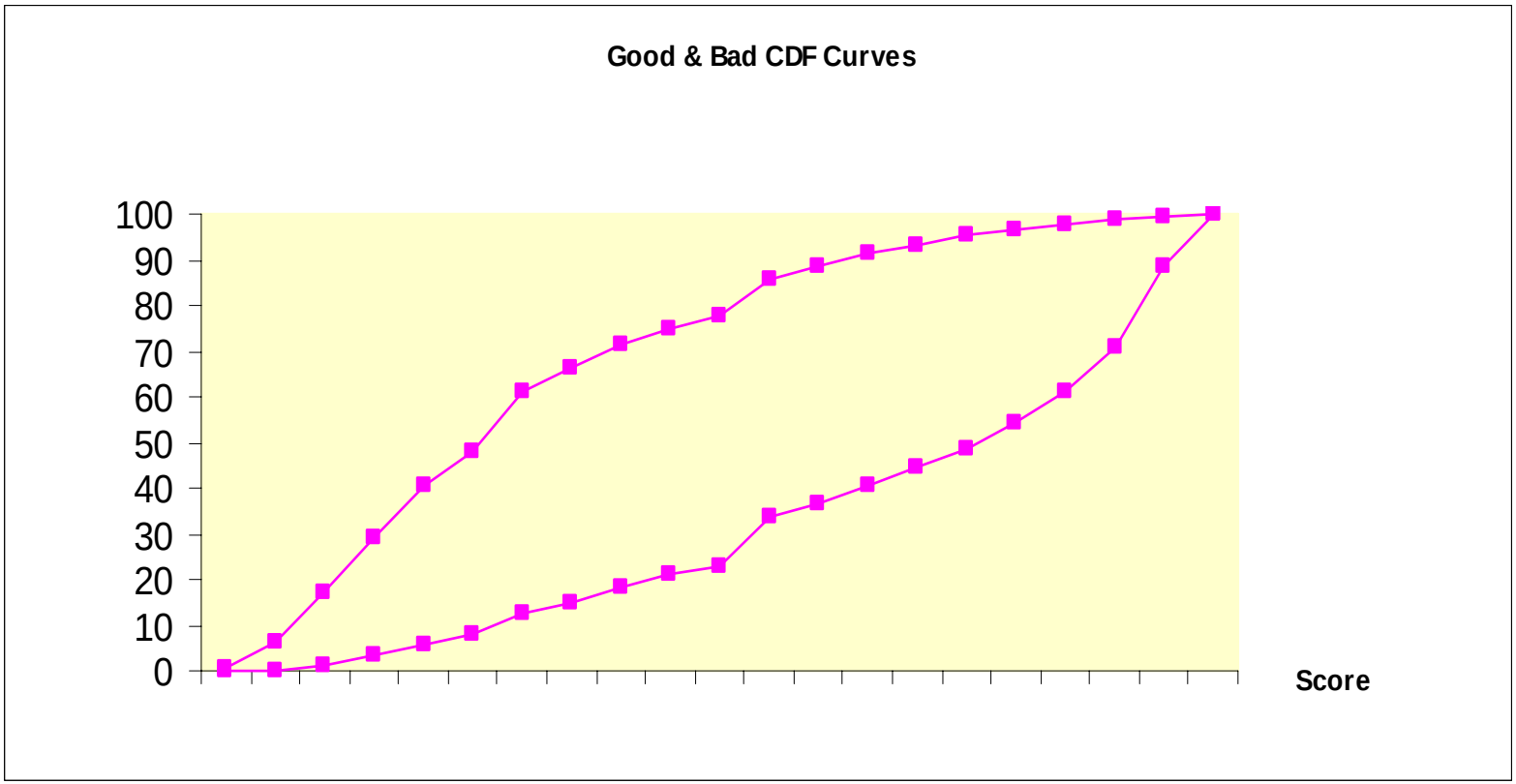
Incorporating Stochastic Dominance Constraints in Credit Scoring Methodology –Getting More from Models

Debashish Sarkar
Management Sciences Group
CIT, New Jersey, USA

Conference on Credit Scoring and Credit Control X, August 29-31, 2007

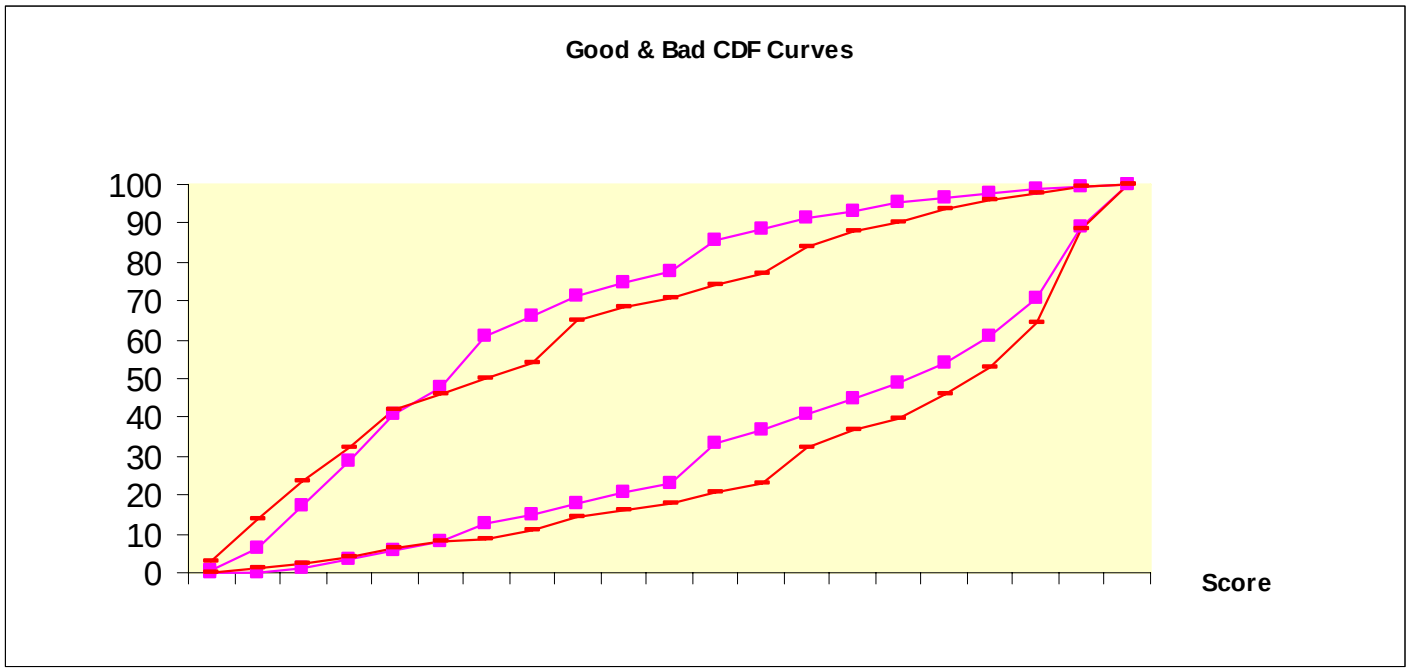
Motivation & Objectives

Scoring Model Output – good & bad accounts cumulative distribution curves



Motivation & Objectives

Two or more scorecards that are pareto-efficient, when used together (e.g., matrix structure), can give better business solutions



Issue: How to generate “pareto-efficient” cards efficiently ?

Stochastic Dominance – definition

FSD - CDF curve for the bad accounts lies above the good CDF curve for all scores: $B(s) \geq G(s) \quad \forall s$

SSD - for any given score, area to the left under the bad CDF curve is greater than or equal to the area under the good curve:

$$\int_{-\infty}^{s} B(t)dt \geq \int_{-\infty}^{s} G(t)dt \qquad \forall \text{s}$$

Other definitions (e.g, TSD) exist. See Hadar & Russell (1969), Whitmore (1970), Hanoch & Levy (1969), Bawa (1982) for review & applications.

Stochastic Dominance – duality results

$S^G(\alpha)$ = α-quantile score for the good accounts (i.e., inverse CDF for goods)

$S^B(\alpha)$ = α-quantile score for the bad accounts

FSD is equivalent to $S^G(\alpha) \geq S^B(\alpha)$ for all $0 \leq \alpha \leq 1$

SSD is equivalent to $\int_0^\alpha S^G(\alpha)d\alpha \geq \int_0^\alpha S^B(\alpha)d\alpha$ for all $0 \leq \alpha \leq 1$

See Ogryczak and Ruszczyski (2002) for references & details.

Stochastic dominance enforcement via Cornish-Fisher Approximation

Cornish-Fisher quantile approximation provides expressions for $S^G(\alpha)$ and $S^B(\alpha)$ as functions of respective mean (μ), standard deviation (σ), standardized central skewness (γ_1), and standardized central excess kurtosis (γ_2):

$$S(\alpha) = \mu + \sigma . x_\alpha$$

$$x_\alpha = z_\alpha + 1/6(z_\alpha^2 - 1)\gamma_1 + 1/24(z_\alpha^3 - 3z_\alpha)\gamma_2 - 1/36(2z_\alpha^3 - 5z_\alpha)\gamma_1^2$$

with $z_\alpha =$ std normal α-quantile value.

So, SD requirements can be written in terms of higher moments of good & bad score

Stochastic Dominance constraints –basis for generating alternative scorecards

- Maximize log-likelihood function subject to only a subset of SD constraints
- Mechanism to control shape/distance between the good and bad CDF curves
- Log-likelihood function, SD constraints are functions of score variable weights (w_j = jth variable weight)

$V_j^i =$ value of jth variable for customer i , and $Q_i = \sum_j w_j V_j^i$ be the predictor variate. So, customer score $e^Q/(1+e^Q)$ is function of w_j.

Accordingly, higher moments of good / bad population scores and quantile functions $S^G(\alpha)$ and $S^B(\alpha)$ are functions of w_j.

Nonlinear Programming Formulation P1

$$\underset{w_j}{\text{Maximize}} \sum_j \sum_{i=good} w_j V_j^i - \sum_j \sum_i \log(1+\exp(w_j V_j^i))$$

Subject to $S^G(\alpha) - S^B(\alpha) \geq \kappa$ (α and κ given)

(Again, $S^G(\alpha)$ and $S^B(\alpha)$ are both functions of w's.)

Nonlinear Programming Formulations P2 & P3

Let $V^G(\alpha)$ and $V^B(\alpha)$ be α-quantile expressions for the good & bad predictor variates respectively.

Problem P2:

$$\underset{w_j}{\text{Maximize}} \sum_j \sum_{i=good} w_j V_j^i - \sum_j \sum_i \log(1+\exp(w_j V_j^i))$$

$$\text{Subject to} \quad \frac{e^{V^G(\alpha)}}{1+ e^{V^G(\alpha)}} - \frac{e^{V^B(\alpha)}}{1 + e^{V^B(\alpha)}} \geq \kappa \quad (\alpha, \kappa \text{ given})$$

Problem P3: $\underset{w}{\text{Maximize}}\ S^G(\alpha) - S^B(\alpha) \quad (\alpha \text{ given})$

Solution to NLP

- Proc NLP within SAS 9.1 PC
- Data = option allows straightforward implementation of the P1-P3 objective functions & constraints
- Total solution times for real problems with 9000+ observations under 5 minutes

Examples

Example 1:
-UK Consumer credit bureau data
-Development sample 4879 good, 4323 bads (bad = 90+DPD ever within 2 yrs, reject inference)
-33 indicator variables from 14 distinct bureau variables
-Validation sample 2073 good, 1825 bad

Example 2:
-North American Commercial credit bureau data
-Development sample 2528 good, 1831 bads (bad = 90+DPD ever within 2 yrs, reject inference)
-39 indicator variables from 11 distinct bureau variables
-Validation sample 5983 good, 4279 bad

Example 1

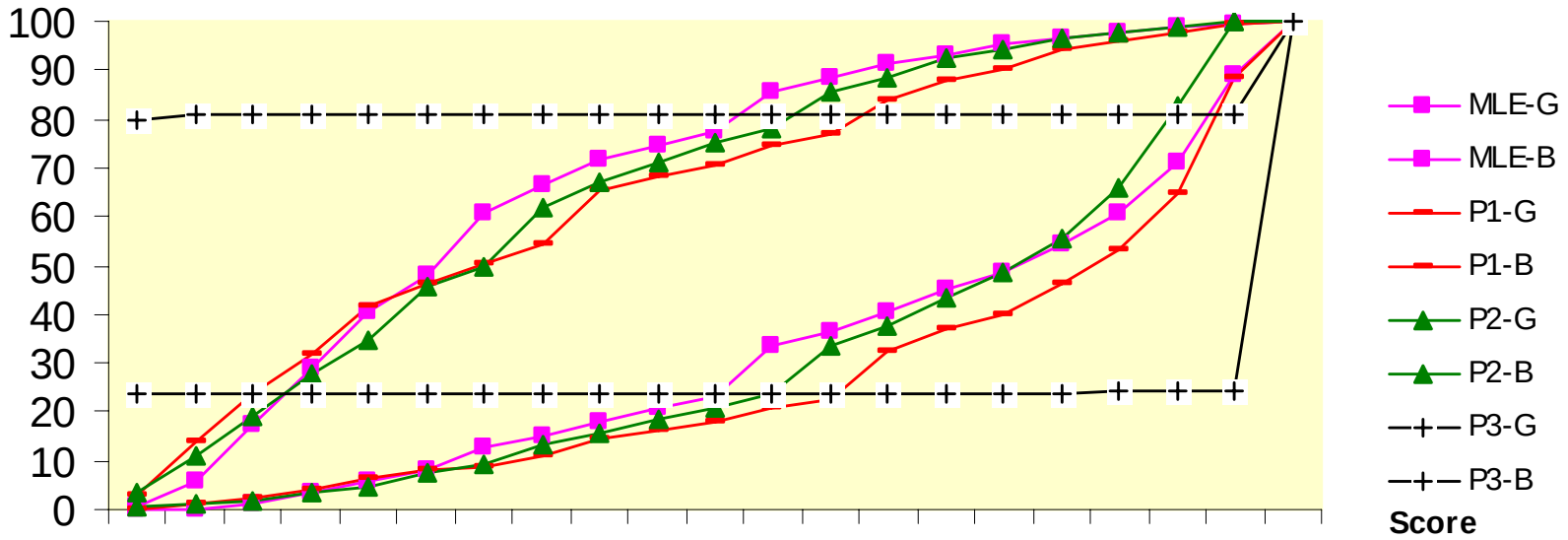
Variable	Logistic regression variable weights (MLE)	MLE with CF: α =0.1, $S^{\alpha}(G) - S^{\alpha}(B) \geq$ 0.4 (score)	MLE with CF: α =0.1, $S^{\alpha}(G) - S^{\alpha}(B) \geq$ 0.4 (variate)	Maximize $S^{\alpha}(G) - S^{\alpha}(B)$ with α =0.5 (score)
W1_1	-0.5772	-0.6792	-0.5836	-95.5685
W1_2	-0.4576	-0.5750	-0.5470	-45.7371
W1_3	0.4056	0.4595	0.3364	-5.2817
W1_4	0.8303	0.9927	0.6631	43.3949
W1_5	1.2155	1.3494	0.9394	111.5168
W2_1	-0.5020	-0.4879	-7.3866	-56.7648
W2_2	-1.1510	-1.4590	-1.3914	-160.1810
W3_1	-0.3324	-0.4807	-0.3060	-7.6186
W3_2	-0.4226	-0.5386	-0.3593	-15.3418
W4_1	-0.3885	-0.5207	-0.3893	-55.8356
W4_2	-0.1440	-0.2535	-0.1026	-36.3506
W5_1	-0.4340	-0.5517	-0.4477	-18.4846
W5_2	-0.9836	-1.4923	-1.0598	-94.6232
W5_3	-1.5920	-2.7095	-1.9979	-132.0503
W5_4	0.5474	0.4704	0.5421	70.5092
W5_5	0.7234	0.5502	0.6742	88.0424
W5_6	0.6296	0.4262	0.6404	39.0659
W5_7	1.6390	1.1107	0.9503	100.3690
Constant	0.5687	1.2716	0.8033	42.8025
KS Value (development sample)	0.544	0.5420	0.5424	0.5744
Validation sample KS (2073 good, 1825 bad)	0.545	0.5470	0.5454	0.5465

Example 1 Continued

Variable	Logistic regression variable weights (MLE)	MLE with CF: α =0.1, $S^{\alpha}(G) - S^{\alpha}(B) \geq 0.4$ (score)	MLE with CF: α =0.1, $S^{\alpha}(G) - S^{\alpha}(B) \geq 0.4$ (variate)	Maximize $S^{\alpha}(G) - S^{\alpha}(B)$ with α =0.5 (score)
W6_1	0.4388	0.5913	0.4779	20.7196
W6_2	0.4901	0.2864	0.3337	48.5796
W6_3	1.0340	0.6168	0.4286	47.4676
W7_1	-0.2799	-0.3277	-0.1910	-8.7908
W7_2	-0.4595	-0.8084	-0.5437	-51.2261
W8	-0.0903	-0.0998	-0.0894	-8.4631
W9_1	-0.3621	-0.3949	-0.3093	-20.5599
W9_2	-0.6188	-0.6081	-0.4327	-54.8516
W10_1	0.4339	0.4456	0.3306	43.9651
W10_2	0.7225	0.7077	0.5424	87.9059
W11_1	-0.2899	-0.2983	-0.2641	-15.1413
W11_2	-0.9690	-1.0914	-0.9061	-92.1342
W12	0.3409	0.3292	0.2060	44.9784
W13	-0.4245	-0.6237	-0.5320	-56.9185
W14	-0.3964	-0.7861	-0.8143	-46.8196
Constant	0.5687	1.2716	0.8033	42.8025
KS Value (development sample)	0.544	0.5420	0.5424	0.5744
Validation sample KS (2073 good, 1825 bad)	0.545	0.5470	0.5454	0.5465

Example 1- continued

Figure 1 : Good & Bad CDF Curves MLE, P1-P3 Models



Example 1- continued

Bad rate Percentages by MLE & P1 Model Score Deciles

		MLE Score Deciles										
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
P1 Model Score Decile	**1**	91.73	77.45	-	-	-	-	-	-	-	-	90.64
	2	83.93	82.16	80.90	0.0	-	-	-	-	-	-	81.75
	3	-	71.70	72.19	69.93	-	-	-	-	-	-	71.82
	4	-	-	61.98	73.59	62.34	-	-	-	-	-	71.86
	5	-	-	-	62.16	51.87	42.45	47.37	-	-	-	51.61
	6	-	-	-	-	52.50	40.15	34.33	-	-	-	40.32
	7	-	-	-	-	0.0	38.84	37.02	22.47	-	-	35.20
	8	-	-	-	-	-	-	24.06	17.26	7.94	-	16.89
	9	-	-	-	-	-	-	-	10.67	5.57	4.11	6.16
	10	-	-	-	-	-	-	-	-	6.64	3.6	4.13
Total		91.08	81.16	74.85	72.53	52.68	40.22	35.14	16.85	6.07	3.69	

Example 1 continued

- Objective - Maximize auto decision rate such that (i) bad rate for auto approvals <= 5.15% and (ii) bad rate for auto declines >= 82.15%
- Results (all samples):
  - 50% for MLE score only
  - 49.7% for P1 model scorecard
  - 53.2% using MLE & P1 scorecards
- 6% more auto decision using both scores

Example 2

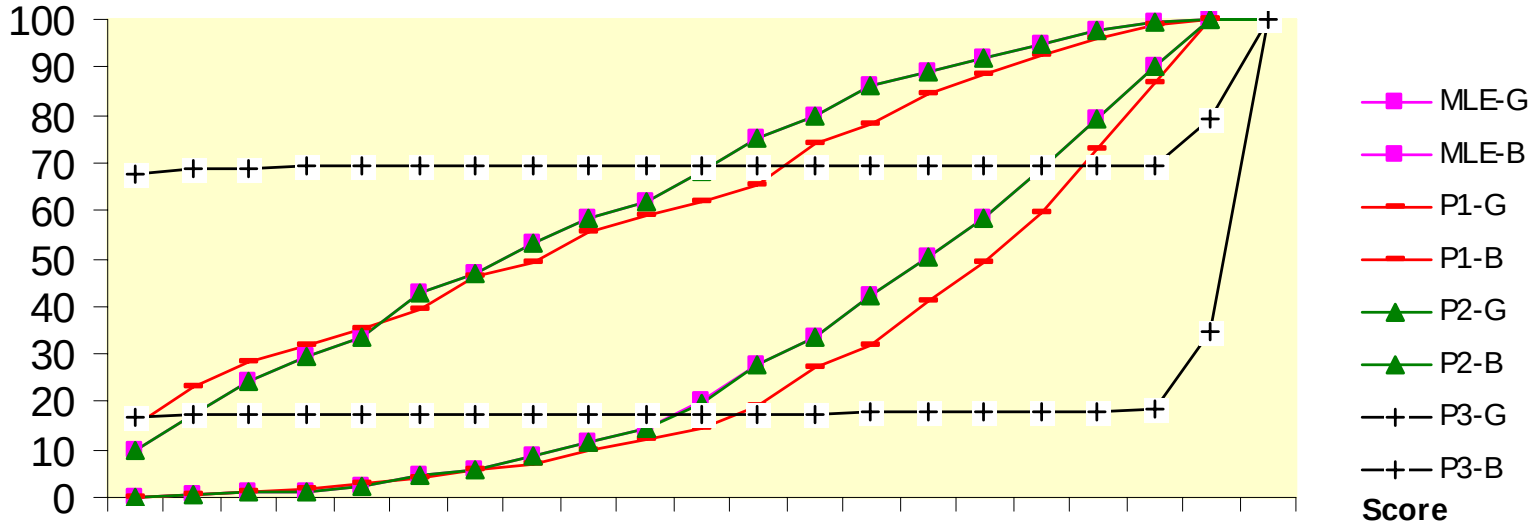
Variable	Logistic regression variable weights (MLE)	MLE with CF: α=0.16, S^{α}(G) - S^{α}(B) $\geq$ 0.5 (score)	MLE with CF: α=0.16, S^{α}(G) - S^{α}(B) $\geq$ 0.5 (variate)	Maximize S^{α}(G) - S^{α}(B) with α=0.5 (score)
W1_1	0.4499	0.4150	0.45	15.1404
W1_2	1.2276	1.3649	1.2294	72.27
W2_1	0.0821	0.0384	0.0819	17.4296
W2_2	0.1369	-0.0201	0.1368	33.955
W2_3	-0.6060	-0.7968	-0.6059	-51.5102
W2_4	-1.0679	-1.5255	-1.0682	-56.8485
W2_5	-1.6864	-2.4085	-1.6873	-75.7523
W3_1	0.1633	0.1622	0.1635	3.2516
W3_2	-0.2334	-0.0453	-0.2340	55.6698
W3_3	-0.1996	-0.2362	-0.1993	3.3786
W3_4	-0.6628	-0.8219	-0.6622	-14.6992
W4_1	0.3711	0.3743	0.3711	25.9623
W4_2	0.9564	0.9372	0.9562	55.9039
W4_3	-0.4742	-0.5466	-0.4745	-26.0535
W5_1	0.388	0.4342	0.3878	25.0167
W5_2	0.8255	0.8832	0.8244	2.9219
W5_3	-0.4354	-0.6314	-0.4377	-45.0829
W6	-0.5408	-0.6606	-0.5400	-27.6243
Constant	0.7011	1.1047	0.7009	18.7720
KS Value (development sample)	0.4802	0.4721	0.4806	0.52
Validation sample KS (5983 good, 4279 bad)	0.48	0.48	0.4802	0.476

Example 2 continued

Variable	Logistic regression variable weights (MLE)	MLE with CF: α=0.16, S^{α}(G) - S^{α}(B) $\geq$ 0.5 (score)	MLE with CF: α=0.16, S^{α}(G) - S^{α}(B) $\geq$ 0.5 (variate)	Maximize S^{α}(G) - S^{α}(B) with α=0.5 (score)
W7_1	-0.0175	0.0196	-0.0170	28.5545
W7_2	0.0603	0.0791	0.0605	4.4645
W8_1	-0.2668	-0.2427	-0.2670	-4.9918
W8_2	-0.333	-0.3168	-0.3339	-17.1341
W8_3	-0.4859	-0.4534	-0.4865	-8.9478
W8_4	-0.2920	-0.1982	-0.2920	-8.1836
W8_5	-0.5996	-0.8201	-0.5998	-59.1433
W8_6	10.2217	6.9583	18.9844	-1.5178
W9_1	-1.1385	-1.2428	-1.1383	-59.191
W9_2	-0.7817	-0.9310	-0.7817	-66.4716
W9_3	-2.252	-2.67	-2.252	-55.7458
W9_4	-3.014	-3.777	-3.014	-119.3731
W10_1	-0.5407	-0.5942	-0.5402	-24.8087
W10_2	-0.6727	-0.7127	-0.6723	-55.3558
W10_3	0.4321	0.4633	0.4316	25.0024
W10_4	0.2742	0.2562	0.2746	21.0402
W10_5	0.7086	0.6582	0.7097	42.0908
W10_6	1.5199	1.4884	1.52	67.5194
W10_7	2.9821	3.2432	2.981	101.2713
W11	-0.2095	-0.3061	-0.2109	-34.4038
Constant	0.7011	1.1047	0.7009	18.7720
KS Value (development sample)	0.4802	0.4721	0.4806	0.52
Validation sample KS (5983 good, 4279 bad)	0.48	0.48	0.4802	0.476

Example 2- continued

Figure 2 : Good & Bad CDF Curves MLE, P1-P3 Models



Conclusions

- Presented methodology to incorporate stochastic dominance constraints efficiently within existing credit scoring model development framework
- Used the framework to formulate nonlinear programming problems whose solutions generate alternate “pareto-efficient” scorecards
- Demonstrated that the nonlinear optimization problems can be readily implemented within SAS making this framework easily accessible to practitioners & researchers
- Solved real-life problems and demonstrated that the use of 2 or more “pareto-efficient” scorecards can lead to better business solutions
- Cornish-Fisher approximation may not work well for extreme quantiles or highly skewed distributions. Quantile approximation of order statistics are being studied (Heidelberger & Lewis (1984), McNeil, Fowler, Mackulak, Nelson (2005))