Journal of Supply Chain Management Science

https://journals.open.tudelft.nl/jscms

http://dx.doi.org/10.18757/jscms.2022.6243

Balancing and variable reduction of firm bankruptcy data

David L. Olson [a*], Bongsug (Kevin) Chae [b],

[a] *James & H.K. Stuart Chancellor's Chair, Department of Supply Chain Management and Analytics, University of Nebraska-Lincoln*

[b] *Department Jon Wefald Chair in Business, Department of Management, College of Business Administration, Kansas State University*

Article history: received 09-01-2022, accepted 23-05-2022, published 30-07-2022

Abstract – Financial stress experienced by supply chain elements causes stress to all members. Predictive data mining is a common tool for predicting bankruptcy. Bankruptcy often involves highly imbalanced datasets with a large number of potential variables, with bankrupt firms being by far the minority case. This study uses data from four studies of firm bankruptcy and examines the impact of data balancing and variable selection on model accuracy. The models used are random forest and gradient boosting based on decision trees, logistic regression, neural networks, and support vector machines. Two machine learning methods are used to trim the number of variables. Stepwise regression and entropy from decision trees are used to generate reduced variable sets. The complexity parameter was used to set levels on number of variables using the entropy (decision tree) option. The impact of reducing variables is examined. Error metrics used were type I and type II error (sensitivity and specificity), overall average error (accuracy), and area under the recall curve (AuC). The average error of extreme gradient boosting and random forest models was found to be better than support vector machines, which had a slight advantage over logistic regression and neural networks. Variable reduction was found to lead to mixed results with respect to relative accuracy. Overall accuracy increased with slight reduction in the number of variables (using stepwise regression), but deteriorated as the number of variables was reduced to the smaller number of variables. The experiments into balancing found that unbalanced data had high error rates, which dropped a great deal with even 10 percent balancing, but balancing beyond 10 percent was found to provide little additional accuracy.

Keywords: Data; Data Pre-Processing; Model Assessment; Variable Importance; Modelling; Data Mining

1. Introduction

Bankruptcy is important for any business, to include supply chain contexts. The interrelationship of supply chain members creates interdependencies where the bankruptcy of one supply chain member can create problems for its supply chain partners. Kolay et al. (2016) differentiated between suppliers facing purely financial distress due to cash flow issues versus economic distress due unprofitable operations. Financially distressed companies were found to be highly likely to reorganize with few spillover costs. Economically distressed firms were found to suffer large losses in market value created costs of replacing bankrupt customers. The research of Hua et al. (2011) indicated that supply chain members could hedge operational risk through financial decisions. Xu et al. (2010) investigated methods to reduce the probability of bankruptcy through coordination of supply chains through information sharing and vendor-managed inventory (VMI). Because manufacturers in VMI and retailers in information sharing might gain less benefit or even suffer losses from cooperation, additional incentive measures were suggested to encourage their efforts at coordination. Sun et al. (2012) evaluated the effectiveness of contractual incentive schemes such as revenue sharing, price discounts, and quantity flexibility contracts.

Should a supply chain member undergo financial distress, Yang et al. (2015) identified three effects that could change firm behavior. The predation effect would result in n creased competition prior to potential bankruptcy as

* Corresponding author. Email address: david.olson@unl.edu

the non-distressed competitor would seek first-mover advantage to increase pressure on the distressed firm. In a more cooperative environment, a supplier might react to bail out the distressed firm through concessions to preserve competition and improve supply chain efficiency. The abatement effect would find a supplier deliberately abetting the competitor's predation, placing increased pressure on the distressed firm. Overall, these effects lead to conditions where a firm's bankruptcy potential can hurt its competitors and benefit its suppliers as well as customers.

Models of supply chain financial risk include data envelopment analysis (DEA) and simulation of outsourcing risks to include foreign exchange, product failure, organizational failure, and political risk. Asheyeri et al. (2014) developed an optimization model to streamline supply chain networks to balance survival probabilities along with long-term profitability.

Thus, financial risk is important to supply chain networks. Data mining classification provides a tool to aid in firm financial bankruptcy. Khemakhem and Boujelbene (2018) examined widely used classification algorithms (decision trees and neural networks) in unbalanced bankruptcy analysis of financial firms. If extremely imbalanced data is not balanced, the minority cases (usually bankrupt firms) are disregarded as a very high accuracy rate is obtained by defaulting to assigning all cases to the non-bankrupt category. Such models are degenerate, providing no help in analyzing cases. We will review basic balancing methods, and our experiments seek to identify the impact of various levels of balancing.

Wang et al. (2017) noted that variable selection is crucial in dealing with high-dimensional data. Bankruptcy classification models typically contain a large number of variables. Three of the datasets we analyze have this characteristic (one with 64 available independent variables; the other two with 96 and 65 respectively). The aim of variable reduction is to reduce irrelevant or redundant information content, focusing on the core information enabling discriminant power in a given dataset (Dash and Liu 1997, Guyon and Elisseeff 2003). Lin et al. (2012) reviewed feature selection methods. It would be attractive to have automated methods (machine learning) to select independent (explanatory) variables. Two automated variable selection methods are stepwise regression, and entropy (used by decision tree algorithms).

Data mining is widely applied to classification problems. One area that has received a great deal of study is bankruptcy prediction. There are three features of this problem class that we examine: balancing datasets that have one class very large relative to the other and selecting variables from the usually large set of financial ratios and other variables commonly found in bankruptcy datasets. We also compare standard data mining classification algorithms applied to bankruptcy data.

This paper examines the impact of data balancing and variable selection on model accuracy in four datasets involving financial failure. A number of classification algorithms are appropriate for analyzing financial failure. We confirm the widely understood relative advantage of random forests and extreme boosting. We also examine the impact of balancing datasets. Financial failure hopefully is rare within such datasets, leading to imbalanced outcomes. We find that even low levels of balancing help a great deal in improving model accuracy.

This introduction discussed the importance of financial failure in supply chains. Section 2 discusses data mining issues related to financial failure datasets. Section 3 presents the four datasets. Section 4 reviews algorithms. Section 5 gives our results, and section 6 our conclusions.

2. Research issues

2.1. Variable selection

Frequently, data mining analysis involves data with a large number of variables. Often some variables contribute more noise than value in predicting the dependent variable. Chan et al. (2007) identified several applications where too many features generated noise, and ignoring redundant variables improved organizational failure prediction, and reducing the set of independent variables can result in better prediction. Tang et al. (2014) surveyed different types of feature selection using filter and embedded methods for classification problems. Tsai (2009) discussed feature selection methods such as stepwise regression and correlation matrix for bankruptcy prediction.

Often independent variables may be highly correlated, containing overlapping information which may distort regression coefficients. Usually, a small subset of these explanatory (independent) variables provide the bulk of the predictive power of a model. It would be beneficial to identify the kernel of explanatory variables that give

most if not all of a model's accurate prediction. The reasons to be able to trim larger sets of variables in data mining to smaller subsets include:

1. Easier analysis
2. Shorter training time
3. Avoiding the curse of dimensionality, which makes problems more complex
4. Reducing overfitting, by reducing variance

Analysts of financial modeling seem to be able to come up with a plethora of ratios to measure firm performance. This creates some problems for regression, in that the correlation across these variables makes it difficult to assess relative contribution of each variable. Stepwise regression is a means to apply machine learning characteristics to logistic regression (which is used in bankruptcy analysis, with a binary output variable). Primarily, however, most of the predictive power comes from a subset of available variables. Ideally a parsimonious model with good predictive power that is easy to implement is preferable (Cui et al. 2020).

The ideal way to select variables is based on deep understanding of the problem. This is the approach of classical statistics, where the ideal regression model is based on selecting variables known to have strong relationships with the dependent variable. This amounts to understanding the system and selecting independent variables the human analyst expects to have a strong relationship with the dependent variable. Machine learning, conversely, uses statistical measures to select variables. Zeng et al. (2009) gave the general idea of attribute relevance analysis to quantify attribute relevance for a given class. Information gain, as reported by random forest in Rattle, is a commonly used measure. Zeng (2017) conversely suggested using boosting to select relevant variables. Both random forest and boosting can be based on decision trees. Thus, a machine learning method to select variables is to apply a decision tree, which has a complexity parameter based on entropy levels to select variables.

The traditional regression method of variable reduction is stepwise, adding variables by their contribution to explaining variance in the dependent variable (Foster 2004). This is done with partial correlation, which considers overlapping content of potential independent variables. The process can start with selecting that independent variable with the highest correlation with the dependent variable. Then, given use of that variable, the partial contribution of independent variables are analyzed to iteratively proceed until added improvement in fit falls below some stated level.

We will apply both the decision tree approach based on entropy as well as stepwise regression as machine learning tools to select explanatory variables. The number of variables selected can be controlled in decision trees through the complexity parameter the minimum improvement in the model needed at each node. We used three levels of this complexity parameter.

2.2. Balancing

In many real applications, imbalanced class distributions are present, confusing many machine learning algorithms (Feng et al. 2019). A major problem in many of these applications is that data is often skewed (Olson 2004). For instance, insurance companies hope that only a small portion of claims are fraudulent. Physicians hope that only a small portion of tested patients have cancerous tumors. Banks hope that only a small portion of their loans will turn out to have repayment problems. Tiwari et al. (2017) found that the presence of imbalanced datasets negatively impacts classifier learning. The most common method to deal with imbalanced data is resampling. That study found that intermediate levels of balancing worked best in their experiments. Babu and Ananthanarayanan (2017) also found that existing classifiers performed poorly on imbalanced datasets.

Zoričák et al. (2020) reviewed balancing approaches, categorized into preprocessing techniques and use of learning algorithms capable of dealing with imbalanced data. These learning algorithms include ensemble classifiers, cost-sensitive learning, and one-class learning. Comprehensive reviews were given by Haibo He and Garcia (2009), Krawczyj (2016), and Haixiang et al. (2017). Preprocessing techniques include undersampling (which makes computation easier, but reduces data content), oversampling (which increases computational burden). One oversampling method is SMOTE (synthetic minority over-sampling technique – Chawla et al. 2020). SMOTE randomly draws observations from the smaller set of outcomes, which can provide exactly the degree of balance desired, but risks some unintentional bias. Singh et al. (2021) found oversampling to outperform undersampling for the methods that they used, which included random forests and gradient boosting. We prefer to replicate the entire minority set multiple times to roughly attain the level of balance desired. In our experiments,

we us 10%, 20%, 30%, 40% and 50% levels measuring the proportion of failed cases to total cases for balancing with the intention of looking at relative elimination of data problems, and relative accuracy performance.

2.3. Process

Because financial failure is heavily weighted with failing firms usually in the minority, dataset balancing is important, and we use five levels reflecting the proportion of failed firms. Every data mining application involves selecting a training set and applying it to a test set. We used a common test set for each of the four national datasets. The third step of our process was to select independent variables. We used stepwise regression as one method, and used single decision tree models with three levels of the complexity parameter as another means to select variables. Single decision trees were not included as they were used to select variables, and random forests and extreme gradient boosting are ensembles of decision tree algorithms. We then applied the five classification algorithms studied. Neural network and support vector machine models can be refined for specific datasets, but that involves a complex and time-consuming process. We run basic neural network models with some variation in node levels, and SVM models with different kernels. The last step of our process was to measure errors. Thus, the five steps of our process:

1. Generate balanced datasets (five levels: 10%, 20%, 30%, 40%, 50%)
2. Partition data (80% training, 0 validation, 20% testing)
3. Identify variables (stepwise, decision trees with complexities of 0.01, 0.02, 0.03)
4. Run algorithms (random forest, gradient boosting, logistic regression, neural network, SVM)
5. Measure errors (sensitivity, specificity, overall error, AuC)

The US dataset was generated as a balanced dataset, so did not need to be balanced. This yielded six datasets from the Poland data, six from the Taiwan data, six for the Slovak data, and one US dataset.

3. Data

We utilized four datasets related to firm bankruptcy:

3.1. Poland Data

Zięba et al. (2016) provided a database of 10,000 observations over 64 financial measures related to firms in Poland. This dataset was highly imbalanced, with 203 bankrupt and 9797 not. Due to data availability, they obtained data on the bankrupt firms over the period 2007-2013 and 2000-2012 for those firms still operating. The 64 financial indicators they selected were determined by availability of data and intensity of occurrence. They tested 16 algorithms, with multiple versions of decision trees, logistic regression, boosting, support vector machines, and random forests.

3.2. Taiwan Data

Liang et al. (2016) presented 6819 observations over 95 explanatory variables for firm bankruptcy in Taiwan. This dataset was also highly imbalanced, with 220 bankrupt and 6599 not. They used three filtering methods (stepwise discriminant analysis (Fisher 1936), stepwise logistic regression (Fisher and Yates 1963), and t-testing (Zimmerman 1997) as well as two wrapper methods (genetic algorithm – Holland 1975; and recursive feature elimination – Guyon et al. 2002). The prediction models used were decision trees, neural networks, support vector machines, naïve Bayes, and K-means clustering.

3.3. Slovak Data

Drotár et al. (2019) presented bankruptcy prediction data for 2013-2016 for Slovak companies in agriculture, construction, manufacturing and retail. This dataset was extremely imbalanced, with 63 bankrupt and 25932 not. The dataset contained 21 distinct financial ratios, along with other variables yielding a total of 63 variables. This data was analyzed from an economic perspective in Zoričák et al. (2020).

3.4. U.S. Data

Olson et al. (2012) used data over the period 2005-2009 of US firms, balancing bankrupt with not bankrupt. This data involved 100 U.S. firms that underwent bankruptcy. All of the sample data are from U.S. companies. About 400 bankrupt company names were obtained using google.com. The companies bankrupted during January 2006 and December 2009 were retained, since it was expected that different results would be obtained after that economic crisis. Financial data ratios during January 2005 to December 2009 were obtained from the Compustat database, yielding the explanatory variables available to predict company bankruptcy. The factors collected were based on the literature. The dataset consists of 1,321 records with full data over 19 attributes, as shown in Table 1. The outcome attribute in bankruptcy has a value of 1 if the firm went bankrupt by 2011 (697 cases) and a value of 0 if it did not (624 cases).

Table 1 recaps the three datasets showing the ratio of bankrupt to total firms. The ratio of bankrupt to total changes as when more bankrupt cases are added, the total number of variables increases.

Table 1. Dataset parameters

Dataset	Explanatory Variables	OK	Bankrupt	Ratio bankrupt/total
Poland	64	9797	203	0.020
		"	1015	0.094
		"	2436	0.199
		"	4263	0.303
		"	6496	0.399
		"	9684	0.497
Taiwan	95	6599	220	0.032
		"	660	0.091
		"	1760	0.211
		"	2860	0.302
		"	4500	0.405
		"	6700	0.504
Slovak	63	25932	189	0.007
		"	2835	0.098
		"	6426	0.199
		"	11151	0.304
		"	17199	0.399
		"	25893	0.500
US	14	624	697	0.528

Balancing yielded different models. Decision trees were used to generate trimmed datasets. The number of rules and variables can be controlled through the complexity parameter. We used three complexity levels: 0.01, 0.02, and 0.03.

4. Algorithms

Zoričák et al. (2020) thoroughly reviewed modeling of bankruptcy prediction. Kumar and Ravi (2007) divided these methods into statistical (regression-based) and intelligent (neural networks, decision trees, support vector machines, and case-based reasoning – which amounted to clustering). While some studies have found clustering to be comparable with other machine learning methods for prediction (Li and Sun 2009, Ahn and Kim 2009, Chuang 2013, Sartori et al. 2016), we agree with Jo et al. (1997) who argued that it was unsuitable for bankruptcy prediction in part because clustering output may not match a clean binary outcome on bankruptcy. Hu et al. (2004) concluded that statistical approaches such as logistic regression in classification modeling are likely to be negatively impacted by unequal sample size (imbalanced) and tend to assign all cases to the majority. We also find that to be true for SVM and neural network models. Kumar and Ravi concluded that intelligent methods outperform

statistical methods, especially in the presence of many variables with complex relationships. Given the highly imbalanced nature of bankruptcy data, some machine-learning approaches were expected to be severely limited. We will balance data systematically to look at which machine learning models are most affected by imbalance.

A decision tree model is one of the most common data mining models. It is popular because the resulting model is easy to understand. The algorithms use a recursive partitioning approach. We used the Rattle rpart package, comparable to CART and ID3/C4. Variables are selected using entropy, a machine learning technique.

A random forest is an ensemble (i.e., a collection) of un-pruned decision trees. Ensemble models are often robust to variance and bias, improving these characteristics in single decision tree models. Random forests are often used when we have large training datasets and particularly a very large number of input variables (hundreds or even thousands of input variables). Because the algorithm iteratively creates subsets of available variables, it is efficient for datasets with large numbers of variables.

Gradient boosting is another ensemble of decision trees. The basic idea of boosting is to associate a weight with each observation in the dataset. A series of models are generated and the weights are increased (boosted) if a model incorrectly classifies the observation. The resulting series of decision trees form an ensemble model. The extreme gradient boosting algorithm builds a gradient boosting model which is an optimal approach to boosting.

A linear regression model is the traditional method for fitting a statistical model to data. It is appropriate when the target variable is numeric and continuous. We used stepwise regression as a means to select variables. Models were built using logistic regression as the output variable for bankruptcy was binary.

Neural network models are based on the idea of multiple layers of neurons connected to each other, feeding the numeric data through the network, combining the numbers, to produce a final answer. Neural network models are well-suited when models require knowledge that is difficult to specify ahead of time, or when data contains high degrees of nonlinearity (Hu et al. 2004). We used the parameter of 10 hidden layers.

Support vector machines (SVM) search for support vectors, data points that are found to lie at the edge of an area in space which is a boundary from one class of points to another. In the terminology of SVM we talk about the space between regions containing data points in different classes as being the margin between those classes. The support vectors are used to identify a hyperplane) that separates the classes. Gür Ali and Yaman (2013) reviewed literature finding that support vector models were often useful in classification models with few input variables and observations, although when applied to large scale problems, memory and time requirements were problematic. Further, the predictive accuracy of SVM models has been found to be negatively affected by irrelevant and redundant variables. Ghaddar and Naoum-Sawaya (2018) developed an iterative variable selection method for SVM models.

5. Results

The process was to partition data with 80 percent used for training and 20 percent used for testing. The test data using the base data was saved and used as a common test set for all models for that dataset. Decision trees were run with parameter settings of 20 for minimum split, 20 for maximum depth, and 7 for minimum bucket. Complexity was varied using levels of 0.01, 0.02 and 0.03 as part of the experiment. Random forests default settings of 500 trees and 10 variables were used. Extreme gradient boosting was used with maximum length of 30, learning rate 0.3, threads set at 2, iterations set at 50, and a binary logistic objective. The neural network models setting for hidden layers was 10. The radial basis kernel was used for SVM models. Except for complexity levels, the other parameter settings were default. It is noted that neural network and SVM models can be fine-tuned to perform better for each data set, but this takes significant exploration, while random forests and extreme boosting yield stable output without such effort due to their structure using multiple runs.

The Taiwan data included only three of the ten variables with the highest correlation with bankruptcy, while two appeared in the six Poland models. The Poland and Slovak datasets were more consistent in reappearance of variables across balanced levels. In all datasets variables with practically zero correlation with bankruptcy often appeared in the models.

For this type of data, with many variables, ensemble models such as random forests or gradient boosting are supposed to do better (Cui et al. 2020). For balanced data, they clearly did. In general, as the number of bankrupt cases increased with balancing, logistic regression and neural network results had more errors. The type of error varied – sometimes type I getting worse, sometimes type II. SVM models were relatively accurate, in line with single decision trees. But random forest and gradient boosting models clearly were better.

5.1. Trimming

We applied balancing to the unbalanced datasets (from Poland, Taiwan and Slovakia), generating new datasets with 10%, 20%, 30%, 40% and 50% proportions of bankrupt cases. The U.S. dataset was balanced by initial design. We applied decision trees with varying entropy (complexity parameter) and stepwise regression to select variables. Table 2 gives averages as proportion of the full number of variables available:

Table 2. Relative proportion of variables by variable generation method

Variable generation method	Average # Variables
Full	1
Stepwise	0.485
Entropy.01	0.144
Entropy.02	0.069
Entropy.03	0.045

Table 2 shows that the variable reduction methods were effective in trimming the number of variables used. Table 3 gives the proportion of variables (relative to the full model) by balancing level.

Table 3. Relative number of variables by balancing level and variable generation method

Balancing level	Step	Entropy.01	Entropy.02	Entropy.03
Base	0.329	0.176	0.109	0.065
10%	0.673	0.212	0.105	0.072
20%	0.504	0.130	0.067	0.045
30%	0.585	0.122	0.049	0.021
40%	0.498	0.134	0.031	0.028
50%	0.367	0.086	0.040	0.033

Balancing level tended to increase the number of variables selected initially, but with little consistent trend.

The algorithms used in addition to decision trees were random forests, gradient boosting, logistic regression, neural networks, and support vector machines. Anzanello et al. (2012) addressed the accuracy measures of sensitivity, specificity, and overall accuracy. Dag et al. (2016) applied sensitivity analysis in comparing classification models using accuracy, sensitivity, specificity, and information gain measures. We compared relative overall accuracy, as well as the maximum of sensitivity (type I) and specificity (type II) errors over the five algorithms. Table 4 gives average errors obtained by algorithm. Figures 1 through 4 display this data visually.

Table 4. Average Errors – Balancing Level versus Algorithm

Sensitivity	Random forest	Gradient boost	Log Regression	Neural net	SVM
Base	0.237	0.457	0.694	0.625	0.722
10%	0.018	0.016	0.695	0.628	0.538
20%	0.016	0.002	0.620	0.461	0.328
30%	0.002	0.007	0.488	0.473	0.202
40%	0.003	0.003	0.339	0.254	0.117
50%	0.002	0.002	0.315	0.369	0.061
Specificity	**Random forest**	**Gradient boost**	**Log Regression**	**Neural net**	**SVM**
Base	0.073	0.023	0.058	0.129	0.035
0.1	0.059	0.010	0.045	0.043	0.014
0.2	0.002	0.007	0.115	0.127	0.036
0.3	0.004	0.013	0.163	0.142	0.064
0.4	0.007	0.016	0.194	0.214	0.104
0.5	0.004	0.024	0.194	0.182	0.161
Overall	**Random forest**	**Gradient boost**	**Log Regression**	**Neural net**	**SVM**
Base	0.035	0.033	0.081	0.129	0.048
0.1	0.008	0.004	0.058	0.054	0.025
0.2	0.003	0.010	0.121	0.129	0.040
0.3	0.006	0.013	0.164	0.096	0.066
0.4	0.055	0.015	0.212	0.214	0.100
0.5	0.028	0.021	0.195	0.186	0.121
AuC	**Random forest**	**Gradient boost**	**Log Regression**	**Neural net**	**SVM**
Base	0.949	0.957	0.832	0.745	0.828
0.1	0.998	0.999	0.803	0.804	0.904
0.2	0.999	0.999	0.851	0.781	0.929
0.3	0.999	0.998	0.827	0.790	0.919
0.4	0.999	0.999	0.825	0.776	0.888
0.5	1.000	0.999	0.799	0.856	0.956

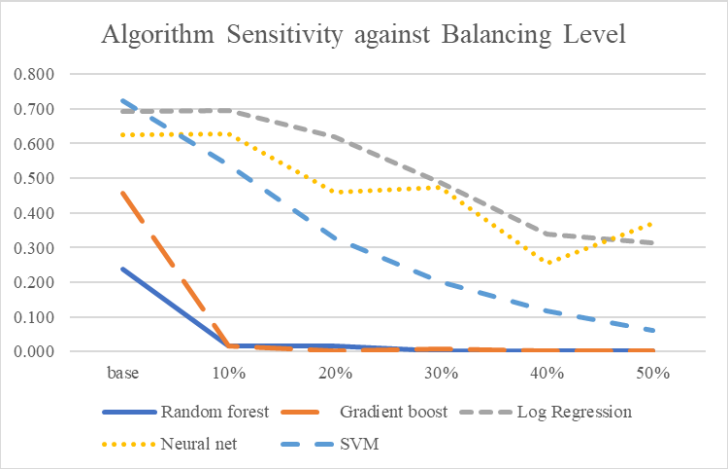

Figure 1. Sensitivity by Algorithm

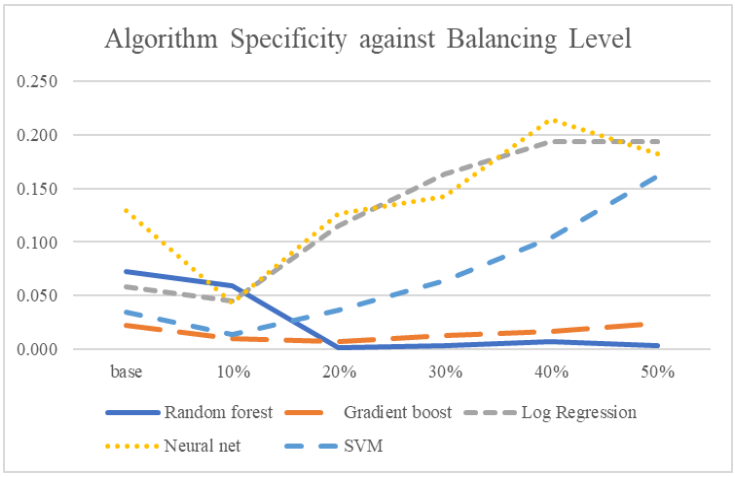

Figure 2. Specificity by Algorithm

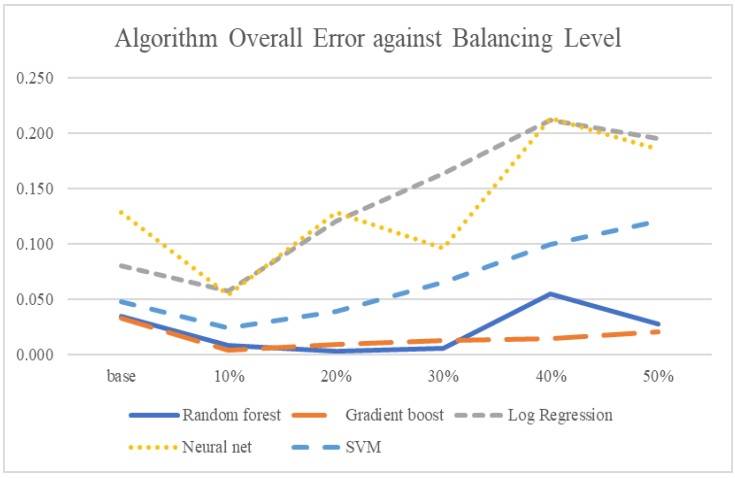

Figure 3. Overall Error by Algorithm

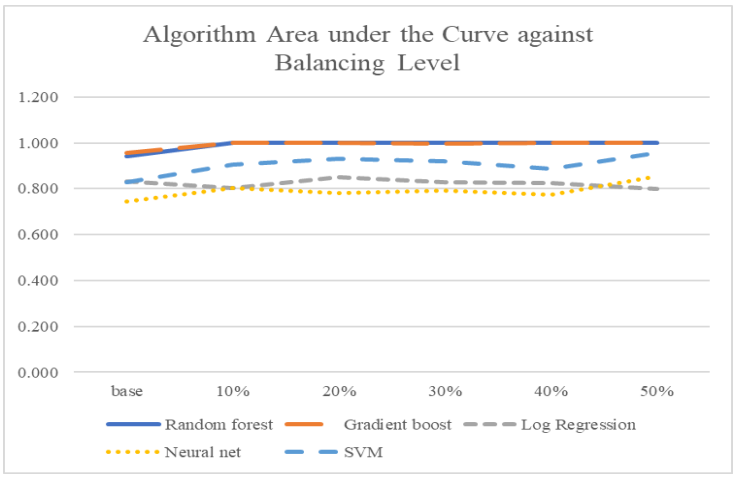

Figure 4. Area Under the Curve by Algorithm

Random forests and boosting ensembles were clearly better than the other algorithms. The gradient boosting method had a slight advantage over random forests in AuC with no balancing, but these algorithms performed nearly identically with balanced data sets. They were also usually better on specific error measures. Support vector machines was a clear third in average error, both for of type I and type II as well as overall average error. Logistic regression and neural networks were less accurate as measured by errors on holdout data. We understand that these last three algorithms include parameters that can be adjusted so that they can perform much better, but that takes quite a bit of searching. It seems more efficient to go to extreme boosting and random forests, which in effect do some of that parameter searching automatically (a form of machine learning).

Degeneracy was identified when models assigned all forecasts to the majority class. This occurred in the Poland dataset for SVM models with the base model, as well as balancing at 10 percent, 20 percent, and 30 percent. Degeneracy occurred in the Taiwan dataset for neural network models balanced at 20 percent when complexity was set at 0.01 and 0.03 (but not of 0.02). In the Slovak data, there was degeneracy for the logistic regression model for data balanced at 10 percent with complexity set at 0.02, and for SVM models in the unbalanced dataset for the base (unbalanced) data, and data balanced at 10 percent, 20 percent, and 30 percent. Thus SVM models had degeneracy occur 8 times out of 95, neural networks twice, and logistic regression once. Random forest and extreme boosting models had no degenerate models. Looking at relative error, Table 5 displays error measure results by variable reduction method:

Table 5. Average measures by Variable Selection Method

Method	Sensitivity	Specificity	Overall error	AuC
Full	0.225	0.078	0.077	0.902
Step	0.309	0.089	0.076	0.921
Entropy.01	0.292	0.059	0.065	0.894
Entropy.02	0.369	0.064	0.068	0.894
Entropy.03	0.383	0.084	0.089	0.884

Data in Table 5 shows that the overall error varied, but with little difference. Area under the curve results, however, showed more difference, with the set of variables generated by step-wise regression having better results over all of the other four methods, to include the base full set of variables. This indicates that reducing the variables slightly seems to be better than reducing the variable set too much.

5.2. Balancing

Average performance by balancing is given in Table 6:

Table 6. Scores by balancing level

Sens	**Full**	**Step**	**Ent.01**	**Ent.02**	**Ent.03**
Base	0.4797	0.6668	0.6039	0.6287	0.65205
10%	0.400579	0.432632	0.397842	0.488053	0.478947
20%	0.173579	0.236684	0.384684	0.369263	0.448842
30%	0.173684	0.250684	0.222526	0.302211	0.419632
40%	0.148	0.138421	0.151526	0.288632	0.241579
50%	0.145474	0.21	0.075526	0.232737	0.135579
Spec	**Full**	**Step**	**Ent.01**	**Ent.02**	**Ent.03**
Base	0.0924	0.07345	0.0455	0.02405	0.08235
10%	0.029421	0.094368	0.029316	0.008789	0.062842
20%	0.086211	0.074789	0.019053	0.072053	0.023263
30%	0.061895	0.102684	0.061737	0.094	0.089632
40%	0.110632	0.102947	0.079421	0.082053	0.173211
50%	0.093105	0.105421	0.138842	0.130684	0.167158
Overall	**Full**	**Step**	**Ent.01**	**Ent.02**	**Ent.03**
Base	0.0697	0.04795	0.06265	0.0435	0.1019
10%	0.037211	0.061895	0.037526	0.016632	0.070421
20%	0.089316	0.074053	0.026158	0.076526	0.026579
30%	0.063842	0.102895	0.063158	0.098	0.056421
40%	0.123789	0.102474	0.083947	0.099526	0.151474
50%	0.095053	0.105316	0.138211	0.101526	0.181895
AuC	**Full**	**Step**	**Ent.01**	**Ent.02**	**Ent.03**
Base	0.88315	0.86845	0.8828	0.8487	0.82895
10%	0.859737	0.885	0.873	0.882895	0.822105
20%	0.901579	0.919263	0.912105	0.891316	0.844737
30%	0.899632	0.915737	0.894842	0.863105	0.879789
40%	0.888684	0.935579	0.828895	0.873263	0.885158
50%	0.891789	0.932684	0.897263	0.885842	0.890684

This information is shown graphically in Figures 5 through 8:

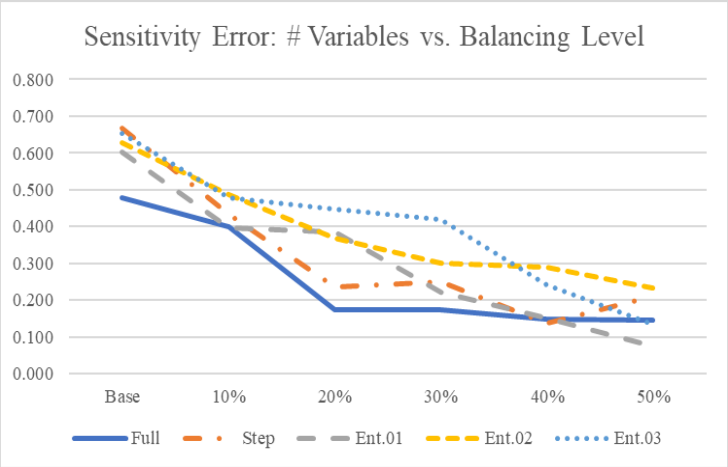

Figure 5. Sensitivity Error by Level of Balancing

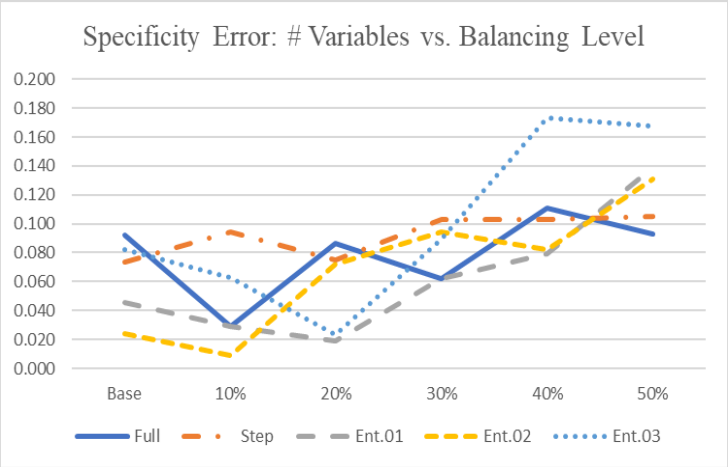

Figure 6. Specificity Scores by Level of Balancing

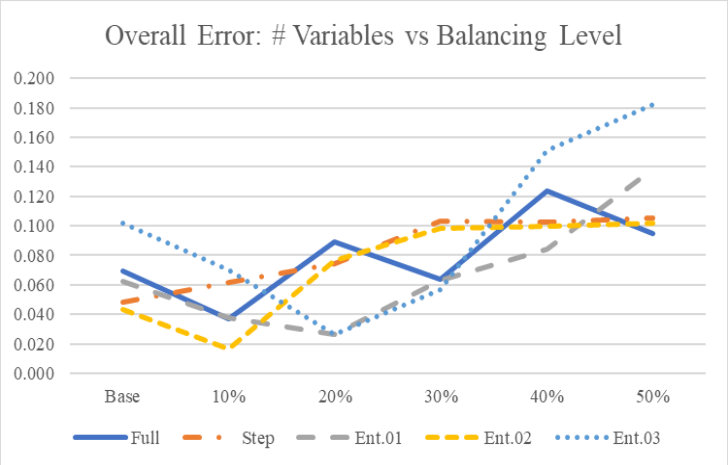

Figure 7. Overall Accuracy by Level of Balancing

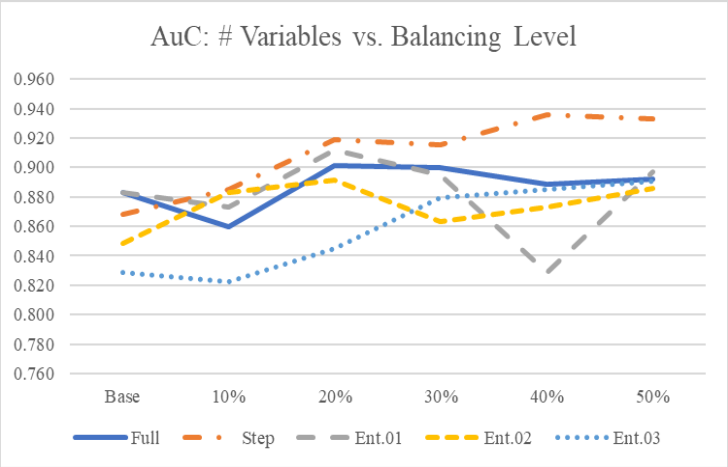

Figure 8. Area under the Curve by Level of Balancing

Viewing Figures 5 and 6, we see that the bias in bankruptcy data gets more extreme with smaller datasets. Sensitivity (type II error) improved with smaller datasets, while specificity (type I error) got worse. Overall error (Figure 7) was best with the Step datasets, but got worse as datasets were further trimmed. The results for Area under the curve were best for Step data, with full datasets next, and generally decreasing accuracy with smaller datasets. We conclude that some trimming of variables is beneficial, but too much is counterproductive.

5. Conclusions

Our results found that in general, the more variables available, the less error, although trimming a few variables improved accuracy performance. The main benefit of balancing was to avoid degenerate models that were obtained with neural network and SVM models (and one case with logistic regression). Our data found little added benefit from balancing to error being more than 10% of total cases. Among methods, we found extreme boosting to be the most beneficial, with random forest models close in relative accuracy. These are models that internally manipulate multiple models. SVM models were next in performance, followed by linear regression and neural networks, the latter two methods yielding very similar results. We note that SVM, linear regression, and neural networks can be fine-tuned to specified data sets, which we did not do, but this fine tuning would take significant computational effort which is not needed by extreme boosting and random forest models.

Our basic conclusions can be itemized:

1. Balancing highly imbalanced datasets has advantage, especially in avoiding degenerate models (which predict no bankruptcy). However, complete balancing is not needed – ten percent balancing gains most of the advantage of balancing.

2. Extreme boosting and random forest models were clearly more accurate in our results. Support vector machines had some advantage over linear regression and neural networks, recognizing that we did not fine tune these last three models. To do so, however, would create more computational burden.

3. Variable selection has some benefit, although there is a slight cost in reduced accuracy. The smaller number of variables reduced from stepwise variable selection improved accuracy slightly. The benefits of trimming datasets is that results are much more focused and clearer to apply, at a small cost in accuracy.

References

Ahn, H., and Kim, K. J. (2009) 'Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach', *Applied Soft Computing*, 9, 599-607.

Anzonello, M. J., Albin, S. L., and Chaovalitwongse, W.A. (2012) 'Multicriteria variable selection for classification of production batches', *European Journal of Operational Research*, 218, 97-105.

Babu, S., and Ananthanarayanan, N. R. (2017) 'EMOTE: Enhanced Minority Oversampling Technique', *Journal of Intelligent & Fuzzy Systems*, 33, 67-78.

Chan, A. B., Vasconcelos, N., and Lanckriet, G. R. (2007) 'Direct convex relaxations of sparse SVM', *Proceedings of the 24th International Conference on Machine Learning*, Corvalis, United States, 145-153.

Chuang, C. I. (2013) 'Application of hybrid case-based reasoning for enhanced performance in bankruptcy prediction', *Information Science*, 236, 174-185.

Cui, H., Rajagopalan, S., and Ward, A.R. (2020) 'Predicting product return volume using machine learning methods', *European Journal of Operational Research*, 281, 612-627.

Dag, A., Topuz, K., Oztekin, A., Bulur, S., and Megahed, F. M. (2016) 'A probabilistic data-driven framework for scoring the preoperative recipient-donor heart transplant survival', *Decision Support Systems*, 86, 1-12.

Dash, M., and Liu, H. (1997) 'Feature selection for classification', *Intelligent Data Analysis*, 1, 131-156.

Drotár, P., Gnip, P., Zoričák, M., and Gazda, V. (2019) 'Small- and medium-enterprises bankruptcy dataset', *Data in Brief*, 25, 1-6.

Feng, L., Wang, H., Jin, B., Li, H., Xue, M., and Wang, L. (2019) 'Learning a distance metric by balancing KL-divergence for imbalanced datasets', *IEEE Transactions on Systems, Man & Cybernetics: Systems*, 49, 2384-2395.

Fisher, R.A. (1936) 'The use of multiple measurements in taxonomic problems', *Annals of Eugenics*, 7, 179-188.

Fisher, R.A. and Yates, F. (1963). *Statistical tables for biological, agricultural and medical research*. London, United Kingdom: Edinburgh: Oliver and Boyd.

Foster, D.P. (2004) 'Variable selection in data mining: Building a predictive model for bankruptcy', *Journal of the American Statistical Association*, 99, 303-313.

Ghaddar, B., and Naoum-Sawaya, J. (2018) 'High dimensional data classification and feature selection using support vector machines', *European Journal of Operational Research*, 265, 993-1004.

Gür Ali, Ő., and Yaman, K. (2013) 'Selecting rows and columns for training support vector regression models with large retail datasets', *European Journal of Operational Research*, 226, 471-480.

Guyon, I., and Elisseeff, A. (2003) 'An introduction to variable and feature selection', *Journal of Machine Learning Research*, 3, 1157-1182.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002) 'Gene selection for cancer classification using support vector machine', *Machine Learning*, 46, 389-422.

He, H., and Garcia, E. (2009) 'Learning from imbalanced data', *IEEE Transactions on Knowledge Data Engineering*, 21, 1263-1284.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., and Bing, G. (2017) 'Learning from class imbalanced data: Review of methods and applications', *Expert Systems with Applications*, 73, 220-239.

Holland, J.H. (1975). *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. Michigan, USA: University of Michigan Press.

Hu, M. Y., Zhang G. P., and Chen, H. (2004) 'Modeling foreign equity control in Sino-foreign joint ventures with neural networks', *European Journal of Operational Research*, 159, 729-740.

Hua, Z., Sun, Y., and Xu, X. (2011) 'Operational causes of bankruptcy propagation in supply chain', *Decision Support Systems*, 51, 671-681.

Jo, H., Han, L., and Lee, H. (1997) 'Bankruptcy prediction using case-based reasoning, neural networks, and discriminant analysis', *Expert Systems with Applications* 13, 97-108.

Khemakhem, S., and Boujelbene, Y. (2018) 'Predicting credit risk on the basis of financial and non-financial variables and data mining', *Review of Accounting and Finance*, 17, 316-340.

Kolay, M., Lemmon, M., and Tashjian, E. (2016) 'Spreading the misery? Sources of bankruptcy spillover in the supply chain', *Journal of Financial and Quantitative Analysis*, 51, 1955-1990.

Krawczyk, B. (2016) 'Learning from imbalanced data: Open challenges and future directions', *Progress in Artificial Intelligence*, 5, 221-232.

Kumar, P. R., and Ravi, V. (2007) 'Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review', *European Journal of Operational Research*, 180, 1-28.

Li, H., and Sun, J. (2009) 'Predicting business failure using multiple case-based reasoning combined with support vector machine', *Expert Systems with Applications*, 36, 10085-10096.

Liang, D., Lu, C. C., Tsai, C. F., and Shih, G. A. (2016) 'Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study', *European Journal of Operational Research*, 252, 561-572.

Lin, W. Y., Hu, Y. H., and Tsai, C. F. (2012) 'Machine learning in financial crisis prediction: A survey', *IEEE Transactions on Systems, Man and Cybernetics – Part C: Applications and Reviews*, 42, 421-436.

Olson, D. L. (2004) 'Data set balancing', *Chinese Academy of Sciences Symposium on Data Mining and Knowledge Management*, Beijing, China, 1-10.

Olson, D. L., Delen, D., and Meng, Y. (2012) 'Comparative analysis of data mining models for bankruptcy prediction', *Decision Support Systems*, 52, 464-473.

Olson, D. L. and Wu, D. (2011) 'Risk management models for supply chain: A scenario analysis of outsourcing to China', *Supply Chain Management: An International Journal*, 16, 401-408.

Sartori, F., Mazzucchelli, A., and Gregorio, A. D. (2016) 'Bankruptcy forecasting using case-based reasoning: The creperie approach', *Expert Systems with Applications*, 64, 400-411.

Singh, A., Ranjan, R. K., and Tiwari, A. (2021) 'Credit card fraud detection under extreme imbalanced data: A comparative study of data-level algorithms', *Journal of Experimental & Theoretical Artificial Intelligence*, In press.

Sun, Y., Xu, X., and Hua, Z. (2012) 'Mitigating bankruptcy propagation through contractual incentive schemes', *Decision Support Systems*, 53, 634-645.

Tang, J., Alelyani, S., and Liu, H. (2014) 'Feature selection for classification: A review', *Data classification: Algorithms and applications*, 37, 1-29.

Tiwari, A. K., Nath, A., Subbiah, K., and Shukla, K. K. (2017) 'Enhanced prediction for observed peptide count in protein mass spectrometry data by optimally balancing the training dataset', *International Journal of Pattern Recognition & Artificial Intelligence*, 31, 1-15.

Tsai, C. F. (2009) 'Feature selection in bankruptcy prediction', *Knowledge-Based Systems*, 22, 120-127.

Wang, S., Wei, J., and Yang, Z. (2017) 'Discrimination structure complementarity-based feature selection', *Computational Intelligence*, 33, 863-898.

Xu, X., Sun, Y., and Hua, Z. (2010) 'Reducing the probability of bankruptcy through supply chain coordination', *IEEE Transactions on Systems, Man, and Cybernetics – Part C (Applications and Reviews)*, 40, 201-215.

Yang, S. A., Birge, J. R., and Parker, R. P. (2015) 'The supply chain effects of bankruptcy', *Management Science*, 61, 2320-2338.

Zeng, F., Li, L., Li, J., and Wang, X. (2009) 'Research on test suite reduction using attribute relevance analysis', *Proceedings of the 8th IEEE/ACIS International Conference on Computer and Information Science*, Shanghai, China, 961-966.

Zeng, J. (2017) 'Forecasting aggregates with disaggregate variables: Does boosting help to select the most relevant predictors?', *Journal of Forecasting*, 36, 74-90.

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, 93-101.

Zimmerman, D. W. (1997) 'A note on interpretation of the paired-samples t test', *Journal of Educational and Behavioral Statistics*, 22, 349-360.

Zoričák, M., Gnip, P., Drotár, P., and Gazda, V. (2020) 'Bankruptcy prediction for small- and medium-sized companies using severely imbalanced datasets', *Economic Modeling*, 84, 165-176.

Author statement

David L. Olson: Conceptualization, Investigation, Data Curation, Writing – Original Draft, Visualization, Project Administration

Bongsug (Kevin) Chae: Conceptualization, Analysis, Writing – Review & Editing, Visualization