2014

A Bayesian Approach to Ranking Private Companies Based on Predictive Indicators

Matthew Dixon
University of San Francisco, mfdixon@usfca.edu

Jike Chong

A Bayesian Approach to Ranking Private Companies based on Predictive Indicators

Matthew Dixon [a,*] and Jike Chong [a]
[a] *Silver Lake Kraftwerk*
1056 Commercial Street
San Carlos, CA 94044, USA

Private equity investors seek to rank potential investment opportunities in growth stage private companies within an industry sector. The sparsity of historical investment transaction data for many growth stage private companies may present a major obstacle to using statistical methods to discern industry specific features associated with successful and failed companies.

This paper describes a Bayesian ranking approach based on (i) extracting and selecting features; (ii) training support vector machine classifiers from feature pairs of labeled companies in an industry; (iii) non-parametric estimation of posterior probabilities of success and failure; and (iv) ranking unlabeled companies within a cohort based on scores derived from posterior probability estimates. We anticipate that this approach will not only be of interest to statisticians and machine learning specialists with an interest in venture capital and private equity but extend to a broader readership whose interests lie in classification methods where missing data is the primary obstacle.

Keywords: Bayesian statistics, machine learning, private equity

1. Introduction

Silicon Valley is currently host to company growth rates and exit valuations of unprecedented levels. Take, for example, the recent purchase of Instagram by Facebook for $1 Billion, representing a 20x return on total investment in only two years. In the absence of any empirically substantiated general formulaic approach to pick the best companies to invest in, investors base investment decisions on a wide set of considerations influenced, in part, by their prior experiences and oftentimes fundamental analysis of companies within an industry sector. Growth stage private companies, however, often have investment transaction histories from which characteristics associated with successful and failed companies (labeled companies) may be discerned using statistical methods. One of the primary challenges in pursuing this approach is the sparsity of the historical data. This papers describes a four step approach based on (i) feature extraction and selection; (ii) classification; (iii) derivation of a score using a novel Bayesian approach to estimate posterior probabilities of success and failure; and (iv) ranking companies within a cohort based on their scores. Focusing on agricultural companies in the cleantech sector, we demonstrate how this approach can be used to rank companies based on a set of features.

Statistical techniques for learning predictive indicators and patterns have been used extensively in the public capital markets. One area of significant traction is in high frequency trading, which uses statistical techniques to generate an estimated $8 billion a year in trading profit in the US alone. The field depends on learning predictive indicators and patterns to automate trading decisions that are made within a fraction of a millisecond and held for no more than a few hours. Many of these trading decisions can be made using a formulaic approach.

In the area of venture capital and private equity investment, investment decisions are made in weeks to months, and the investments are often held for 3 or more years. There are significantly more factors influencing an investment over its lifetime, and there is a scarcity of historical information for the companies involved. Moreover, investment analysts in the industry, as a whole, have not themselves applied the concept of using predictive analytics to support investment decisions but have rather developed intuition and drawn conclusions from the business literature and educational infrastructure supporting this field.

*Corresponding author email: mfdixon@usfca.edu

AI Communications
ISSN 0921-7126, IOS Press.

Given the scarcity of information, the complexity of selecting successful portfolio companies, together with the absence of an empirically substantiated general formulaic approach to venture capital and private equity investments, we turn to classification methods to discern attributes associated with successful or failed companies. We demonstrate a methodology for arriving at a score which can be used to rank a company within an industry sector or sub-sector.

2. Literature Review

Over the last 40 years, there has been much research published in the business literature on models for evaluating potential VC investment opportunities. The literature is too extensive to list but the evolution in this approach can be traced at decadal intervals through seminal contributions by Wells [13] in the 1970s, Tyebjee and Bruno [12], followed by Hall and Hofer [6]. In the absence of databases on global venture capital investment histories, researchers relied more on subjective criteria delineating successful from unsucessful investments rather than statistical inference of the most prominent factors.

Then came the realization in the aftermath of the dot com bubble by Zacharakis and Shepherd [14] that VCs are prone to over-confident decision making and that the decision factors purported to be critical to the investment decision process by investors may be too circumstantial and not conducive to building models of the investment decision process. This realization has led to arguably more credence being given to models based on systemic factors affecting investment decision processes. However, relatively few works have provided a general approach based on statistical modeling. One explanation for this paradox is that some of the most important systemic factors are qualitative in nature and difficult to quantify. Another major obstacle has been that regression methods tend to be data intensive and the absence of sufficient historical investment data precludes their application.

Recently, Gompers, Kovner, Lerner and Scharfstein [5] show that entrepreneurs with a track record of success are much more likely to succeed again over entrepreneurs with a poor track record. The authors use a logistic regression model for estimating the likelihood of a company succeeding given the prior experience of a company's management team and other characteristics such as company age, stage and location. Most crucially for the more technical discussion here is that their analysis draws upon an extensive archive of transaction and company summary data from Dow Jones VentureSource, which the authors estimated to be approximately 90% complete. Using all funding transactions between 1975 and 2003 across all industry sectors, they arrive at regression coefficients with a 1% significance level. This model represents entrepreneurs' track records using between 3,831 companies and 19,617 companies with known outcome. In this study, the choice of regression coefficients is heavily based on answering a specific hypothesis and drawing upon the author's own extensive knowledge of the venture capital industry.

Guided only loosely by the statistical approach taken by Gompers et al. [5], we set out to answer the different question of how private equity investors can use a quantitative approach to rank companies in a particular industry sector, rather than across all sectors. This point of departure confronts the immediate obstacle of there being much less historical data available in any given sector - so little in fact that logistic regression is generally out of the question.

In this paper, we go further still and develop an approach which is robust even to sparse datasets such as nascent industry sectors, in which the number of known outcomes is relatively sparse compared to more mature sectors such as IT and Communications. For nascent sectors, there are hundreds and not thousands of companies with known outcome and we were unable to obtain statistically significant regression coefficients other than the age of the company. In the absence of sufficient numbers of objective performance measures for sector wise analysis, we look to novel approaches to discern patterns in historical transaction data.

Bhat and Zaelit [1] approach the problem of predicting private company exits from qualitative data using machine learning methods. By applying the random forest algorithm to the first three rounds of company information, they are able to predict whether a company will be successful or fail. Diagnosis of their model applied across nine industry sectors reveals a 75% average rate and the average area under the ROC curve is 0.83. One valuable contribution of their work is that they are

able to rank which features of a company and its transaction history offer the most predictive power for late stage investment decision making. While their approach, like that of Gompers et al. [5], is at a broader scope than an industry sector, the authors attempt to capture industry specific effects through encoding a unique sector ID as a feature in the model.

An additional point of interest is that Bhat and Zaelit [1] are able to incorporate a measure of the strength of an investor's social network into their analysis which Hochberg, Ljungqvist and Lu [8] show to be an important measure of the strength of an investor. The authors use the degree of centrality of the investor in a social network as an attribute. While we believe that knowledge of the investors in the company is valuable, it is not apparent how a company shall fail or succeed by virtue of the social network ranking of the investors. A further issue which the authors raise, but do not address, is that investors' social networking strength may often be specific to particular industry sectors where their portfolio is most concentrated.

Our approach differs from that of Bhat and Zaelit [1] in a number of ways. Firstly, we take a more granular approach to identifying companies, instead extracting features and training models using only the investment transaction history pertaining to only one industry sector. In doing so, we are able to learn how a company will perform based on failed and successful companies from a cohort of companies which are more closely related. Secondly, we classify the outcomes of a company using SVMs rather than the Random Forest method. Thirdly, instead of using the degree of centrality of an investor in a social network, we follow the approach of Farmer [4] who revisited the widely held notion that follow-on investments are seen by the investment community as a vote of confidence for earlier investments.

Farmer [4] posed the hypothesis that a visionary early investment will attract a sequence of follow-on investments from a more "prestigious" investor. To measure the concept of prestige, Farmer [4] used graph eigenvector centrality, an approach which is best known for its use in Google's page rank algorithm. The approach assigns all investors equal prestiges and increases the prestige of a particular investor each time other more prestigious investors invest in a portfolio company in a following round, as described in Section 3.1. We further extend this approach by distinguishing an investor's prestige with respect to a particular industry sector in order to more meaningfully characterize their deal syndication patterns.

The inherent challenge in training models from a subset of labeled companies within a particular industry sector is the sparsity of historical transaction data. We partially address this problem by training models from two dimensional feature spaces instead of high-dimensional feature spaces. This approach follows from our observation that the number of complete labeled n-tuples of features decreases significantly as n is increased. Bayes' Theorem is then applied to estimate whether a company will succeed or fail given the results from all of the models. These estimates yield a score which is used to rank companies in a sector or sub-sector.

This paper describes a four step predictive approach for ranking private companies within a cohort. This approach can be applied to sparse industry specific historical data. We begin in the next section by describing the private company transaction data that was used for our analysis before describing the first step- the process of feature extraction and selection.

Section 4 describes the infrastructure and configuration necessary for the next step - training and evaluating the performance of classification methods. We present results showing the performance of logistic regression and SVM classifiers on labeled feature pairs extracted from all companies in the cleantech sector with complete observations for that pair. These results indicate that SVM has a performance advantage over logistic regression when the feature space is not linearly separable.

Section 5 presents a methodology for scoring a company based on estimates of posterior probabilities of success and failure given a set of SVM model outputs. The methodology is best explained by first considering the simplest case, described in Section 5.1, where a single minimum Euclidean distance of an unlabeled point in a feature space to the separating hyperplane is used to estimate the conditional probability of that company succeeding or failing. Then in Section 5.2 we present a more general approach which combines the distances of unlabeled points in multiple feature spaces to the separating hyperplanes to estimate the conditional probability of a particular company failing or succeeding. The method culminates

in Section 6 with the demonstration of a ranking approach for all agricultural companies in the cleantech sector which uses a score based on the estimated posterior probabilities. While the results presented herein are specific to the cleantech industry, the ranking methodology is applicable to any industry sector. Section 7 concludes.

3. Step 1: Feature Extraction and Selection

Dow Jones VentureSource has a good coverage of domestic and international venture-backed and private equity-backed companies. It contains records for 67,000 companies and 19,000 active investors. About 50% of the companies (over 33,000 companies) in the database are international companies. There are 3,775 companies we have identified as cleantech related companies, out of which 10.9% has data that can be used to infer company successes and failures. We define a successful company as one which has reached or is filing an IPO, or has been acquired for at least 1.5x the total money invested in the company. Conversely, we define a failed company as any company which has either been acquired at less than 1.5x of the total money invested, filed Chapter 11 or is labeled in the database as being bankrupt or out-of-business.

Each company profile in the database contains overview details such as the date in which the company was founded, its location, status and an abstract describing the company. Each company profile is also tagged with various keywords which we use to identify which industry sectors it belongs to. The profile also lists the company's investment transaction history- each transaction shows the date, the amount invested and the investment syndicate. In only approximately 10% of the transactions for the cleantech sector did we observe post money valuations. Post money valuations are regarded as one of the most sensitive details on the performance of a private company and are oftentimes not disclosed to the public.

We separate the process of obtaining features for our models into two steps: (i) feature extraction; and (ii) feature selection. The first step involves extracting metrics from the investment transaction histories which are considered to be indicative to investors. Examples of these include estimating the average transaction frequency between investment rounds for each company, calculating the total money invested in a company and counting the total number of investors that have invested in the company since the first round. These features are all examples of primitive quantities which are directly extracted from the data and their sparsity measured. The concept of an investor's strength is not readily available in the database and we describe an approach for estimating this in the next section. Some of the features that we extracted are listed in Table 1 together with a brief explanation and the % of labeled companies with missing observations.

3.1. Investor Prestige

The quality of the investors is often considered one of the most important factors contributing to a company's success. Conversely, the success and failure of companies in an investor's portfolio demonstrates the capabilities of the investor. In a nascent sector such as cleantech, where there aren't a significant number of exits, the quality of an investor needs to be assessed by an alternative metric.

We implemented a feature that assesses early-stage investor prestige. It uses the "investor rank" metric first proposed by Farmer [4] for use in the venture investment space. The technique is based on the peer-evaluation model which was also used in Google's Page rank algorithm to index the websites on the Internet. For a growth stage investor who is looking to identify investable growth stage opportunities, companies with prestigious investors are often more likely to be investigated as they could be considered as pre-qualified by prestigious peers.

In venture investing, the peer-evaluation model characterizes peer-approval by the willingness and action of an investor to provide follow-on investment. The fundamental assumption is that the best evaluation of the quality of a company is how peers in the same domain would like to be associated with it. Since the peer approval relationships are established through investments in the same company, we can aggregate the peer approval relationships within specific sectors to infer an investor's prestige within specific sectors. Figure 1 illustrates the concept using a funding graph for the funding rounds of Enphase Energy Inc. (a cleantech company). Consider for example Applied Ventures, the lead investor in the first funding round.

As new investment partners join the syndicate in subsequent funding rounds, the prestige of Applied Ventures is increased by an amount proportional to the prestige of those investors. The addition of Bay Partners in Round 3 increases the prestige of all investors in Round 2 by an amount proportional to the prestige of Bay Partners which, by backward induction, increases the prestige of Applied Ventures.

To determine the importance of an investor in the network of investors, we construct a graph of investors where the nodes are investors and the edges are the peer approval relationships between the investors. The edges are directed, pointing from each later-stage investor to each early stage investor, as illustrated in Figure 1. The graph is represented as a transistion matrix and passed to the pagerank algorithm [11]. This celebrated algorithm is an iterative approach for computing the measure of eigenvalue-centrality which can be interpreted here as the relative importance of investors.

We use a standard damping factor of 0.85 to represent how likely investors are to invest in a follow-up round. We extend this approach by distinguishing an investor's prestige with respect to particular industry sectors by including only the companies in a specific industry sector in the graph construction process for the computation of an industry sector specific investor prestige. The net result is that investors may have multiple sector specific investor prestiges if their portfolio spans multiple sectors. This extension more intuitively characterizes deal syndication patterns and investor prestige in specific industry sectors.

Investment history is a feature that is often publicly available. The main advantage of this approach is that the data in VentureSource is well populated with funding information. For example, 94% of the cleantech company records in VentureSource contain some information on its funding rounds. The approach is not with out issues. For instance, it is biased towards early stage investors; An investor that focuses on early-stage investment will score higher than an equally capable later-stage investor. Also, the investor is rewarded for funding successful early round investments but not penalized for early-stage investments that did not generate returns and failed.

Once the investor's prestiges are estimated based on all of its transactions within a particular industry sector, we use the maximum prestige value of all investors in a company to arrive at a feature which represents the prestige of the most prestigious investor involved in one or more transactions with that company. For investors in the cleantech sector, we refer to this company specific quantity as "Cleantech Prestige".

The choice of whether to use all extracted features for supervised learning experiments is guided by statistical experiments which are described in the next section.

3.2. Feature selection

The feature selection process is guided by the availability of sufficient quantitative data which can be extracted from the database. One approach to determine whether a feature has a sufficient number of labeled observations for a trend to be discerned is to test the null hypothesis that samples in two or more groups are drawn from the same population, implying that the samples are indistinguishable and exchangeable. Because of the small population of labeled data, samples are not assumed to be drawn from a parametric distribution. This non-parametric distributional representation favors the application of a Wilcoxon rank sum test.

The Wilcoxon-Mann-Whitney (WMM) rank-sum test assesses whether one of the two groups of independent observations tend to have larger medians than the other. The groups need not be the same size but are assumed to be symmetrically distributed about their medians and the response values are ordinal. The responses are chosen to be features of labeled companies in a particular industry sector. Response values representing monetary amounts are first rescaled by the log_{10} function and all response values are further normalized to have zero mean and unit variance. This scaling is chosen for consistency with feature scaling in the SVM classifier and is explained in Section 4. Stating a null Hypothesis of significance more formally:

Definition 3.2.1 (Null Hypothesis: H_0) *let Y_1 denote the median of the normalized response values in the group with the largest median and Y_2 the median of the response values from the other group. The null hypothesis of significance $H_0 : Y_1 - Y_2 \leq 0$ is one-tailed in the direction of the observed effect.*

As a control for the test, we introduce the random variable $Z \sim N(0,1)$ and draw random samples totaling the number of labeled companies composed of 221 (54%) successful and 190 (46%) failures. We use the implementation of the WMM test provided in version 1.0-20 of the `coin` R package [9]. This implementation allows for ties in the response values and computes exact p-values either by the shift algorithm or by the split-up algorithm.

Table 2 shows the resulting p-values from the WMM rank-sum test which represent the probability that the true value of the effect is of sign opposite to the observed value. For the features listed in Table 1, all result in rejection of H_0 at a significance level of 10%. In fact, all features except "Cleantech Prestige" result in rejection of H_0 at the 0.01% significance level. Over a set of 100 experiments, the control Z is found to have a p-value with mean $\mu = 0.5$ and standard deviation $\sigma = 0.29$.

4. Step II: Classification

The standard approach to predicting dichotomous dependent variables in finance and economics is to use a linear method such as logistic regression or linear discriminant analysis. Logistic regression employs the use of a logit function in the generalized linear model (GLM) to estimate the probability of success from the logarithm of the odds ratio (see Hilbe [7] for a discourse on logisitic regression methods). Under the assumption of independence of the explanatory variables, a Gaussian distribution of the errors and the complete set of explanatory variables, logistic regression can be employed here to predict the likely success or failure of a company.

We reiterate that our focus on prediction of successful companies in a nascent industry sector is limited by relatively few numbers of labeled observations. In the pan-industry study of Gompers et al. [5], a multivariate logistic regression model is viable since companies with missing explanatory variables can be discarded with likely marginal impact on the statistical significance of the regression coefficients. Simply discarding incomplete observations in our study results in too limited a dataset for estimating statistically significant regression coefficients, even if the test statistic is bootstrapped. An alternative linear approach referred to as linear disciminant analysis requires fewer observations by assuming that the explanatory variables are normally distributed, an approach which we do not pursue on account of the variables failing standard normality tests.

We instead apply logistic regression models with boostrapping to pairwise combinations of features. We iterate over all 28 combination of pairs from the seven features shown in Table 1 and record the features representing the x and y co-ordinates in the feature space. All features representing monetary amounts are rescaled by the log_{10} function because the values vary by several orders of magnitude and lead to clustering of training data which is observed to increase the error rate of the classifier. All features are further normalized to have zero mean and unit variance in order to avoid disparate scaling which results in one feature dominating in the model.

All of our results are summarized in Table 3 and are obtained using R version 2.14.1 and the GLM implementation is provided in version 2.15.2 of the `stats` package. For each feature pair, the logistic regression model is applied to 1000 bootstrap replicates drawn from 70% of the dataset. The number of complete pairs in the training set is shown in the Table 3 as *size* and varies for each pair. The trained model is applied to the remaining 30% of observations in the dataset and summary statistics of the error rate are shown as *err* and *stderr* in the table. The Z-statistic (the ratio of the estimated coefficient to the standard error) and the p-value, for the two-tailed test of the null Hypothesis that the true value of the regression coefficient is zero, are additionally estimated.

For many feature pairs, the null Hypothesis can be rejected at the 10% significance level (p-value < 0.1). p_x and p_y are the p-values corresponding to the regression coefficients for the features represented by the x and y co-ordinates respectively. Despite the "Last Round PMV" being a much sparser feature than the others, it appears from the p-values to be an important feature in discerning a company's success or failure regardless of the choice of the other feature. The other features are found to be only important factors when paired with one of a subset of the other features.

From scatter-plots of the two dimensional feature space similar to one shown in the top-left plot of Figure 2, we observe that the line separating the failed and successful companies is non-linear

in the transformed co-ordinates. The absence of linear separability in some of the feature spaces motivates the use of SVMs, which are favored in the machine learning community for their ability to characterize non-linear separating hyperplanes by selecting from a set of non-linear kernels. Furthermore, SVMs do not impose distributional assumptions on the features.

We use the SVM classifier implemented in version 1.6 of the R package `e1071` [10], an interface to `libSVM` [3]. For performance comparison with the logisitic regression model, we use the same 70/30 dataset partition rule to respectfully train and test the classifier. `libSVM` provides a choice of four kernels: (i) a linear kernel; (ii) a polynomial kernel; (iii) a sigmoid kernel ;and (iv) a radial kernel. The classifier is trained on a sample of no more than 273 labeled cleantech companies. The choice of kernel is based on the error rate using default parameters applied to the test set consisting of up to 124 labeled companies.

Table 3 further shows the comparative performance of bootstrapped SVMs with logistic regression. The mean and standard deviation of the error rate across the bootstrap replicates is shown for each method. Feature pairs are eliminated from the table if the error rate from both methods is above 0.3. The kernel type which minimizes the SVM error rate is shown in the far-right column. For the feature pairs where the best kernel type is non-linear, we note that the error rate from the SVMs is consistently lower than for the logistic regression models. This provides some evidence that the ability to represent non-linear separating hyperplanes reduces the error rate. We further note, however, that the evidence is not entirely conclusive - in many cases the logistic regression outperforms the linear SVM. The standard errors are comparable across methods indicating that they are equally sensitive to noise and outliers in the training set.

Our experiments show that the choice of kernel substantively affects the accuracy prediction rate over the test set. In a separate study, we partitioned the data by a 70/15/15 rule to train, tune and test. We found only marginal effects by tuning the SVM parameters about their default values and thus used the following parameter values: tolerance of termination criterion (0.001), ϵ in the insensitive-loss function (0.1), cost of constraints violation[1], the degree of the polynomial (3), γ (1/(data dimension)) and a coefficient needed for polynomial and sigmoid kernels C_0 (0).

Having trained models from pairs of feature, the remaining part of this paper shall partially address the problem of how to combine each model in order to predict rankings of company's likelihood to succeed. Despite finding a preference for SVM classifiers, we emphasize that the Bayesian ranking approach described hereforth can be applied to any classifier which produces a score for each company. In logistic regression, for example, this score is the estimate of the log odds ratio for each feature pair. SVMs output a distance of a company to the separating hyperplane. Without loss of generality, we continue with demonstration of a Bayesian ranking methodology applied to SVM classifiers trained over all combinations of feature-pairs.

5. Step III: A Bayesian Approach for Estimating Posterior Probabilities

The trained SVM classifier takes as input the rescaled co-ordinates of a point in a feature space and outputs the (signed) shortest Euclidean distance of the point to the separating hyperplane. A positive distance indicates that the point lies in the success region and a negative distance indicates otherwise. For a particular feature pair, the absolute value of this distance could be interpreted as a relative scale from which to score and hence rank companies. The problem with this interpretation is that the complete labeled data for any given feature pair is biased and thus contains different number of failed and successful companies. A further concern is that this distance is not strictly comparable between different feature spaces and is difficult to compare how a company scores with respect to different factors.

We therefore turn to a Bayesian approach to estimate the posterior probability [2] of success or failure of a company given the observed distance of a company. For ease of exposition, we begin by considering the simplest case where the posterior probability estimate depends on the output of single model and refer to this as the "univariate" approach.

[1]The cost of constraints violation is the C constant of the regularization term in the Lagrange formulation [10].

5.1. A univariate Bayesian approach

Definition 5.1.1 (Univariate observed data) *The observed data is the set of minimum Euclidean distances between the observed points in a feature space and the separating hyperplane given by a model trained from labeled features. Denote* $X \in \mathbf{X}$ *as the minimum distance from an observed point to the separating hyperplane and* $\mathbf{X}$ *the set of all observed distances in the feature space corresponding to companies in a particular industry sector. Let* $x \in \mathbb{R}$ *denote a value on the infinite line intersecting an observed point and the nearest point of intersection with the separating hyperplane. We follow the convention that* $x = 0$ *coincides with the point of intersection with the separating hyperplane and the sign of* x *depends upon the region that it lies (positive for success).*

By Bayes' Theorum, the uncertainty in the hypothesis $\mathcal{H} \in \{S, F\}$ that a company will succeed (S), given that the minimum observed distance $X > x$ from an observed point to the separating hyperplane takes the form of the conditional posterior probability

$$P(\mathcal{H} = S|X > x) = \frac{P(X>x|\mathcal{H}=S)P(\mathcal{H}=S)}{P(\mathcal{H}=S)P(X>x|\mathcal{H}=S)+(1-P(\mathcal{H}=S))P(X>x|\mathcal{H}=F)} \quad (1)$$

in which $P(X > x \mid \mathcal{H} = S)$ is a function of $x \in \mathbb{R}$ referred to generally as the likelihood function and is the conditional probability that $X > x$ given the hypothesis $\mathcal{H} = S$. $P(\mathcal{H} = S)$ is referred to as the prior. The denominator is the probability that $X > x$ and normalizes the product of the likelihood function and the prior. Equivalently, the uncertainty in the hypothesis $\mathcal{H} = F$ is given by replacing 'S' by 'F' and conditioning on the set $X \leq x$ in the above expression.

The prior probability of success is estimated from the labeled data by computing the ratio of the size N_s of the set of successful company data points $\mathbf{X}_s$ against the size N of the set of all labeled data points $\bar{\mathbf{X}} \subset \mathbf{X}$. The set of all failed companies is treated in a similar way. The likelihood function is estimated over a grid of values of $x \in \Omega_h := \{x \mid x = ih,\ i := -m \to m\}$ for some grid size h and bound m. For each x the proportional of $X \in \mathbf{X}_s$ satisfying $X > x$ is computed.

The univariate approach is illustrated in Figure 2 for a particular feature pair using all labeled cleantech companies. The top left plot shows the feature space of the log of the "Total Amount Invested" against the "Number of Investors" for all labeled companies in the cleantech sector. This feature space is divided into two regions by the separating hyperplane. The red region represents failed companies and the green region represents successful companies. Circles falling within the red region or squares within the green region indicate misclassification. The top right plot shows the corresponding posterior probabilities of success or failure as a function of x as denoted by the green dashed line and the red dotted line respectively. The posterior probability that the nearest distance X from a failed company data point to the separating hyperplane is less than or equal to x is observed to decrease from left to right. Conversely, the posterior probability that the nearest distance X from a successful company data point to the separating hyperplane is greater than x is observed to increase. The bottom left plot shows the histogram of observed minimum distances from the hyperplane using all successful cleantech companies. Mis-labeled successful companies appear to the left of $x = 0$ and can be observed as the solid circles in the red region of the feature space in the top left plot. Finally, the bottom right plot shows the histogram of observed minimum distances from the hyperplane using all failed cleantech companies. Mis-labeled failed companies appear to the right of $x = 0$ and can be observed as the empty squares in the green region of the feature space in the top left plot.

Example 1: To further illustrate the univariate ranking approach, consider the following pedagogical example in which an unlabeled company has a corresponding point whose minimum Euclidean distance is 0.1 from the separating hyperplane. The simple labeled training set is shown in Table 4 below and includes two points which are misclassified. In this example, each prior is 0.5 and the likelihood functions $P(X > 0.1 \mid \mathcal{H} = S)$ and $P(X \leq 0.1 \mid \mathcal{H} = F)$ are estimated as the proportion of successful observations for which $X > 0.1$ and the proportion of failed observations for which $X \leq 0.1$ respectively. From Equation 1,

$$P(\mathcal{H} = S \mid X > 0.1) = \frac{\frac{4}{5} \cdot \frac{1}{2}}{(\frac{1}{2} \cdot \frac{4}{5} + \frac{1}{2} \cdot \frac{1}{5})} = \frac{4}{5} \quad (2)$$

and

$$P(\mathcal{H} = F \mid X \leq 0.1) = \frac{\frac{4}{5} \cdot \frac{1}{2}}{(\frac{1}{2} \cdot \frac{4}{5} + \frac{1}{2} \cdot \frac{1}{5})} = \frac{4}{5} \quad (3)$$

The score is given by the difference of the two posterior probabilities and is 0, which indicates that the predicted outcome is equally likely to be either class and the result is thus uncertain. The score increases in magnitude as the minimum signed Euclidean distance of an unlabeled point from the hyperplane can be predicted with increased certainty. Note that the effect of mis-classification cancelled through. Note also that although the priors are equal here, by construction the approach generalizes to non-equal priors.

5.2. A multivariate Bayesian approach

The proceeding discussion has thus far only considered models trained from feature pairs. From each set of minimum Euclidian distances from points in a two dimensional feature space, the previous section demonstrated a Bayesian approach for estimating the posterior probability of a company succeeding or failing. Through applying this approach separately to each combination of feature pairs listed in Table 3, a scoring system can be derived for evaluating the relative effect of feature pairs on the likelihood of a company succeeding or failing. Because each score is associated with a particular feature pair, it is not meaningful however to compare companies using such a scoring system. We now propose a more general approach, referred to as the "multivariate approach", for evaluating the combined effect of many features and arriving at a scoring system for ranking companies in a cohort.

Recall that Table 1 lists a set of seven features extracted from investment histories of 3,776 cleantech companies provided in VentureSource. The right-hand column of the table shows the percentage of missing data for each feature and the last row shows that 89% percent of the cleantech companies have an unknown outcome (unlabeled). This leaves an upper bound of only 411 labeled cleantech companies from which to train and test the model. The actual number of labeled data points available depends on the feature set used to train the model. One severe limitation is that the number of labeled data points, for which the feature set is complete, diminishes as more features are included in the set. So, for example, a feature set of the seven features listed in the Table 1 results in fewer than 90 labeled complete observation points, whereas for example a feature set consisting of the "Number of Investors", the "Number of Rounds" and the "Total Money In" has 346 labeled complete observation points for cleantech companies. This data sparsity pattern severely limits the application of SVM classifers to higher dimensional feature spaces in nascent industry sectors.

To partially address this problem we extend the Bayesian approach described in the previous section by combining all qualifying SVM classifiers in a set $\mathcal{M}$. Put informally, we attempt to estimate the probability that a company will succeed or fail based on a multivariate set of Euclidian minimum distances obtained by training separate SVM classifiers on feature pairs. Each SVM classifier is trained from a complete labeled data for a feature pair to yield the univariate set of minimum Euclidian distances of points to the separating hyperplane. Each univariate set of distances is then aggregated to provide a panel of distances, where each column corresponds to a feature pair and each row to a labeled company. This approach can be stated more formally:

Definition 5.2.1 (Multivariate observed distances) *The observed data is the set of minimum distances between observed points in feature spaces and their separating hyperplanes, each given by a separate model trained with different features. Let $\mathcal{M}$ denote the set of model identifiers whose prediction accuracy on data is at least 0.8. Denote $X_i \in \mathbf{X}$ as the set of minimum distances from a point to the separating hyperplane in feature space $i \in \mathcal{M}$ and $\mathbf{X}$ the panel of all* complete *observed distances in the feature spaces corresponding to companies in a particular industry sector. Define the event $E_i := (X_i > x_i)$ so that $\mathcal{E} := \cap_{i \in \mathcal{M}} E_i$ is the realized combined event from which the probability of success is estimated. Conversely, define the event $\bar{E}_i := (X_i \leq x_i)$ so that $\bar{\mathcal{E}} := \cap_{i \in \mathcal{M}} \bar{E}_i$ is the realized event from which the probability of failure is estimated.*

The posterior probability of success conditional on each distance $X_i > x_i$, is given by

$$P(\mathcal{H}=S|\mathcal{E}) = \frac{P(\mathcal{E}|\mathcal{H}=S)P(\mathcal{H}=S)}{P(\mathcal{H}=S)P(\mathcal{E}|\mathcal{H}=S)+(1-P(\mathcal{H}=S))P(\mathcal{E}|\mathcal{H}=F)} \quad (4)$$

and conversely, the posterior probability of failure conditional on each distance $X_i \leq x_i$ is given by

$$P(\mathcal{H}=S|\mathcal{E}) = \frac{P(\mathcal{E}|\mathcal{H}=S)P(\mathcal{H}=S)}{P(\mathcal{H}=S)P(\mathcal{E}|\mathcal{H}=S)+(1-P(\mathcal{H}=S))P(\mathcal{E}|\mathcal{H}=F)} \quad (5)$$

The final score $s(\{x_i\}_{i\in\mathcal{M}})$ for a company is obtained by subtracting $P(\mathcal{H}=F|\bar{\mathcal{E}})$ from $P(\mathcal{H}=S|\mathcal{E})$ and is a function with a range in $[-1,1]$. This score function is illustrated in Figure 3 for the case when two models are combined where x_1 and x_2 represent threshold distances on two different features spaces.

Example 2: To illustrate the multivariate ranking approach, we extend Example 1 to include two models and let the unlabeled company have corresponding points whose minimum Euclidean distance from each of the separating hyperplanes is 0.1 and 0.2. The simple labeled training set is shown in Table 5 below and includes two points which are mis-classified. In this example, each prior is 0.5 and the likelihood functions $P(X_1 > 0.1, X_2 > 0.2 \mid \mathcal{H}=S)$ and $P(X_1 \leq 0.1, X_2 \leq 0.2 \mid \mathcal{H}=F)$ are estimated as the proportion of successful observations for which $X_1 > 0.1$ and $X_2 > 0.2$, and the proportion of failed observations for which $X_1 \leq 0.1$ and $X_2 \leq 0.2$ respectively. From Equation 4,

$$P(\mathcal{H}=S \mid X_1 > 0.1, X_2 > 0.2) = \frac{\frac{4}{5}\cdot\frac{1}{2}}{(\frac{1}{2}\cdot\frac{4}{5}+\frac{1}{2}\cdot\frac{1}{5})} = \frac{4}{5}, \quad (6)$$

and from Equation 5

$$P(\mathcal{H}=F \mid X_1 \leq 0.1, X_2 \leq 0.2) = \frac{\frac{4}{5}\cdot\frac{1}{2}}{(\frac{1}{2}\cdot\frac{4}{5}+\frac{1}{2}\cdot\frac{1}{5})} = \frac{4}{5}. \quad (7)$$

The score is given by the difference of the two posterior probabilities and is again 0.

It should be made clear that the complete panel of distances $\mathbf{X}$ is aggregated only from the labeled companies with observed distances for all feature pairs corresponding to the set of qualifying models $\mathcal{M}$. So if a labeled company is missing one or more of the features used to train any model in $\mathcal{M}$, then that company can no longer be included as a row in the panel $\mathbf{X}$. Clearly, as the number of models increases, this rule poses an increasingly severe restriction on the number of data points used to provide a non-parametric estimate of the posterior probabilities and we illustrate this point in Section 6. This restriction can be overcome with the assertion of a parametric posterior probability, but this is beyond the scope of this paper.

5.3. Unlabeled companies with missing features

The above approach describes the simplest scenario when all required features of an unlabeled company with index j are available from which to estimate the posterior probability. In practice, most unlabeled companies do not have a complete feature set and it is necessary to define the subset of models $\mathcal{M}^j \subseteq \mathcal{M}$, corresponding to the available features, from which to estimate the posterior probabilities. This means that the set of combined models may be different for each unlabeled company and hence the number of labeled companies used to estimate the posterior probability will vary too. The exact number will depend on which labeled companies have the complete set of features corresponding to the subset of models $\mathcal{M}^j$. Denoting the set of labeled company indices for which the i^{th} Euclidian distance is available as $\mathcal{I}_i$, then the reduced complete panel $\mathbf{X}^j \subseteq \mathbf{X}$ is aggregated only from the companies whose indices are in the set $\mathcal{I}^j := \cap_{i\in\mathcal{M}^j}\mathcal{I}_i$.

For the i^{th} model in $\mathcal{M}^j$, the corresponding feature pair of the j^{th} unlabeled companies are provided as input to predict the minimum Euclidean distance x_i from the separating hyperplane. If $\mathcal{I}_s$ and $\mathcal{I}_f$ denote the respective set of indices of all labeled successful and failed companies, then the posterior probabilities of success and failure are separately estimated using a panel of all labeled companies with respective indices $\mathcal{I}^j \cap \mathcal{I}_s$ and $\mathcal{I}^j \cap \mathcal{I}_f$.

This scoring approach bypasses the more severe limitation of being unable to train models on higher dimensional feature sets of labeled company data because of insufficient numbers of labeled companies with complete features. The requirement to only include a labeled company with complete features in a training set results in a great number of companies with missing features being disregarded. The proposed approach instead attempts to use as much of the available historical information as possible to arrive at a score. Each model is trained from a feature pair and must meet a prediction accuracy rate threshold (≥ 0.80) on

data in order for the model output to be included as a component of the panel $\mathbf{X}_j$. Furthermore, any differences between the set sizes $\left|\mathcal{I}^j \cap \mathcal{I}_s\right|$ and $\left|\mathcal{I}^j \cap \mathcal{I}_f\right|$ are corrected for through the prior probability estimates:

$$\tilde{P}(\mathcal{H} = S) := \tfrac{\left|\mathcal{I}^j \cap \mathcal{I}_s\right|}{\left|\mathcal{I}^j\right|}, \qquad \tilde{P}(\mathcal{H} = F) := \tfrac{\left|\mathcal{I}^j \cap \mathcal{I}_f\right|}{\left|\mathcal{I}^j\right|}. \tag{8}$$

The likelihood function estimates $\tilde{P}(\mathcal{E}|\mathcal{H} = S)$ and $\tilde{P}(\bar{\mathcal{E}}|\mathcal{H} = F)$ are also respectively estimated as the proportion of the set size $\left|\mathcal{I}^j \cap \mathcal{I}_s\right|$ and $\left|\mathcal{I}^j \cap \mathcal{I}_s\right|$ over which event $\mathcal{E}$ and $\bar{\mathcal{E}}$ is observed. We emphasize that unlike a Naive Bayes' classifier, we do not impose class conditional independence on the events.

6. Step IV: Ranking Unlabeled Companies

We now demonstrate how the above Bayesian scoring approach is applied to a set of 81 agricultural cleantech companies using models trained from labeled cleantech companies. The top and bottom five unlabeled companies in the agricultural cleantech sector are respectfully shown in Tables 6 and 7.

Each table lists a particular model output for each company whose rank decreases from from left to right. The model ID is shown in the first column and has been selected based on its prediction accuracy and the availability of the feature pair for at least one of the five companies shown in each table. A missing entry indicates that one or both of the features used to train the model were unavailable for the company. The final score for each company together with the set sizes $\left|\mathcal{I}^j \cap \mathcal{I}_s\right|$ and $\left|\mathcal{I}^j \cap \mathcal{I}_f\right|$ are shown at the bottom of each table. We observe that these sizes vary between each company based on its available features and hence the number of models used. In some cases, only seven models are used and in others up to 16 models. The reader should refer to Table 8 for the specification of the model and feature pairs corresponding to each model ID.

It is recognized that there is some redundancy in combining models trained with the same features appearing in multiple feature pairs, however, it is our objective here to use as many of the models with high prediction rate for the purpose of ranking the companies. It is further recognized that the final ranking order is sensitive to the choice of the prediction accuracy threshold constant because this determines which models qualify for inclusion in $\mathcal{M}$. Returning to the remark at the end of the Section 5.2 concerning scalability of this approach, we observe that $\left|\mathcal{I}^j \cap \mathcal{I}_s\right|$ and $\left|\mathcal{I}^j \cap \mathcal{I}_f\right|$ in Table 6 are smaller when the score is based on a larger set of qualifying model outputs. This means that the estimate of the posterior probabilities, but not the SVM models, inevitably deteriorates as the population size decreases.

This should however be contrasted with an approach based solely on estimating a score from the output of a SVM classification model trained on a higher dimensional feature set. Due to missing values, the quality of the output from the SVM model itself will inevitably deteriorate as the dimensionality of the feature set is increased because there are far fewer labeled companies with complete higher dimensional feature sets than there are with at least a pair of features. Our approach is able to maximize the number of labeled companies that can be used to train SVM classification models and the quality of each model does not depend on the number of features which an investor would like to include in the ranking estimate. An additional advantage of this approach is that it requires no restrictive distributional assumptions on the model output and no subjective set of weights for each model (which can often be misleading when the model outputs are correlated) but instead accounts for correlation between model outputs through the representation of the likelihood function as a joint probability estimate.

7. Conclusion

Private equity investors seek to identify potential investment opportunities in growth stage private companies and rank their prospects relative to a cohort of companies such as an industry sector. Growth stage private companies often have investment transaction histories from which industry specific characteristics associated with successful and failed companies may be discerned using statistical methods. In general, one of the primary challenges in pursuing this approach is the sparsity of historical data on private companies which is exacerbated in nascent sectors by the relatively few number of observed exits. Furthermore, the

labeled historical data is not always linearly separable and we turn to Support Vector Machine (SVM) classifiers to represent non-linear boundaries in low dimensional feature space. The data sparsity, however, prohibits this approach scaling to higher dimensional feature sets and we introduce a non-parametric Bayesian approach to combine the SVM classifiers and yield an overall ranking of likely success for an unlabeled company.

This paper describes a four step predictive approach for ranking private companies within a cohort which can be applied to sparse industry specific historical data. Each step is illustrated using a set of seven company features extracted from a database of 411 labeled cleantech companies. In Section 3, we described the first step of feature extraction and then feature selection based on Hypothesis testing and in Section 4, we proceeded to specify the configuration of SVM classification models applied to feature pairs and presented results on the performance accuracy of each model. In Section 5 we defined a novel non-parametric Bayesian approach for scoring companies using results from a set of qualified SVM models trained on different feature pairs. Finally in Section 6, we demonstrated this approach by ranking a set of 81 unlabeled agricultural companies in the cleantech industry sector using up to 16 SVM classifiers each trained on a different pair of features.

The main advantage of this ranking approach is that it includes labeled companies with missing features which would otherwise be excluded if the approach was based on the output of a single classifier trained from seven dimensional feature sets. Furthermore, the approach does not require parametric representation of the likelihood function nor does it require independence of the model results. The rankings do, however, depend on the threshold prediction accuracy for the model to be qualified and the population size from which to estimate the posterior probabilities decreases as more models are aggregated to yield the score, although the number of labeled companies used to train the models remains constant.

Being able to include labeled companies with missing data is a critical step towards a machine learning based methodology for ranking companies relative to an industry sector. We anticipate that this approach will not only be of interest to statisticians and machine learning specialists with an interest in venture capital and private equity but extend to a broader readership whose interests lie in classification models applied to finance and economics where missing data is the primary obstacle.

Acknowledgements

The authors would like to thank the anonymous referees for their useful advice which substantially improved the accessibility and readibility of this paper.

References

[1] H.S. Bhat and D. Zaelit, Predicting private company exits using qualitative data, *Advances in Knowledge Discovery and Data Mining*, Lecture Notes in Computer Science, 6634, 2011, 399–410.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*, Information Science and Statistics, Springer-Verlag, New York, 2006.

[3] C.C. Chang and C.J. Lin, LIBSVM : a library for support vector machines, *ACM Transactions on Intelligent Systems and Technology* **2**(3)(2011), 1–27.

[4] C. Farmer, The Top Ten VC Firms According to Investor Rank, Presentation at Disrupt NYC, USA, 2011.

[5] P. Gompers, A. Kovner, J. Lerner and D. Scharfstein, Performance Persistence in Entrepreneurship, *Journal of Financial Economics* **96**(1)(2010), 18–32.

[6] J. Hall and C. W. Hofer, Venture capitalists decision criteria in new venture evaluation, *Journal of Business Venturing* **8**(1)(1993), 25–42.

[7] J.Hilbe, *Logistic Regression Models*, Chapman & Hall/CRC Texts in Statistical Science, 2009.

[8] Y. V. Hochberg, A. Ljungqvist and Y. Lu, Whom You Know Matters: Venture Capital Networks and Investment Performance, *Journal of Finance* **62**(1)(2007), 251–301.

[9] T. Hothorn, H. Hornik, M. van de Wiel and A. Zeileis, coin: A Computational Framework for Conditional Inference, 2010.

[10] D. Meyer, Support Vector Machines: The Interface to libsvm in package e1071, Technical report, Technische Universität Wien, Austria, 2011.

[11] L. Page, S. Brin, R. Motwani. and T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web, Technical Report, Stanford InfoLab, USA, 1999.

[12] T.T. Tyebjee and A. V. Bruno, A Model of Venture Capitalist Investment Activity, *Management Science* **30**(9)(1984), 1051–1066.

[13] W. A. Wells, Venture capital modelling: Evaluation criteria for the appraisal of investments, *The Financier ACMT* **1**(2)(1974), 54–64.

[14] A.L.Zacharakis and G. D. Meyer, The potential of actuarial decision models, *Journal of Business Venturing* **15**(4)(2000), 323–346.

Tables

Feature	Explanation	%Missing
Number of Rounds	The number of funding rounds a company has entered	6%
Number of Investors	The total number of investment institutions that have funded the company	13%
Cleantech Transaction Frequency	The relative frequency of transactions compared to the cleantech sector average	73%
Total Money In (US $ M)	The amount of all investment in a company	40%
Last Round Investment (US $ M)	The most recent investment in the company	40%
Last Round PMV (US $ M)	The most recent valuation of the company following its last funding round	86%
Cleantech Prestige	The "prestige" of a company based on the "prestige" of its investors in the cleantech industry (see Section 3.1)	40%
Label	The type of the labeled company's exit event (either success or failure)	89%

Table 1

A list of some of the features extracted from VentureSource and the % of labeled companies with missing observations for each feature.

Feature	%p-value
Total Rounds	< 0.001
Number of Investors	< 0.001
Cleantech Transaction Frequency	< 0.001
Total Money In (US $ M)	< 0.001
Last Round Investment (US $ M)	< 0.001
Last Round PMV (US $ M)	< 0.001
Cleantech Prestige	< 0.1
Z	$\mu = 0.50,\ \sigma = 0.29$

Table 2

The null Hypothesis for each of the features shown here is rejected at the 10% significance level or lower.

				GLM				SVM	
x-feature	y-feature	size	err	stderr	p_x	p_y	err	stderr	kernel
Total Money In	Total Rounds	238	0.164	0.01	0	0.383	0.162	0.009	linear
Total Money In	Cleantech Transaction Freq.	137	0.27	0.019	0	0.062	0.259	0.026	linear
Last Round Investment	Total Rounds	237	0.154	0.009	0	0.016	0.152	0.012	linear
Last Round Investment	Cleantech Transaction Freq.	137	0.274	0.03	0	0.077	0.255	0.036	radial
Last Round Investment	Total Money In	237	0.161	0.016	0.001	0.472	0.152	0.019	linear
Number of Investors	Total Rounds	273	0.365	0.029	0.28	0	0.296	0.011	radial
Number of Investors	Total Money In	233	0.157	0.018	0.285	0	0.129	0.021	radial
Number of Investors	Last Round Investment	232	0.167	0.025	0.62	0	0.166	0.022	linear
Last Round PMV	Total Rounds	129	0.113	0.015	0	0.008	0.122	0.022	linear
Last Round PMV	Cleantech Transaction Freq.	72	0.256	0.052	0.009	0.097	0.233	0.056	radial
Last Round PMV	Total Money In	129	0.124	0.024	0.061	0.344	0.131	0.022	linear
Last Round PMV	Last Round Investment	129	0.111	0.025	0.009	0.869	0.119	0.023	linear
Last Round PMV	Number of Investors	126	0.153	0.018	0	0.076	0.137	0.016	linear
Cleantech Prestige	Total Rounds	212	0.291	0.013	0.013	0	0.272	0.021	radial
Cleantech Prestige	Total Money In	192	0.162	0.009	0.001	0	0.156	0.011	linear
Cleantech Prestige	Last Round Investment	192	0.155	0.015	0.009	0	0.148	0.015	linear
Cleantech Prestige	Last Round PMV	101	0.142	0.04	0.316	0	0.132	0.02	linear

Table 3

This table compares the performance of bootstrapped logistic regression (GLM) with SVM classification over all combinations of feature pairs drawn from a set of seven features. From left to right, the name of the features represented by the x and y co-ordinates in each feature space are shown in the first two adjacent columns. The third column shows the number of labeled columns with complete data for each feature pair. The mean and standard deviation of the error rate for the logistic regression classifier and the SVM classifier are shown in the adjacent columns. The kernel type which yields the SVM classifier with the lowest error rate is shown in the far right column.

ID	X	$\mathcal{H}$
1	1	S
2	0.8	S
3	0.6	S
4	0.4	S
5	-0.2	S
6	0.2	F
7	-0.4	F
8	-0.6	F
9	-0.8	F
10	-1	F

Table 4

The labeled training set for Example 1.

ID	X_1	X_2	$\mathcal{H}$
1	1	0.4	S
2	0.8	0.5	S
3	0.6	0.7	S
4	0.4	0.3	S
5	-0.2	0.1	S
6	0.2	-0.1	F
7	-0.4	-0.4	F
8	-0.6	-0.8	F
9	-0.8	-0.6	F
10	-1	-0.5	F

Table 5
The labeled training set for Example 2.

Rank	1	2	3	4	5
Model ID	Absorbent Technologies	Guangxi Fenglin Group	Targeted Growth Inc.	Marrone Bio Innovations Inc.	GAT Microencapsulation
3	0.37	0.55	0.51	0.46	0.13
4	-0.33	0.01	0.48	0.40	0.17
6	0.89	0.94			0.98
12	0.90	0.91			0.63
14	0.83	0.98	0.93	0.94	-0.14
15	-0.46	-0.15	0.77	0.84	-0.33
16	0.87	0.81	0.81	0.70	-0.21
17	1.02	0.90			0.38
18		1.01	0.86	0.58	0.19
19	-0.28	-0.11	0.39	0.44	-0.01
20	-0.30	-0.09	0.47	0.60	0.03
21	0.79	0.61			0.31
22		-0.11	0.39	0.38	0.01
24	0.90	0.89			0.88
26	0.97	0.70			0.22
27		0.22			0.08
Score	0.29	0.28	0.27	0.22	0.18
$\lvert\mathcal{I}^j \cap \mathcal{I}_s\rvert$	89	73	153	153	73
$\lvert\mathcal{I}^j \cap \mathcal{I}_f\rvert$	21	21	129	129	21

Table 6
The top five ranked agricultural companies in the cleantech sector as predicted by the Bayesian ranking methodology. Each row in the top portion of the table corresponds to a qualified model whose ID can be used to trace the feature pairs listed in Table 8. Each column shows the model output for each company which is a signed minimum Euclidean distance from the observed point in the feature space to the nearest point on the separating hyperplane.

Rank	77	78	79	80	81
Model ID	MYCOSYM International AG	Ecovegetal SAS	Hydroprotect	Aqua -Biokem	OrganicOcean Inc.
3	-0.25	-0.55	-1.27	-1.27	-1.27
4	-0.86	-0.50	-1.19	-1.19	-1.19
14	-0.90	-0.95	-0.98	-0.98	-0.98
15	-0.76	-0.94	-0.96	-0.96	-0.96
16	-0.90		-1.01	-1.01	-1.01
18	-0.84		-0.97		
19	-1.11	-0.39	-1.14	-1.14	-1.14
20	-1.10		-1.16	-1.16	-1.16
22	-1.19		-1.19		
Score	-0.23	-0.24	-0.30	-0.30	-0.30
$\left\|\mathcal{I}^j \cap \mathcal{I}_s\right\|$	153	203	153	199	199
$\left\|\mathcal{I}^j \cap \mathcal{I}_f\right\|$	129	143	129	140	140

Table 7

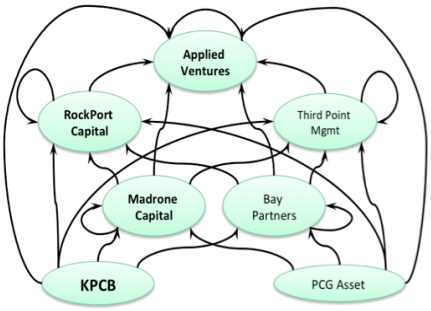
The bottom five ranked agricultural companies in the cleantech industry sector as predicted by the Bayesian ranking methodology. Each row in the top portion of the table corresponds to a qualified model whose ID can be used to trace the feature pairs listed in Table 8. Each column shows the model output for each company which is a signed minimum Euclidean distance from the observed point in the feature space to the nearest point on the separating hyperplane.

Model ID	x-feature	y-feature
3	Total Rounds	Total Money In (US$ M)
4	Total Rounds	Last Round Investment (US$ M)
6	Total Rounds	Last Round PMV (US$ M)
12	Cleantech Transaction Frequency	Last Round PMV (US$ M)
14	Total Money In (US$ M)	Total Money In (US$ M)
15	Total Money In (US$ M)	Last Round Investment (US$ M)
16	Total Money In (US$ M)	Number of Investors
17	Total Money In (US$ M)	Last Round PMV (US$ M)
18	Total Money In (US$ M)	Cleantech Prestige
19	Last Round Investment (US$ M)	Last Round Investment (US$ M)
20	Last Round Investment (US$ M)	Number of Investors
21	Last Round Investment (US$ M)	Last Round PMV (US$ M)
22	Last Round Investment (US$ M)	Cleantech Prestige
24	Number of Investors	Last Round PMV (US$ M)
26	Last Round PMV (US$ M)	Last Round PMV (US$ M)
27	Last Round PMV (US$ M)	Cleantech Prestige

Table 8

This table shows the mapping of model IDs to the feature pair.

Figures



Round 1	Applied Ventures
Round 2	Applied Ventures, RockPort Capital Partners, Third Point Management
Round 3	Applied Ventures, RockPort Capital Partners, Third Point Management, Madrone Capital Partners, Bay Partners
Round 4	Applied Ventures, RockPort Capital Partners, Third Point Management, Madrone Capital Partners, Bay Partners, KPCB, PCG Asset management LLC
Round 5	IPO

Fig. 1. An illustrative funding graph for the funding rounds of Enphase Energy Inc. (a cleantech company) which exemplifies Farmer's peer-evaluation model [4].

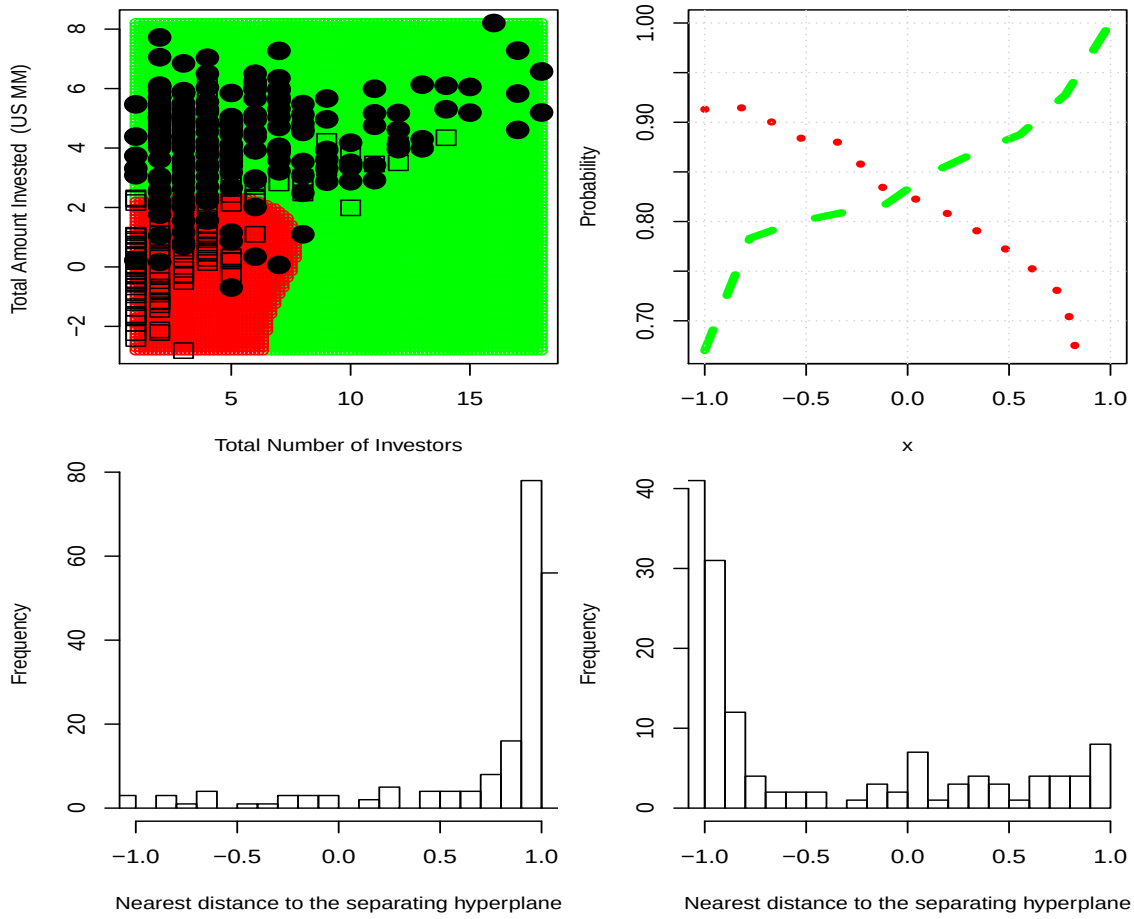


Fig. 2. (Top left) The figure shows the feature space of the log of the Total Amount Invested (US MM) against the Number of Investors for all labeled companies in the cleantech sector. This feature space is divided into two regions by the separating hyperplane. The red region represents failed companies and the green region represents successful companies. Circles falling within the red region or squares within the green region indicate mis-classification. (Top right) The posterior probabilities of success or failure as a function of x are shown with the green dashed line and the red dotted line respectively. The posterior probability that the nearest distance X from a failed company data point to the separating hyperplane is less than or equal to x is observed to decrease from left to right. Conversely, the posterior probability that the nearest distance X from a successful company data point to the separating hyperplane is greater than x is observed to increase. (Bottom left) The histogram of observed minimum distances from the hyperplane is shown for all successful cleantech companies. Mis-labeled successful companies appear to the left of $x = 0$ and can be observed as the solid circles in the red region of the feature space in the top left figure. (Bottom right) The histogram of observed minimum distances from the hyperplane is shown for all failed cleantech companies. Mis-labeled failed companies appear to the right of $x = 0$ and can be observed as the empty squares in the green region of the feature space in the top left figure.

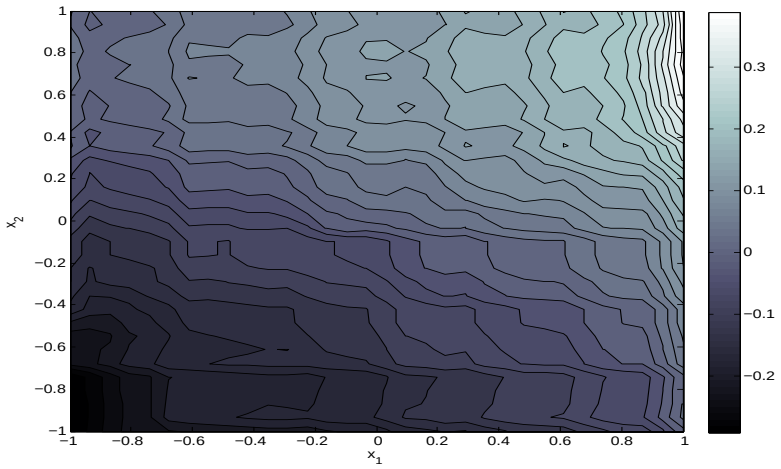


Fig. 3. A contour plot of the score surface over the domain $\Omega := [-1, 1] \times [-1, 1]$ in which x_1 and x_2 respectfully represent the threshold value of X_1 and X_2. X_1 denotes the shortest Euclidean distances from a point to the separating hyperplane in a feature space in which the x-feature is the "Total Number of Investors" and the y-feature is "Total Amount Invested". X_2 denotes the same quantity for a feature space in which the x-feature is the "Number of Financing Rounds" and the y-feature is "Total Amount Invested".