Predicting Private Company Exits using Qualitative Data

H. S. Bhat[1] and D. Zaelit[2]

[1] University of California, Merced, 5200 N. Lake Rd., Merced, CA 95343, USA,
hbhat@ucmerced.edu
[2] SVB Analytics, 555 Mission St., San Francisco, CA 94105, USA,
dzaelit@svb.com

Abstract. Private companies backed by venture capitalists or private equity funds receive their funding in a series of rounds. Information about when each round occurred and which investors participated in each round has been compiled into different databases. Here we mine one such database to model how the private company will exit the VC/PE space. More specifically, we apply a random forest algorithm to each of nine sectors of private companies. Resampling is used to correct imbalanced class distributions. Our results show that a late-stage investor may be able to leverage purely qualitative knowledge of a company's first three rounds of funding to assess the probability that (1) the company will not go bankrupt and (2) the company will eventually make an exit of some kind (and no longer remain private). For both of these two-class classification problems, our models' out-of-sample success rate is 75% and the area under the ROC curve is 0.83, averaged across all sectors. Finally, we use the random forest classifier to rank the covariates based on how predictive they are. The results indicate that the models could provide both predictive and explanatory power for business decisions.

1 Introduction

Venture capitalists (VC's) face the challenge of choosing a few outstanding investments from a sea of thousands of potential opportunities. A VC funds a startup company with cash in exchange for an equity stake. From this point of view, it may appear that the dynamics of the transaction are similar to that of an investor buying shares of a publicly traded company. Such appearances are false. When a VC funds a startup, the VC often takes an active role in managing the startup, providing expertise and advice in both managerial and technical areas. In this way, the experience and wisdom of the VC's who invest in a startup directly influence the startup's trajectory.

When confronted with a company they have not seen before, one question that potential investors would like to be able to answer is: how will this company eventually *exit* the private equity space? In this study, we assume that the final outcome of a private company will be one of five outcomes. The private company can (1) go bankrupt, (2) proceed via an initial public offering (IPO) to become a

publicly traded company, (3) be subject to a leveraged buyout (LBO), (4) merge with or be acquired by another company, or (5) stay private.

Our goal in this paper is to use information about *who* invests in a private company and *when* these investments are made to predict how the company will exit. Our prediction is generated by a statistical model inferred from data available through the Private Equity module of ThomsonONE, a data set formerly known as VentureXpert. Numerous academic libraries have access to this database, and it is often used as a source of data for research papers in the VC/PE space—see [2,5,6,12,13]. Here we develop a method for converting each VC- or PE-backed company in the database into a list of numerical and nominal attributes with a class label corresponding to one of the five possible states; we then use this list of labeled instances to train and test a classifier.

The random forest algorithm [3], [10, Chap. 15] and other machine learning algorithms such as support vector machines and boosting are available as free codes, implemented in a variety of languages and environments. Such algorithms have proven useful in a wide variety of applications. Though it seems very natural to leverage machine learning algorithms and large databases to model the exits of VC-/PE-backed companies, to the best of our knowledge, this is the first study to do so. As such, this paper represents a first attempt at solving the problem.

Our analysis shows that a late-stage investor may be able to use knowledge of a company's first three rounds of funding together with a random forest classifier to assess the probability that the company will not go bankrupt, and also to assess the probability that the company will eventually make an exit of some kind (and no longer remain private). For both of these two-class classification problems, our models' average success rate across all sectors is 75% and the average area under the ROC curve is 0.83.

In what follows, we discuss the details of our procedure, starting from the data, proceeding to issues of representation, investor ranking and instance resampling, and then on to specific working models and their associated results.

2 Data Extraction and Representation

For this study, we focused on the following attributes:

- The year in which the company was founded. See the right panel of Figure 1 for a histogram of the inception years for all companies in Sector 6 (energy) used in this study—the most populated decade is the decade from 2000 to the present. Other sectors' distributions are similar.
- The company's sector, encoded as a four-digit number.
- The *rounds*, i.e., dates on which the company received funding.
- A list of historical investors in the company. This list includes each investor's name, type, and a list of the rounds in which that investor participated.
- The company's exit status.

	Exit					
Sector	**Bankrupt**	**IPO**	**LBO**	**M&A**	**Private**	**Totals**
1xxx: Communications	1540	826	295	2073	3289	8023
2xxx: Computer	3197	1541	653	4872	9941	20204
3xxx: Electronics	511	604	215	853	1852	4035
4xxx: Biotech/Pharma	283	458	67	509	1652	2969
5xxx: Medical/Health	704	740	412	1228	2787	5871
6xxx: Energy	204	287	150	238	850	1729
7xxx: Consumer	1317	865	1715	1395	4997	10289
8xxx: Industrial	614	473	1009	974	2763	5833
9xxx: Other	2374	1675	2636	2307	9215	18207
Totals	10744	7469	7152	14449	37346	77160

Table 1. Summary of exit types by broad sector category for 77,160 private companies. All companies are either formerly or currently VC- or PE-backed.

All of these attributes have to do with *who* invested in the private company and *when* the investment was made. Notably, the amount of money invested by the investor in each round of funding, as well as the pre- and/or post-money valuations of the companies are absent. In short, our study makes use of qualitative rather than quantitative features of a private company's investment history.

Let us elaborate on a few of the attributes mentioned above. The company's exit status is, in the original data set, a nominal attribute with 12 possible values. Since exit status is the class variable, we group a few of these categories together to reduce the number of classes from 12 to five. We list here the five class labels in italics together with the original exit types contained in each class:

1. *Bankrupt:* Defunct, Bankruptcy - Chap. 7, Bankruptcy - Chap. 11
2. *IPO:* Went Public
3. *LBO:* LBO
4. *M&A:* Acquisition, Merger, Pending Acquisition
5. *Private:* Active, Other, In Registration, and Private Company (Non-PE)

The company's market sector is encoded as a four-digit number. The first digit of this four-digit number gives us a broad sector categorization, as seen in the left-most column of Table 1, which also shows the breakdown of exits by sector. One can readily see two trends. First, the classes are imbalanced, necessitating the use of a resampling procedure described in Section 3.1.

Second, different sectors behave differently:

- For sector 2xxx, only 3.23% of companies had an LBO exit, while for sector 8xxx, 17.30% of companies exited via LBO.
- For sector 6xxx, 16.60% of exits are IPO and 13.77% of exits are M&A. For sector 1xxx, 10.30% of exits are IPO and 25.84% of exits are M&A.

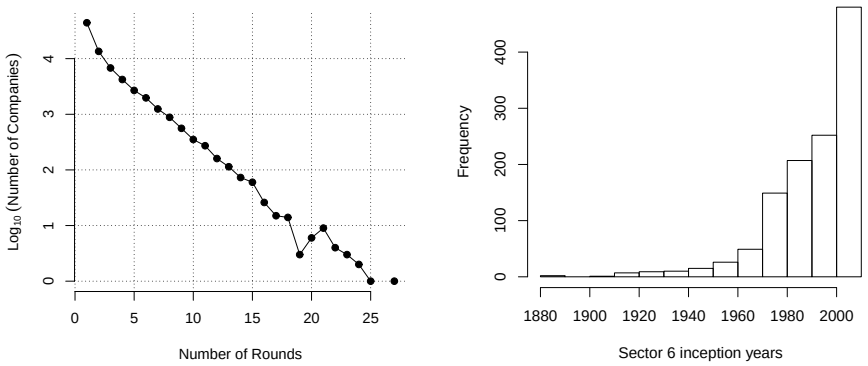

Fig. 1. In the left panel, we plot $\log_{10}$ of the number of companies in the database with x rounds of funding, as x goes from 1 to 27. Linear decay in this plot shows that the same plot with a non-logarithmic y-axis would feature exponential decay. In the right panel, we plot the distribution of inception years for Sector 6 (energy) companies.

It is plausible that the reason the percentages differ so much from one sector to the next is that the factors influencing success/failure differ greatly from one sector to another. For these reasons, in the present study, we shall segregate the data by the broad sectors indicated in Table 1.

2.1 Social Network Ranking

Here we explain how we turn the investor name into an attribute. The social network of coinvestment plays a key role. There are 9545 unique investors in our data set, so we seek a mapping from the set of all investor names to the set of integers $P = \{1, 2, 3, \ldots, 9545\}$.

Let each investor be a node, and join two nodes by an edge if the two investors both invested in the same company at some point of time. To repeat, the coinvestment need not occur at the same time. Once we form the adjacency matrix for this social network, we sort investors by degree. Ties are broken simply by using the order in which we encounter the investor as we parse the data. Once the investors are sorted by degree, we have our mapping: the investor is mapped to its position $p \in P$ in the sorted list.

In the sorted list, the top two investors, Undisclosed Firm and Individuals, are placeholders that do not correspond to any one firm. The next 10 investors are: J. P. Morgan Partners (FKA: Chase Capital Partners), New Enterprise Associates, Inc., Intel Capital, Kleiner Perkins Caufield & Byers, Oak Investment Partners, Sequoia Capital, Goldman, Sachs & Co., Mayfield Fund, HarbourVest Partners LLC, and Bessemer Venture Partners. These names should be familiar to those

who follow the VC/PE space, indicating that even a rough social network ranking does correspond to intuitive/anecdotal rankings of VC's and PE funds.

Next we turn to the investor type. This is a nominal attribute with 18 possible values: Development, Buyouts, Seed Stage, Balanced Stage, Recap, Unknown, Energy, Early Stage, Expansion, Fund of Funds, Mezzanine Stage, Later Stage, Turnaround, Distressed Debt, Real Estate, Other Private Equity, Secondary Funds, and Generalist. There is effectively a 19th possible value when the type of the investor is not listed, i.e., the datum is missing. We let Q be the set of 19 possible investor types.

2.2 Mapping Companies to N-tuples

In what follows, we use the term vector to mean an N-tuple $\mathbf{x} = (x_1, x_2, \ldots, x_n)$; one N-tuple represents one company. All the ingredients are in place to define a function that maps companies to vectors. One issue is that the number of rounds of funding enjoyed by a private company varies from one company to the next. Let $n(x)$ be the number of companies that have received precisely x rounds of funding; Figure 1 shows $\log_{10} n(x)$ versus x. The graph is approximately linear, indicating an exponential decay of $n(x)$. There are two considerations to make:

- The maximum number of rounds for any company is 27, yet 92.5% of companies have at most 5 rounds of funding.
- From the point of view of applicability, a model that predicts exit type accurately using *fewer* rounds worth of investor information is preferable.

Based on both considerations, we develop a round-by-round representation of the data. We find that there are a maximum of 31 investors in any round of funding. One round of funding then corresponds to a vector $(\mathbf{p}, \mathbf{q}) \in P^{31} \times Q^{31}$, where P and Q were both defined in Section 2.1. We have $\mathbf{p} = (p_1, p_2, \ldots, p_{31})$, and each p_j is the result of mapping the j-th investor name to P using social network ranking. We also have $\mathbf{q} = (q_1, q_2, \ldots, q_{31})$, and each $q_j \in Q$ is the investor type for investor j.

We see, then, that the representation of a company consists of a number of distinct rounds. We use superscripts to denote the round number. Then, in a model where we retain only the first five rounds of funding, a company C is represented by a vector $C = (\mathbf{h}, \mathbf{p}^1, \mathbf{q}^1, \mathbf{p}^2, \mathbf{q}^2, \mathbf{p}^3, \mathbf{q}^3, \mathbf{p}^4, \mathbf{q}^4, \mathbf{p}^5, \mathbf{q}^5)$.

Besides the information contained in the investor lists, we have a relatively small amount of information that we represent by a vector $\mathbf{h}$. For the model with k rounds of funding, we have $\mathbf{h} \in \mathbb{Z}^{3+k}$, with h_1 equal to the precise four-digit sector code, h_2 equal to the year in which the company was founded, $h_3 = k$, and h_4 through h_{3+k} equal to integer representations of the dates on which the k rounds of funding occurred.

Note that entry-wise addition of two N-tuples generally results in an N-tuple that is not a meaningful representation of any possible company. This lack of linearity excludes a host of statistical methods. This is in contrast to a "bag of words" representation of our data, where we would represent the

investor list for one company by a vector $\mathbf{v} \in \mathbb{R}^{9545}$, where v_k represents the number of times that investor k participated in a round of funding for that company. This representation of the data does have a linear structure and lends itself to models based on matrix factorizations such as the SVD, yielding latent semantic analysis-type models [7]. Though linear models based on the SVD have performed very well on other problems, we found through detailed testing that for our problem, such models suffered from poor predictive power and extremely long computation times. The latter was due to the higher-dimensional spaces incurred by the bag of words representation. For this reason, we moved away from a linear representation of the data to the $(\mathbf{p}, \mathbf{q})$ structure described above.

2.3 Missing Entries

The most striking thing about our representation of the data set is the relatively large number of missing entries incurred. To see how this arises, consider that one company's first round may involve three investors while another company's first round may involve 30. In the first instance, only the first three components of $\mathbf{p}^1$ and $\mathbf{q}^1$ would be populated with meaningful information—the remaining 28 components are missing. In the second instance, there would be only one missing component in each of $\mathbf{p}^1$ and $\mathbf{q}^1$. As long as we wish to retain as much information per round as we have on hand, our representation of the data will lead to missing entries.

Both the missing entries and the lack of vector space structure point to random forests as an appropriate class of models for this problem. Classification/decision tree algorithms upon which random forests are based contain natural methods for estimating missing data. Breiman's tests [3] indicate that random forests can yield accurate models even with 80% missing data.

3 Model Development and Results

In this work, we focus on two two-class problems: distinguishing companies labeled as "bankrupt" from those that are not, and distinguishing companies labeled as "private" from those that are not.

We develop models that make predictions using only the first three funding rounds. We discard all rounds later than round three, and we cap the "number of rounds" entry h_3 of $\mathbf{h}$ so that it is at most equal to three.

3.1 Resampling and Cross-Validation

As can be seen from Table 1, the bankrupt vs. non-bankrupt problem will be highly imbalanced regardless of sector. The imbalance causes problems for all classifiers that we have tried. The problem manifests in a classifier that always predicts "non-bankrupt", yielding an area under the ROC curve close to 0.5, i.e., a perfectly useless model, even if its overall accuracy is anywhere from $70-90\%$. To avoid this issue, for any training set that we feed to the random forest, we

sample with replacement from the training set to form a new training set with uniform class distribution. We do not touch the test set. To summarize:

1. Let X = collection of all labeled vectors for one of the two-class problems for one of the sectors.
2. Randomly partition X into K disjoint subsets $\{S_i\}_{i=1}^{K}$.
3. For $i = 1 : K$,
   (a) Let $\mathbf{S} = \bigcup_{j \neq i} S_j$. Sample with replacement from $\mathbf{S}$ to form a new training set $\tilde{\mathbf{S}}$ with uniform class distribution.
   (b) Train a random forest with training set $\tilde{\mathbf{S}}$.
   (c) Test the random forest on S_i.
4. Aggregate the test results from all K folds of cross-validation.

Similar approaches have been discussed by Breiman et al [4]. The imbalance is not as acute but still exists for the private vs. non-private problem, so we employ the resampling procedure for that problem as well.

3.2 Results

All results will be for random forests with 80 trees per forest and 25 randomly chosen attributes per tree. The results are computed using Weka RandomForest [9]. We have found that the test results—both overall correctness and area under the ROC curve—are relatively insensitive to the parameters chosen. For example, varying the number of trees from 35 to 160 in steps of 10 yields ROC areas and overall correctness within 5% of the results quoted below.

In Weka, we have built models using a number of different classifiers appropriate for the attributes and instances in our data set. Even with the resampling procedure described above, the following methods yielded models with poorer predictive power than random forests: logistic regression, support vector machines (with standard kernels), and neural networks. Meta-classifiers such as boosting and bagging performed well and deserve investigation in future work.

Bankrupt vs. Non-Bankrupt Problem. Across all sectors, we find our model performs best for companies in the energy sector (sector 6). The classifier's overall accuracy is 83.4%. The confusion matrix in this case is as follows:

	predicted bankrupt	predicted non-bankrupt
truly bankrupt	167	37
truly non-bankrupt	250	1275

Here the positive class is "bankrupt" and the negative class is "non-bankrupt." Let T/F denote true/false and P/N denote positive/negative. Then we define

$$\text{Positive Precision} = \frac{TP}{TP+FP}, \quad \text{Positive Recall } = \frac{TP}{TP+FN} \tag{1}$$

$$\text{Negative Precision} = \frac{TN}{TN+FN}, \quad \text{Negative Recall} = \frac{TN}{TN+FP} \tag{2}$$

We compute these metrics to assist with decision-making. Each quantity is an estimate of a conditional probability:

$$\text{Pos Precision} = P(\text{truly } + \mid \text{predict } +) \quad \text{Pos Recall} = P(\text{predict } + \mid \text{truly } +)$$
$$\text{Neg Precision} = P(\text{truly -} \mid \text{predict -}) \quad \text{Neg Recall} = P(\text{predict -} \mid \text{truly -})$$

What we notice for Sector 6 is high negative precision, i.e., when the model says that a company is not going to be bankrupt, there is a 97.2% chance it will not go bankrupt. However, when the model says that a company is going to go bankrupt, there is only a 40% chance that it will truly go bankrupt. There are two reasons why this happens:

First, the original data set is rich with examples of non-bankrupt companies, and poor with examples of bankrupt companies. This is purely a function of how the data was gathered—companies that have gone bankrupt already have no incentive to give their historical details to ThomsonONE, and because information on private companies need not be reported publicly, ThomsonONE has no way of finding out about all past bankrupt companies.

Second, given that positive and negative recall are above 0.8, it may well be the case that if our trained models were merely tested on data sets with a much larger number of bankrupt companies, the performance would be much improved. Right now our algorithm predicts bankrupt in over 80% of the cases where the company truly is bankrupt (positive recall = 0.819), but unfortunately, our data set is only 11% bankrupt.

Added together, the two reasons just presented indicate that if the model were trained on a data set that included a more rich set of bankrupt companies, the positive precision would increase.

In addition to the above metrics, there is the ROC curve, formed by accounting for not only the classifier's prediction but also the value of its margin function for each instance—for more details about the construction of ROC curves, see [8]. The curve indicates that a practical decision-making system can be designed based on the margin. When the margin is high, i.e., when we are at the part of the ROC curve near $(0, 0)$, the classifier is consistently correct, giving the curve a large positive slope. This implies that when the margin is high, the classifier gives a useful and trustworthy prediction.

Similar results can be noted across all sectors, as shown in Table 2. ROC curves for all 9 sectors are plotted in the left and right panels of Figure 2. We have separated the ROC curves into two panels merely to enable the reader to distinguish them.

Private vs. Non-Private Problem. Here the model performs much more uniformly across all sectors. This time, let us examine the performance for the largest sector, Sector 2, comprising companies in the general area of computers. The confusion matrix is:

	predicted private	predicted non-private
truly private	2312	767
truly non-private	765	2217

We use the same definitions as given above in (1-2), but now the positive class is "private" and the negative class is "non-private." All four metrics are very close to each other: positive precision = 0.751, positive recall = 0.751, negative precision = 0.743 and negative recall = 0.743. The classifier correctly classifies 74.7% of all instances, and the area under the ROC curve is 0.828.

Very similar results can be noted across all sectors, as shown in Table 3. ROC curves for all 9 sectors are plotted in the left and right panels of Figure 3. Again, we have separated the ROC curves into two panels merely to enable the reader to distinguish them.

Ranking the Covariates. As detailed by Breiman [3], random forests provide estimates of variable importance. In Weka, the built-in RandomForest module does not include this feature; we have utilized an extension of the module developed by Livingston [11]. In the table below, we rank our attributes (or covariates) by their importance in the random forest. The importance is given as a RawScore averaged across 10 rounds of cross-validation. For reasons of space, we include in Table 4 only the top 10 attributes for Sector 6: results for other sectors show the same general grouping of attributes.

There are several clear trends to discern from the ranking. Early rounds of funding matter more than later rounds of funding. The type of an investor matters just as much if not more than its identity. Finally, the rankings for both two-class problems show remarkable similarity, both in terms of the order of the ranking and the clustering of the RawScore values in certain intervals. The top four most important attributes are the same for both two-class problems.

4 Discussion and Conclusion

Having performed this study, we see three main ideas for improving the model using currently available data.

First, from Table 4, we see that the identity of the first investor in round one is one of the most important attributes for the random forest models built in this paper. Since this identity consists of the investor's social network ranking, we are left to believe that a more informative social network may lead to better predictions of company exits. The network used in this study ignores temporal details such as the fact that investor A may be completely divested from a startup company by the time that investor B decides to invest. In this case, our network prescribes a connection between the two investors that is not present in reality. Another point is that we have formed *one* network for all investors/companies; forming different networks for each sector may yield better models.

Second, based on our knowledge of the data set, when we view the rankings in Table 4, we infer that attributes that have very low percentages of missing entries (such as the dates of the rounds of funding, which are *never* missing) are much more important for the model's predictive power. We therefore believe that a better understanding of missing entries may yield a more predictive model. Indeed, there may be something significant to learn from (a) the number of

investors in each round and (b) which investors do and do not participate in a given round of funding. This is analogous to wisdom from the winners of the Netflix prize, who found that modeling which movies *were* and *were not* rated by a user improved their predictions [1].

Finally, as hinted above, we have only begun to explore other ensemble classifiers such as boosting and bagging. It is likely that combining random forests with other models will yield a model that beats our current results—the only question is how to search for this combination in a principled fashion.

We form two main conclusions: (1) applying resampling and random forests to qualitative data in the VC/PE-space does indeed yield models with useful predictive and explanatory power; and (2) a late-stage investor who has purely qualitative knowledge of a company's first three rounds of funding can use this information to improve his/her understanding of that company's future trajectory. Overall, the results indicate that data mining can be used to provide both predictive and explanatory power for VC decisions.

References

1. Bell, R.M., Koren, Y.: Lessons from the Netflix prize challenge. SIGKDD Explorations 9(2), 75–79 (2007)
2. Bottazzi, L., Rin, M.D., Hellmann, T.: Who are the active investors? Evidence from venture capital. Journal of Financial Economics 89, 488–512 (2008)
3. Breiman, L.: Random Forests. Machine Learning 45(1), 5–32 (2001)
4. Chen, C., Liaw, A., Breiman, L.: Using Random Forest to Learn Imbalanced Data. Tech. Rep. 666, University of California, Berkeley (July 2004)
5. Clercq, D.D., Dimov, D.: Explaining venture capital firms' syndication behaviour: a longitudinal study. Venture Capital 6(4), 243–256 (2004)
6. Das, S.R., Jagannathan, M., Sarin, A.: Private equity returns: An empirical examination of the exit of venture-backed companies. Journal of Investment Management 1(1), 152–177 (2003)
7. Eldén, L.: Matrix Methods in Data Mining and Pattern Recognition, Fundamentals of Algorithms, vol. 4. SIAM, Philadelphia (2007)
8. Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Letters 27(8), 861–874 (2006)
9. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. SIGKDD Explorations 11(1), 10–18 (2009)
10. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer Series in Statistics, Springer, second edn. (2009)
11. Livingston, F.: Implementation of Breiman's Random Forest Machine Learning Algorithm. Tech. rep., North Carolina State University ECE 591Q (Fall 2005), http://www4.ncsu.edu/∼fjliving/docs/JournalPaper_Livingston.pdf
12. Milner, F., Vos, E.: Private Equity: a Portfolio Approach. Journal of Alternative Investments 5(4), 51–65 (2003)
13. Ueda, M., Hirukawa, M.: Venture Capital and Productivity. In: First Banca d'Italia/CEPR Conference on Money, Banking and Finance. Rome (Oct 2003), http://www.cepr.org/meets/wkcn/5/591/papers/ueda.pdf

	Random Forest Results					
Sector	**+Precision**	**+Recall**	**-Precision**	**-Recall**	**AUC**	**Accuracy**
1: Communications	0.331	0.742	0.913	0.644	0.765	0.663
2: Computer	0.306	0.816	0.950	0.651	0.801	0.677
3: Electronics	0.264	0.634	0.933	0.743	0.770	0.729
4: Biotech/Pharma	0.241	0.614	0.952	0.797	0.803	0.780
5: Medical/Health	0.277	0.756	0.956	0.731	0.824	0.734
6: Energy	0.400	0.819	0.972	0.836	0.880	0.834
7: Consumer	0.352	0.838	0.970	0.774	0.876	0.782
8: Industrial	0.297	0.726	0.961	0.798	0.850	0.790
9: Other	0.348	0.835	0.969	0.765	0.881	0.774

Table 2. Sector-by-sector results for the binary classification problem with "Bankrupt" as the positive (+) class and "Non-Bankrupt" as the negative (-) class. "AUC" stands for area under the ROC curve. Results show metric-wise consistency and sector-wise variation. Averaged across all sectors, the AUC is 0.83 and the accuracy is 0.75.

	Random Forest Results					
Sector	**+Precision**	**+Recall**	**-Precision**	**-Recall**	**AUC**	**Accuracy**
1: Communications	0.812	0.726	0.657	0.758	0.809	0.739
2: Computer	0.751	0.751	0.743	0.743	0.828	0.747
3: Electronics	0.793	0.719	0.702	0.779	0.838	0.746
4: Biotech/Pharma	0.774	0.721	0.789	0.832	0.860	0.783
5: Medical/Health	0.795	0.771	0.755	0.780	0.847	0.776
6: Energy	0.760	0.696	0.711	0.773	0.825	0.734
7: Consumer	0.755	0.804	0.777	0.723	0.840	0.765
8: Industrial	0.758	0.764	0.735	0.729	0.821	0.747
9: Other	0.747	0.743	0.750	0.754	0.828	0.748

Table 3. Sector-by-sector results for the binary classification problem with "Private" as the positive (+) class and "Non-Private" as the negative (-) class. AUC stands for area under the ROC curve. Results are consistent across both metrics and sectors. Averaged across all sectors, the AUC is 0.83 and the accuracy is 0.75.

RawScore	Attribute	RawScore	Attribute
99.9	Date of rnd 1	89.1	Date of rnd 1
58.1	Type of inv 1 in rnd 1	52.5	Type of inv 1 in rnd 1
32.0	Identity of inv 1 in rnd 1	34.4	Four-digit sector code
27.5	Four-digit sector code	29.2	Identity of inv 1 in rnd 1
17.4	Total # of rnds	22.4	Inception year
17.2	Inception year	15.0	Type of inv 2 in rnd 1
13.1	Type of inv 2 in rnd 1	13.3	Total # of rnds
12.4	Type of inv 1 in rnd 2	12.1	Type of inv 1 in rnd 2
10.2	Date of rnd 2	10.3	Date of rnd 2
6.76	Type of inv 1 in rnd 3	6.79	Type of inv 3 in rnd 1

Table 4. Ranking of attributes for sector 6 random forest models, with bankrupt vs. non-bankrupt results on the left and private vs. non-private results on the right.

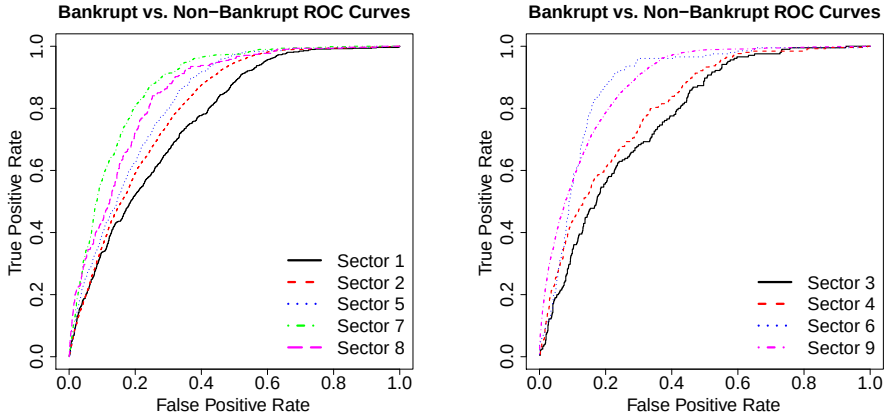

Fig. 2. ROC curves for the Bankrupt vs. Non-Bankrupt classification problem. Each curve corresponds to test set results for a sector-specific random forest model.

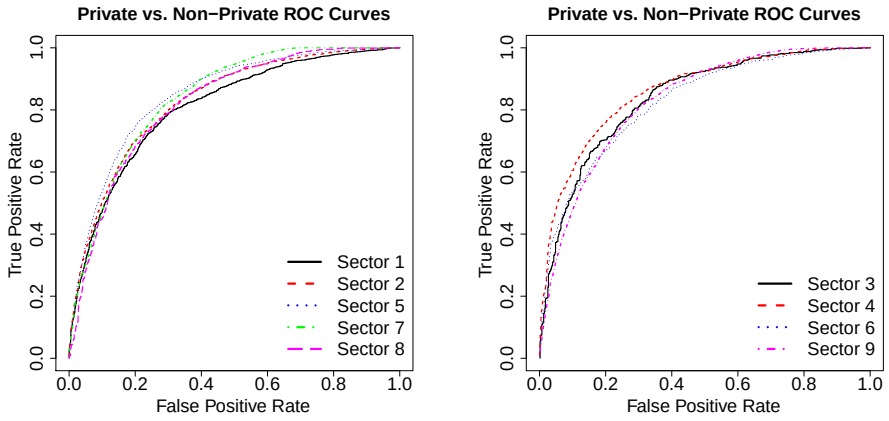

Fig. 3. ROC curves for the Private vs. Non-Private classification problem. Each curve corresponds to test set results for a sector-specific random forest model.