Session I

Building and Validating Credit Rating and Scoring Models

Dennis Glennon

Deputy Director – Credit Risk Modeling
Risk Analysis Division
Washington, DC 20219
Dennis.Glennon@occ.treas.gov

Outline

- Introduction
- Developmental Evidence: Building statistical-based rating/scoring models
- Performance Evaluation: Verifying the model works

Introduction

- Is there a supervisory concern?
  - Sound modeling practices
- How do we approach the supervision of model risk?
  - Focus on two fundamental properties of a valid modeling process:
    - logically consistent model/sample design
    - valid statistical methods

Introduction

- Sound modeling practices
  - There are generally accepted, or industry-accepted, methods of building and validating models.
  - These methods incorporate procedures developed in the statistics, econometrics, information-theory, and operations research literature.
  - Although these methods are valid, they may not be appropriate in all applications.
    - A model selected for its ability to discriminate between high and low risk may perform poorly at predicting the likelihood of default.

Developmental Evidence

- Model development is a process.
- Simply stated:
  - define the purpose – discrimination vs prediction;
  - select a sample that reflects/represents the targeted population – a reference data set;
  - select a modeling technique consistent with the purpose;
  - identify risk factors that reflect the lender's knowledge and historical experience;
  - fit the model and check for model mis-specification or overfitting of the data;
  - develop methods of verifying that the model works – outcome-based methods.

Developmental Evidence

- Sample-design issues
  - Missing data
    - not available
    - censored data (i.e., reject inference)
    - truncated data (i.e., prepayment/attrition)
  - Omitted variables (implicitly held constant)
    - product terms (e.g., price, payment options)
    - economic conditions (e.g., interest rates, employment, business/industry conditions)
  - Pooling time-sensitive data

Developmental Evidence

- Modeling techniques
  - Expert systems
  - Regression
    - logit, probit, least squares, neural network
  - Decision-tree methods
    - CHAID, CART
  - Linear programming

Developmental Evidence

- *Step 1:* Univariate analysis used to reduce the set of potential risk factors to a subset of feasible risk factors
  - correlation
  - weight of evidence
- *Step 2:* Multivariate analysis used to capture the combined effect of multiple factors on expected performance
  - regression approach

Developmental Evidence

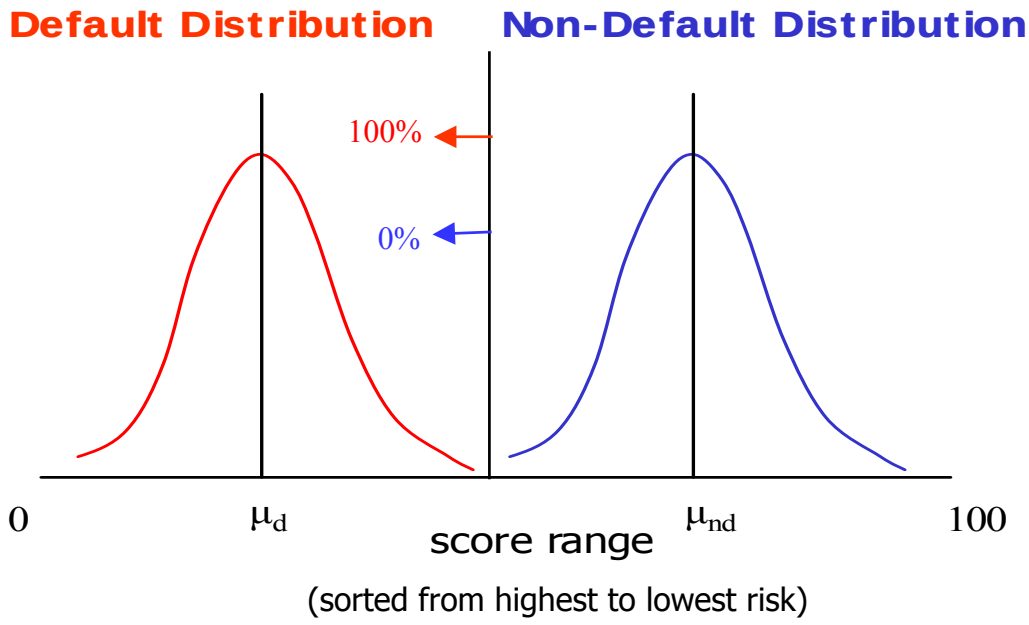
- Selecting the "best" model
  - Business sense
  - Diagnostic test
  - Out-of-sample analysis
    - In-time sample (i.e., hold-out)
    - Out-of-time sample
  - Cross-validation/benchmark analysis

Performance Evaluation

- Common performance measures
  - Kolmogorov-Smirnov (K-S)
  - Gains charts/cumulative accuracy profiles (CAP)
  - Divergence
  - Log-odds

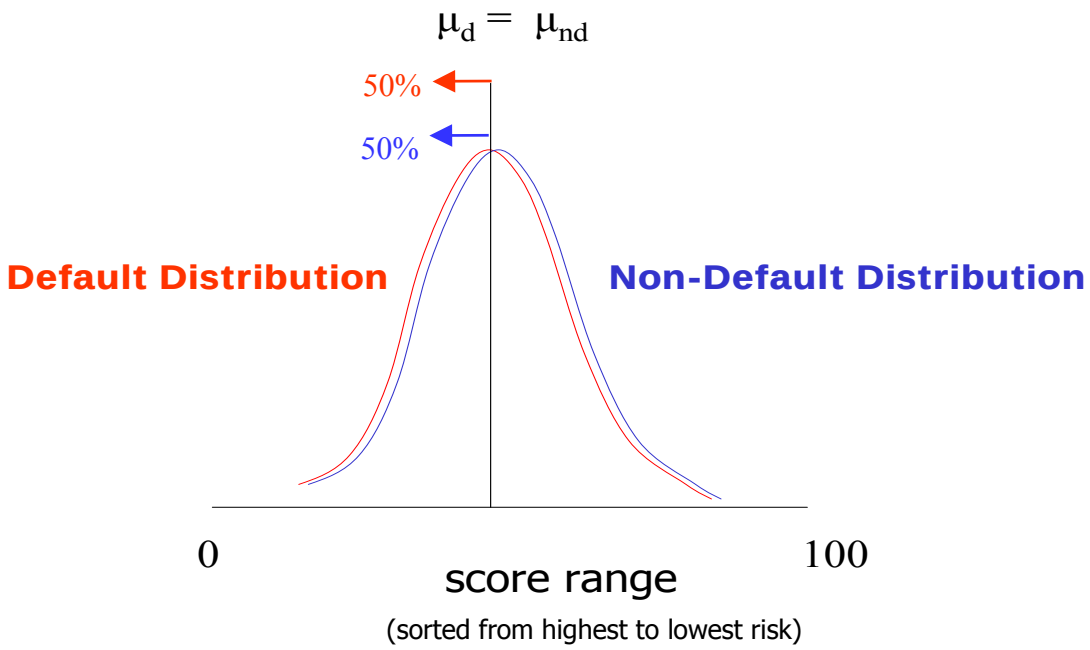
Performance Evaluation

- Model Performance Measures - K-S
  - *Upper Bound:* If the scores partition the population into two separate groups in which one group contains all the defaulted accounts and the other all the non-defaulted accounts, then the K-S is 100.



Performance Evaluation

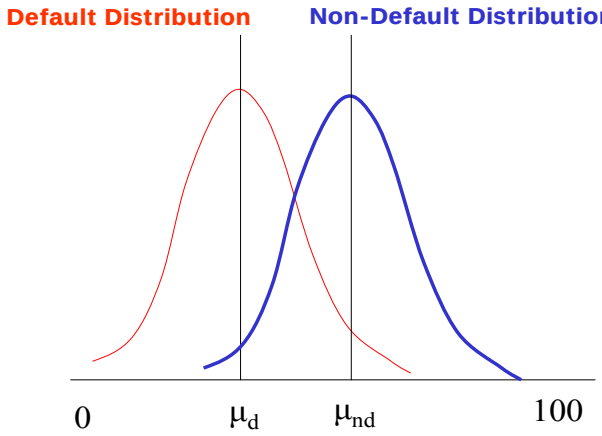
- Model Performance Measures - K-S
  - *Lower Bound:* If the model can not differentiate between non-defaulted and defaulted accounts, then it is as if the model selects individuals randomly from the population. There would be no difference in the location of the distributions. The K-S would be 0.



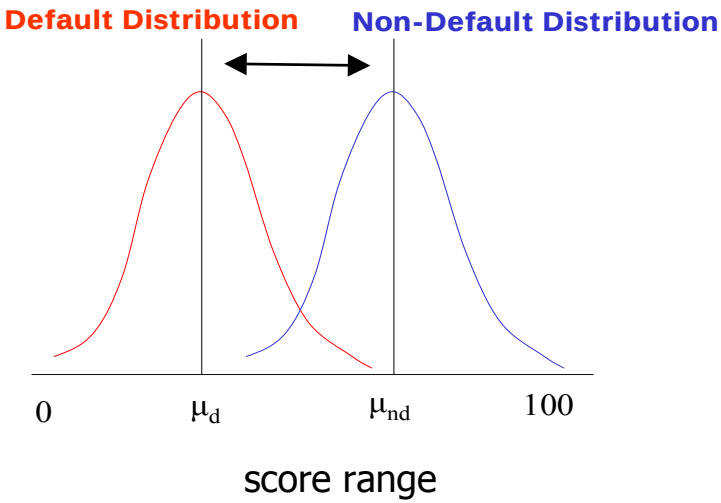
Performance Evaluation

- Model Performance Measures - K-S
  - These results suggests that the K-S value will fall between 0 and 100, and that the higher the value the better the model is at separating the non-defaulted from defaulted accounts.

Model 1



Model 2



Performance Evaluation

Model performance: Kolmogorov-Smirnov (K-S)

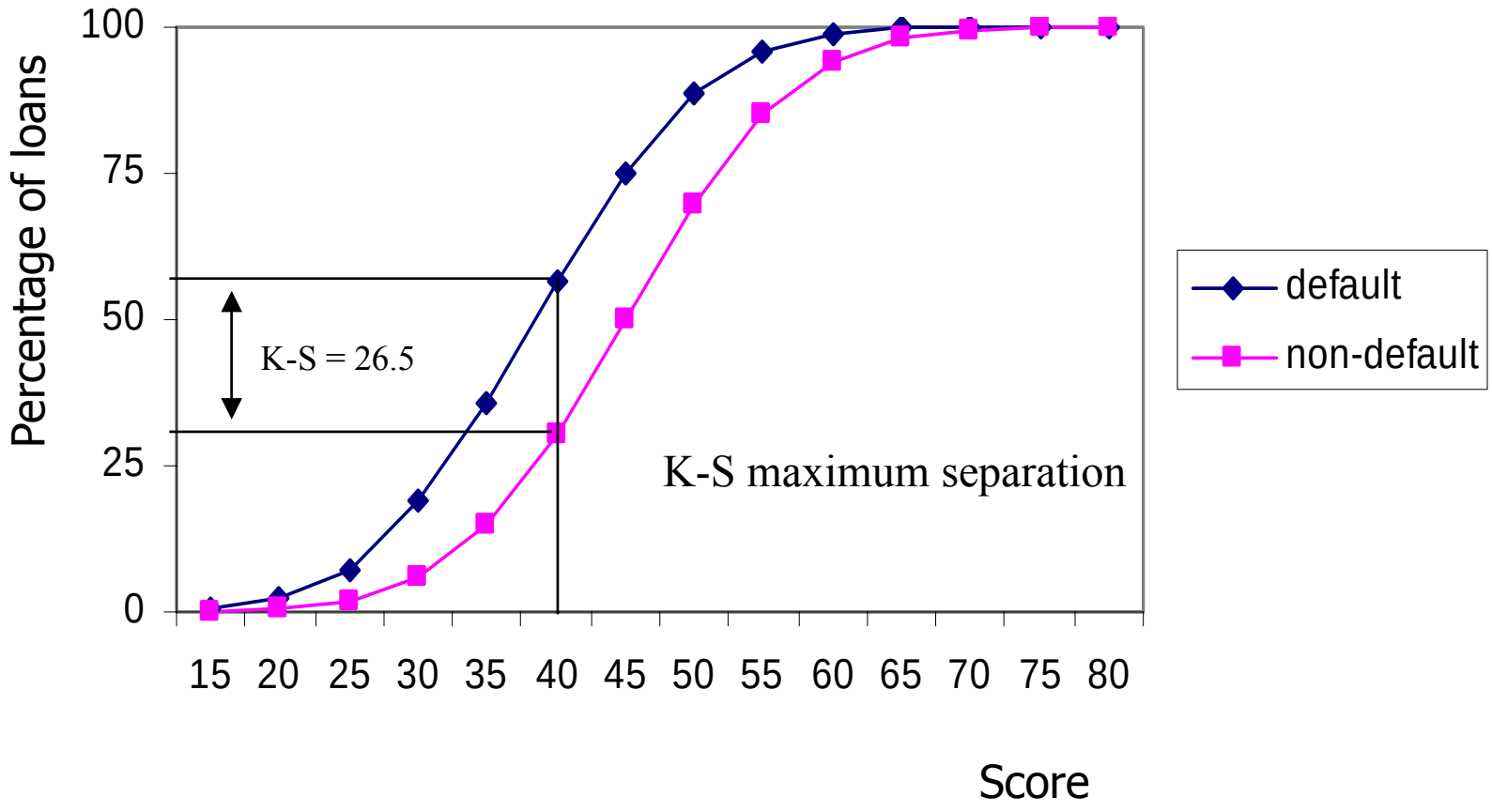
Obs (i)	Score Range lower	upper	Distributions Default (#)	Non-Def (#)	Cumulative Distributions Default (%)	Non-Def (%)	K-S
1	0	15	82	458	0.34	0.05	0.29
2	15	20	428	3205	2.12	0.37	1.75
3	20	25	1235	13886	7.24	1.75	5.49
4	25	30	2778	41657	18.77	5.92	12.85
5	30	35	4074	91645	35.69	15.09	20.60
6	35	40	5092	152741	56.82	30.36	**26.46**
7	40	45	4365	196381	74.94	50.00	24.94
8	45	50	3274	196381	88.53	69.64	18.89
9	50	55	1698	152741	95.58	84.91	10.67
10	55	60	764	91645	98.75	94.08	4.67
11	60	65	232	41657	99.71	98.24	1.47
12	65	70	58	13886	99.95	99.63	0.32
13	70	75	9	3205	99.99	99.95	0.04
14	75	80	1	458	100	100	0.00
15	80	100	1	31	100	100	0

(82+428)/24092

Total Bad = 24092

Performance Evaluation

- Cumulative non-default and default distributions



Performance Evaluation

- Separation as a modeling objective
  - *Comment:* The K-S statistic is *not* a measure derived from the difference between the actual and predicted values of the dependent variable; as such, it is *not* an R^2-type measure of model accuracy.
  - *Comment:* For that reason, in practice, the K-S test is used to evaluate the model as a segmentation or classification tool. As a result, this test does not necessarily identify the model that is best at predicting the probability of default.

Performance Evaluation: K-S Test

- *Hypothesis Test:* the difference between two distributions
  - Test statistic (K_α)

$$K_\alpha = 100 \{ D \left[\frac{(\#non\text{-}defaults + \#defaults)}{(\#non\text{-}defaults)(\#defaults)} \right]^{1/2} \}$$

where α = significance level (e.g., .95)

D = critical value (table value)

Performance Evaluation: Hypothesis Test

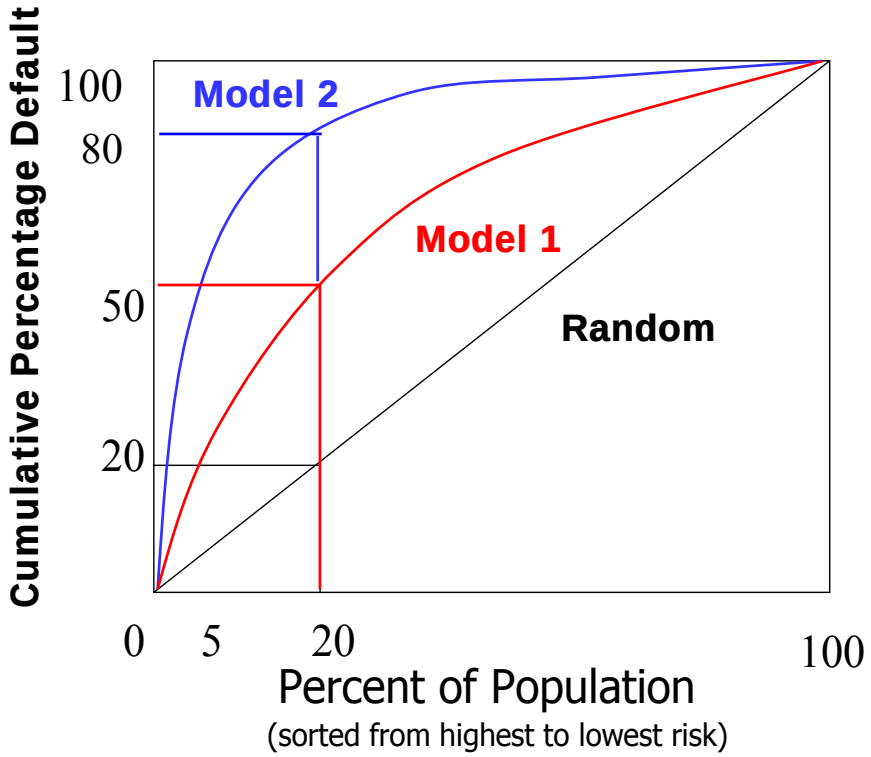
- Difference between two distributions
  - *Example 1 (from above):* Default Rate = 2.35%
    - # Defaults = 24,091
    - # Non-defaults = 999,977
    - $K_{\alpha=.95}$ = 0.80% < KS = 26.5%

- *Example 2:* Default Rate = 4.41%
  - # Defaults = 441
  - # Non-defaults = 9,559
  - $K_{\alpha=.95}$ = 5.94%

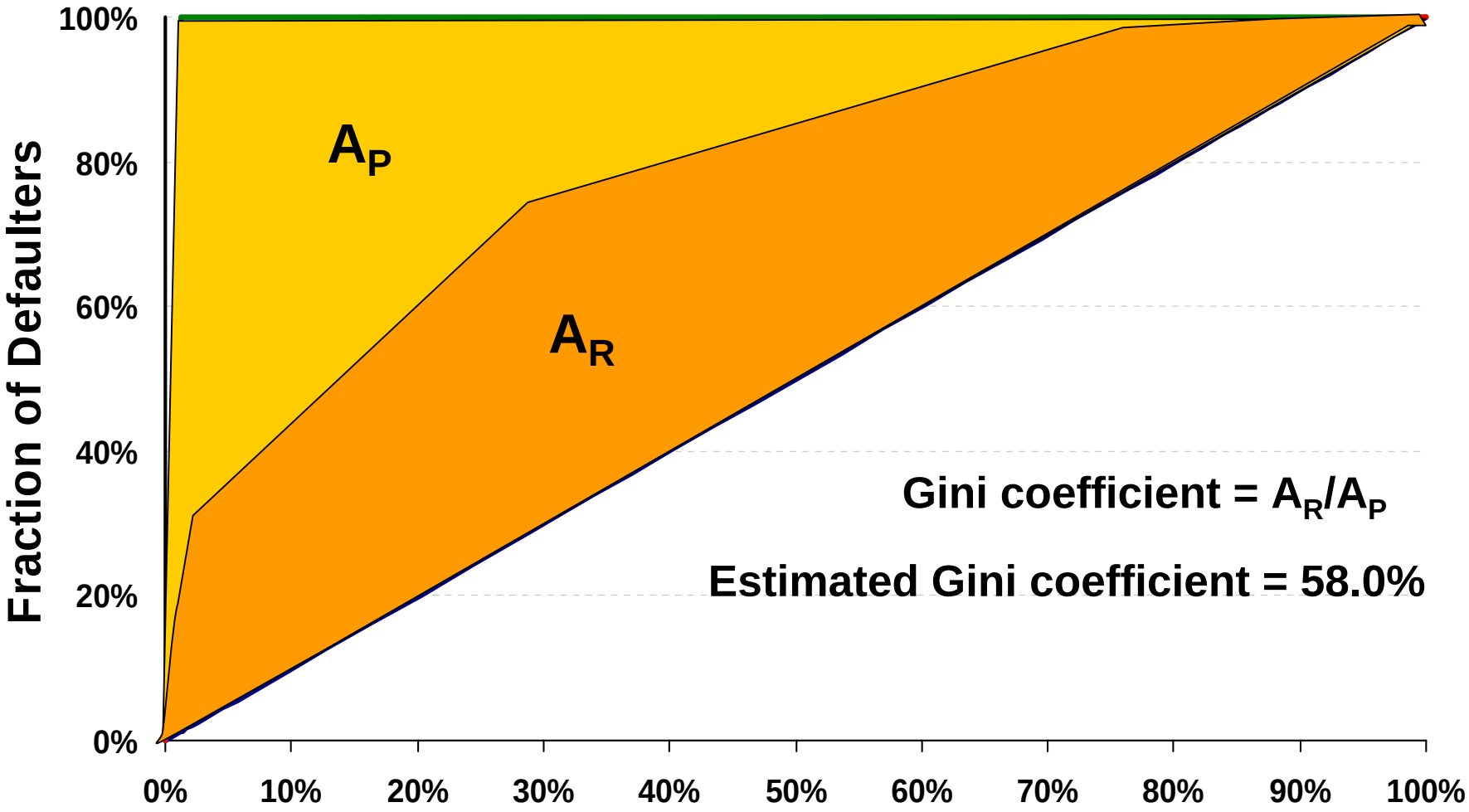
- *Example 3:* Default Rate = 50.0%
  - # Defaults = 1,500
  - # Non-defaults = 1,500
  - $K_{\alpha=.95}$ = 4.45%

Performance Evaluation: Gains Chart

- Model Performance Measure



Performance Evaluation: Gini Coefficient



Performance Evaluation: Gini Coefficient

- There is no magic number
  - Higher is better, but there is no ratio that says a scoring or rating system is “good” or “bad”
- All errors are *not* created equal
  - Gini coefficient treats “false negatives” and “false positives” as equally bad
- Be careful about making comparisons
  - Dispersion of credits across score ranges or grades
  - Number of defaulters in sample
  - Portfolio composition

Performance Evaluation: Diagnostics

- Goodness-of-fit measures
  - R^2-type measure of goodness-of-fit are generally not used.
- Robustness test: out-of-sample analysis
  - *In-time sample*: Observations randomly selected from the development or reference data.
  - *Out-of-time sample*: Observations randomly selected from a population with observation and performance periods different from those of the reference sample.

Performance Evaluation

- In-time and out-of-time analyses
  - The models are evaluated in terms of their ability to maintain:
    - stable parameter estimates across the different validation samples, and
    - a given level of separation between the good and bad distributions (i.e., stable K-S statistics).

Performance Evaluation

- Are these tools really useful?
- *Illustrative Example:* Developing a Credit Scoring Model for Risk Segmentation Purposes
  - *Sample:* A simulated random sample of 10,000 observations.
  - *Performance:* Derived from the following data generating process.

Illustrative Example: Data

- *Data Generating Process*
  - y = 0 if Y < 0
    y = 1 otherwise
  - $Y = \mathbf{-52.5} + \mathbf{1.0}\ r_1 + \mathbf{1.0}\ r_2 + \mathbf{1.0}\ r_3 + \mathbf{2.0}\ r_4 + \mathbf{2.0}\ r_5 + \mathbf{2.0}\ r_6 - e$
  - where
    - $e \sim \text{logistic}(0, \pi^2/3)$
    - $r_1 - r_6$ are uncorrelated continuous random variables
  - pr(default) = pr(y=1)
  - mean of y = 0.0866

Illustrative Example: Univariate

Variables	Estimated Parameters β	P-Values Pr > ChiSq	Divergence Index
r1	0.1442	0.0001	0.3746
r2	0.1415	0.0001	0.4875
r3	0.1339	0.0001	0.4488
r1	0.3729	0.0001	3.746
r2	0.2917	0.0001	1.383
r3	0.2911	0.0001	1.345
w1	0.0356	0.3172	0.0048
w2	-0.0795	0.0256	0.0192
w3	-0.0197	0.5792	0.0039
w4	0.0523	0.1019	0.7826
w5	0.2237	0.0001	0.2839
w6	0.0412	0.2465	0.0033

The estimated parameters β are derived from the univariate (logit) regression models:

$y = b_0 + b_1 x$ where x = {r1, r2, r3, …, w6}; and y = 1 if default, 0 otherwise.

The Divergence Index is: $D = \Sigma^{10}_{g=1} (p_g - q_g) \ln(p_g/q_g)$, where p_g (q_g) is the percentage of non-default (default) accounts in the g^{th} decile.

Illustrative Example: Logit Model

		Exact	Over-specified		Mis-specified	
Variables	Development Parameters	Estimated Parameters	Estimated Parameters	Estimated Parameters	Estimated Parameters	
Intercept	-52.5	-53.1642	-53.2803	-53.2157	-11.7397	<.0001
r1	1.0	1.0127	1.0141	1.0128	0.3476	<.0001
r2	1.0	0.9777	0.9781	0.9779	0.3255	<.0001
r3	1.0	0.9745	0.9775	0.9746	0.3201	<.0001
r4	2.0	2.0327	2.0312	2.0330	0.6712	<.0001
r5	2.0	1.9921	1.9715	1.9673	1.3224	<.0001
r6	2.0	2.0185	2.0304	2.0254		
w1	0		-0.0575			
w2	0		-0.0769			
w3	0		0.0166			
w4	0		0.4387	0.4401	-9.7901	<.0001
w5	0		0.0312			
w6	0		0.0307			

Illustrative Example

Parameter Stability

Variables	Exact Specification		Over-Identified		Mis-Specified	
	Development	Validation	Development	Validation	Development	Validation
Intercept	-53.1642	-52.8708	-53.2157	-52.8258	-11.7397	-10.3896
r1	1.0127	1.0195	1.0128	1.0197	0.3476	0.3161
r2	0.9777	1.0070	0.9779	1.0072	0.3255	0.3108
r3	0.9745	1.0214	0.9746	1.0217	0.3201	0.3188
r4	2.0327	1.9487	2.0330	1.9494	0.6712	0.6006
r5	1.9921	2.0400	1.9673	2.0669	1.3224	1.0034
r6	2.0185	2.0384	2.0254	2.0315		
w4			0.4401	-0.1527	-9.7901	-1.8131

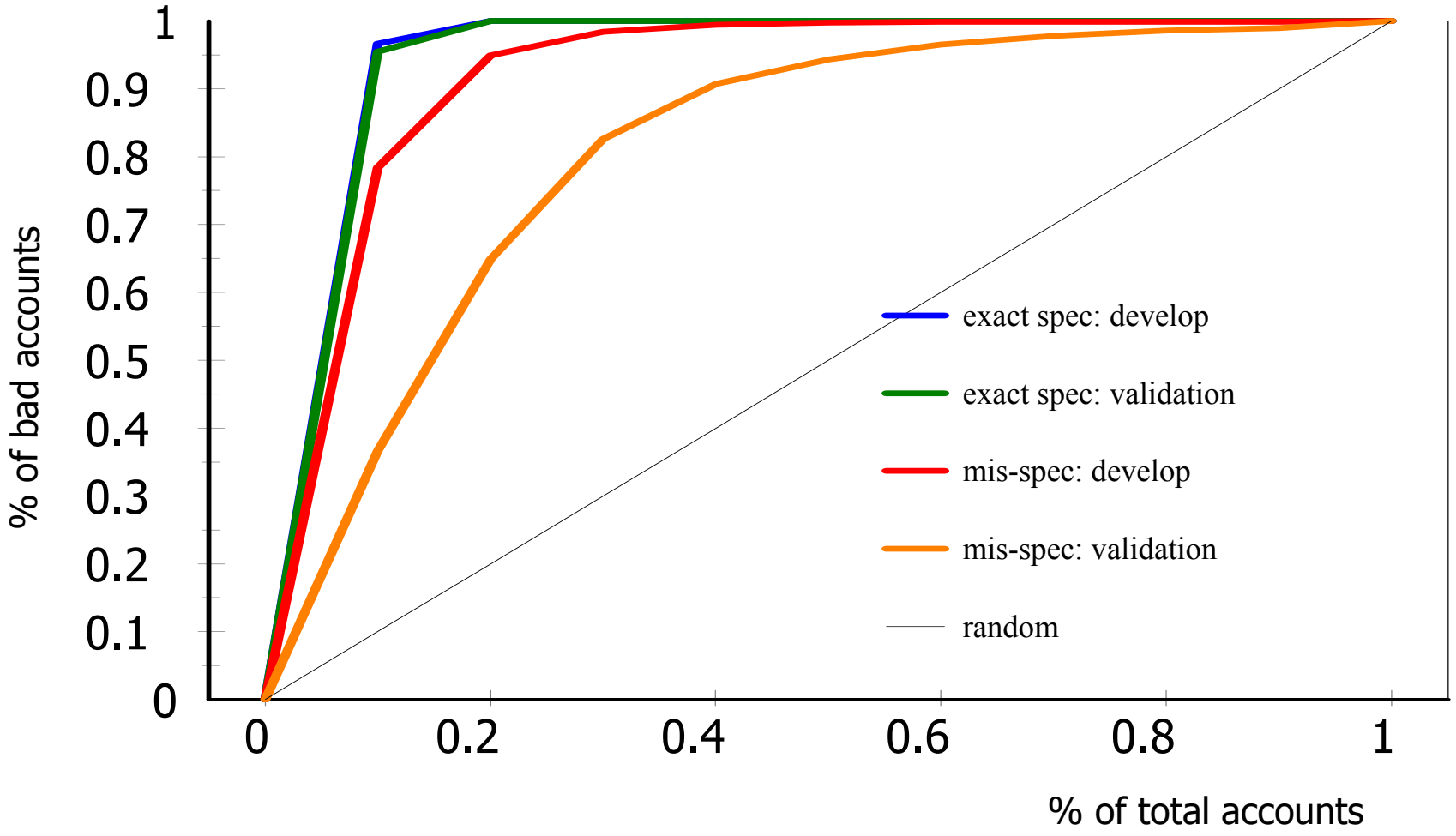
black – statistically significant at the 1% level
blue – statistically significant at the 10% level
red – statistically insignificant at the 10% level

Illustrative Example

- K-S Test
  - The K-S stat is a measure of the degree of separation between the non-default and default distributions.

Model	Sample		
	Development	Validation	Total
Exact	94.9	93.6	94.1
Over-Identified	94.7	93.2	94.0
Mis-Specified	82.0	57.6	66.2

Illustrative Example: Gains Charts



Performance Evaluation: Other Issues

- Developing benchmarks for performance monitoring and early-read/early-warning analysis.
  - Benchmark values and distributions constructed at time of model development are used to differentiate between
    - temporary shifts due to “random” shocks
    - permanent drift due to structural changes

Conclusion

- Model development is a process.
- Models should be developed using sound modeling practices.
- Model verification is an integral part of the model development process.