

Predictive Model of Insolvency Risk for Australian Corporations

Rohan A. Baxter, Mark Gawler, Russell Ang

Analytics, Office of the Chief Knowledge Officer,
Australian Taxation Office,
P.O. Box 900, Civic Square, ACT 2608
{firstname.lastname@ato.gov.au}

Abstract

This paper describes the development of a predictive model for corporate insolvency risk in Australia. The model building methodology is empirical with out-of-sample future year test sets. The regression method used is logistic regression after pre-processing by quantisation of interval (or numeric) attributes. We show that logistic regression matches the performance of ensemble methods, such as random forests and ada boost, provided that pre-processing and variable selection is performed.

A distinctive feature of the insolvency risk model described in this paper is its breadth; since we are using income tax return data we are able to risk score one million companies across all industries, all corporation types (public, private) and all sizes, as measured either by assets or number of employees. This is an application paper that uses standard credit scoring methodology on a new data source. The contribution is to demonstrate that insolvency risk can be estimated using income tax return data. The corporate insolvency prediction model is still in development and so we welcome suggestions for improvements on the current methodology.

Keywords: corporate insolvency prediction, logistic regression, random forests, ada boost.

1 Introduction

We define corporate insolvency risk as the probability that a company will become insolvent in the next 12 months. Corporate insolvency risk is used, often in tandem with credit risk scores, to identify debtors who are at risk of becoming insolvent. Debt collection strategies can then be selected with the insolvency risk in mind. For example, an important debt collection strategy is early intervention to avoid an insolvent company increasing its debt, thus avoiding an increase in the eventual legal write-off of debt at insolvency.

The project described in this paper had a number of goals. The first was to test whether corporate insolvency prediction was possible using the available income tax return data. The second was to test the feasibility of a model designed to risk score across the full spectrum of

companies (as opposed to constraining the target field to industry sector, for example). The third goal was to identify a preferred regression method after assessing logistic regression, ada boost, and random forests.

We should clarify goals that we consider are beyond the scope of this paper, although they are of interest for future work. First, we are not comparing the relative effectiveness of tax return data and financial statement data. Note that publically available financial statement data is only available for a tiny fraction of Australian companies, whereas this paper is focussed on all Australian companies that are registered in the tax system. Second, we are not comparing stratified models to a single all-company model. We intend to perform and describe such a comparison in future work. We also intend to test a multi-level model where both company-level and industry-level effects are jointly included.

Section 2 puts the current work in the context of a long history of insolvency prediction models and of recent work in Australia. Section 3 describes the data we obtained for building the model. In Section 4, we provide our particular predictive evaluation model-building methodology. All model performance evaluation is done on out-of-sample, future year test datasets. This means that not only are our test datasets distinct from the training datasets, they are also constrained to be test data from future years. Section 5 gives some descriptive data understanding results, then describes and evaluates the predictive models. Preliminary results on Financial Year 2006/2007 are then given. Section 6 discusses issues arising from the present work and possible future directions. We give our conclusions in Section 7.

2 Related Work

We have developed an empirical model of insolvency, as opposed to a structural model. An empirical model is data-driven and is built and assessed using predictive performance as the criterion, whereas structural models use an explicit function based on a theory of companies and insolvency. In the mid-1960s, Altman (Altman, 1993) developed the Z-Score model which uses 6 ratios and a linear discriminant model. There have been many variants since then by Altman and others. We use Altman's ratios, as well as a further 8 financial ratio variables defined by Ohlson (Ohlson, 1980), who used a logistic regression model.

Shin et al (Shin, 2006) compare ensemble models (bagging, boosting) with logistic regression, decision trees, neural networks and nearest neighbour. They also compare different feature selection methods. Their dataset

Copyright © 2007, Australian Computer Society, Inc. This paper appeared at the Sixth Australasian Data Mining Conference (AusDM 2007), Gold Coast, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 70. Peter Christen, Paul Kennedy, Jiuyong Li, Inna Kolyshkina and Graham Williams, Ed. Reproduction for academic, not-for-profit purposes permitted provided this text is included.

is restricted to 76 Turkish banks and their evaluation test data does not use future year sample datasets. They conclude that neural networks with appropriate feature selection is competitive with ensemble models and logistic regression on their Turkish bank dataset.

2.1 Recent Australian models

Jones and Hensher (Jones and Hensher, 2004; Hensher, Jones and Greene, 2007) have published a series of papers using new methods for models of predicting corporate insolvency. Their main focus is predictive performance at a group level, rather than at the individual company level. They note that predictive performance of models has not improved greatly since the 1960s (Hensher, Jones and Greene, 2007, p88). They observe that a type II error rate (where a solvent company is predicted to go insolvent) of 20% is typical for in-sample modelling results and even higher for out-of-sample future year results. In our project, a 20% type II error rate is acceptable as long as the results are still actionable to improve the operational efficiency of our business. One reason to support this contention is that many of the solvent companies in the 20% type II error category may not be technically or legally insolvent, but instead, may be financially distressed or even trading while insolvent. We have confirmed this hypothesis using surrogate variables for financial distress, such as level of indebtedness.

Similarly to Hensher and Jones, Hossari's PhD thesis (Hossari, 2006), focuses on improving the methodology in model building for predicting corporate collapse. Hossari uses multi-level models with financial ratio data extracted from audited financial statements. The data selected is a balanced sample, matched by industry classification, with fewer than 100 companies in the dataset. Model assessment is done on the single sample. Hossari found that available software for the multi-level models didn't scale well to larger datasets.

Moody's has an existing corporate default model for private companies, with 27K private companies in the dataset (Moody's, 2000; Moody's, 2000a). Our methodology mirrors Moody's *RiskCalc* approach. However, instead of audited financial statements we use income tax return fields as data sources for financial ratios. This allows us to score approximately one million private companies. Moody's early Australian model in 2000 achieved an Area under the ROC curve (AUC) (Fawcett, 2004) of 0.7, compared with an AUC score of 0.93, as generated by the most recent version of our model.

3 Data

Our population consists of active Australian companies, which we define as those companies that have had at least one income tax return since 2003. This covers about one million companies from all industry sectors, all size ranges and all corporation types, such as public and private.

3.1 Insolvency Target Variable

We obtain insolvency data from the Australian Securities and Investments Commission (ASIC 2007). This data is

publicly available. There are at least seven different stages in the insolvency process, ranging from voluntary administration to liquidation. In our modelling process we use the widest possible definition of insolvency and so define a company to be insolvent if it enters any stage of insolvency, even if only temporarily, during a financial year.

Our principal interest is in predicting financial distress in general rather than insolvency specifically. However, the use of insolvency as the target variable has the advantage of definiteness and objectivity. Nonetheless, it is a broad target; there will many companies trading while insolvent that do not actually go into administration. This is consistent with David Hand's hypothesis that financial and customer modelling often involves ambiguous target concepts (Hand, 2006).

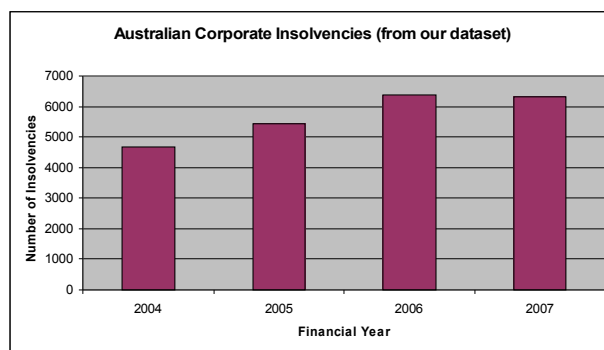


Figure 1: Insolvencies by Financial Year

3.2 Financial Ratio Variables

We adopt the financial ratio variables used by Altman (Altman, 1993) and Ohlson (Ohlson, 1990). Their financial data are obtained from audited financial statements provided by companies to the relevant corporate regulator (i.e. Securities Commission in the U.S.). Since we are scoring both public and private corporations, we need to exclude ratios using variables that apply only to public companies, such as *market value of equity*; data sources like those used by Altman and Ohlson are inadequate for our purposes. Instead, we have taken company income tax return data and adapted them for use as financial ratios.

Given that tax financial data are different from accounting financial statements, the question arose as to whether they would be suitable for accurate insolvency risk prediction. We shall see in the Results section that the results are roughly comparable with those using audited financial data. This is a useful and, as far as we know, novel finding.

The non-ratio financial variables used are:

1. Total Assets.
2. Net Income.

The two sets of financial ratio variables that we have used are:

1. Altman variables (Altman, 1968):
 - i. Earnings before Tax and Interest / Total Assets.

- ii. Net Sales / Total Assets.
- iii. Market Value of Equity / Total Liabilities [No income tax return label equivalent of this is available, so we are unable to use it.]
- iv. Working Capital / Total Assets.
- v. Retained Earnings / Total Assets.

2. Ohlson Variables (Ohlson, 1980)

- vi. TltoTA: Total Liabilities / Total Assets.
- vii. CLtoCA: Current Liabilities / Current Assets.
- viii. NltoTA: Net income / Total Assets.
- ix. FFOtoTL: Pre-tax income plus depreciation and amortization costs / Total Liabilities.
- x. INTWO: Flag that is 1 if cumulative net income over previous two years is positive.
- xi. OENEG: Flag that is 1 if Owner's Equity is negative [Not available in income tax return data.]
- xii. CHIN: Change in Income from previous year to current year.
- xiii. TA: Size as ln (Total Assets/ GDP price growth) [This definition not available in income tax return]
- xiv. Berry Ratio (Gross Profit / Operating Expenses)

3.3 Financial Distress Indicator Variables

Our modelling has also included input variables derived from the lodgment and payment behaviour of companies. Does a company lodge returns and pay taxes on time? If it is late, then how late is it? It is not surprising that issues such as these have proven sound indicators of financial corporate distress. Intuitively, a company at high risk of insolvency with cash flow problems or ongoing lack of profitability will be a poor debtor. However, there are counter-examples which must be managed; profitable companies with disorganised book-keeping may also be late lodgers and make late tax payments.

The precise definitions of these variables is not critical to the main thrust of this paper and so are not provided here.

3.4 Company Demographic Variables

Two company demographic variables that have been included are:

1. Age of company according to ASIC registration.
2. Industry classification using ABS industry codes.

The financial distress indicator variables, when added to the financial ratio variables, greatly improve our model's predictive performance. Company Demographic variables have a relatively minor effect relative to the other two input variable categories.

4 Predictive Model Methodology

We developed training and test datasets using the fundamental design principle that test data should be in the future relative to the training data. As mentioned in Section 2, this approach is not always used for model assessment, thereby bringing test results into question. For our business needs, real out-of-sample performance is what determines long-run utility of the model for the client and hence, client acceptance of the model (Moody's, 2000). Therefore, out-of-sample performance is the key assessment criterion for the insolvency risk model.

Our out-of-sample, future-year approach has been to train and test the model using data from consecutive financial years and then scoring a dataset derived from a later year. Table 1 shows the time frames for the extraction of training, testing and scoring datasets.

Dataset Type	Input Variable Year	Insolvency Target Year
Training	FY 2005 or before	FY 2006
Test	FY 2006 or before	FY 2007
Score	latest available data	not applicable

Table 1: Training/Test/Score Dataset Design

We use pair sampling (i.e. for every insolvent company, we find a solvent company), thus training and testing our model with balanced datasets. Pair sampling introduces a bias that causes an overestimate of variable significance in the model (Zmijewski 1984). This might be problematic if we were to interpret model parameters for explanatory purposes, but is less so in our current context of maximising predictive performance. As yet we have not tried matched pair sampling, where insolvent and solvent companies are compared based on size, industry classification and private/public status.

5 Results

5.1 Data Exploration

In order to check data quality and that the variable relationships are consistent with commercial practice, univariate plots of input variables versus the insolvency rate were produced. For example, the Berry Ratio (Gross Profit/Operating Expenses) is shown in Figure 2, where a value of zero (the wide x-axis value labelled '03:0-0.68') or a high value (the right-hand most x-axis value labelled '08:1.27-high') indicate the least risk of insolvency.

These observations are consistent with business knowledge. A Berry Ratio value of zero applies to companies with no operating profit and often with no operating expenses. Such companies, which include passive investment companies relying solely on investment income, carry little risk of becoming insolvent. A high Berry Ratio value indicates companies with large profit margins, where operating profits are much greater than expenses.

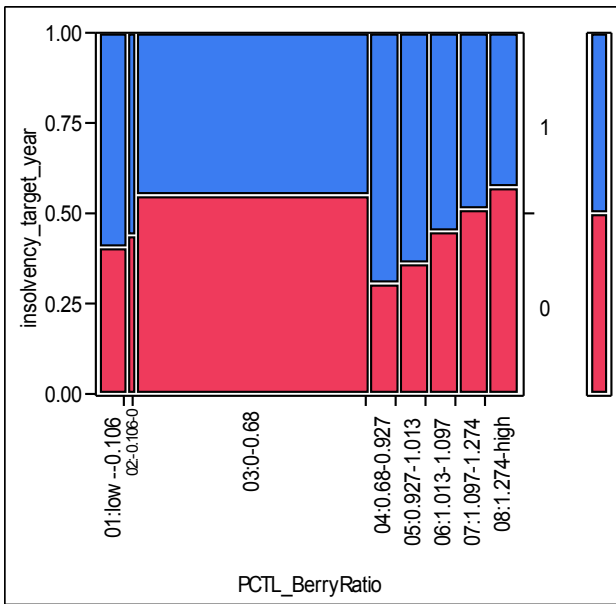


Figure 2: Discretised Berry Ratio vs Proportion of Insolvencies (for balanced sample). The lower part (red) of each category indicates of the proportion of solvent companies.

We performed another data quality check on the stability of the univariate relationships across Financial Years (FY). An example of this is shown for Net Income in Figure 3, which presents the rate of insolvency against net income deciles for two financial years (2004 and 2005). The overall rate of corporate insolvency approximates to 0.006 (roughly six in 1000 companies) for FY 2004 and 0.010 for FY 2005. The left-most net income category (labelled 1) is for negative income less than \$-32K. The insolvency rate is highest for this category. While there is variance across the financial years, the insolvency rate pattern shares roughly the same trend. Note that the pattern of maximum and minimum values (at deciles 1 and 4) are fairly consistent across years.

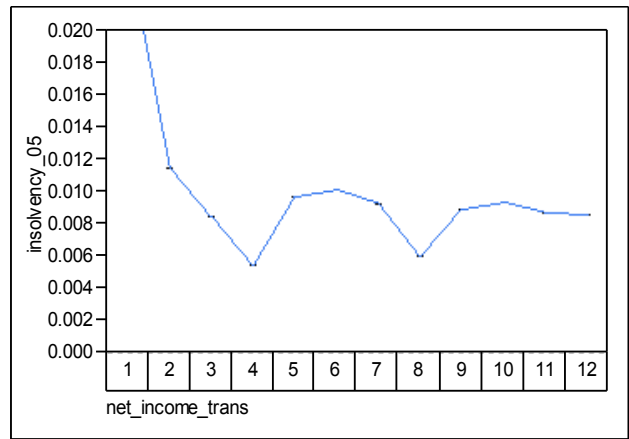
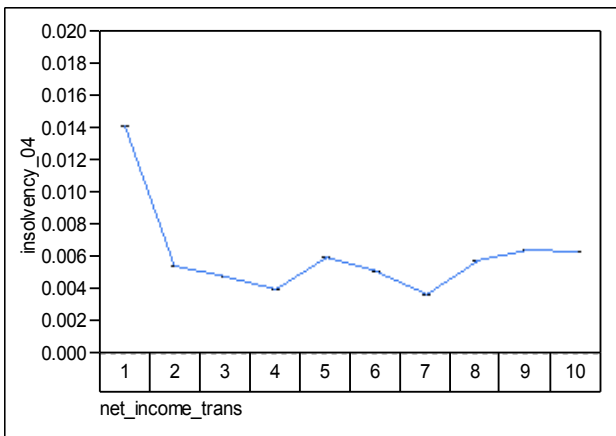
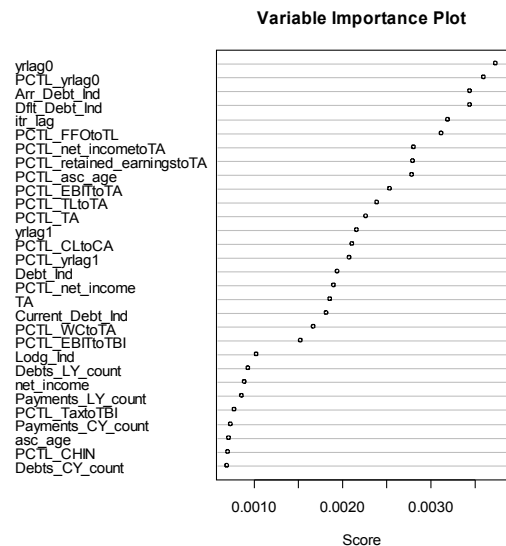


Figure 3: Net Income (quantised into deciles) vs. Insolvency Rate (for unbalanced sample) for Financial Years 2004 and 2005 . The y-axis gives the insolvency rate.

5.2 Variable Selection and Importance

Figure 4 plots variable importance based on the ada boost, while Figure 5 plots variable importance for the random forest model (showing two measures of importance). The variable importance measures used in these figures are defined in their respective R packages. They are based on the average % change in predictive accuracy when the variable is included and then excluded from the model.

There are significant differences in the variable importance rankings. The Ada Boost model flags its first four variables as being of much higher importance than the rest, while in Figure 5 (Random Forest model) these same variables appear in position 14 at the highest. This shows that variable importance ranks can be very model specific. It suggests that no single variable, operating alone, is highly predictive of insolvency and so rankings of importance are not definitive (Hand 2006). We also found that variable importance rankings are dataset sample specific. Resampling the training dataset and retraining the model leads to major changes in the variable ordering and minor changes in predictive performance.



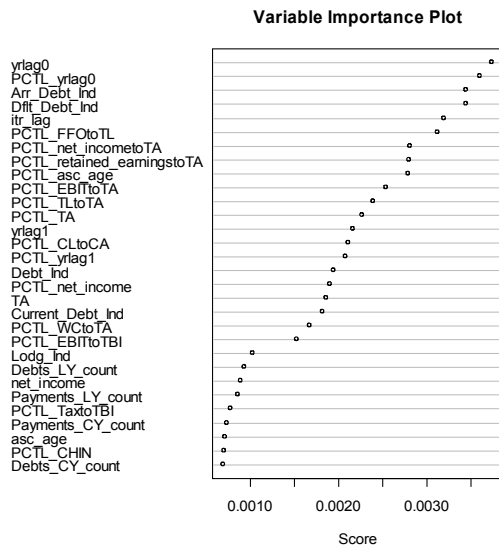


Figure 4: Variable Importance according to the Ada Boost model

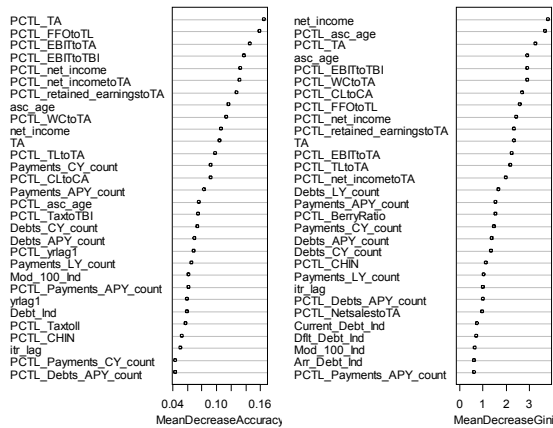


Figure 5: Variable Importance according to the Random Forest model

5.3 Model Building

Our production models are built using SAS Enterprise Miner 5.1. In our data preprocessing, we quantise interval variables into up to 10 quantiles. The quantisation of interval (continuous) variables helps prevent over-fitting by the regression model. It also helps with extreme values by allocating them to a single bin such as 'lowest quantile' or 'highest quantile'. Handling extreme values in this way improves the regression model's robustness, making it less sensitive to a particular data sample. We used a logistic regression model with variable selection. The optimisation target is validation misclassification cost and the cost ratio between insolvency and solvency is 50 to 1 (i.e. It is 50 times more beneficial to correctly identify an insolvent company than it is to correctly identify a solvent company). This cost was used because correct identification of insolvency is more important to our decision making than identifying true solvent companies.

SAS Enterprise Miner does not have the recent ensemble

methods readily available. We were interested in benchmarking our SAS results with the results achieved using R and an R package, Rattle (Rattle, 2006), both of which are freely available open source software. A direct comparison could not be made because the R package is currently incapable of handling datasets that are of the scale of our company dataset (close to one million companies). Instead we sampled down to datasets of about 20K companies for both training and test datasets.

The Rattle classifier methods that we use include Random Forests, Ada Boost, SVM, Decision Tree (rpart) and Logistic Regression (glm). Note that we use classifiers to predict the probability of insolvency, normally a regression-like task. Classifiers that predict only a categorical class outcome rather than a probability are not applicable here. It should be noted that R's logistic regression package does not have variable selection (available in other R packages). It also does not have the ability to optimise cost using a validation dataset instead of the training dataset.

Classifier	(AUC)	AUC with transformed interval variables	(AUC) with transformed interval variables, variable selection
rpart	0.80	0.84	0.81
ada boost	0.88	0.88	0.89
rf	0.88	0.87	0.87
ksvm	0.84	0.88	0.88
glm	0.84	0.86	0.86

Table 4: Area under the ROC curve (AUC). It should be noted that there is a significant variance in the AUC estimates when the sampling of the test dataset is decreased from 1m to 20K. We intend to incorporate this source of variance into the model once we have computed it (our best guess is ± 0.03).

We have chosen to present our results using the Area Under Curve (AUC) measure derived from ROC curves. We present AUC results for each classifier on test data for a number of different samples:

1. sample without pre-processing
2. sample with interval variables quantised (following the SAS EM approach)
3. sample with quantisation and using variables only selected by SAS EM logistic regression.

The question that arises is: are the results returned by the various classifiers affected by pre-processing or by the variable selection step?

Figure 6 and Table 4 give the classifier results for the discretised interval variable dataset, using the variables as selected in the SAS EM model. The two ensemble methods (ada boost and random forests) are consistent

across the different data pre-processing steps. Logistic regression and SVM improve with the discretised interval variable dataset. As can be seen in both Figure 6 and, Table 4 the performance of the decision tree (rpart) is consistently lower than that of other models. This is as expected, given that, relative to the ensemble methods, decision trees generally have a high variance and low bias (Hastie, et al, 2001).

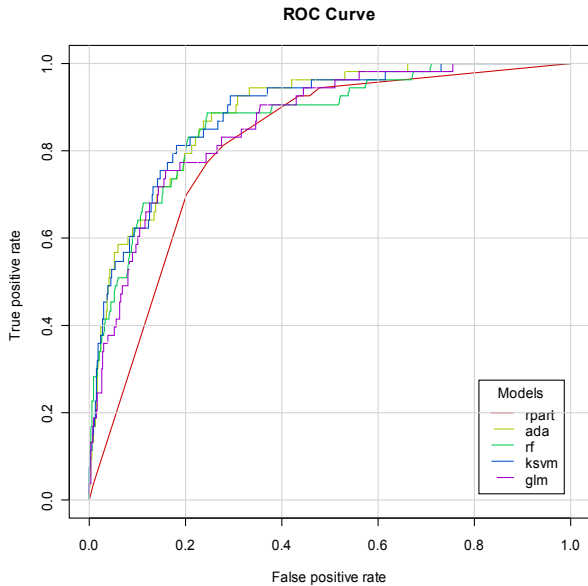


Figure 6: ROC Curves for classifiers estimating the probability of insolvency: rpart, ada, rf, ksvm and glm for the sampled test dataset with quantised interval variables and variable selection as done in SAS EM. Note these curves are based on a single test dataset sample and so we expect they will have relatively large confidence intervals on the curves (see Table 4 caption).

5.4 Results on Test Financial Year 2007

Figure 7 shows the predictive performance of the model when applied to Financial Year 2007. Note that this year has just ended so all of the 2007 insolvency data is not yet available. The trend across the quantiles (low risk on left, high risk on right) shows a general trend upwards as we would expect if the model were predictive.

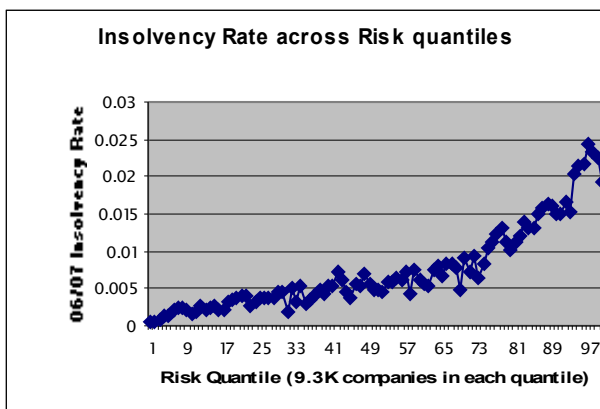


Figure 7: Result on FY 2007 test

The model places 15% of all insolvencies in the top 5%

risk quantile, 27% of insolvencies in the top 10% risk quantile and 50% of insolvencies in the top 25% risk quantile. These results are comparable with those achieved by similarly large-scale commercial models making future year insolvency predictions. Personal bankruptcy prediction have even better results than company bankruptcy prediction, with 50% of bankruptcies being placed in the top 10% risk quantile (Experian, 2007).

6 Discussion

6.1 Stratification Models

We elected not to stratify the set of companies into sub-segments, despite the likelihood that it would improve model predictive performance, due to pragmatic, operational resource reasons. The first phase of the modelling process has been a proof-of-concept. Should the accuracy of the broad, single company type model prove sufficient, there will be no need to develop sub-segment models. We have briefly explored segmentation by:

1. public vs. private company
2. industry sector : Finance and Property sectors have very different financial ratio behaviour to Retail, Manufacturing and Construction.
3. size (as measured by total assets)

It is evident that predictive performance is improved by developing models for particular segments.

6.2 Related Entities

The first version of our model treats each company as independent of other companies. In reality, there are many types of corporate groups, involving interrelated companies. An extension of our model would identify related entities and include some form of risk score aggregation.

For small companies (<\$100K assets), the credit risk (ability to pay debts on time) of the proprietor plays a significant role in the company's insolvency risk (ability to pay debts at all). In some cases, the proprietor's risk is as critical as the financial status of the company. For these smaller firms, the bankruptcy risk of business owners should be assessed and, where necessary, combined with the insolvency risk of the company entity (Moody's, 2000a). We plan to incorporate this relationship in future versions of our model.

6.3 Hazard Models

For this prototype, we have adopted a single insolvency period (one year) as a target. Some authors have postulated that hazard models, which utilise time-series data, are more accurate than static models (Shumway, 1999). However, in practice, hazard models have not been found to improve predictive accuracy significantly. Even so, with a view to optimising performance, we intend to extend our model to include some time-series data in future work.

7 Conclusion

We have built a corporate insolvency risk model for one million active Australian corporations using income tax return data and data from the Australian Securities and Investments Commission. The predictive performance of this model matches that achieved by commercial models whose scope is restricted to particular industries or public companies. Our data sources have been found to be suitable for corporate insolvency prediction and a single predictive model can be built for all corporations. The ensemble methods slightly outperform logistic regression at this stage (we do need to check test data variability issues). At this stage, we prefer logistic regression for its convenience of deployment as SQL in a data warehouse environment.

8 Acknowledgements

We thank Brian Irving, David Kuhl, and Stuart Hamilton for several helpful discussions. We thank Anthony Siouclis for his expert economist advice on the definition and use of tax label ratios. We also thank the referees for comments that improved the clarity of the paper.

9 References

- Altman, E.I. (1968): Financial Ratios, Discriminant analysis and the prediction of corporate bankruptcy, *Journal of Finance*, **23**:4, pp589-609.
- Altman, E.I.; Haldeman, R.G; and Narayanan, P. (1977) Zeta Analysis: A new Model to Identify Bankruptcy Risk of Corporations. *Journal of Banking and Finance*, **1**, 9-24
- Altman, E.I. (1993): *Corporate Financial Distress and Bankruptcy: A Complete Guide to Predicting and Avoiding Distress*. New York: Wiley
- ASIC (2007): Australian Securities and Investments Commission, <http://www.asic.gov.au/>.
- Baxter, R.A. (2006) Finding Robust Models Using a Stratified Design, *AI2006: Advances in Artificial Intelligence, Lecture Notes in Computer Science*, **4304**, pp 1064-1068, Springer.
- Experian (2007) Harben, S. and Curtis,C., *Modelling Personal Bankruptcy in the UK*, White Paper, Experian-Scorex, <http://www.experian-scorex.com/>.
- Fawcett, T. (2004). ROC Graphs: Notes and Practical Considerations for Researchers. Technical report, Palo Alto, USA: HP Laboratories.
- Hand, D. (2006) Classifier Technology and the Illusion of Progress. *Statistical Science* **21**(1). pp 1-15.
- Hastie, T., Tibshirani, R., and Friedman, J. (2001) *The Elements of Statistical Learning*, Springer.
- Hensher, D.A., Jones, S. and Greene, W.H. (2007): An Error Component Logit Analysis of Corporate Bankruptcy and Insolvency Risk in Australia.
- Hossari, G. (2006): A Ratio-Based Multi-Level Modelling Approach for Signalling Corporate Collapse: A Study of Australian Corporations. PhD Thesis, Swinburne University of Technology.
- Jones, S. and Hensher, D.A. (2004) Predicting Firm Financial Distress: A Mixed Logit Model. *The Accounting Review*, **79**(4), pp. 1011-1038.
- Kaski, S. (2001) Bankruptcy Analysis with Self-Organizing Maps in Learning Metrics. *IEEE Transactions on Neural Networks*, **12**:4, 2001.
- Keasy, K; and Watson, R. (1991): Financial Distress Prediction Models: A Review of their Usefulness. *British Journal of Management*, **2**, 89-102.
- Lin, L; and Piesse, J. (2004): The Identification of Corporate Distress in UK Industrials: A Conditional Probability Analysis Approach. Research Paper 024 The Management Research Papers. Kings College London. University of London.
- Moody's (2000): RiskCalc For Private Companies: Moody's Default Model. Rating Methodology. May 2000, Moody's Investor Service.
- Moody's (2000a): RiskCalc For Private Companies II: More Results and The Australian Model. Dec. 2000, Moody's Investor Service.
- Ohlson, J.S. (1980): Financial Ratios and the Probabilistic Prediction of Bankruptcy, *Journal of Accounting Research*, **19**, pp109-31.
- Rattle (2006). Rattle Software, An R Package, <http://rattle.togaware.com/>, R software, <http://r-project.org/>.
- Shumway, T. (1999): Forecasting Bankruptcy More Accurately: A Simple Hazard Model.
- Shin, S.W., Lee, K.C. and Kilic, S.B. (2006) Ensemble Prediction of Commercial Bank Failure Through Diversification of Bank Features, *AI2006: Advances in Artificial Intelligence, Lecture Notes in Computer Science*, **4304**, pp887-896, Springer.
- Sung, T.K., Chang, N. and Lee, G. (1999) Dynamics of Modeling in Data Mining: Interpretative Approach to Bankruptcy Prediction. *Journal of Management Information Systems*, **16**:1, pp. 63-86.
- Wilson, R.L; and Sharda, R. (1994): Bankruptcy Prediction using Neural Networks. *Decision Support Systems*, **11**, 545-557
- Zmijewski, M. (1984) Methodological Issues Related to the Estimation of Financial Distress Prediction Models. *Journal of Accounting Research*, **22**, 59-62.
- CreditRisk: Credit Risk Website, <http://www.creditrisk.com/>. Accessed 29 Jun 2007.

