE.2008.239).

Hobert, S. 2018. *Empirische Erkenntnisse und Gestaltungsansätze zum Einsatz von Wearable Computern im Industriesektor*, Göttingen: Cuvillier Verlag.

Huang, S.-C. 2009. "Integrating nonlinear graph based dimensionality reduction schemes with SVMs for credit rating forecasting," *Expert Systems with Applications* (36:4), pp. 7515-7518 (doi: 10.1016/j.eswa.2008.09.047).

J. Brocke, A. Simons, Björn Niehaves, K. Riemer, Ralf Plattfaut, and A. Cleven 2009. "Reconstructing the giant: On the importance of rigour in documenting the literature search process," in *ECIS*.

Kim, M.-J., and Kang, D.-K. 2012. "Classifiers selection in ensembles using genetic algorithms for bankruptcy prediction," *Expert Systems with Applications* (39:10), pp. 9308-9314 (doi: 10.1016/j.eswa.2012.02.072).

Kwon, J., Choi, K., and Suh, Y. 2013. "Double Ensemble Approaches to Predicting Firms' Credit Rating," in *17th Pacific Asia Conference on Information Systems, PACIS 2013, Jeju Island, Korea, June 18-22, 2013*, J.-N. Lee, J.-Y. Mao and J. Y. L. Thong (eds.), p. 158.

Lachnit, L., and Müller, S. 2017. *Bilanzanalyse*, Wiesbaden: Springer Fachmedien Wiesbaden.

Lessmann, S., Baesens, B., Seow, H.-V., and Thomas, L. C. 2015. "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *European Journal of Operational Research* (247:1), pp. 124-136 (doi: 10.1016/j.ejor.2015.05.030).

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. 2018. "Hyperband: A novel bandit-based approach to hyperparameter optimization," *Journal of Machine Learning Research* (18), pp. 1-52.

Liang, D., Tsai, C.-F., Lu, H.-Y., and Chang, L.-S. 2020. "Combining corporate governance indicators with stacking ensembles for financial distress prediction," *Journal of Business Research* (120), pp. 137-146 (doi: 10.1016/j.jbusres.2020.07.052).

Liang, D., Tsai, C.-F., and Wu, H.-T. 2015. "The effect of feature selection on financial distress prediction," *Knowledge-Based Systems* (73), pp. 289-297 (doi: 10.1016/j.knosys.2014.10.010).

Lin, W.-C., Lu, Y.-H., and Tsai, C.-F. 2019. "Feature selection in single and ensemble learning-based bankruptcy prediction models," *Expert Systems* (36:1), e12335 (doi: 10.1111/exsy.12335).

Montavon, G., Orr, G. B., and Müller, K.-R. (eds.) 2012. *Neural Networks: Tricks of the Trade*, Berlin, Heidelberg: Springer Berlin Heidelberg.

Obermann, L., and Waack, S. 2016. "Interpretable Multiclass Models for Corporate Credit Rating Capable of Expressing Doubt," *Frontiers in Applied Mathematics and Statistics* (2) (doi: 10.3389/fams.2016.00016).

Olson, D. L., Delen, D., and Meng, Y. 2012. "Comparative analysis of data mining methods for bankruptcy prediction," *Decision Support Systems* (52:2), pp. 464-473 (doi: 10.1016/j.dss.2011.10.007).

Österle, H., Becker, J., Frank, U., Hess, T., Karagiannis, D., Krcmar, H., Loos, P., Mertens, P., Oberweis, A., and Sinz, E. J. 2010. "Memorandum zur gestaltungsorientierten Wirtschaftsinformatik," *Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung* (62:6), pp. 664-672 (doi: 10.1007/BF03372838).

Pai, P.-F., Tan, Y.-S., and Hsu, M.-F. 2015. "Credit Rating Analysis by the Decision-Tree Support Vector Machine with Ensemble Strategies," *International Journal of Fuzzy Systems* (17:4), pp. 521-530 (doi: 10.1007/s40815-015-0063-y).

Peemöller, V. H. 2013. *Bilanzanalyse und Bilanzpolitik*: *Einführung in die Grundlagen: Rechnungslegung, Jahresabschluß, Bilanzierung und Bewertung, Bilanzpolitik, Bilanzanalyse, Analyseinstrumente*, Wiesbaden: Gabler Verlag.

Raschka, S. 2014. *About Feature Scaling and Normalization*. https://sebastianraschka.com/Articles/2014_about_feature_scaling.html. Accessed 3 June 2021.

Shearer, C. 2000. "The CRISP-DM Model: The New Blueprint for Data Mining," *Journal of Data Warehousing* (5:4).

Tian, S., Yu, Y., and Guo, H. 2015. "Variable selection and corporate bankruptcy forecasts," *Journal of Banking & Finance* (52), pp. 89-100 (doi: 10.1016/j.jbankfin.2014.12.003).

Webster, J., and Watson, R. T. 2002. "Analyzing the Past to Prepare for the Future: Writing a Literature Review," *MIS Q* (26:2), pp. xiii-xxiii.

Zięba, M., Tomczak, S. K., and Tomczak, J. M. 2016. "Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction," *Expert Systems with Applications* (58), pp. 93-101 (doi: 10.1016/j.eswa.2016.04.001).