

The CIBC logo is displayed in yellow, bold, sans-serif capital letters. It is positioned in the upper left corner of a dark red banner that spans the top of the slide. A thick yellow curved line separates the banner from the white background below.

Modeling Private Firm Default: PFirm

Grigoris Karakoulas
Business Analytic Solutions

May 30th, 2002

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

CIBC 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

CIBC 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

CIBC 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

- **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

CIBC 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

Private Firm Data Mining

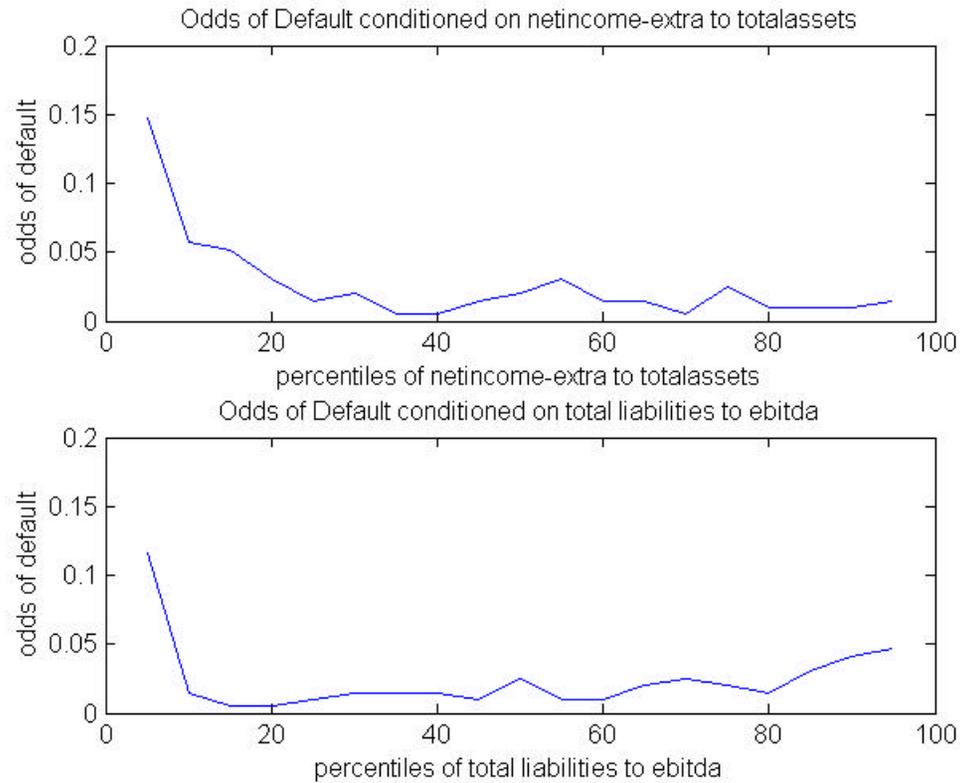
- **Training/Test split of dataset**
 - Test (out-of-sample) set contains obligors with F/S dates since 1998/02 (temporal constraint)
 - 454 obligors; 760 F/S records
 - Training set contains obligors not in test set (cross-sectional constraint) and with F/S dates prior to 1998/02
 - 1446 obligors; 4495 records
 - temporal + cross-sectional constraints = *true out-of-sample testing*

Private Firm Data Mining

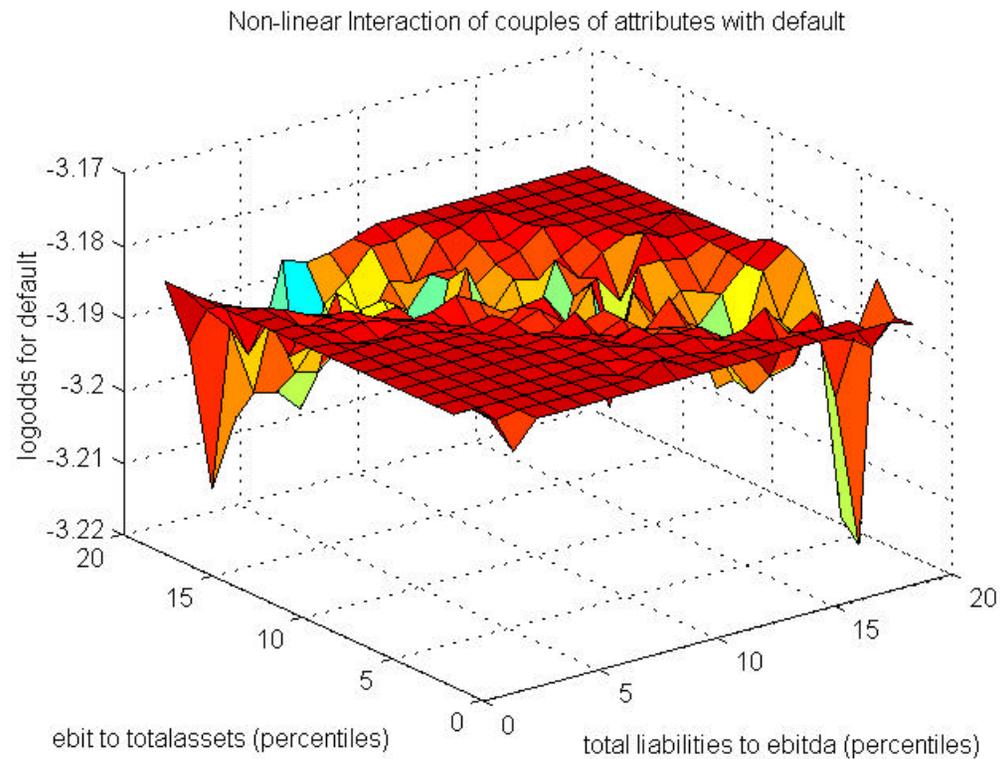
- **Descriptive statistics of some financial ratios in training set**

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

Private Firm Data Mining



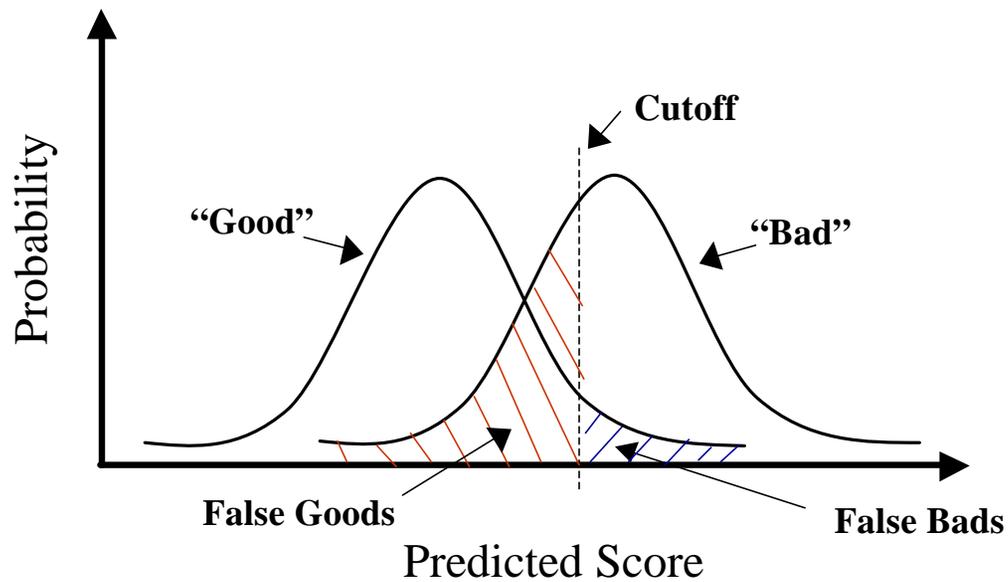
Private Firm Data Mining



CIBC 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**

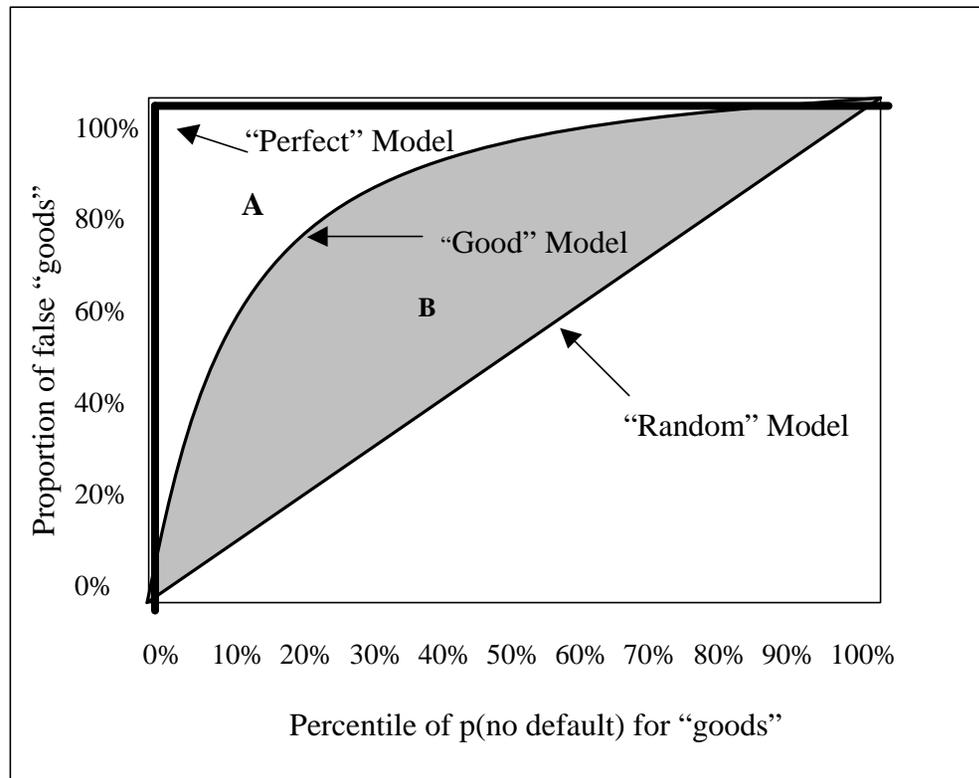
Model Development



How can we induce from data “good” and “bad” distributions with little overlap?

CIBC Model Development

- Receiver-Operating Characteristic (ROC) curve



CIBC Model Development

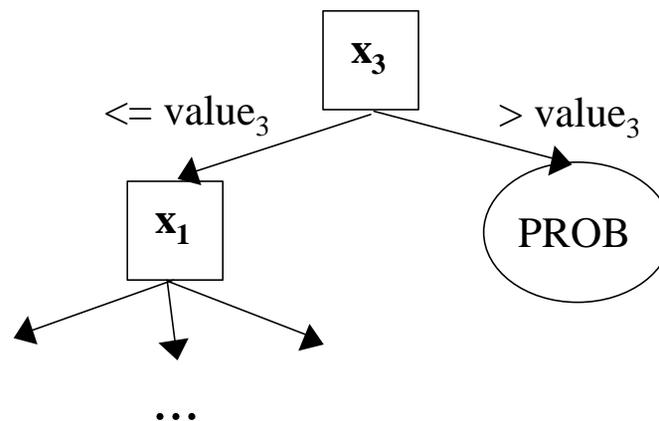
- **Area under ROC curve is the probability that a randomly selected “bad” obligor will have predicted score of no-default less than that of a randomly selected “good obligor**
 - a measure of separability of two distributions
- **Use the area under ROC curve as the performance criterion in an algorithm that learns a model from data**
 - criterion = $2 \cdot \text{AUROC} - 1$

CIBC 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 $X = [x_1, x_2, \dots, x_n]$ be the vector of input variables (financial ratios) and Y the output binary variable (default event)

CIBC Model Development

- To compute probability of default $P\{Y=1|x_1, x_2, \dots, x_n\}$ one needs to make assumptions for independence amongst input variables
- NBTree learns these assumptions from data by recursively building a decision tree



PROB =
 $P\{Y|X', x_3 > \text{value}_3\}$
where X' denotes the
variables excluding x_3

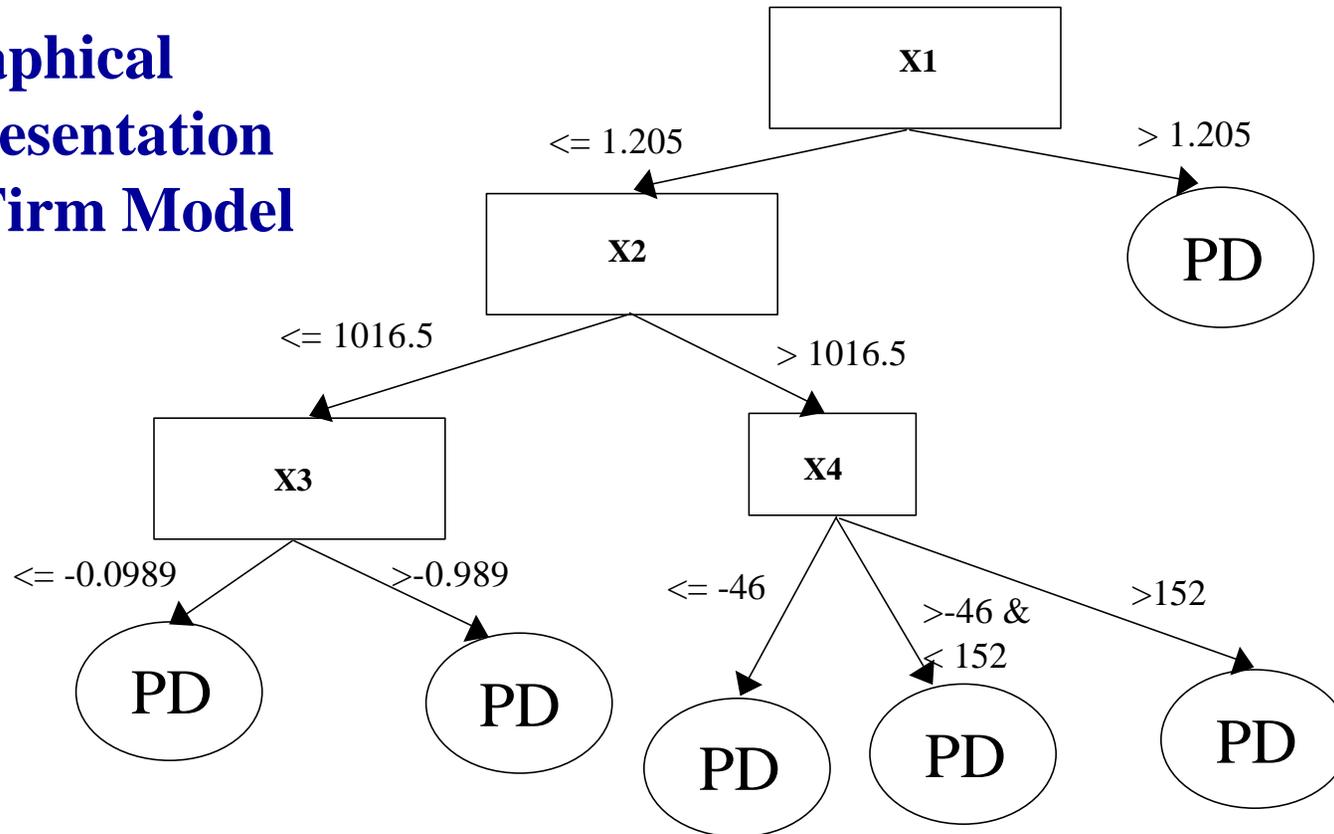
CIBC Model Development

- **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**

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

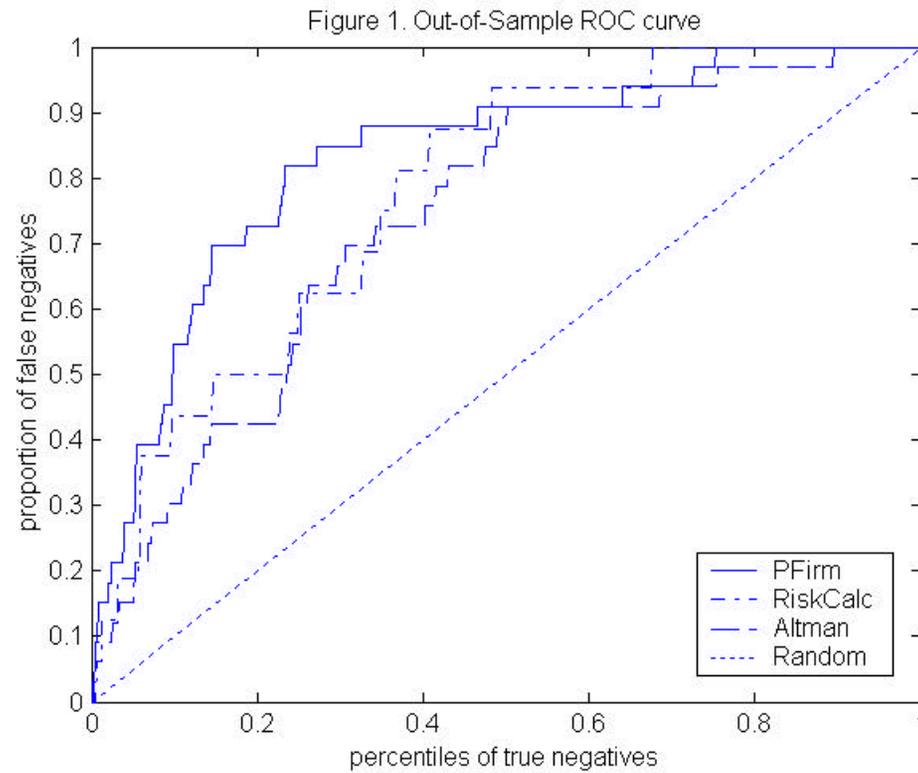
Model Development

- Graphical Representation of PFirm Model



- **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



Model Evaluation

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

	PFirm	NI/TA- TL/TA	RiskCalc	Altman	Shumway
AUROC	0.6628	0.4358	0.5539	0.4774	0.4605
Accuracy Ratio	0.6542	0.4324	0.5396	0.4736	0.4578

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

	May-98		May-99		May-00	
	Percentile	Rel. Contr.	Percentile	Rel. Contr.	Percentile	Rel. Contr.
profitability1	53.00%	0.73%	21.00%	-4.75%	3.00%	-26.63%
profitability2	33.00%	0.00%	14.00%	0.00%	3.00%	-0.01%
liquidity1	40.00%	-0.01%	41.00%	0.00%	42.00%	0.00%
liquidity2	52.00%	0.00%	30.00%	0.00%	43.00%	0.00%
leverage1	39.00%	-67.70%	61.00%	58.20%	55.00%	23.42%
profitability3	34.00%	-31.46%	15.00%	-32.66%	8.00%	-39.21%
leverage2	81.00%	0.09%	99.00%	0.64%	100.00%	2.67%
leverage3	45.00%	0.00%	18.00%	-0.01%	14.00%	-0.01%
growth1	NaN	NaN	95.00%	1.81%	8.00%	-0.70%
growth2	NaN	NaN	19.00%	-1.92%	6.00%	-7.35%
PD	0.008518	0.008518	0.017634	0.017634	0.692126	0.692126

- **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**
 - $Z = b_1 * (\text{WorkingCapital}/\text{TotalAssets}) + b_2 * (\text{RetainedEarnings}/\text{TotalAssets}) + b_3 * (\text{EBIT}/\text{TotalAssets}) + b_4 * (\text{bookEquity}/\text{TotalLiabilities}) + b_5 * (\text{Sales}/\text{TotalAssets})$
- **Shumway's: logistic regression model**
 - $S = b_1 * (\text{NetIncome}/\text{TotalAssets}) + b_2 * (\text{TotalLiabilities}/\text{TotalAssets}) + b_3 * (\text{CurrentAssets}/\text{CurrentLiabilities})$

The CIBC logo is displayed in a bold, yellow, serif font against a dark red background. The background is a horizontal bar that tapers to a point on the left side, with a yellow curved line separating it from the white background below.

Modelling Private Firm Default: PFirm

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