Bankruptcy Prediction of Greek Small and Medium-Sized Enterprises Using Imbalance Data

Vassiliki Papadouli, Elias Houstis, and Manolis Vavalis *

Department of Electrical and Computer Engineering, University of Thessaly, Volos, Greece
Email: vpapadouli@uth.gr (V.P.); enh@uth.gr (E.H.); mav@uth.gr (M.V.)
*Corresponding author

***Abstract*—Detecting financial distress in businesses that lead to bankruptcy has been studied for a century. Building large labeled bankruptcy data sets is non-trivial and challenging. We produce an imbalanced data set of bankrupt and non-bankrupt Greek Small and Medium-sized Enterprises (SMEs) covering three years before the bankruptcy data and utilize it to test the bankruptcy predictive ability of well-known statistical and several supervised classifiers. A set of machine learning classifiers has been utilized demonstrating good predictive ability. The AutoML supervised classifier applied to the entire imbalanced data set shows worth noticing performance. We implement several supervised algorithms in a semi-supervised framework to remedy the imbalance of the data set and observed better overall performance than the supervised ones. To measure the effect of combining data from compatible European and Greek markets, we developed customized and AutoML-based transfer deep learning classifiers to predict the bankruptcy of Greek SMEs. Our findings justify transfer learning as an alternative methodology for studying bankruptcy prediction-related problems.**

***Keywords*—bankruptcy prediction, statistical models, hazard models, supervised machine learning, self-training, semi-supervised, transfer learning**

I. Introduction

Small and Medium-sized Enterprises (SMEs) comprise many nations' backbone of the economy. SMEs are highly vulnerable to economic changes, and the risk of default is very high, especially for those of smaller size. Given their economic importance, research for predicting their default risk is essential for both SME stakeholders and policymakers.

According to the requirements of the regulatory frame-works, a company is characterized as "defaulted" when either the obligor cannot pay its credit obligations or is past due more than 90 days on any material credit obligation. We consider bankruptcy the legal process whereby financially distressed firms resolve their debts and we interpret a firm's financial distress as temporary cash flow difficulty leading to insolvency, default, and bankruptcy. Predicting firms' bankruptcy early can assist the firm's management in taking measures to limit the downturn's effects or even avoid default. From the creditor's perspective, SMEs are considered high-risk customers because of a higher failure rate and ambiguous information reporting.

Detecting financial distress in businesses that lead to bankruptcy has been studied for a century, and building large labeled bankruptcy data sets for the Greek market is non-trivial and challenging. In this study, we develop a Greek SME data set that includes data of derogative information derived from different resources. This information includes companies that are bankrupt and inactive. A company is characterized as "non-defaulted" or "non-bankrupt" if there is: no event of bankruptcy, no event of the bankruptcy petition, and no delay of payments. In the case of the Greek SME data set, the active companies are considered "non-bankrupt".

Based on these distinct characteristics, the demand for predicting SMEs' bankruptcy risk has increased the need for research on this topic. Since Altman introduced the breakthrough bankruptcy prediction model in 1968, a large body of research has focused on predicting corporate financial distress that leads to bankruptcy.

This paper evaluates statistical and machine learning models on imbalanced and unlabeled data. Building large labeled bankruptcy data sets for the Greek market is non-trivial and challenging. We produce a well-defined and publicly available financial data set of 170 bankrupt and 1424 non-bankrupt Greek SME firms for three years before the bankruptcy. Note that previous bankruptcy prediction Greek studies are based on very limited, and often unavailable data sets.

The variables in our data set include the financial ratios used in many studies [1]. Identifying the statistically significant variables involved in the Greek SME data set is crucial for applying statistical and machine learning classifiers for bankruptcy prediction. We apply statistical inference tests and correlation analysis to select significant ones. Then, the data set's remaining variables (features) are ranked using the permutation importance technique [2]. The resulting top features are incorporated into the statistical and supervised algorithms for estimating the solution of the binary bankruptcy prediction classification problem.

The main objective of this study is to assess the performance of statistical and machine learning

Manuscript received August 24, 2023; revised September 25, 2023; accepted March 19, 2024; published August 15, 2024.

classifiers on the imbalanced bankruptcy Greek SME data set produced, based on existing [3] and emerging means. Specifically, we implement and test the following statistical and machine learning algorithms: Altman's Z”-score, Springate and Taffler's models, Linear Discriminant Analysis (LDA), Logistic Regression (LR), Naive Bayes (NB), Artificial Neural Networks (ANN), Support Vector Machine (SVM), Decision Trees (DT), Random Forest (RF) classifiers, and Extreme Gradient Boosting (XGBoost). In implementing these classifiers, we identify the significant feature by applying statistical and permutation importance techniques and handle the data imbalance with synthetic sampling [4]. The AutoML framework from scikit-learn utilizes all the features of the imbalanced Greek data set without synthetic sampling with worth noticing performance.

We employ semi-supervised learning models to assess their predictive ability on the Greek imbalanced data set. Their superior performance suggests that we can exploit commercially available unlabeled data sets for bankruptcy prediction, including the one from the ICAP company.

Lastly, we apply the transfer learning framework to train machine learning models to address the Greek market's lack of large bankruptcy label data sets. We develop a pre-trained customized-based deep learning model and AutoML on a large public domain data set for a European market with similar features to the Greek market and fine-tuned them on our Greek SME bankruptcy data set.

The rest of the paper is organized as follows. Section II elaborates on the collection of the data set utilized for training and testing the bankruptcy prediction models considered in our study, predictive statistical and supervised machine learning models whose computational performance is analyzed and the performance metrics assessed, describes the variable selection procedure applied using univariate analysis, correlation analysis, and permutation importance techniques, and restates well-known performance metrics applied to evaluate the bankruptcy prediction classifiers. Section III presents the performance results of well-established Multiple Discriminant Analysis (MDA) bankruptcy models and describes our supervised algorithms and discuss their predictive ability. The same classifiers were implemented in a semi-supervised framework, and their performance is also reported in Section III. In addition, Section III presents two new deep learning classifiers utilizing the transfer learning paradigm, which are pre-trained on a large public domain Polish data set and fine-tuned on the Greek data set. Our concluding remarks can be found in Section IV.

II. METHOD

The produced dataset consists of 170 bankrupt SME companies and 1424 non-bankrupt ones with about the same characteristics. The data set is generated by mining financial reports of Greek firms from 2002–2015. These data sets are obtained from several sources, including the commercial company iMENTOR, the archive of the Greek financial journal Naftemporiki, and Datastream as well as from other sources of unstructured data [5].

In our study, firms are considered financially distressed only when declared bankrupt by a court decision and other complied means [6]. The validity of each firm's status is cross-checked through the Hellenic Statistical Authority. The variables included in our data set monitor firms' soundness, stability, and performance. They include those variables utilized by Altman's Z-score [7] and Taffler's model [8]. Some complementary financial indicators were utilized as suggested by Hofer [9]. The financial ratios considered in this study are listed in https://doi.org/10.24432/C5F600.

The data set produced contains three years' worth of data for each company. The year a company declares bankruptcy is denoted as the benchmark year “t”. This means that (t−1), (t−2), and (t−3) represent 1, 2, and 3 years before the bankruptcy occurred, and cumulative data represent a total of 3 years. It is worth mentioning that companies in financial difficulty that eventually go bankrupt cease publishing their financial statements until the failure date. Thus, our study assumes that their last published balance sheet refers to one year before bankruptcy (t−1).

In this section, we introduce the statistical and machine learning models applied to study the bankruptcy prediction problem for Greek SMEs utilizing the data set described above. The statistical models are functional forms linearly relating to the various financial ratios. We considered and tested the well-established Altman's Z”-score, Springate, and Taffler's models. Despite their simplicity and applicability, these models are restricted by assumptions such as linearity, normality, and independence among predictor variables.

Further, we implement several supervised, semi-supervised, and transfer learning techniques to study the bankruptcy prediction problem as a binary classification problem with imbalanced data. The machine learning algorithms considered include Linear Discriminant Analysis (LDA), Logistic Regression (LR), Naive Bayes (NB), Artificial Neural Networks (ANN), Support Vector Machine (SVM), Decision Trees (DT), Random Forest (RF), Extreme Gradient Boosting (XGBoost) classifiers, and AutoML framework from the scikit-learn library [10]. The Automated Machine Learning (AutoML) framework automates the machine learning phases of selection, composition, and parameterization [11, 12].

For training and testing the supervised algorithms, we split the Greek SME data set into 77% and 23% subsets, respectively. The training set consists of 130 bankrupt and 1089 non-bankrupt companies, and the test set of 40 bankrupts and 335 non-bankrupt ones for each of the three time periods. In the case of supervised classifiers except for AutoML, we apply the Synthetic Minority Oversampling Technique (SMOTE) to over-sample the minority class labels in the training set. Thus, each period contains 1089 bankrupt and non-bankrupt companies for training and 40 bankrupts and 335 non-bankrupt companies for testing.

All the above predictive models have been implemented in python utilizing various packages including Pandas, Scikit-Learn, tensorflow, and Auto-sklearn, Auto-keras to realize AutoML frameworks. We identified the optimal combinations of hyper-parameters required using 5-fold cross-validation applying Gridsearchcv, and Keras Tuner.

Finally, we extract the most important variables of the models using ELI5, a Python library that computes feature importance for any black-box estimator.

The initial selection of variables in the Greek SME data set aims to detect the variables that individually have significant predictive power for estimating a firm's default probability. Next, we describe three variable (feature) preselection procedures that result in an initial group of variables used in machine learning bankruptcy prediction classifiers implemented in this study.

TABLE I. STATISTICAL TESTS RESULTS FOR EACH VARIABLE ON THE TRAINING DATA SET FOR EACH PERIOD WITH 5% (UNDERLINED) AND 1% (STARED) SIGNIFICANT LEVELS

Category (ratios)	**Code**	**P-values**			
		***t*-test**			**g-rank**
		***t*−1**	***t*−2**	***t*−3**	**Cumulative Data**
Liquidity	x1	0	0	0	0
	x2	0	0	0	0
	x3	0.322	0.43	0.397	1
	x4	0.002	0	0	0
	x5	0	0	0.195	0
	x6	0	0	0	0
	x7	0.007	0.002	0.001	0
	x8	0.182	0	0.001	0
Profitability	x9	0	0	0	0
	x10	0.0001	0.438	0	0.54
	x11	0	0.0002	0	0
	x12	0	0	0	0
	x13	0.0001	0	0.027*	0.03*
	x14	0.0001	0.759	0	0.28
Leverage	x15	0	0	0	0
	x16	0	0	0	0
	x17	0	0	0	0
	x18	0	0	0	0
	x19	0.6	0.002*	0	0
	x20	0.1	0.001	0.348	0.01*
Activity	x21	0.008	0.008*	0	0.08
	x22	0.0007	0	0	0
	x23	0.056	0.01*	00	0.28
Efficiency	x24	0.011*	0.231	0	0
	x25	0.013*	0.023*	0	0
	x26	0.051	0.003	0.006	0.08
	x27	0.0002	0	0	0
Growth	x28	0	0.06	0.387	0
	x29	0.92	0.75	0.002	0.09
Size Other	x30	0	0	0	0
	x31	0	0	0	0
	x32	0.016*	0	0.023	0
	x33	0.3	0.31	0.313	0
	x34	0.493	0.599	0.066	0
	x35	0	0	0	0
	x36	0	0	0	0

The first procedure conducts a univariate analysis to identify the financial indicators (variables) that distinguish between default and non-default firms. This procedure is based on an inferential statistics t-test under the assumptions of normal distribution and equal variances. The statistical significance level for selecting a variable is set at 5%, so variables below this level are ignored. Based on this univariate analysis, an independent variable could show high statistical significance in predicting the default probability. The results of this procedure on our data set variables are shown in Table I for each period (t−1, t−2, t−3).

The second procedure calculates the correlation among the variables in the Greek SME data set. Highly correlated variables are not desirable to be included in the bankruptcy prediction models and are disregarded.

The third procedure ranks the remaining data set variables (features) obtained from the above steps using the permutation importance technique. According to the research of Hutter *et al*. [13], permutation feature importance measures the increase in the model's prediction error after randomly permuting the feature's values, breaking the relationship between the feature and the true outcome. The permutation importance procedure is implemented by ELI5 python library on the testing data set for computationally efficient justification [13], and we keep those features with a positive score.

The selected variables in our implementation of supervised and semi-supervised LR and LDA models are those obtained by applying all the above three procedures. The rest of the algorithms, including the transfer learning ones, use those features obtained by correlation and permutation importance analysis.

One of the major objectives of this study is to assess the performance of statistical and machine learning classifiers for assessing the bankruptcy risk of firms trained on imbalanced data. For this, we need to utilize several performance measures. In this section, we restate well-known performance metrics for binary classifiers.

A confusion matrix can be used in classification to evaluate the model's performance by identifying what type of errors are made. For binary classification, a confusion matrix is defined in Table II.

TABLE II. CONFUSION MATRIX FOR BANKRUPTCY PREDICTION CLASSIFIERS

		Predicted	
		Non-Bankrupt	Bankrupt
Actual	Non-Bankrupt	TNR	FPR
	Bankrupt	FNR	TPR

Balanced accuracy is a machine learning error metric for binary and multi-class classification models most suitable for imbalanced data sets. It turns out that this metric is the mean of sensitivity and specificity, where sensitivity (True Positive Rate) is the probability of a positive case being accurately classified as being positive, and specificity (True Negative Rate) is the probability of a negative case being accurately classified as negative. It is, therefore, often seen as a better alternative to standard accuracy computed by Eq. (1):

$$\text{balanced accuracy} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \qquad (1)$$

An extension of the confusion matrix is the Receiver Operating Characteristic (ROC) curve. The associated AUROC gives the overall performance of a classifier, summarized over all possible thresholds. AUROC is the area under the ROC curve, and a ROC curve shows the trade-off between true positive rate and false-positive rate.

It essentially shows how much the classifier is capable of distinguishing between groups. The value of AUROC usually lies from 0.5 to 1 and indicates its discriminating power. We use AUROC for group membership instead of probability scores.

III. RESULTS AND DISSCUSSIONS

Next, we summarize the main characteristics and present the performance of three well-known statistical models on the Greek SMEs data set. They include Altman's Z"-score, Springate's S-model, and Taffler's Z-model presented in detail in many publications [14].

One of the most popular MDA models was proposed by Altman *et al.* [15]. According to a discrimination analysis, 22 potential variables from the firms' annual reports considered and kept the five variables with the highest significance. These variables involve profitability ratios, coverage ratios, liquidity ratios, capitalization ratios, and earnings variability of a firm. We specifically consider the Z''-score extension of the original Altman's model, applicable for manufacturing and non-manufacturing firms, including private and public firms. According to that model, the firms are categorized into one of the following three zones: If $Z'' > 2.9$—Safe zone, $1.23 < Z'' < 2.9$—Grey zone, and $Z'' < 1.23$—Distress zone. For its performance evaluation, we tested this model for each Greek SMEs data set period. An extension of the above model was developed in 1978 [16]. The Springate S-model involves 4 out of the 19 financial ratios that are frequently used in the literature to assess the firm's bankruptcy risk. It involves a standard calculation in which the firm is healthy if the Springate overall index $S > 0.862$ while it is classified as potentially bankrupt if $S < 0.862$.

The third model considered is Taffler's Z-score MDA model [8] which is based on an extensive survey of the vast array of data. The cut-off value of Taffler's Z-score value equals 1.95. Thus, the firm will likely go bankrupt if Z-score is less than 1.95. If it exceeds 1.95, the firm is solvent and unlikely to go bankrupt the following year.

The specific variables and formulas involved in the above three models are:

$$\text{Altman: } Z'' = 6.56X_1 + 3.26X_2 + 6.72X_3 + 1.05X_4$$

$$\text{Springate: } S = 1.03X_1 + 3.07X_2 + 0.66X_3 + 0.4X_5$$

$$\text{Taffler: } Z = 3.2 + 12.18X_6 + 2.5X_7 - 0.68X_8 + 0.029X_9$$

where $X_1 = \frac{\text{Working Capital}}{\text{Total Assets}}$, $X_2 = \frac{\text{Retained Earnings}}{\text{Total Assets}}$, $X_3 = \frac{\text{Earnings Before Interest}-\text{Taxes}}{\text{Total Assets}}$, $X_4 = \frac{\text{Book Val.Equity}}{\text{Total Liabilities}}$, $X_5 = \frac{\text{Sales}}{\text{Total Assets}}$, $X_6 = \frac{\text{Profit before Tax}}{\text{Current Liabilities}}$, $X_7 = \frac{\text{Current Assets}}{\text{Total Liabilities}}$, $X_8 = \frac{\text{Current Liabilities}}{\text{Current Assets}}$, and X_9=No-credit days (Note: no-credit interval in days = (liquid current assets / daily cash operating expenses) or (quick assets – current liabilities) / ((sales – profit before tax) / 365)).

Table III displays the confusion matrix of the three statistical models considered on the Greek SME data set over three years. Considering the average performance of the three models over all periods, we observe that a)

1) Altman classifies correctly 66% of bankrupt firms and incorrectly 36% of the healthy firms,
2) Springate classifies correctly 89% of the bankrupt firms, but 75% of healthy firms are categorized as bankrupt for the three years,
3) Taffler classifies bankrupt companies correctly with an average rate of more than 76%

The average balance score of the three classifiers indicates that Taffler's model performs slightly better than the rest.

TABLE III. THE ACCURACY (%) OF ALTMAN'S, TAFFLER'S AND SPRINGATE'S CLASSIFIERS FOR THE GREEK SMES DATA SET CONSISTING OF 1434 ACTIVE AND 170 BANKRUPT FIRMS

Period	Status	Non Bankrupt (NB)			Bankrupt (B)			Total
		Altman	Taffler	Springate	Altman	Taffler	Springate	
t−1	NB	500	610	313	571	814	1111	1424
	%	35.1	42.8	21.9	40.1	57.1	78.1	
	B	15	32	12	124	138	158	170
	%	8.8	18.8	7	72.9	81.1	93	
t−2	NB	490	621	335	529	803	1089	1424
	%	34.4	43.6	23.5	37.1	56.3	76.4	
	B	19	44	18	108	126	152	170
	%	11.1	25.8	10.5	63.5	74.1	89.5	
t−3	NB	520	684	401	459	740	1023	1424
	%	36.5	48	28.1	32.2	52	71.8	
	B	21	42	28	103	128	142	170
	%	12.3	24.7	16.4	60.5	75.2	83.6	%

Next, we briefly describe the supervised models and present their performance for Greek SMEs data.

Linear Discriminant Analysis (LDA) is a supervised method that draws one hyperplane and projects the data onto this hyperplane in a way that maximizes the separation of the two categories. It identifies a hyperplane by maximizing the distance between the means of two classes and minimizing the variation between each category. LDA can be viewed as using the variables obtained from the univariate analysis, removing the highly correlated ones, and keeping the most important variables based on permutation importance. To achieve maximal separation, the optimal weights are found, and the cutoff value is obtained by calculating the average of the means of the discriminant scores of each group.

An alternative to LDA models is the class Logit or Logistic regression (LR) model which does not assume normality on the predictor variables. It is essentially a linear model, using a sigmoid function f(x) = 1/ (1 + e−x) to conduct the classification, and the output is between 0 and 1. For a firm, Logit models compute the probability that it will default given a set of variables. Hence, the probability of default depends conditionally on these variables [17]. Unlike the model of Altman, the output of LR is the probability of default, while MDA generates a score adopted to classify an observation between the no-default and default classes. We apply a Logit model in each of the three time periods using the statistically significant variables identified by our preliminary analysis, removing highly correlated ones, and selecting those ranked high according to the permutation importance methodology.

The Gaussian Naive Bayes (NB) classifier is a probabilistic algorithm based on the Bayes Theorem that assumes that the features involved in the training data set are independent and contribute equally to the prediction of the target class. It predicts the target class for a particular instance by computing the conditional probabilities involved in the Bayes formula, it is not sensitive to noise and performs well when faced with outliers.

The Support Vector Machine (SVM) algorithm computes a hyperplane that best separates the two classes for a separable binary training data set. For non-linearly separable data, the original feature space is transformed into a higher-dimensional space using the so-called kernel functions and then finds support vectors to maximize the separation (margin) between two classes.

The Decision Tree (DT) algorithm iteratively divides the features of prediction space according to some purity criterion providing a tree structure consisting of nodes that represent the features of a data set, branches that depict the decision rules, and a leaf node that indicates the outcome (class label). Several separation criteria have been proposed [18], including miss-classification error rate and Gini index. Our models were tuned based on various criteria and different values of the maximum depth of trees.

The Random Forest (RF) is an ensemble of decision trees, where each tree utilizes a bootstrap sample drawn from a training set with replacement. Each tree splits out a class prediction aggregated using the majority vote to return the final prediction. For each iterative split, the selection is shuffled at random forcing the tree to consider all alternatives equally, contributing to a robust ensemble. Extreme Gradient Boost (XGB) is a distributed, scalable Gradient-Boosted Decision Tree (GBDT) machine learning algorithm for classification. XGBoost [19, 20] combines regression trees and gradient boosting reducing modeling complexity and preventing the problems associated with overfitting.

Neural Networks (MLP) have been proposed to build bankruptcy prediction models that are at least as accurate as discriminant analysis [21]. Our study uses a Multilayer Perceptron (MLP), a forward neural network connecting multiple layers in a directed graph. Apart from the input nodes, each node has a nonlinear activation function. An MLP can use backpropagation in a supervised learning mode and solve problems that are not linearly separable. Since our data set is small, we apply simple neural networks with no more than two hidden layers.

Automated Machine Learning (AutoML) is the process of automating the tasks of applying machine learning. Auto-sklearn [10] automatically searches for the right learning algorithm for a new machine learning data set and optimizes its hyperparameters. It comprises of 15 classification algorithms and 14 feature pre-processing algorithms for feature selection or transformation of features into a different space. It takes care of data scaling, encoding categorical parameters, and missing values. We apply the sklearn implementation of AutoML using all financial ratios in the Greek data set without handling its imbalance with synthetic sampling.

The performance results of the statistical and machine learning models on the selected testing sets for each considered year are shown in Table IV.

The performance of the supervised algorithms described above for the Greek SME data set is depicted in Tables IV and V. The training of AutoML was done on the original imbalanced data set, and the rest supervised classifiers were trained on balanced data set obtained with synthetic sampling from the Greek SMEs data set.

The experimental results suggest that most supervised algorithms perform exceptionally well except Naive Bayes. The performance of the classifiers varies in different data periods. In the case of $t-1$ period data, the Multi-layer Perceptron (MLP) achieves 95% balanced accuracy. Extreme Gradient Boosting and Random Forest methods (RF) closely follow, achieving an overall performance of 91%, and 90%, respectively. However, Naive Bayes (NB) produces poor results. For two years before bankruptcy ($t-2$) data, XGBboost outperforms the remaining methods, achieving a total predictiveness equal to 89%. Finally, the AutoML applied to the data for period $t-3$ outperforms the rest with 93% balanced accuracy on the imbalanced data set. Decision Trees, Random Forest, Extreme Gradient Boosting, and Multi-layer Perceptron algorithms produce equally good results for the $t-3$ period, except for Naive Bayes. We observe that DT, RF, XGB, MLP, and ATM reach approximately 90% performance when computing the average accuracy achieved over the three years of training data. It’s worth noticing that AutoML achieves significant accuracy on imbalanced data without any explicit pre-processing of the train data.

TABLE IV. THE CONFUSION MATRIX OF THE MACHINE LEARNING ALGORITHMS CONSIDERED ON A LABELED SUBSET OF THE GREEK SME DATASET FOR EACH OF THE 3 YEARS

Year	Status	Non Bankrupt (NB)									Bankrupt (B)									Total
		LDA	LR	NB	SVM	DT	RF	XGB	AML	MLP	LDA	LR	NB	SVM	DT	RF	XGB	AML	MLP	
t−1	NB	281	284	197	316	320	321	330	332	330	54	51	138	19	15	14	5	3	5	335
	%	83.8	84.7	58.8	94.3	95.5	95.8	98.5	99.1	98.5	16.1	15.2	41.2	5.7	4.5	4.2	1.5	0.9	1.5	
	B	8	2	8	11	4	6	6	8	3	32	38	32	29	36	34	34	32	37	40
	%	20	5	20	27.5	10	15	15	20	7.5	80	95	80	72.5	90	85	85	80	92.5	
t−2	NB	281	281	319	313	316	320	323	324	324	54	54	16	22	19	15	12	11	11	335
	%	83.8	84.7	95.2	93.4	94.3	95.5	96.4	85.3	96.7	16.1	15.2	4.7	6.6	5.7	4.5	3.6	14.6	3.3	
	B	7	4	19	11	9	9	7	8	10	33	36	21	29	31	31	33	32	30	40
	%	17.5	10	47.5	27.5	22.5	22.5	17.5	20	25	82.5	90	52.5	72.5	78	77.5	82.5	80	75	
t−3	NB	279	265	90	317	322	35	325	322	313	56	70	245	18	13	5	10	13	22	335
	%	83.2	79.1	26.8	94.6	96.1	87.5	97	96.1	93.4	16.7	20.8	73.1	5.3	3.9	12.5	3	3.9	6.6	
	B	6	9	3	16	7	7	9	4	9	34	31	37	24	33	33	31	36	31	40
	%	15	22.5	7.5	40	17.5	17.5	22.5	10	22.5	85	77.5	92.5	60	82.5	82.5	77.5	90	77.5	

TABLE V. BALANCED ACCURACY (ACC) AND AUC SCORES OF SUPERVISED ML MODELS ON TESTING SETS FOR EACH PERIOD OF THE GREEK SMEs DATA SUBSET

Year	Score	Supervised Learning Algorithms								
		LDA	LR	NB	SVM	DT	RF	XGB	MLP	AML
t−1	ACC	83%	89%	69%	83%	92%	90%	91%	95%	89%
	AUC	0.91	0.93	0.76	0.91	0.93	0.98	0.98	0.99	0.97
t−2	ACC	83%	86%	73%	82%	87%	86%	89%	85%	88%
	AUC	0.92	0.93	0.87	0.92	0.87	0.96	0.97	0.93	0.96
t−3	ACC	84%	78%	59%	77%	90%	895	87%	85%	93%
	AUC	0.92	0.89	0.80	0.92	0.90	0.96	0.97	0.94	0.98

Collecting large, labeled data for bankruptcy prediction is non-trivial and challenging, particularly for the Greek market. In contrast, finding unlabeled data is relatively easy and cheap. Unlabeled data contains useful information that can be extracted using Semi-Supervised Learning (SSL) methods, and their contribution has proven effective in solving the bankruptcy prediction classification task [22].

Semi-supervised classification is an extension of the supervised classification problem [23]. The training data consists of both l labeled instances $\{(x_{i,}y_i\}_{i=1}^{l}$ and u unlabeled instances $\{x_j\}_{j=l+1}^{l+u}$ One typically assumes that there is much more unlabeled data than labeled data, i.e., $u \gg$ l. Semi-supervised classification aims to train a classifier f from both the labeled and unlabeled data such that it is better than the supervised classifier trained on the labeled data alone.

Semi-Supervised Learning (SSL) can potentially utilize imbalanced labeled and unlabeled data to achieve better (or at least the same) performance than supervised learning with fewer labeled instances. SSL works because it can link the distributions of unlabeled and the target label data [23]. There are several semi-supervised learning methods, and each makes slightly different assumptions about this link. These methods include self-training, probabilistic generative models, co-training, graph-based models, semi-supervised support vector machines, and many others.

We have implemented a Self-training method variation of SSL described by the pseudo-algorithm.

Algorithm 1. Pseudo-algorithm

Input: labeled $\{(x_{i,}y_i\}_{i=1}^{l}$ and unlabeled $\{(x_{i,}y_i\}_{i=1}^{l}$ data

1. Set L = $\{(x_{i,}y_i\}_{i=1}^{l}$ and U = $\{(x_{i,}y_i\}_{i=1}^{l}$
2. Repeat:
3. Train *f* from *L* using supervised learning.
4. Apply *f* to the unlabeled instances in *U*.
5. Remove a subset *S* from *U*; add $\{(\boldsymbol{x}, f(\boldsymbol{x}))|\boldsymbol{x} \in S$ to *L*.

The main idea of the algorithm, which is characterized by the fact that the learning process uses its predictions to teach itself, is to first train f on labeled data. The function f is then used to predict the labels for the unlabeled data. A subset S of the unlabeled data and their predicted labels are then selected to augment the labeled data. Typically, S consists of the few unlabeled instances with the most confident f predictions. The function f is re-trained on the larger labeled data set, and the procedure repeats. It is also possible for S to be the whole unlabeled data set. In this case, L and U cover the entire training sample, but the assigned labels on unlabeled instances might vary from iteration to iteration.

We selected LDA, LR, NB, SVM, DT, RF, XGB, and MLP as supervised classifiers f inside the self-training algorithm. The performance of the self-training classifiers is summarized in Table VI.

TABLE VI. THE CONFUSION MATRIX OF SELF-TRAINING CLASSIFIERS ON GREEK SME DATA SET

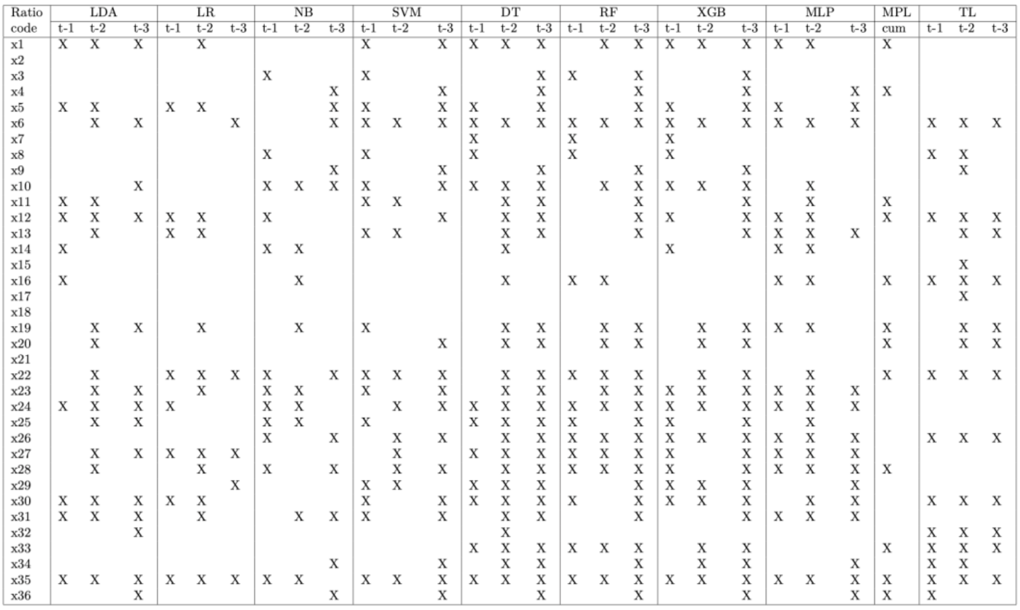
Year	Status	Non Bankrupt (NB)								Bankrupt (B)								Total
		LDA	**LR**	**NB**	**SVM**	**DT**	**RF**	**XGB**	**MLP**	**LDA**	**LR**	**NB**	**SVM**	**DT**	**RF**	**XGB**	**MLP**	
t−1	NB	265	286	195	317	320	320	330	323	70	49	140	18	15	15	5	12	335
	%	79.1	85.3	58.2	94.6	95.5	95.5	98.5	96.4	20.9	14.6	41.8	5.4	4.5	4.5	1.5	3.6	
	N	5	5	11	11	4	5	5	4	35	35	29	29	36	35	35	36	40
	%	12.5	12.5	27.5	27.5	10	12.5	12.5	10	87.5	87.5	72.5	72.5	90	87.5	87.5	90	
t−2	NB	247	276	300	314	317	322	321	323	88	59	35	21	18	13	14	12	335
	%	73.7	82.3	89.5	93.7	94.6	96.1	95.8	96.4	26.3	17.6	10.5	6.3	5.4	3.9	4.2	3.6	
	B	2	3	20	10	8	9	8	11	38	37	20	30	32	31	32	29	40
	%	5	7.5	50	25	20	22.5	20	27.5	95	92.5	50	75	80	77.5	80	72.5	
t−3	NB	280	262	52	289	324	323	322	323	55	73	283	37	11	12	13	12	335
	%	83.5	78.2	15.5	86.2	96.7	96.4	96.1	96.4	16.5	21.7	84.5	13.8	3.3	3.6	3.9	3.6	
	B	9	8	4	10	5	8	7	9	31	32	36	30	35	32	33	31	40
	%	22.5	20	10	25	12.5	20	17.5	22.5	77.5	80	90	75	87.5	80	82.5	77.5	

In Table VII, we depict the balance accuracy and AUC scores for the considered semi-supervised classifiers for the Greek SMEs data set in each of the three periods. The data in Table VII and the average performance of the semi-classifiers suggest that the self-learning classifiers based on DT, RF, XGBoost, and MLP models exhibit slightly better computation behavior than the supervised ones. It's worth noticing that the SSL logistic regression performs significantly better than its supervised counterpart.

TABLE VII. BALANCED ACCURACY (ACC) AND AUC SCORES ON TESTING SETS FOR EACH PERIOD

Year	Score	Semi-supervised Learning							
		LDA	**LR**	**NB**	**SVM**	**DT**	**RF**	**XGB**	**MLP**
t−1	ACC	83%	90%	56%	83%	92%	91%	93%	93%
	AUC	0.86	0.93	0.77	0.89	0.93	0.97	0.97	0.96
t−2	ACC	84%	87%	69%	84%	87%	86%	87%	84%
	AUC	0.91	0.94	0.80	0.93	0.88	0.96	0.96	0.93
t−3	ACC	80%	79%	52%	81%	92%	88%	89%	87%
	AUC	0.87	0.89	0.78	0.88	0.93	0.94	0.96	0.95

Another class of methods for dealing with limited-sized data sets for bankruptcy prediction is the so-called Transfer Learning (TL) methodology. The basic idea of our learning scheme is to transfer knowledge from pre-trained classifiers using large labeled data to classify "similar" in nature data sets. Specifically, we used a pre-trained deep learning classifier on the well-known "Polish" data set [24] after fine tune it and testing it on our Greek data set of bankrupt firms. In this classifier, the lower layers (those closer to the input) refer to general features (problem-independent variables), and the upper layers refer to specialized features (problem-dependent variables). Since the testing data set of the current task is small and similar to the pre-trained data, more layers are frozen to avoid over-fitting. Specifically, the base of MLP is frozen, and only the classifier (top layer) is trained. Furthermore, the learning rate is a hyper-parameter that controls how much you adjust the weights of your neural network. In the case of pre-trained model, a small learning rate is preferred since high learning rates increase the risk of losing previous knowledge.



Ratio	LDA			LR			NB			SVM			DT			RF			XGB			MLP			MPL	TL		
code	t-1	t-2	t-3	t-1	t-2	t-3	t-1	t-2	t-3	t-1	t-2	t-3	t-1	t-2	t-3	t-1	t-2	t-3	t-1	t-2	t-3	t-1	t-2	t-3	cum	t-1	t-2	t-3
x1	X	X	X		X					X		X	X	X	X		X	X	X	X	X	X	X		X			
x2																												
x3							X			X					X	X		X			X							
x4									X			X			X			X			X			X	X			
x5	X	X		X	X				X	X		X	X		X			X	X		X	X		X				
x6		X	X			X			X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X
x7													X			X			X									
x8							X			X			X			X			X							X	X	
x9									X			X			X			X			X						X	
x10			X				X	X	X	X		X	X	X	X		X	X	X	X	X		X					
x11	X	X								X	X			X	X			X			X		X		X			
x12	X	X	X	X	X		X					X		X	X			X	X		X	X	X		X	X	X	X
x13		X		X	X					X	X			X	X			X			X	X	X	X			X	X
x14	X						X	X						X					X			X	X					
x15																											X	
x16	X							X						X		X	X					X	X		X	X	X	X
x17																											X	
x18																												
x19		X	X		X			X		X				X	X		X	X		X	X	X	X		X		X	X
x20		X										X		X	X		X	X		X	X				X		X	X
x21																												
x22		X		X	X	X	X		X	X	X	X		X	X	X	X	X		X	X		X		X	X	X	X
x23		X	X		X		X	X		X		X		X	X		X	X	X	X	X	X	X	X				
x24	X	X	X	X			X	X			X	X	X	X	X	X	X	X	X	X	X	X	X	X				
x25		X	X				X	X		X			X	X	X	X		X	X		X		X					
x26							X		X		X	X		X	X	X	X	X	X	X	X	X	X	X		X	X	X
x27		X	X	X	X	X					X		X	X	X	X	X	X	X		X	X	X	X				
x28		X			X		X		X		X	X		X	X	X	X	X	X		X	X	X	X	X			
x29						X				X	X		X	X	X			X	X	X	X			X				
x30	X	X	X	X	X					X		X	X	X	X	X		X	X	X	X		X	X		X	X	X
x31	X	X	X		X			X	X	X		X		X	X			X			X	X	X	X				
x32			X											X												X	X	X
x33													X	X	X	X	X	X		X	X				X	X	X	X
x34									X			X		X	X			X		X	X			X		X	X	
x35	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
x36			X						X			X			X			X			X			X	X	X		

Fig. 1. The features used in each algorithm for each time period.

In each period, we used the common features of the Polish and our data set. We dropped the highly correlated ones, and then removed features with zero or negative permutation importance while checking improvement in the model performance. The algorithmic parameter values used in our transfer learning experiments are given in Table VIII while the features (ratios) used in each algorithm for each of the three periods are given in Fig. 1.

TABLE VIII. THE ALGORITHM PARAMETER VALUES USED IN THE TRANSFER LEARNING EXPERIMENTS

Parameter	*t*–1	*t*–2	*t*–3
hidden layers	2	2	2
units	8,10	5,10	9,12
learning rate	0.0001	0.001	0.001
hidden activation function	relu	relu	relu
output activation function	sigmoid	sigmoid	sigmoid

In addition, we have implemented an AutoML transfer learning classifier utilizing the AutoKeras [25] framework, trained on the "Polish" data set. In this deep learning classifier, the base of MLP is frozen, and only the classifier (top layer) is trained. We fine-tune it and then test it on our Greek data set of bankrupt firms.

A detailed confusion matrix for two deep transfer learning classifiers is given in Tables IX and X, we depict the associated balanced accuracy and AUC scores for these two transfer deep learning classifiers.

TABLE IX. THE CONFUSION MATRIX OF THE TWO DEEP TRANSFER LEARNING CLASSIFIERS ON THE GREEK BANKRUPTCY DATA SET CONSISTING OF A TOTAL OF 375 SMES WITH 335 OF THEM CLASSIFIED AS NON-BANKRUPT (NB) AND THE REST 40 AS BANKRUPT (B)

Year	Status	Predicted			
		Non-bankrupt (NB)		Bankrupt (B)	
		TL	TL-AML	TL	TL-AML
t–1	NB	270	240	65	95
		80.6%	71.6%	19.4%	28.4%
	B	4	5	36	35
		10%	12.5%	90%	87.5%
t–2	NB	324	274	11	61
		85.3%	81.7%	14.6%	15.3%
	B	8	15	32	25
		20%	37%	80%	62.5%
t–3	NB	272	313	63	22
		81.1%	93.4%	18.9%	6.6%
	B	6	17	34	23
		15%	42.5%	85%	57.5%

TABLE X. BALANCED ACCURACY (ACC) AND AUC SCORES OF THE TWO TRANSFER LEARNING CLASSIFIERS FOR EACH PERIOD OF THE GREEK SMES DATA SET

Period	Score	TL	TL-AML
t–1	ACC	85%	79%
	AUC	0.91	0.86
t–2	ACC	83%	72%
	AUC	0.91	0.75
t–3	ACC	83%	75%
	AUC	0.88	0.86

IV. SYNOPSIS AND CONCLUSIONS

In this study, we presented a comprehensive literature review, constructed a data set of 170 bankrupt and 1424 non-bankrupt Greek SMEs data, carried out a feature selection analysis, demonstrated several algorithmic alternatives for effectively forecasting firms' bankruptcy with imbalanced data, and compared their performance.

Our experiments indicate that the well-established traditional bankruptcy prediction models, such as Altman's *Z*-score, Springate, and Taffler's models, correctly predict a satisfactory number of bankrupt firms but classify most healthy firms as bankrupt. Taffler's model achieves slightly better performance based on average balance score. Tables IV–VI summarize the average balance and AUC scores for supervised and semi-supervised classifiers for the Greek SMEs data set. Our results are in accordance with recent studies [26] and related applications [20, 27–29].

Based on these results, Fig. 2 displays and compares the average balance score of the supervised and semi-supervised classifiers, indicating a slightly better performance of semi-supervised ones on the imbalanced Greek SMEs data set. Please note that most of the data sets used in similar studies are not publicly available while our data set and software are freely available upon request.

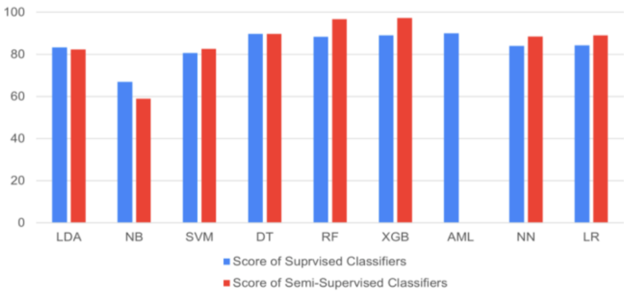


Fig. 2. Average balanced graph of the performance of supervised and semi-supervised machine learning algorithms over the 3 years before bankruptcy.

We further note that to perform comparative experiments against bankruptcy prediction data previously presented in publication for the Greek case we have been extensively searching through the peer-review literature. We could only find a short table in [22] indicating bankruptcy predictions with accuracy less than 0.627, 0.664, 0.732 for the *t*−1, *t*–2, and *t*−3 periods. This accuracy is much lower than most of the cases presented in this paper. Similar results are given in [26].

Specifically, from this graph, we concluded that the supervised classifiers DT, RF, XGB, and AML with synthetic over-sampling are effective alternatives for bankruptcy forecasting. It is worth noticing that the sklearn implementation of AutoML classifier utilizing all financial ratios in the Greek data set and ignoring the data imbalance outperforms the supervised ones. The corresponding semi-supervised classifiers performed better on the Greek SMEs data set. Their performance suggests we can exploit commercially available unlabeled data sets for bankruptcy prediction, including the one from the "ICAP" company. Finally, we developed and tested two transfer learning algorithms utilizing pre-trained deep learning models on an existing labeled data set for non-Greek markets. After fine-tuning them on the Greek data set, Fig. 1 indicates satisfactory performance in each bankruptcy period. The performance of the two transfer learning classifiers justifies transfer learning as an alternative methodology for studying bankruptcy prediction-related problems considering information from similar external markets.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

VP collected the data, performed the experiments and contributed in their analysis; EH proposed the concept, conducted the research, analyzed the data and reviewed every version of the paper he co-authored; MV conducted the research, co-authored and supervised all revisions. All authors had approved the final version.

Funding

This research has been co-financed by the European Regional Development Fund of the European Union and Greek National funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH-CREATE-INNOVATE (project code: T1EDK-02161). Part of the work presented in this paper is based on the first author's Diploma Thesis.

References

[1] W. Kusonkhum, K. Srinavin, N. Leungbootnak, P. Aksorn, and T. Chaitongrat, "Government construction project budget prediction using machine learning," *Journal of Advances in Information Technology*, vol. 13, no. 1, 2022.

[2] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[3] L. W. Siew, L. W. Hoe, M. A. Bakar, and L. P. Fun, "Data driven decision analysis on the performance of electronic companies with TOPSIS model," *Journal of Advances in Information Technology*, vol. 13, vol. 2, 2022.

[4] T. Daghistani and R. Alshammari, "Comparison of statistical logistic regression and random forest machine learning techniques in predicting diabetes," *Journal of Advances in Information Technology*, vol. 11, no. 5, 2020.

[5] I. Aattouchi and M. A. Kerroum, "A new framework for analyzing news in the financial markets to enhance the investor's perception," *Journal of Advances in Information Technology*, vol. 13, no. 1, 2022.

[6] S. Ruchi and S. Pravin, "Improved opinion mining for unstructured data using machine learning enabling business intelligence," *Journal of Advances in Information Technology*, vol. 14, no. 1, 2023.

[7] E. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *The Journal of Finance*, vol. 1, p. 21, 1968.

[8] R. J. Taffler, "The assessment and performance using a of company solvency statistical model," *Accounting and Business Research*, vol. 113, no. 52, pp. 295–308, 1983.

[9] C. Hofer, *Measuring Organizational Performance*, Edward Elgar Publishing, 2006.

[10] M. Feurer, K. Eggensperger, S. Falkner, M. Lindauer, and F. Hutter, "Auto-sklearn 2.0: Hands-free automl via meta-learning," *Journal of Machine Learning Research*, vol. 23, no. 261, pp. 1–61, 2022.

[11] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, The MIT Press, 2012.

[12] F. Hutter, L. Kotthoff, and J. Vanschoren, *Automated Machine Learning: Methods, Systems, Challenges*, 1st ed. Springer Publishing Company, 2019.

[13] C. Molnar, *Interpretable Machine Learning*, 2nd ed. Munich: Christoph Molnar, 2022.

[14] E. Altman, M. Iwanicz-Drozdowska, E. K. Laitinen, and A. Suvas, "Financial distress prediction in an international context: A review and empirical analysis of Altman's Z-score model," *Journal of International Financial Management & Accounting*, vol. 28, no. 2, pp. 131–171, 2017.

[15] N. Bărbuță-Mișu and M. Madaleno, "Assessment of bankruptcy risk of large companies: European countries evolution analysis," *Journal of Risk and Financial Management*, vol. 13, no. 3, pp. 1–28, 2020.

[16] J. Sinarti and T. M. Sembiring, "Bankruptcy prediction analysis of manufacturing companies listed in Indonesia stock exchange," *International Journal of Economics and Financial Issues*, vol. 5, pp. 354–359, Sep. 2015.

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

[18] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer, 2009.

[19] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc. KDD '16: the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[20] S. González, S. García, J. Del Ser, L. Rokach, and F. Herrera, "A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities," *Information Fusion*, vol. 64, pp. 205–237, 2020. doi: 10.1016/j.inffus.2020.07.007

[21] M. Odom and R. Sharda, "A neural network model for bankruptcy prediction," in *Proc. IJCNN International Joint Conference on Neural Networks*, IEEE, 1990, vol. 2, pp. 163–168.

[22] S. Karlos, S. Kotsiantis, N. Fazakis, and K. Sgarbas, "Effectiveness of semi-supervised learning in bankruptcy prediction," in *Proc. 2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA)*, IEEE, 2016, pp. 1–6.

[23] Z. Xiaojin and A. Goldberg, *Introduction to Semi-Supervised Learning*, San Rafael, USA: Morgan & Claypool Publisher, 2009.

[24] S. Tomczak. (2016). Polish companies' bankruptcy data set. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data

[25] H. Jin, Q. Song, and X. Hu, "Auto-keras: An efficient neural architecture search system," in *Proc. the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ACM, 2019, pp. 1946–1956.

[26] S.-A. Alexandropoulos, C. Aridas, S. Kotsiantis, and M. Vrahatis, "A deep dense neural network for bankruptcy prediction," in *Proc. EANN 2019, Engineering Applications of Neural Networks*, 2019, pp. 435–444.

[27] S. Chen, B. Mulgrew, and P. M. Grant, "A clustering technique for digital communications channel equalization using radial basis function networks," *IEEE Trans. on Neural Networks*, vol. 4, pp. 570–578, July 1993.