This is an Accepted Manuscript of an article published by Taylor & Francis in INFOR: Information Systems and Operational Research in 2017, available online: https://doi.org/10.1080/03155986.2017.1282290

Data Envelopment Analysis may Obfuscate Corporate Financial Data: Using Support Vector Machine and Data Envelopment Analysis to Predict Corporate Failure for Nonmanufacturing Firms

Xiaopeng Yang[a], Stanko Dimitrov[b]

[a] *Centre for Management of Technology and Entrepreneurship, University of Toronto, 200 College Street, Toronto, ON M5S 3E5, Canada*

[b] *Department of Management Sciences, University of Waterloo, 200 University Avenue West, Waterloo, ON, N2L 3G1, Canada*

Abstract:

Corporate failure prediction has drawn numerous scholars' attention because of its usefulness in corporate risk management, as well as in regulating corporate operational status. Most research on this topic focuses on manufacturing companies and relies heavily on corporate assets. The asset size of manufacturing companies play a vital role in traditional research methods; Altman's Z score model is one such traditional method. However, a limited number of researchers studied corporate failure prediction for nonmanufacturing companies as the operational status of such companies is not solely correlated to their assets. In this paper we use support vector machines (SVMs) and data envelopment analysis (DEA) to provide a new method for predicting corporate failure of nonmanufacturing firms. We show that using *only* DEA scores provides better predictions of corporate failure predictions than using the original, *raw*, data for the provided dataset. To determine the DEA scores, we first generate efficiency scores using a slack-based measure (SBM) DEA model, using the recent three years historical data of nonmanufacturing firms; then we used SVMs to classify bankrupt and non-bankrupt firms. We show that using DEA scores as the only inputs into SVMs predict corporate failure more accurately than using the entire raw data available.

Keywords: support vector machine (SVM); data envelopment analysis (DEA); corporate failure predictions; nonmanufacturing firms; data obfuscation.

1. Introduction

Corporate failure prediction is an attractive research topic in the sense that it can provide useful information about the operational status of a company, and it may affect a management team's decision-making process. Information on corporate stress or failure may also, in turn, affect the stock market, customers' choice, business partners, and even competitors' policy. All of these factors lead to intense research efforts within both industry and academia. However, firms, especially private firms, may not necessarily want to reveal all financial information to an auditor that is required to predict its likelihood of corporate failure. Such financial information, if made publicly available, may provide competitors with a competitive advantage. As such, there are two competing factors in place, one is predicting corporate failure, and the other is having the financial data available to predict corporate failure. In this paper, we show that we a firm may indeed help an auditor determine its likelihood of failure without revealing all of its financial

information via data envelopment analysis (DEA).

A number of methods have are used in corporate failure prediction, most of which use several financial ratios from the financial statements of a company to evaluate the corporate stress or possibility of failure. The methods that are of interest to us are those that use financial ratios and those that use data envelopment analysis. Among all these methods, Altman's method is predominant and referred to in all other studies (Altman E. I., 1968). Altman used multiple discriminant analysis to create a model that utilizes several ratios in a linear formula to generate a score. This score can classify a company into three categories: at the risk of failure, healthy, and the middle status, a "gray area." However, most methods, either Altman's method or other ratio analysis methods, use financial ratios including asset size, and assume it as a crucial factor relative to other factors. For manufacturing companies, this is a valid assumption as many factors need to match the scale of the company asset, such as debt, sales, working capital, earnings, etc., and these factors are important in judging whether a firm may run into stress. In particular, for manufacturing firms where the initial investment occupies a large part of the total asset and is a precondition to ensure other factors are operating properly, discussing the problem of corporate failure prediction without considering assets is meaningless. However, the total assets of a nonmanufacturing firm usually are not decisive since such firms, to enhance their competitiveness, pay more money in working capital such as salary, short-term consumables, etc. to provide better service and make more profit. Therefore, using Altman's traditional method to predict corporate failure for nonmanufacturing firms may result in inaccurate conclusions.

Based on his original model, used mainly for predicting bankruptcy for manufacturing firms, which was named the Altman's Z score (Altman E. I., 1968), Altman then proposed another method that he named the Altman Z'' model (Altman E. I., 2002) to cover the nonmanufacturing industry. Then he assigned appropriate coefficients to variables after determining Z'' score on nonmanufacturing firms in order to allow his previous method to be applicable for both manufacturing and nonmanufacturing companies. Nevertheless, he did not change the status quo, and his method still strongly relies on assets. Unfortunately, most nonmanufacturing companies mainly focus on services, and their most important asset is their people and they do not have a large real asset base (Growth of the Service Sector, 2011). It follows that a new outlet needs to be explored to predict corporate failure for the nonmanufacturing sector. In addition, some firms, private firms in particular, may not want to reveal their operational and financial conditions, due to security and competitive concerns. The main contribution of this study is providing a method to predict corporate failure while preserving firm privacy.

Since first proposed in 1978 by Charnes et al. (1978), DEA has developed into a prevalent non-parametric approach that is used to evaluate the relative efficiencies of a group of peer units which have the same productive process and inputs/outputs, i.e., decision making units (DMUs). As the first DEA model, Charnes, Cooper and Rhodes (CCR) model extended Farrell's (Farrell, 1957) prototype model about technical and allocative efficiency. Following this, DEA became a powerful tool which is active in various research fields such as management, finance, agriculture, military, non-profit organizations and many others (Emrouznejad et al. 2008; Paradi & Zhu, 2013; Liu et al. 2013; Yang & Morita, 2013; Sutton & Dimitrov, 2013).

Comparing to other methods in corporate failure prediction for nonmanufacturing firms, the main benefits to using DEA in our research can be found in the following aspects: (1) It allows us to select inputs/outputs depending on actual needs, which can eliminate or at least mitigate the influence of the asset factor. (2) DEA is easier to use since it is a nonparametric method and users do not need to handle complicated parameters. Meanwhile, DEA offers more objective analysis results. (3) DEA divides attributes into inputs and outputs and relates them to each other. The efficiency score generated based on such an assumption is more informative compared to barely using raw data. It follows that we propose a method combining DEA and support vector machines (SVM) together, which uses the efficiency scores calculated by DEA model to classify healthy and bankrupt firms without using firm generate, raw, data. In particular, we show that if we use *only* the DEA scores associated with firms, we are able to better classify healthy and bankrupt firms, than when using the raw firm data, financial data that is private to each firm. This differs from existing work as existing work considers DEA scores in addition to the raw data or DEA relative to other methods, and not DEA scores only by themselves to classify firms.

The remainder of this article is structured as follows: Section 2 is the literature review of previous studies in predicting corporate failure. Section 3 introduces the DEA model we are using in this research, and how to combine DEA and SVM. Section 4 provides an application about nonmanufacturing firms covering a number of industries. Section 5 summarizes the research and provides additional discussion.

2. Literature Review

A number of methods and related applications are broadly studied in the field of bankruptcy prediction. In order to compare our method with others and make a distinct contribution in this field, we summarize and review the main methodologies in the previously published papers in this section.

2.1. Ratio Analysis Methods

William Beaver proposed a method in 1967 (Beaver, 1967) to predict bankruptcy. In the paper, the author defined failure as "the inability of a firm to pay its financial obligations as they mature" and a financial ratio as "a quotient of two numbers, where both numbers consist of financial statement items." The application in Beaver's study used the data from Moody's industrial manual between 1954 and 1964. For each bankrupt firm from Moody's, a healthy firm with the same asset size in the same industry was matched. Beaver claimed that accurate comparison between firms with different asset sizes could not be made (Alexander, 1949). Based on this assumption, he compiled 30 ratios and picked 14 of them to be the most effective in determining the likelihood of bankruptcy, which are *cash flow/total debt, current assets/current liabilities, net income/total assets, quick assets/current liabilities*, etc. Then he claimed that "*cash flow/total debt"* and "*total debt/total assets"* were the best two indicators for bankruptcy prediction. As such univariate methods neglect many other ratios which might affect the results in estimating the corporate failure, Edward Altman applied the first multivariate approach, multiple discriminant analysis (MDA) (Altman E. I., 1968), to bankruptcy prediction in 1968. At that time, MDA is usually used in classifying an observation into several previously defined groups. Its main merit was allowing for the entire profile of variables to be analyzed simultaneously rather

than individually (Altman E. I., 2002).

Using a similar method to Beaver's, Altman paired the healthy firms with bankrupt ones, and there were 66 corporations half of which were bankrupt and half were non-bankrupt in Altman's study. Eventually, the five most influential ratios, as determined by Altman, in determining the likelihood of bankruptcy, were selected as the main indicators used to predict corporate failure including *working capital / total assets, retained earnings / total assets, earnings before income & taxes / total assets, market value of equity / total liabilities, sales / total assets*. Altman selected the ratios based on: (1) the relative contribution of each individual variable as measured by various potential functions, (2) the inter-correlation between the variables, (3) the predictive accuracy of various profiles and (4) judgment of the analysis (Altman E. I., 1968).

In the same study, Altman next assigned appropriate coefficients to these five ratios and defined the sum of the weighted ratios as the Z score, which relied heavily on the asset size and was considered to be only suitable for the manufacturing industry. Based on Altman's Z score method, a large number of related studies were developed by employing different ratios (Deakin, 1972; Ohlson, 1980; Zmijewski, 1984; Hsieh, 1993; Grice & Dugan, 2001; Shumway, 2001; Grice & Ingram, 2001), of which the majority still focused on manufacturing companies. Then Altman proposed his perhaps lesser known Z'' score method, in which he revised the coefficient and ratio items to make them fit nonmanufacturing industries. Unfortunately, the Z'' score method is still affected by asset size, which motivates us to investigate the corporate failure prediction problem using DEA and SVM in this research.

2.2. Privacy in Data Mining

As we are interested in predicting whether a firm will be bankrupt or not, it is natural to view our study as a data-mining classification problem. In fact, we use support vector machines (SVMs), a type of data-mining classification method, in our study. As we are dealing with financial data, we are not only interested in accurate predictions, but also ensuring that the data we use can be privatized. We are by no means the first to consider the role of privacy in data mining, in fact, the area is richly explored by researchers in privacy in data mining with Vaidya et al. (2006) providing an excellent reference into the area. The field is heavily focused on extracting meaningful conclusions via private data. Works consider clustering (Oliveira & Zaïane, 2004), as well as classification (Lindell & Pinkas, 2000; Agrawal & Srikant, 2000) appear in the literature. In addition to modifying data mining methods in order to come to meaningful conclusions, the works describe methods to obfuscating/privatizing data. As a general simplification, the methods consider adding noise to data such that the original data cannot be determined. To our knowledge, we are the first to consider DEA as a way of obfuscating/privatizing data, and we show in this paper the privatized data is more suitable for predicting the class of the associated variable than the original, raw, data. It is worth noting that Misiunas et al. (2016) use DEA to reduce the number of point (DMUs) considered in an Artificial Neural Network, we on the other hand use DEA to reduce and eliminate dimensions to consider in classification.

2.3. Data Envelopment Analysis and Corporate Bankruptcy

Since first introduced via the CCR model, DEA is now a prevalent method in predicting

corporate stress and is used in many studies (Premachandra et al. 2011; Li et al. 2014; Shetty et al. 2012; Xu & Wang, 2009). Cielen et al. (2004) concluded that DEA and linear programming models can outperform decision tree methods based on the result of comparing the three methods, though the authors did not indicate if DEA is more accurate than linear programming models. On the other hand, Sueyoshi & Goto (2009) proposed DEA-DA (discriminant analysis) based on DEA models and applied it to bankruptcy prediction. Their research showed that DEA-DA is more appropriate for longitudinal data (Sueyoshi & Goto, 2009). Another study integrated rough set theory (RST) into SVM, which is used to increase the accuracy of predicting corporate failure (Yeh et al. 2010). The above research compares various DEA methods to one another. However, to our knowledge none of the research in using DEA for bankruptcy prediction explicitly compare using the same classification method, SVMs in this paper, with different data, raw data or only DEA scores.

In addition to comparing DEA and other methods, most current research shows that DEA is a better method to use for corporate failure prediction. However, an additional key difference between prior work and that presented here is no study covered predicting the failure of nonmanufacturing firms with very small asset sizes besides the research conducted by Paradi et al. (2014). In their study, the DEA scores of the firms in the recent five years are calculated, a cut-off point for each year was also calculated to classify bankrupt and healthy firms. Generally, there are two shortcomings in Paradi et al. (2014) that we address in this paper. The first shortcoming is that the cut-off point distinguishing bankrupt/healthy firms is based on each year, which means multiple such cut-off points are needed to predict corporate failure in different stages. The second one is that the cut-off point for different years is also different, which makes it impossible to find a unique value to classify firms. In our current research, we utilize SVMs to avoid the process of calculating cut-off points; moreover, the SVM method uses all the DEA scores of the several recent years, then we do not need to provide multiple cut-off points.

Work on predicting corporate failure, regardless of the method, is of paramount interest to not only banks but also venture capitalist prior to making any investments. Unlike banking, a firm may be more averse to providing its financial and operational data to an unknown venture capitalist. As such, works on using DEA may allow a firm to only release its DEA score to help a venture capitalist make an investment decision, and not have the firm release all of its closely held information.

2.4. Support Vector Machines

SVMs, for our application, are used for classification purposes and employs supervised learning. In our case, we only consider two classes, bankrupt and not bankrupt, and an SVM finds a best-fit function such that most points on one side of the function belong to one class, and all other points belong to the other. The core idea of an SVM is given a set of data, with l elements in the training data, inputs $x \in \Re^n$ and outputs $y \in \{-1, +1\}$, we would like to find a hyper plane that separates inputs based on their outputs and maximizes the distance between itself and the closest point in the set $\{x_i |\ y_i = -1\}$ and the closest point in the set $\{x_i |\ y_i = +1\}$. This is done by solving the following mathematical program (Deng et al. 2012)

$$\min_{\alpha} \quad \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j (x_i \cdot x_j)\alpha_i \alpha_j - \sum_{j=1}^{l} \alpha_j$$

$$s.t. \quad \sum_{i=1}^{l} y_i \alpha_i = 0 \qquad (1)$$

$$0 \le \alpha_i \le C \qquad i = 1, \dots, l$$

In order to fit non-linear functions that separate the points in the two classes, x is mapped to a potentially higher dimension via the mapping $\phi(\cdot)$ and (1) is solved simply by replacing $x_i \cdot x_j$ with $\phi(x_i) \cdot \phi(x_j) = \kappa(x_i, x_j)$, $\kappa(x_i, x_j)$ is referred to as the kernel function of a SVM and in our study we experiment with various kernel functions. SVMs, using financial data have been used to predict corporate failures in the past (Min & Lee, 2005).

3. Using DEA within an SVM

An SVM is a powerful tool for extracting information from data sets; however, sometimes it may not be an effective method when there are noisy observations or the data is distributed uniformly on the feature space, independent of class. On the other hand, the data points may have multi-attributes, and it is very common that these attributes are correlated or influence each other; therefore, information mining via SVM alone may neglect the inner connection between such attributes. This observation inspires us to use DEA at first to analyze each data point as a decision making unit (DMU), which consists of input and output attributes and considers the internal transformation from inputs to outputs. Then we use the efficiency scores obtained via DEA to continue extracting further information about the changing trend of these scores. In other words, DEA is a projection-like method that reduces dimensionality for SVMs. Eventually, we use SVM methods to predict corporate failure based only on DEA scores. In a method combing DEA and SVMs, we can utilize the merits of both methods. Also, such an idea provides us more accurate results for corporate failure prediction.

3.1. Generating DEA Scores via the Slacks-Based Measure

As we use the same the same data as Paradi et al. (2014), we restrict our attention to the Slack-Based Measure (SBM) model. As the authors point out radial DEA models, like "CCR and BCC (Banker et al. 1984) models are limited by the fact that they do not account for mix inefficiencies. In this case, the company under examination is not limited to 'proportional attributes change', but is evaluated by the general deviation from best firms. It follows that the SBM model (Tone K. , 2001), which accounts for mix inefficiencies is more suitable for the current study." We now introduce the SBM model. The model considers a set of *n* DMUs with input vectors, represented by an ($m \times n$) matrix $\boldsymbol{X}$, and output vectors, represented by a ($s \times n$) matrix $\boldsymbol{Y}$. There is a total of *m* and *s* inputs and outputs, respectively. Thus, the efficiency score of DMU_o (the DMU under evaluation) is determined via the following fractional programming SBM model:

$$\rho = \min_{\lambda, s^-, s^+} \frac{1 - \frac{1}{m}\sum_{i=1}^{m} \frac{s_i^-}{x_{io}}}{1 + \frac{1}{s}\sum_{r=1}^{s} \frac{s_r^+}{y_{ro}}}$$
$$s.t. \quad \boldsymbol{x}_o - \boldsymbol{s}^- = \boldsymbol{X}\boldsymbol{\lambda} \quad , (1)$$
$$\boldsymbol{y}_o + \boldsymbol{s}^+ = \boldsymbol{Y}\boldsymbol{\lambda}$$
$$\boldsymbol{\lambda} \geq 0, \boldsymbol{s}^- \geq 0, \boldsymbol{s}^+ \geq 0$$

in which $\boldsymbol{x}_o = (x_{1o}, x_{2o}, \ldots, x_{mo})^T$ is the input vector and $\boldsymbol{y}_o = (y_{1o}, y_{2o}, \ldots, y_{so})^T$ is the output vector for DMU_o. Slack vectors $\boldsymbol{s}^- \in \boldsymbol{R}^m$ and $\boldsymbol{s}^+ \in \boldsymbol{R}^s$ are explained as input excesses and output shortfalls, respectively. The production possibility set $\boldsymbol{P}$ is defined as:

$$\boldsymbol{P} = \{(\boldsymbol{x}, \boldsymbol{y}) \mid \boldsymbol{x} \geq \boldsymbol{X}\boldsymbol{\lambda}, \boldsymbol{y} \leq \boldsymbol{Y}\boldsymbol{\lambda}, \boldsymbol{\lambda} \geq 0\} . (2)$$

It can thus be concluded that the combination $(\boldsymbol{X}\boldsymbol{\lambda}, \boldsymbol{Y}\boldsymbol{\lambda})$ formed by a non-negative vector $\boldsymbol{\lambda}$ outperforms $(\boldsymbol{x}_o, \boldsymbol{y}_o)$. Tone (2001) noted that the SBM model, as written above satisfies the following properties: (P1) *Units Invariance*: The objective function values is independent of the measurement units of the inputs and outputs. (P2) *Monotonicity*: DMU efficiency is monotonically decreasing with the slack for either the inputs or outputs. (P3) *Reference Set Dependence*: It is sufficient to determine the efficiency of a DMU only through its corresponding reference set. (P4) *Charnes-Cooper Transformation*: The Charnes-Cooper transformation may be used to linearize model (1).

The objective function in (1) is the ratio of the mean input and output mix inefficiencies, bounded above by 1. Let $(\rho^*, \boldsymbol{\lambda}^*, \boldsymbol{s}^{-*}, \boldsymbol{s}^{+*})$ be the optimal solution of an inefficient DMU_o as determined by (1). This DMU_o can be made efficient by reducing its input excesses and augmenting its output shortfalls as follows:

$$\hat{\boldsymbol{x}}_o = \boldsymbol{x}_o - \boldsymbol{s}^{-*}$$
$$\hat{\boldsymbol{y}}_o = \boldsymbol{y}_o + \boldsymbol{s}^{+*} \quad . (3)$$

The new DMU, $(\hat{\boldsymbol{x}}_o, \hat{\boldsymbol{y}}_o)$, is considered as an improving target for the original DMU, $(\boldsymbol{x_0}, \boldsymbol{y_0})$. The reference set of DMU_o is composed of all the positive elements in vector $\boldsymbol{\lambda}^*$. In the cases where only the slacks in the inputs are needed for investigation, the input-oriented SBM model tends to be used. The input-oriented SBM model is the numerator of the SBM model with corresponding proper modification to constraints that can be expressed as follows:

$$\rho = \min_{\lambda, s^-, s^+} 1 - \frac{1}{m}\sum_{i=1}^{m} \frac{s_i^-}{x_{io}}$$
$$s.t. \quad \boldsymbol{x}_o - \boldsymbol{s}^- = \boldsymbol{X}\boldsymbol{\lambda} \quad \cdot (4)$$
$$\boldsymbol{y}_o \leq \boldsymbol{Y}\boldsymbol{\lambda}$$
$$\boldsymbol{\lambda} \geq 0, \boldsymbol{s}^- \geq 0$$

The output-oriented SBM can be similarly be obtained using a similar approach. In addition to the input and output-oriented SBM models, there are other variants concerning returns to scale, super efficiency, Russell measure, etc. An interested reader should see Cooper et al. (2007) for more details.

Unlike traditional ratio methods and Altman's Z'' score model, we use the DEA *efficiency score* generated by SBM as the material data of SVM to classify banks as bankrupt and non-bankrupt. We classify all data into two groups, bankrupt and healthy. Since we have already known whether a firm is bankrupt or not, we can determine the accuracy of our classification method. The inputs and outputs of SBM model are extracted from Altman's five ratios. All of the numerators of the ratios are considered to be outputs and the denominators are defined as inputs in the model. The ratios are split rather than being input directly, because nominal DEA models may not handle ratios directly, however there is work in incorporating ratios in DEA that is beyond the scope of the current work (Emrouznejad & Amin, 2009).

3.2. Data Preparation

As we are using the same data as Paradi et al. (2014) we use the same approach as the authors to be able to use the SBM model, we rephrase the discussion the authors use in the remainder of this section. A purpose of this research is to see how accurately bankruptcy can be predicted regardless of the asset size. Altman's research inspires all of the indicators utilized; but due to data availability, some of the indicators are not available, such as Earnings before Interest and Tax (EBIT). Therefore, we need to reorganize the indicators. In this research, EBIT is substituted for Operating Income, which is also considered to be a very valuable indicator of corporate health. Moreover, the attribute "Total Liabilities" was removed, though present in Altman's method. As we do not have the data for "Working Capital," this indicator was split into "Current Assets" and "Current Liabilities."

Unlike manufacturing companies, in the model presented here for nonmanufacturing firms, we include the number of employees and the number of shareholders. The number of employees was added to introduce the measure of human capital (the most important "asset" in a nonmanufacturing firm) as a contributor to the efficiency of a company. The number of shareholders was added because, for many smaller nonmanufacturing firms, shareholders may act as decision-makers and invest time and money to facilitate a firm's success. From this perspective, the number of shareholders may be viewed as the public's perception on the financial well-being of a company.

A negative value is a common problem in the DEA literature. In our research we have negative values in Retained Earnings (RE), Operating Income (OI) and Book Value of Equity (BVE), thus making the SBM model not applicable. In order to use the SBM model, we split each output into their respective positive and negative components. These three variables are originally categorized as outputs. As some negative values exist in these three outputs for some firms, we manually add three more corresponding inputs. The positive components are treated as outputs; for the negative components, we use their absolute values as inputs. For example, RE was split into RE^+ and RE^-, with RE^+ becoming an output, but RE^- (which is the absolute value for the negative one) becomes an input. This approach results in making RE^+, or any positive component, as large as possible. However, RE^-, or any negative component, its absolute value is

viewed as an input that should be minimized. After conducting all of the transformations, we now write the inputs/outputs of our revised model in Table 1, which is a reproduction of Table 1 in Paradi et al. (2014).

Table 1: Inputs/Outputs Classification

Outputs	Inputs
Current Assets (CA)	Current Liabilities (CL)
Positive Retained Earnings (RE^+)	Negative Retained Earnings (RE^-)
Positive Operating Income (OI^+)	Negative Operating Income (OI^-)
Positive Book Value of Equity (BVE^+)	Negative Book Value of Equity (BVE^-)
The Number of Shareholders (SH)	The Number of Employees (EM)

3.3. Using DEA scores within SVMs

As discussed in Section 2.4 we test a variety of kernel functions, $\kappa(x_i, x_j)$, in this study. We list the kernel functions used in Table 2, their implementations come from the R (Team, 2015) kernel library (Karatzoglou et al. 2004). We note that we want to use standard textbook SVM methods to show the viability of our method, as these are the methods implemented in most third-party vendor software available in practice.

Table 2: List of Kernel Functions Used

Kernel Name	Kernel Generating Functions	Parameters
Gaussian RBF	$\kappa(x_i, x_j) = e^{\left(-\sigma\lvert x_i - x_j\rvert^2\right)}$	σ
Polynomial	$\kappa(x_i, x_j) = \left(s \cdot x_i^T \cdot x_j + c\right)^d$	s, c, d
Hyperbolic tangent	$\kappa(x_i, x_j) = \tanh(s \cdot x_i^T \cdot x_j + c)$	s, c
Laplacian	$\kappa(x_i, x_j) = e^{(-\sigma\lvert x\_i - x\_j\ \rvert)}$	σ
Bessel	$\kappa(x_i, x_j) = -Bessel_{v+1}^n \sigma\lVert x_i - x_j\rVert$	σ, n, v
Spline	$\kappa(x_i, x_j) = 1 + x_i \cdot x_j + x_i \cdot x_j \min(x_i, x_j) - \frac{x_i + x_j}{2}\min(x_i, x_j)^2 + \frac{\min(x_i, x_j)^3}{3}$	N/A

We discuss the parameters we consider in our study in Section 4.

As we only have 68 firms with known values in the first 3 years of operations with known outcome, either bankrupt or non-bankrupt (Table 5), we use 10-fold cross validation to separate or data into training and testing data. Further, to statistically compare the accuracy of using the *raw data*, the firm attributes the first three years of operations, and the *DEA data*, the SBM values computing from the first three years of operations, we bootstrap the 10-fold cross validation by creating at most 500 instances of the 10-fold cross validation (we use 500 10-fold cross validation instances for all of the kernels, except for the polynomial kernel which is the most computationally intensive of all the kernels considered and we only use 500 instances when

we are not able to make a statistically significant comparison using either 50 or 100 instances). Our 500 replications with 10-fold cross-validation mean that for most instances we use 5,000 test data instance with approximately 6.9 observations in each instance. When comparing the accuracy resulting from each dataset, we use the Wilcoxon rank-sum test (Wilcoxon, 1945) to see if on average, the SVM using the DEA data is statistically more accurate than the SVM using the raw data. We consider the number of test data instances that are accurately predicted along with the p-value from the rank-sum test.

4. Application to Bankruptcy Prediction for Nonmanufacturing Firms

We start this section from data collection in the nonmanufacturing industry of North America. From a large number of candidate data points, we select the data which has full records of the recent 3 years. Then we use these records to calculate the DEA efficiency scores for the recent 3 years, and based on this, we classify bankrupt and non-bankrupt firms by different SVMs. By comparing the results of different SVMs, we conclude that using DEA data as inputs into an SVM not only preserves a firm's privacy, but it enables better corporate failure prediction relative to using raw data for classification alone.

4.1. Data Acquisition

In this research, we collected the data through Mergent Online database (Mergent, 2011) and a third-party company focusing on corporate bankruptcies in North America as of the 1980s selected by SIC (Standard Industrial Classification) codes. In this study, we only consider firms classified as nonmanufacturing or service-based firms. These companies must also have filed for bankruptcy between the years of 2000 and 2006, primarily due to data availability consideration. Due to the economic recession taking place, bankruptcy filings from 2007 to present were not selected, as there may be external factors leading to firm bankruptcies during that period.

We used the most recent 3 years data before bankruptcy as we consider such data can reflect the recent trend of the operational status changing of a company, and older data may not be significant in the prediction of bankruptcy. When possible, we excluded firms that filed for bankruptcy but did not fail. Many such companies, file for bankruptcy for factors other than insolvency. For example, some liquidations were driven by legal considerations, and others due to financial distress, all such bankruptcies were filed in an attempt to reorganize and restructure and alleviate debt. We collected data from the following sources: Balance Sheets, Income Statements, Cash Flow Statements and Retained Earnings. Current assets, total assets, current liabilities, total liabilities, retained earnings and shareholders' equity values were extracted from the Balance Sheet. The operating profit was calculated using the formula Net Sales – Cost of goods – Expenses, from the Income Statement. We also have data on the number of employees and shareholders.

We next separated the firms into bankrupt and healthy. We selected and matched a healthy company for every bankrupt company based on SIC number and the number of healthy years. For each bankrupt company, the corresponding healthy company must still be in existence 5 years after the bankruptcy of the bankrupt company, and must not have filed for bankruptcy during the time each firm is compared. The same financial dimensions were considered for both

healthy and bankrupt firms each year. For example, if a bankrupt company filed for bankruptcy in 2002, financial data was collected for 1997-2001. The healthy firm matched with the bankrupt firm must be in existence and not have filed for bankruptcy during the entire period of 1996-2006. Unfortunately, we are not able to find a healthy match for each bankrupt company, meaning that there are more bankrupt companies than non-bankrupt companies.

We summarize the steps we take in our analysis in the algorithm below:

1. Acquire data from Mergent with paired health and bankrupt firms, the set of all firms and their financial data is D.
2. Compute the SBM DEA scores for all firms in D, let $S(D)$ be the DEA scores.
3. For each kernel in Table 2:
   - 3.1. Initialize the kernel with parameter values
     - 3.1.1. Randomly make 10 folds out of D
       - 3.1.1.1. Train the SVM initialized in 3.1 on 9 out of the 10 folds built in 3.1.1
       - 3.1.1.2. Test the SVM on the one fold not included in 3.1.1.1, keeping track of the number of correct predictions
       - 3.1.1.3. Repeat 3.1.1.1 for all 10 folds
     - 3.1.2. Repeat 3.1.1 and all sub-steps for $S(D)$
     - 3.1.3. Repeat steps 3.1.1 and 3.1.2 500 times
   - 3.2. Statistically compare the number of correct predictions of the SVM using D and the SVM using $S(D)$
   - 3.3. Go to 3.1 with updated kernel parameter values and repeat until either computation time per instance (one fold) is greater than 5 hours, the precision limits of the machine is reached or until the grid is exhausted

4.2. Results Analysis

The kernels described in Table 2 each have a set of parameters associated with them, as listed in the third column of the same table. In our study, we conducted a grid search, as suggested and used in the literature (Hsu et al. 2010; Duan & Keerthi, 2005; Min & Lee, 2005), over the set of parameters to find the best values of those considered. Grid search simply means that we enumerate all possible parameter values, however as the parameters themselves are continuous we have steps of varying size over the parameters and we try all possible combinations of these discretized parameters, as suggest by the literature we attempted integer values 1 through 10 for degrees and for other parameters they were drawn from the set $\{2^{-5}, 2^{-5}, \dots, 2^{13}, 2^{15}\}$. We also considered finer parameters in the range of 0 to 10 with varying step size, ranging from 1 to 0.01, we were not able to reach values of 10 for all kernels (when computation time was over 5 hours per instance or if the machine precision is reached in our computations), for example the polynomial kernel, we only considered degree, scale and offset (d, c, s) pairs that took less than 5 hours to compute and for larger values, the machine percision limits were reached. Most of the expriments were carried out on 24 core machines with sufficient RAM, either 128 GB or 64 GB. For each set of parameters we considered, we used 10-fold cross-validations, and kept track of the number of times each the trained kernel correctly predicted the class, bankrupt or not, of the firms that were held out. As there is an exponential number of ways 10-folds may be created, we generated 500, 10-folds for each parameters setting on all kernels except for the polynomial kernel. The reason we limited the number of replications for the polynomial kernel is that it

took in the order of hours to complete all 500 replications for some parameter values, as such we reduced the number of replications to 50 and if we observed a statistically inconclusive outcome, then we increased the number of replications to at most 500. This means that if there is no statistical comparison found for any parameter values, then it means that we ran 500 replications of our 10-fold cross validation. After we ran all of the replications, we then conducted a Wilcoxon rank-sum test on the number of correct prediction, comparing the number of correct predictions using the raw data to the number of correct predictions using only the DEA data. In Table 3 below we show for each kernel, with names in the first column, the number parameter configurations we attempted for each kernel, in the second column, the fraction of tests in which the SVM using DEA data performed statistically better at the 95% confidence level (as measured by the Wilcoxon rank-sum test), in the third column, and the SVM using the raw data performed better, at the same confidence level, in the fourth column. For example, for the Gaussian RBF kernel, the first row of the table, we conducted 172 experiments, of those experiments 0.94 (94%) of the parameter configurations we tried during our grid search resulted in the SVM using only the DEA data to more accurately predict the class of the testing data than the SVM using the raw data at the 95% confidence level (CL). Conversely, 0.05 (5%) of all parameter configurations resulted in the raw data SVM providing more accurate predictions on the testing data than the DEA data SVM with 95% confidence. The remaining 0.01 (1%) of instances lead to no statistical difference in the prediction accuracy between the two data sources used by the SVMs.

Table 3 fixed the parameters to use for each SVM and alternated the data source for each SVM, and then the accuracy of each SVM is compared to one another. A natural extension of Table 3 is to compare the best performing raw data SVM to the best performing DEA data SVM (i.e., we do not use the same parameters for both of the SVMs). In Table 4, for each kernel, listed in the first column of the table, the parameters of the best performing DEA SVM, column two, the parameters of the best performing raw SVM, column three, and finally the p-value, column four, of the rank-sum test with the alternative hypothesis that the DEA SVM is more accurate than the raw SVM. As seen in the fourth column as the p-values are all less than 0.01, thus with 99% confidence for all parameters considered, the DEA SVM is more accurate that the raw SVM. These results also suggest that the best performing DEA data SVM, across all kernels, will also perform better than the best performing raw data SVM across all kernels. Our results suggest that DEA values, derived from raw data may be more informative, at least for this application than the raw data available in the same application.

Table 3: SVM Performance Depending on Training Data

Kernel	**Number of Experiments**	**Fraction DEA better at 95% CL**	**Fraction raw better at 95% CL**
Gaussian RBF	172	0.94	0.05
Polynomial	1749	0.59	0.33
Hyperbolic tangent	872	0.67	0.28

Laplacian	130	0.98	0.02
Bessel	798	0.81	0.08
Spline	1	1	0

Table 4: Comparing Best performing SVMs

p-values test if DEA SVM is more accurate than raw data SVM. Meaning we are checking if the number of correctly classified companies using DEA only is greater than the number of correctly classified companies using the raw data.

Kernel	DEA Parameters	Raw Parameters	p-value
Gaussian RBF	$\sigma = 4$	$\sigma = 32$	0.00
Polynomial	$s = 8, c = 8, d = 8$	$s = 3, c = 0, d = 4$	0.00
Hyperbolic Tangent	$s = 2, c = 9$	$s = 5480.15, c = 9946.68$	0.00
Laplacian	$\sigma = 1.85$	$\sigma = 6.06$	0.00
Bessel	$\sigma = 4, \nu = 0, n = 1$	$\sigma = 4, \nu = 0, n = 1$	0.00
Spline	N/A	N/A	0.00

5. Discussion and Conclusion

Our results suggest that DEA may indeed be used to not only predict corporate failures, but also obfuscate corporate financial and operational data as the data used to formulate the predictions was only DEA SBM values, and not the original, raw, data. Note that in the SVM setting only using the DEA SBM values were better predictors of corporate failure than using the original, raw, data, please see Table 4. Note that only releasing DEA SBM values may be preferred for private firms that do not want to make their initial financial and operational information made publicly available. We envision private firms releasing their DEA SBM values computed against a set of public firms that must make their financial and operational information publicly available. Private firms, usually with some form of venture capital backing, have two exit opportunities in practice: acquisition or an initial public offering (IPO). In order to have either exit, the firm must be vetted by multiple third parties. Third parties must have a sense as to how the firm is doing in order to know the best way to move forward with respect to the exit, an IPO or acquisition. We envision making DEA SBM values publicly available will enable third parties to know earlier how the firm is doing without having full access to their financial data. Our

proposed approach will enable more third parties to know about the private firm earlier without the overheads and trust issues that are involved when financial data is handed over to a third party. We think that the proposed method will be of use to people in the financial, venture capital, and entrepreneurial sector, by allowing each to evaluate firms while keeping financial and operational data private for a longer period of time.

Our research at first surveyed the related studies in bankruptcy prediction, stretching from ratio models to Altman's Z'' model, and then proposed the approach of combining DEA and SVM to predict corporate failure. We split the negative factors into positive and negative component, which could be a viable option when needed in DEA analyses. Then the DEA scores were generated via SBM model, and then as the inputs for classification via SVM. From the result comparison, we can conclude that using only DEA scores and SVM is apparently a more appropriate method in predicting corporate failure relative to using raw data as inputs to a SVM.

Although our research provides some meaningful findings, there is still a number of suggestions for subsequent future work which includes: (1) we may want to consider additional DEA models or constraint conditions, for example, Assurance Region models may be used to place additional restrictions on variable weights potentially leading to different results; (2) considering alternate input and output dimensions, we selected a particular set of dimension as inputs and outputs, selecting different dimensions may lead to improved prediction accuracy; (3) due to the available data, we were limited in the number of DMUs in this study, additional data will enable a more comprehensive assessment; (4) the selection of kernels in SVM effects analysis result. Therefore, we may need to do consider additional kernels that may improve the accuracy of SVMs. To expand the future direction in point (4), a natural question to ask is if our results are an artifact of SVMs. To explore this point further, we did the same comparative study, though not as detailed as that presented here using the caret package available in R (Kuhn, et al., 2016). The package provides a large set of classification and regression methods one may use in their project. Using ten-fold cross-validation, we used 22 of the binary classification methods available in the package and found that the best fit DEA data only methods outperformed the best-fit raw data methods, both best fits were determined within the package, in 16 of 22 settings. These preliminary results suggest that our results are not unique to SVMs. It is still unclear if there is something common across the various classification methods that make some methods more amicable to DEA scores than the raw data.

Acknowledgements

We would like to thank Dr. Brian Cozzarin for allowing us to use his computing resources, allowing us to complete our experiments in a timely fashion. Stanko Dimitrov would like to thank the Natural Sciences and Engineering Research Council of Canada for partially supporting this research.

Table 5: SBM Scores of Companies

DMU	Company	Year 1	Year 2	Year 3	Bankrupt/Non-Bankrupt
1	1-800 Flowers.com Inc	0.316	1	0.759	Non-bankrupt
2	A.C. Moore Arts & Crafts Inc	0.318	0.503	0.323	Non-bankrupt
3	AccuHealth	0.268	0.269	0.142	Bankrupt
4	ACG Holdings	0.145	0.273	0.439	Non-bankrupt
5	AHT Corp	0.606	1	1	Bankrupt
6	All Star Gas Corp	0.021	0.072	1	Bankrupt
7	AMC Entertainment	0.13	0.17	0.139	Non-bankrupt
8	American Banknote	0.11	0.144	1	Bankrupt
9	American Consumers	1	1	0.427	Non-bankrupt
10	Ames Department Stores	0.547	0.63	0.417	Bankrupt
11	Arden Group	0.366	0.659	0.817	Non-bankrupt
12	Ascena Retail Group	0.536	0.635	1	Non-bankrupt
13	Avado Brands	0.132	0.181	0.217	Bankrupt
14	Big Buck Brewery & Steakhouse	0.005	0.196	0.083	Bankrupt
15	BioScrip Inc	0.484	0.266	0.27	Non-bankrupt
16	Bon-Ton Stores	0.331	0.722	0.646	Non-bankrupt
17	Borders Group Inc	0.798	0.683	0.637	Non-bankrupt
18	Briazz Inc	0	0.004	0.142	Bankrupt
19	Caliber Learning Network	0.165	0.634	0.714	Bankrupt
20	Carmike Cinemas Inc	0.242	0.41	0.414	Non-bankrupt
21	Carrols Corp	0.074	0.113	0.454	Non-bankrupt
22	Casual Male Corp	0.382	0.766	0.682	Bankrupt
23	CD Warehouse	0.531	1	0.282	Bankrupt
24	Children's Place Retail Stores Inc	0.337	0.565	0.538	Non-bankrupt
25	Cinemaster Luxury Theaters Inc	0.032	0.318	0.077	Bankrupt
26	Commodore Applied Technologies	0.013	0.064	0.138	Non-bankrupt
27	Computer Learning Centers	0.188	0.313	0.48	Bankrupt
28	Converse	0.155	0.177	0.362	Bankrupt
29	Cooker Restaurant Corp	0.054	0.286	1	Bankrupt
30	Crown Books Corp	0.191	1	0.493	Bankrupt
31	Dairy Mart	0.107	0.224	0.18	Bankrupt
32	Drug Emporium Inc	0.128	0.154	0.185	Bankrupt
33	Eat At Joes Ltd	1	1	1	Non-bankrupt
34	ELXSI Corp	0.556	0.798	1	Non-bankrupt
35	eToys Inc	1	1	1	Bankrupt
36	Express Scripts Inc	1	1	1	Non-bankrupt
37	Family Room Entertainment Corp	1	1	1	Non-bankrupt
38	Florsheim Group Inc	0.205	0.346	0.541	Bankrupt
39	Fresh Choice Inc	0.133	0.25	0.079	Bankrupt
40	Furr's Restaurant Group Inc	0.036	0.058	0.17	Bankrupt
41	Gadzooks Inc	0.316	0.469	0.43	Bankrupt
42	Gerald Stevens Inc	0.449	0.835	0.358	Bankrupt
43	Grand Union Company Inc	0.299	1	0.374	Bankrupt

44	Hastings Entertainment Inc	0.218	0.381	0.42	Non-bankrupt
45	HCI Direct Inc	0.289	0.685	1	Bankrupt
46	Healthcare Integrated Services	0.308	1	0.519	Bankrupt
47	Heilig-Meyers Company	1	1	1	Bankrupt
48	Home Depot	1	1	1	Non-bankrupt
49	Homeland Holding Corp	0.225	0.281	0.172	Bankrupt
50	Horizon Pharmacies Inc	0.153	0.37	0.248	Bankrupt
51	House2Home Inc	0.534	1	1	Bankrupt
52	Integra Inc	0.031	0.142	0.218	Bankrupt
53	Integrated Health Services Inc	0.138	1	0.469	Bankrupt
54	Jacobson Stores Inc	0.371	1	1	Bankrupt
55	Jennifer Convertibles Inc	0.191	0.137	0.088	Non-bankrupt
56	Jos. A Bank Clothiers Inc	0.316	0.505	0.402	Non-bankrupt
57	Kasper ASL Ltd	0.17	0.229	1	Bankrupt
58	Kushner-Locke International Inc	1	1	0.801	Bankrupt
59	LaCrosse Footwear	0.534	0.48	0.596	Non-bankrupt
60	Lamonts Apparel Inc	0.149	0.21	0.161	Bankrupt
61	Lechters Inc	1	1	1	Bankrupt
62	Med/Waste Inc	0.102	0.26	0.335	Bankrupt
63	Meritage Hospitality Group Inc	0.094	0.213	0.101	Non-bankrupt
64	Mexican Restaurants Inc	0.308	0.312	0.175	Non-bankrupt
65	New York Health Care Inc	0.144	0.219	0.067	Non-bankrupt
66	RadNet Inc	0.141	0.183	0.119	Non-bankrupt
67	Rocky Brands Inc	1	1	0.626	Non-bankrupt
68	Sagemark Companies Ltd	1	1	1	Non-bankrupt

References

Agrawal, R., & Srikant, R. (2000). Privacy-preserving data mining. *ConferenceProceedings of the 2000 ACM SIGMOD international conference on Management of data - SIGMOD '00. 29*, pp. 439-450. New York, New York, USA: ACM Press.

Alexander, S. S. (1949). The Effect of Size of Manufacturing Corporation on the Distribution of the Rate of Return. *Review of Economics and Statistics*, 229-235.

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *Journal of Finance, 23*(4), 589-609.

Altman, E. I. (2002). *Bankruptcy, Credit Risk and High Yield Junk Bonds.* Malden, Massachusetts: Blackwell Publishers Inc.

Banker, R. D., Charnes, A., & Cooper, W. W. (1984, sep). Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis. *Management Science, 30*(9), pp. 1078-1092.

Beaver, W. H. (1967). Financial Ratios as Predictor of Failure. *Journal of Accounting Research, 5*, 71-111.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research, 2*, 429-444.

Cielen, A., Peeters, L., & Vanhoof, K. (2004). Bankruptcy prediction using a data envelopment analysis. *European Journal of Operational Research, 154*, 526–532.

Cooper, W. W., Seiford, L. M., & Tone, K. (2007). Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software (Second Edition). Springer.

Deakin, E. (1972). A Discriminant Analysis of predictors of Business Failure. *Journal of Accounting Research*, 167-179.

Deng, N., Tian, Y., & Zhang, C. (2012). *Support Vector Machines: Optimization Based Theory, Algorithms, and Extensions.* Chapman and Hall/CRC.

Duan, K.-B., & Keerthi, S. S. (2005). Which Is the Best Multiclass SVM Method? An Empirical Study. Springer Berlin Heidelberg.

Emrouznejad, A., & Amin, G. R. (2009). DEA models for ratio data: Convexity consideration. *Applied Mathematical Modelling, 33*(1), 486-498.

Emrouznejad, A., Parker, B. R., & Tavares, G. (2008). Evaluation of research in efficiency and productivity: A survey and analysis of the first 30 years of scholarly literature in DEA. *Socio-Economic Planning Sciences*, *42*(3), 151–157.

Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society Series A, 120*, 253-90.

Grice, J. S., & Dugan, M. (2001). The Limitations of Bankruptcy Prediction Models: Some Cautions for the Researcher. *Review of Quantitative Finance and Accounting, 17*, 151–166.

Grice, J. S., & Ingram, R. W. (2001). Tests of the generalizability of Altman's bankruptcy prediction model. *Journal of Business Research, 54*, 53– 61.

Growth of the Service Sector. (2011, 12). Retrieved from http://www.worldbank.org/depweb/beyond/beyondco/beg_09.pdf

Hsieh, S.-J. (1993). A Note on the Optimal Cutoff Point in Bankruptcy Prediction Models. *Journal of Business Finance and Accounting, 30*(3), 457–464.

Hsu, C.-w., Chang, C.-c., & Lin, C.-j. (2010). A practical guide to support vector classification .

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab -- An S4 Package for Kernel Methods in R. *Journal of Statistical Software, 11*(9), 1-20.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., . . . Candan., C. (2016).

caret: Classification and Regression Training.

Li, Z., Crook, J., & Andreeva, G. (2014). Chinese companies distress prediction: an application of data envelopment analysis. *Journal of the Operational Research Society, 65*, 466–479.

Lindell, Y., & Pinkas, B. (2000). Privacy Preserving Data Mining. *JOURNAL OF CRYPTOLOGY, 15*, pp. 36-54.

Liu, J. S., Lu, L. Y., Lu, W.-M., & Lin, B. J. (2013). Data envelopment analysis 1978–2010: A citation-based literature survey. *Omega, 41*(1), 3-15.

Meisel, S., & Mattfeld, D. (2010). Synergies of Operations Research and Data Mining. *European Journal of Operational Research, 206*(1), pp. 1-10.

Mergent, I. (2011). *Mergent Online.* Retrieved from http://www.mergentonline.com/

Min, J. H., & Lee, Y.-C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications, 28*(4), pp. 603-614.

Misiunas, N., Oztekin, A., Chen, Y., & Chandra, K. (2016). DEANN: A healthcare analytic methodology of data envelopment analysis and artificial neural networks for the prediction of organ recipient functional status. *Omega, 58,* pp. 46-54.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research, 18,* 109–131.

Oliveira, S. R., & Zaïane, O. R. (2004). Achieving Privacy Preservation when Sharing Data for Clustering. Springer Berlin Heidelberg.

Paradi, J. C., & Zhu, H. (2013). A survey on bank branch efficiency and performance research with data envelopment analysis. *Omega, 41*(1), 61-79.

Paradi, J., Wilson, D., & Yang, X. (2014). Data envelopment analysis of corporate failure for non-manufacturing firms using a slacks-based measure. *Journal of Service Science and Management, 7*(4), 277-290.

Premachandra, I., Chen, Y., & Watson, J. (2011). DEA as a tool for predicting corporate failure and success: A case of bankruptcy assessment. *Omega, 39,* 620–626.

Shetty, U., Pakkala, T., & Mallikarjunappa, T. (2012). A modified directional distance formulation of DEA to assess bankruptcy: An application to IT/ITES companies in India. *Expert Systems with Applications, 39,* 1988–1997.

Shumway, T. (2001). Forecasting bankruptcy more accurately: a simple hazard model. *Journal of Business, 74*, 101–124.

Sueyoshi, T., & Goto, M. (2009). Methodological comparison between DEA (data envelopment analysis) and DEA–DA (discriminant analysis) from the perspective of bankruptcy assessment. *European Journal of Operational Research, 199,* 561–575.

Sutton, W., & Dimitrov, S. (2013). The U.S. Navy explores detailing cost reduction via Data Envelopment Analysis. *European Journal of Operational Research, 227*(1), 166–173.

Team, R. C. (2015). *R: A Language and Environment for Statistical Computing.* Vienna.

Tone, K. (2001). A slacks-based measure of efficiency in data envelopment analysis. *European Journal of Operational Research, 130*(3), pp. 498-509.

Vaidya, J., Clifton, C. W., & Zhu, M. (2006). *Privacy preserving data mining.* Springer.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bull., 1*(6), 80–83.

Xu, X., & Wang, Y. (2009). Financial failure prediction using efficiency as a predictor. *Expert Systems with Applications, 36*, 366–373.

Yang, X., & Morita, H. (2013). Efficiency improvement from multiple perspectives: An application to Japanese banking industry. *Omega, 41*(3), 501-509.

Yeh, C.-C., Chi, D.-J., & Hsu, M.-F. (2010). A hybrid approach of DEA, rough set and support vector machines for business failure prediction. *Expert Systems with Applications, 37*(2), 1535–1541.

Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research, 22*, 59–82.