Modeling Private Firm Default: PFirm

Grigoris Karakoulas
Business Analytic Solutions

May 30th, 2002
Outline

• Problem Statement
• Modelling Approaches
• Private Firm Data Mining
• Model Development
• Model Evaluation
• Explaining Model Prediction
• Discussion
Problem Statement

• A loan is commonly considered to be in default if any of the following occur:
  – a loan is classified as non-accrual
  – a borrower is 90 days or more past due in its principal or interest payments
  – a borrower has filed for bankruptcy protection
  – a loan is partially or fully written off

• Only a few quantitative models for private Middle Market firms
  – most banks use *judgmental* models
Problem Statement

• Growing interest for private firm models due to Basel II Accord and loan securitization

• Quantitative models can be used as a decisioning tool to:
  – automate mechanical tasks such as financial assessment of a company
  – analyze multidimensional interactions
  – simulate complex what-if scenarios
  – provide early warning signals
Problem Statement

• Given historical data from annual financial statements of defaulted and non-defaulted firms estimate
  - probability $P\{y_{t+k} | X_t\}$, that a firm will default ($y=1$) within the next $K$ months from the date of financial statements $T$
  - for a short term horizon model $K=12$ months
Problem Statement

• **Independent variables from the literature**
  – Coverage ratios
    • EBIT / interest
    • EBITDA / interest
  – Profitability ratios
    • (net income - extraordinary items) / total assets
    • EBIT / total assets
  – Leverage ratios
    • total liabilities / net worth
    • total liabilities / total assets
Problem Statement

• Independent variables (cont.)
  – Liquidity
    • working capital / total assets
    • current assets / current liabilities
    • cash / total assets
  – Activity ratios
    • accounts payable
    • accounts receivable
  – Growth ratios (net sales, net income)
  – Financial size (assets)
Modelling Approaches

• Discriminant Analysis for estimation of generative models

• Limitations of DA
  – assumes explanatory variables have a multivariate normal distribution
  – requires the proportion of default/non-default in the sample to be the same in the population
  – linear classification rule
Modelling Approaches

• Probit and Logit (*discriminative*) models
  - \( y^*_{t+k} = bX_t + u_t \)
    - \( y=1 \) if \( y^*_{t+k} \geq 0 \); \( y=0 \) otherwise
    - assumptions about distribution of \( u_t \)
  - pros: estimation of expected probability of default
  - violation of assumption about distribution of defaults in the population makes parameter estimates biased
Modelling Approaches

• Instead of $y_i$ being the (0/1) random variable, suppose the length of time $t_i$ that firm $i$ survives is the random variable
  – each firm either defaults during the sample period, survives the sample period, or leaves the sample for some other reason

• The hazard function $h_d(t;x,b)$ gives the instantaneous probability of the length of time $t$ ending with default conditional on surviving up to that time
Modelling Approaches

• With hazard models there is no need to assume independence between firm-year observations as with previous approaches.

• All the above modelling approaches are **parametric**
  - a lot of effort for crafting the form of the model
  - difficult to capture interactions amongst variables
Private Firm Data Mining

- History of financial statements of Canadian companies since 1991
- Exclude real estate firms, financial institutions and government as obligors
- Data cleansing
- Database of private firms
  - 2,177 obligors
  - 8,757 financial statements
Private Firm Data Mining

• **Candidate Input Variables:**
  – 34 financial variables
    - debt service coverage, profitability, liquidity, leverage, activity, growth, financial size
  – type of financial statement
    - 1 for audited and unqualified; 2 for reviewed and compiled; 0 otherwise

• **Target Variable:** 0/1 (=default) in the next 12 months from the F/S date
Private Firm Data Mining

- **Construct the dataset of observations**
  - for each defaulted (“bad”) obligor construct one observation of the input variables from financial statements with date
    - at least 12 months prior to default and
    - no more than 24 months prior to default
  - for each “good” obligor and for each financial statement date in our database construct an observation of the input variables
• **Training/Test split of dataset**
  – Test (out-of-sample) set contains obligors with F/S dates since 1998/02 (temporal constraint)
    • 454 obligors; 760 F/S records
  – Training set contains obligors not in test set (cross-sectional constraint) and with F/S dates prior to 1998/02
    • 1446 obligors; 4495 records
  – temporal + cross-sectional constraints = *true out-of-sample testing*
Private Firm Data Mining

- Descriptive statistics of some financial ratios in training set

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Median</th>
<th>25% Quartile</th>
<th>75% Quartile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Assets ($M)</td>
<td>3.947</td>
<td>1.896</td>
<td>10.271</td>
</tr>
<tr>
<td>Inventory/COGS</td>
<td>0.1648</td>
<td>0.0861</td>
<td>0.279</td>
</tr>
<tr>
<td>Liabilities/Assets</td>
<td>0.693</td>
<td>0.4928</td>
<td>0.85</td>
</tr>
<tr>
<td>Net Income Growth</td>
<td>6.235</td>
<td>-38.14</td>
<td>77.5</td>
</tr>
<tr>
<td>Net Income/Assets</td>
<td>0.0795</td>
<td>0.037</td>
<td>0.1428</td>
</tr>
<tr>
<td>Quick Ratio</td>
<td>0.9107</td>
<td>0.5752</td>
<td>1.4496</td>
</tr>
<tr>
<td>RE/A</td>
<td>0.2359</td>
<td>0.0786</td>
<td>0.4147</td>
</tr>
<tr>
<td>Sales Growth</td>
<td>7.625</td>
<td>-1.28</td>
<td>20.94</td>
</tr>
<tr>
<td>Cash/Assets</td>
<td>0.0675</td>
<td>0.0148</td>
<td>0.1774</td>
</tr>
<tr>
<td>EBIT/Interest</td>
<td>3.33</td>
<td>1.56</td>
<td>8.78</td>
</tr>
</tbody>
</table>
Private Firm Data Mining

**Odds of Default conditioned on net income-extra to total assets**

**Odds of Default conditioned on total liabilities to EBITDA**
Private Firm Data Mining

Non-linear interaction of couples of attributes with default

-3.17
-3.18
-3.19
-3.2
-3.21
-3.22
20

ebit to total assets (percentiles)

0
5
10
15
20

total liabilities to ebitda (percentiles)
Model Development

• Predictive performance
  – *true bads*: actual defaults *(bads)* correctly predicted as defaults
  – *true goods*: actual good obligors correctly predicted as good
  – *false bads*: actual good obligors incorrectly predicted as defaults (*Type II Error*)
  – *false goods*: actual defaults incorrectly predicted as good (*Type I Error*)

• In a probabilistic model there is tradeoff between true goods and false goods
How can we induce from data “good” and “bad” distributions with little overlap?
Model Development

- Receiver-Operating Characteristic (ROC) curve
Model Development

- Area under ROC curve is the probability that a randomly selected “bad” obligor will have predicted score of no-default less than that of a randomly selected “good obligor
  - a measure of separability of two distributions

- Use the area under ROC curve as the performance criterion in an algorithm that learns a model from data
  - criterion = 2*AUROC-1
Model Development

• **NBTree** is an in-house technique for learning a probabilistic model from data
  – a decision tree (*discriminant model*) where internal nodes are partitioning the data into subsets and each leaf node contains a *generative model* for estimating conditional probability using variables **not** in the path to that leaf

• Let $X = [x_1, x_2, \ldots, x_n]$ be the vector of input variables (financial ratios) and $Y$ the output binary variable (default event)
To compute probability of default 
\[ P\{Y=1|x_1, x_2, \ldots, x_n\} \] one needs to make assumptions for independence amongst input variables.

NBTree learns these assumptions from data by recursively building a decision tree.

\[ \text{PROB} = P\{Y|X', x_3 > \text{value}_3\} \] 
where \( X' \) denotes the variables excluding \( x_3 \).
Model Development

- Feature Selection is a hard problem
  - various heuristic approaches, e.g. forward selection, backward selection
- Use an in-house feature selection technique based on genetic algorithms for searching for a “best” subset of input variables such that the NBTree model has the biggest area under the ROC curve
Model Development

• Our feature selection technique selected a “best” set of model variables (PFirm)
  
  • Profitability1
  • Profitability2
  • Liquidity1
  • Liquidity2
  • Leverage1
  • Profitability3
  • Leverage2
  • Leverage3
  • Growth1
  • Growth2
Model Development

- Graphical Representation of PFirm Model

```
x1

x2

x3

x4

PD <= 1.205
<= 1016.5
<= -0.0989
<= -46

PD > 1.205
> 1016.5
>-0.989
> -46
>-46 & < 152
> 152
```
Model Evaluation

- Four benchmark models (Appendix) on the out-of-sample (test) dataset:
  - RiskCalc 10-variable model
    - NB. Since RiskCalc is continuously recalibrated by Moody’s its performance is in-sample rather out-of-sample
  - Altman’s 5-variable model by refitting it on our training data
  - Shumway’s model by refitting it on our training data
  - NI/TA - TL/TA (naïve predictor)
Model Evaluation

Figure 1. Out-of-Sample ROC curve
Model Evaluation

- Summary of comparisons based on area under ROC curve in previous graphs and accuracy ratio for area under CAP curve

<table>
<thead>
<tr>
<th></th>
<th>PFirm</th>
<th>NI/TA-TL/TA</th>
<th>RiskCalc</th>
<th>Altman</th>
<th>Shumway</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUROC</td>
<td>0.6628</td>
<td>0.4358</td>
<td>0.5539</td>
<td>0.4774</td>
<td>0.4605</td>
</tr>
<tr>
<td>Accurary Ratio</td>
<td>0.6542</td>
<td>0.4324</td>
<td>0.5396</td>
<td>0.4736</td>
<td>0.4578</td>
</tr>
</tbody>
</table>
Model Evaluation

• **Main points from evaluation:**
  – PFirm seems to be robust in changes in the cycle since it is trained on expansion years and tested on recession years
  – Altman’s, Shumway’s and naïve-predictor models have almost the same performance
    • they are linear models in contrast to PFirm and RiskCalc that are **non-linear** and perform better
  – One of the reasons that PFirm is performing better than RiskCalc is because PFirm is capturing co-dependencies amongst variables
Explaining Model Prediction

- **Case Study: XYZ Corp.**
  - classified date: July 2001

<table>
<thead>
<tr>
<th></th>
<th>May-98</th>
<th>May-99</th>
<th>May-00</th>
</tr>
</thead>
<tbody>
<tr>
<td>profitability1</td>
<td>53.00%</td>
<td>0.73%</td>
<td>21.00%</td>
</tr>
<tr>
<td>profitability2</td>
<td>33.00%</td>
<td>0.00%</td>
<td>14.00%</td>
</tr>
<tr>
<td>liquidity1</td>
<td>40.00%</td>
<td>-0.01%</td>
<td>41.00%</td>
</tr>
<tr>
<td>liquidity2</td>
<td>52.00%</td>
<td>0.00%</td>
<td>30.00%</td>
</tr>
<tr>
<td>leverage1</td>
<td>39.00%</td>
<td>-67.70%</td>
<td>61.00%</td>
</tr>
<tr>
<td>profitability3</td>
<td>34.00%</td>
<td>-31.46%</td>
<td>15.00%</td>
</tr>
<tr>
<td>leverage2</td>
<td>81.00%</td>
<td>0.09%</td>
<td>99.00%</td>
</tr>
<tr>
<td>leverage3</td>
<td>45.00%</td>
<td>0.00%</td>
<td>18.00%</td>
</tr>
<tr>
<td>growth1</td>
<td>NaN</td>
<td>NaN</td>
<td>95.00%</td>
</tr>
<tr>
<td>growth2</td>
<td>NaN</td>
<td>NaN</td>
<td>19.00%</td>
</tr>
<tr>
<td>PD</td>
<td>0.008518</td>
<td>0.008518</td>
<td>0.017634</td>
</tr>
</tbody>
</table>
Discussion

• PFirm is built on in-house techniques for feature selection and model development

• NBTree is a non-parametric modeling technique that combines the advantages of discriminant and generative techniques

• The evaluation results show that PFirm performs better than benchmark models including Riskcalc

• Work underway for incorporating industry factors into PFirm
Appendix: Benchmarks

• **RiskCalc: a three stage model**
  – total assets
  – net income/assets
  – net income growth
  – interest coverage
  – quick ratio
  – cash & equivalents/assets
  – inventories/GOCS
  – sales growth
  – liabilities/assets
  – retained earnings/assets
Appendix: Benchmarks

• Two linear models for predicting the probability of default for 1 and 5 years

• Each model is estimated in three stages:
  – (i) transform the input data of the model variables into percentiles (binning)
  – (ii) build univariate default models by separately fitting each transformed model variable to the target variable
  – (iii) use the output of the above model to fit a linear probit model for predicting default
Appendix: Benchmarks

- **Altman’s: logistic regression model**
  - \( Z = b_1 \times (\text{WorkingCapital/TotalAssets}) + b_2 \times (\text{RetainedEarnings/TotalAssets}) + b_3 \times (\text{EBIT/TotalAssets}) + b_4 \times (\text{bookEquity/TotalLiabilities}) + b_5 \times (\text{Sales/TotalAssets}) \)

- **Shumway’s: logistic regression model**
  - \( S = b_1 \times (\text{NetIncome/TotalAssets}) + b_2 \times (\text{TotalLiabilities/TotalAssets}) + b_3 \times (\text{CurrentAssets/CurrentLiabilities}) \)
Modelling Private Firm Default: PFirm

Grigoris Karakoulas
Business Analytic Solutions
grigoris.karakoulas@cibc.ca