

デフォルト確率推定モデル作成における WOE変換の役割とその利用

氏名：大勢待 利明(おおせまち としあき)

所属名：株式会社 数理技研

部署属名：金融工学センター

The use of WOE in a model to estimate PD

Name : Toshiaki Ohsemachi

Department, Corporate

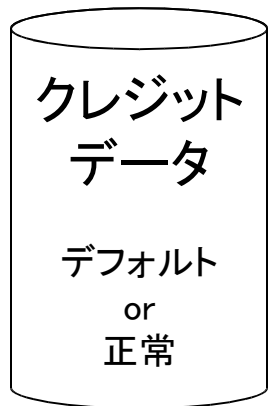
Financial Engineering Center, SURIGIKEN CO.,LTD.

本日の発表の流れ

もくじ

- 1 WOE変換の意義と重要性
- 2 発想の原点
- 3 WOEとは
- 4 WOEの利用手順
- 5 モデル作成・検証でのWOEの利用方法
- 6 モデル・モニタリングでのWOEの利用方法
- 7 まとめ

1 WOE変換の意義と重要性



変数加工
(前処理)

デシジョンツリー

ロジステック回帰モデル

$$y = \frac{1}{1 + e^{-(a_1x_1 + \dots + a_nx_n + b)}}$$

ニューラルネットワーク

遺伝的プログラミング

サポートベクターマシン

$$F(X) = W^T * X_i - h$$

生存時間モデル

$$S(t) = S_0(t) \exp(a_1x_1 + \dots + a_nx_n)$$

モデル作成に有効な変数を準備できるかどうか勝負

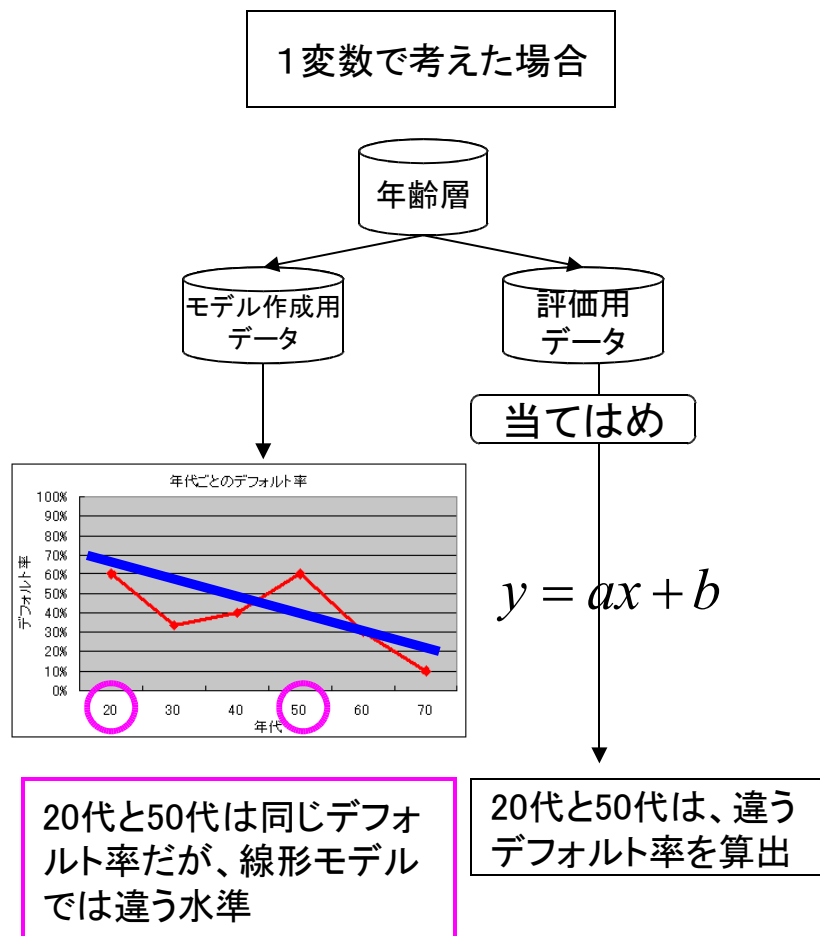
WOE変換

どのロジックで解くか？

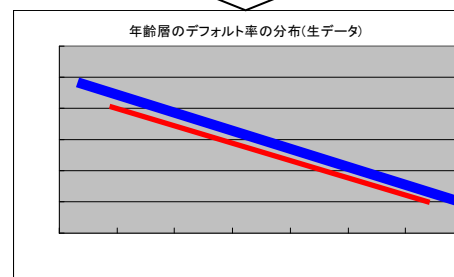
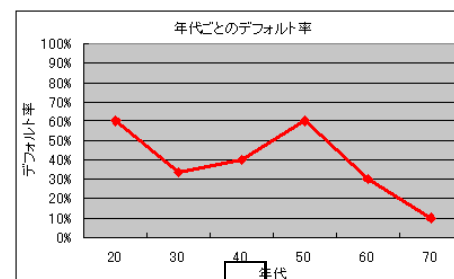
結果に大きな違いはない・・・

- ①スコアリングに使用するモデル自体は飽和状態
- ②ロジック選択に依存したスコアリング・モデルの精度向上の時代は終焉を迎えつつある

2 発想の原点



変数の値とデフォルト率の関係を線型にできれば...



WOE変換

生データの場合より
精度up

3 WOEとは

由来

1950年、ジョン・グッド(統計学者)が提唱、地質学の分野で応用。
その後、医療の分野でも利用。
近年、信用リスクの分野でも外資系の金融機関で利用され始めている。

WOE(Weights Of Evidence) ターゲットとする変数が、デフォルトか正常かの2値の場合

$$\text{Weights Of Evidence}_{\text{attribute}} = \log\left(\frac{\text{P_Default}_{\text{attribute}}}{\text{P_Normal}_{\text{attribute}}}\right)$$

$$\text{P_Normal}_{\text{attribute}} = \frac{\text{Number Of Normal}_{\text{attribute}}}{\text{Number Of Normal}_{\text{total}}}$$

$$\text{P_Default}_{\text{attribute}} = \frac{\text{Number Of Default}_{\text{attribute}}}{\text{Number Of Default}_{\text{total}}}$$

年齢層	カテゴリ値1	count	Normals	P_Normal attribute	Defaults	P_Default attribute	Default rate	WOE値	カテゴリ値2
missing	missing	50	42	2.3%	8	4.1%	16.0%	0.579	0.579
20-22	20	200	152	8.4%	48	24.9%	24.0%	1.084	1.084
23-26	23	300	246	13.6%	54	28.0%	18.0%	0.720	0.720
27-29	27	450	405	22.4%	5	2.6%	1.1%	-2.158	-2.158
30-34	30	500	475	26.3%	25	13.0%	5.0%	-0.708	-0.708
35-44	35	350	339	18.8%	11	5.7%	3.1%	-1.191	-1.191
44+	44	150	147	8.1%	3	1.6%	2.0%	-1.655	-1.655
合計		2000	1807		193		9.7%		

4 WOEの利用手順

- 4.1 連続変数をカテゴリ変数に変換
- 4.2 カテゴリ変数を連続変数に変換
- 4.3 安定性のある変数の作成
- 4.4 有効な変数の選択

連続変数を
カテゴリ変数へ

カテゴリ変数を
WOE変換

CAP曲線差
変数安定

IVで
変数選択

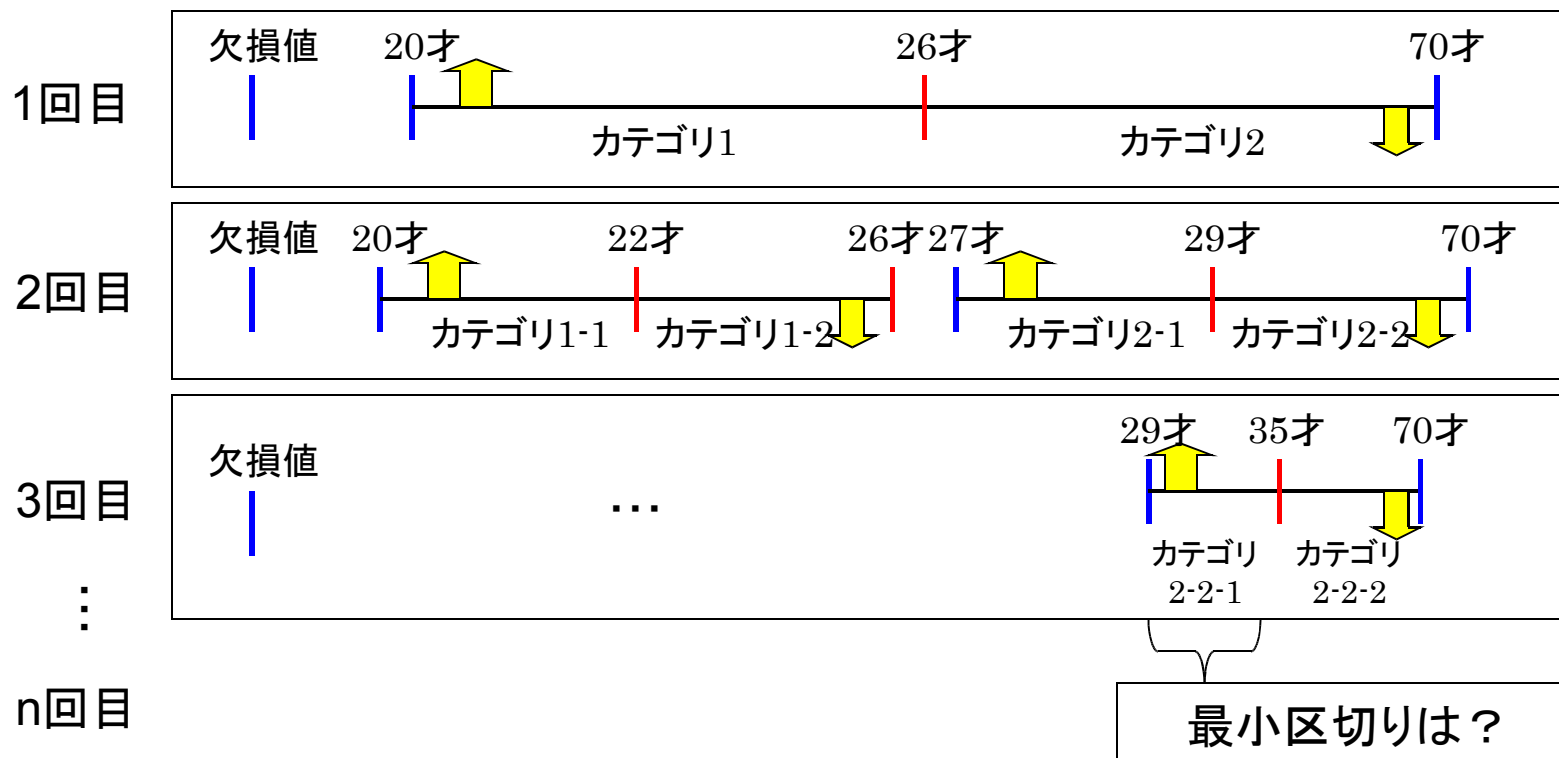
4.1 連続変数をカテゴリ変数に変換

連続変数

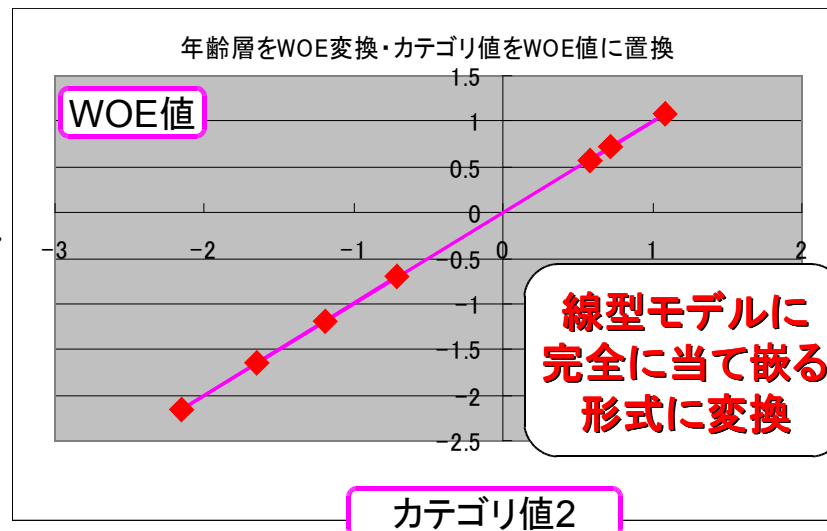
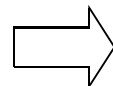
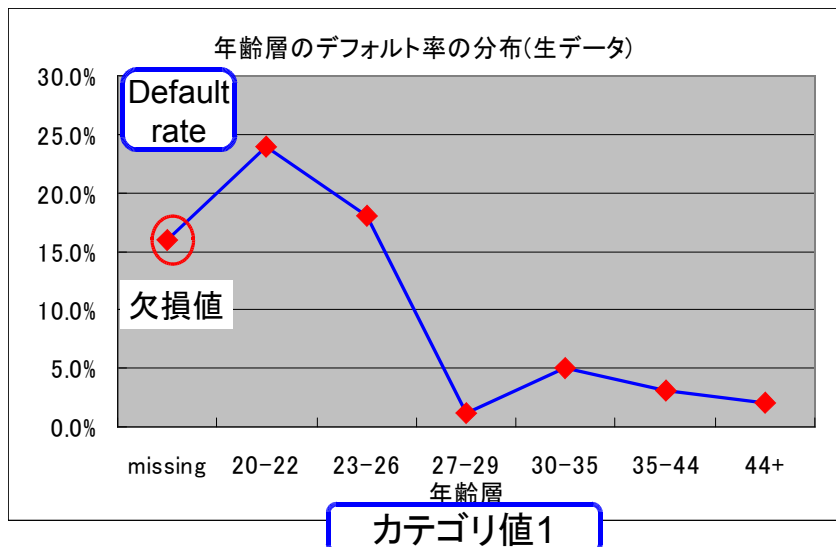
分散分析(=ANOVA(=Analysis of Variance))で連続変数をカテゴリ化

カテゴリ化

デフォルト率の差が最大になる年齢で、2つのグループに分離



4.2 カテゴリ変数をWOE変換



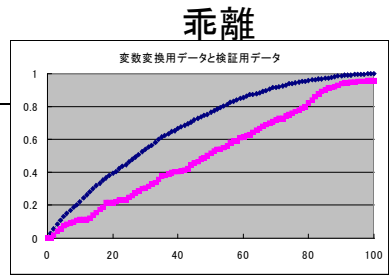
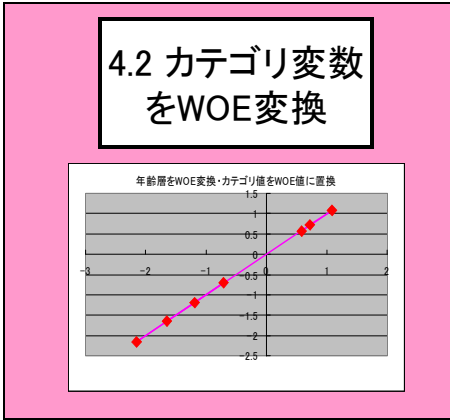
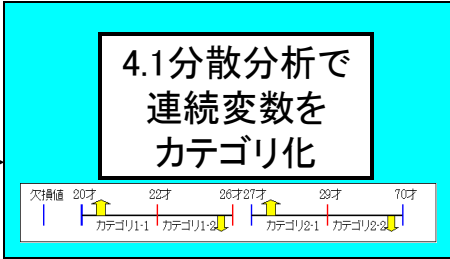
カテゴリ値1は、名目値にすぎないので、
カテゴリ値2それぞれをWOE値に置き換える

年齢層	カテゴリ値1	count	Normals	P_Normal attribute	Defaults	P_Default attribute	Default rate	WOE値	カテゴリ値2
missing	missing	50	42	2.3%	8	4.1%	16.0%	0.579	0.579
20-22	20	200	152	8.4%	48	24.9%	24.0%	1.084	1.084
23-26	23	300	246	13.6%	54	28.0%	18.0%	0.720	0.720
27-29	27	450	405	22.4%	5	2.6%	1.1%	-2.158	-2.158
30-34	30	500	475	26.3%	25	13.0%	5.0%	-0.708	-0.708
35-44	35	350	339	18.8%	11	5.7%	3.1%	-1.191	-1.191
44+	44	150	147	8.1%	3	1.6%	2.0%	-1.655	-1.655
合計		2000	1807		193		9.7%		

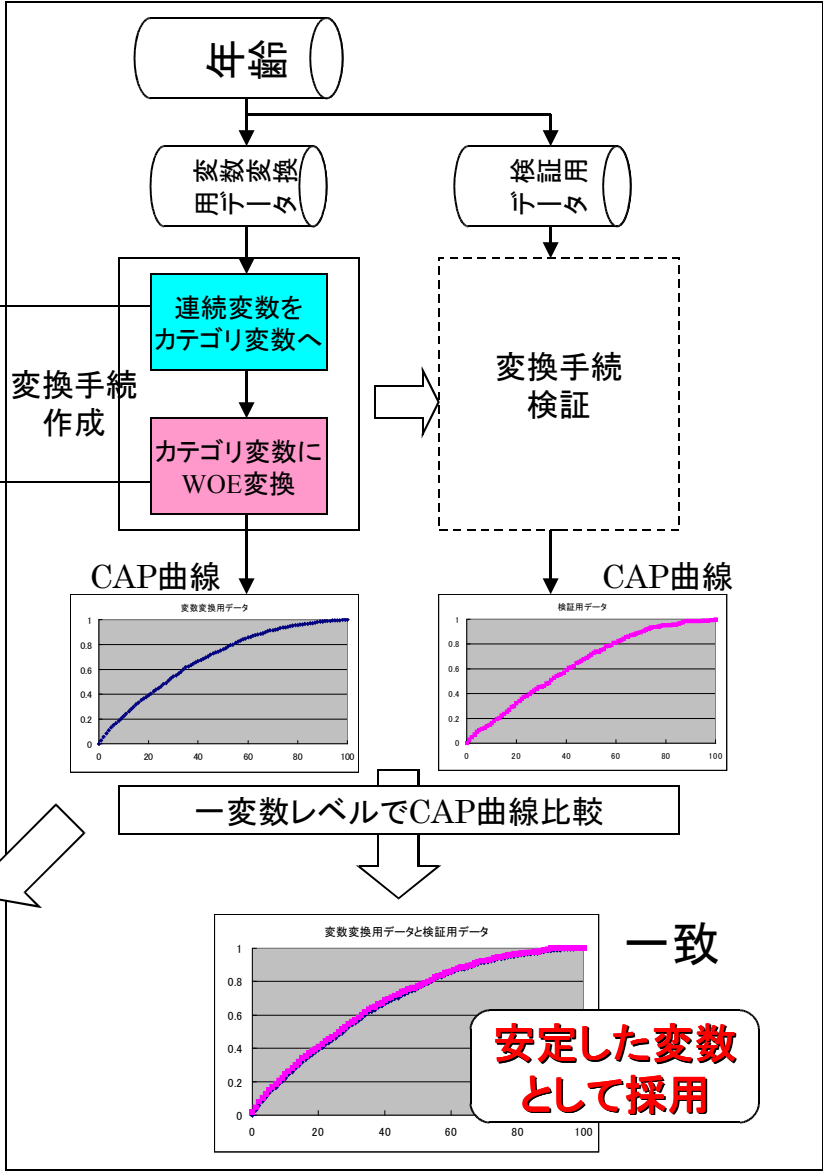
4.3 安定性のある変数の作成

(区切りの調整)

最小区切り
を調整



CAP曲線による変数の安定性の確認



4.4 有効な変数の選択方法

Information Value (=IV)による変数の有効性の確認

$$IV = \sum \left((P_Default_{\text{attribute}} - P_Normal_{\text{attribute}}) * WOE_{\text{attribute}} \right)$$

年齢層	カテゴリ値1	count	Normals	P_Normal attribute	Defaults	P_Default attribute	Default rate	WOE値	カテゴリ値2
missing	missing	50	42	2.3%	8	4.1%	16.0%	0.579	0.579
20-22	20	200	152	8.4%	48	24.9%	24.0%	1.084	1.084
23-26	23	300	246	13.6%	54	28.0%	18.0%	0.720	0.720
27-29	27	450	405	22.4%	5	2.6%	1.1%	-2.158	-2.158
30-34	30	500	475	26.3%	25	13.0%	5.0%	-0.708	-0.708
35-44	35	350	339	18.8%	11	5.7%	3.1%	-1.191	-1.191
44+	44	150	147	8.1%	3	1.6%	2.0%	-1.655	-1.655
合計		2000	1807		193		9.7%		

Information Value = 0.066

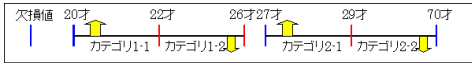
変数の影響力

< 0.02: unpredictive
 0.02 - 0.1 : weak
 0.1 - 0.3 : medium
 0.3 + : strong

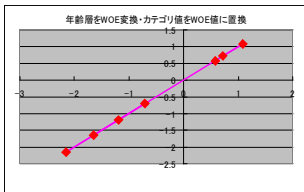
5 モデル作成・検証でのWOEの利用方法

変数加工

分散分析で連続変数をカテゴリ化

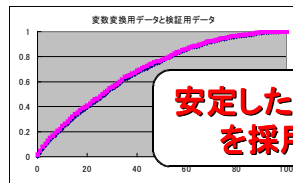


カテゴリ変数をWOE変換



変数選択

CAP曲線

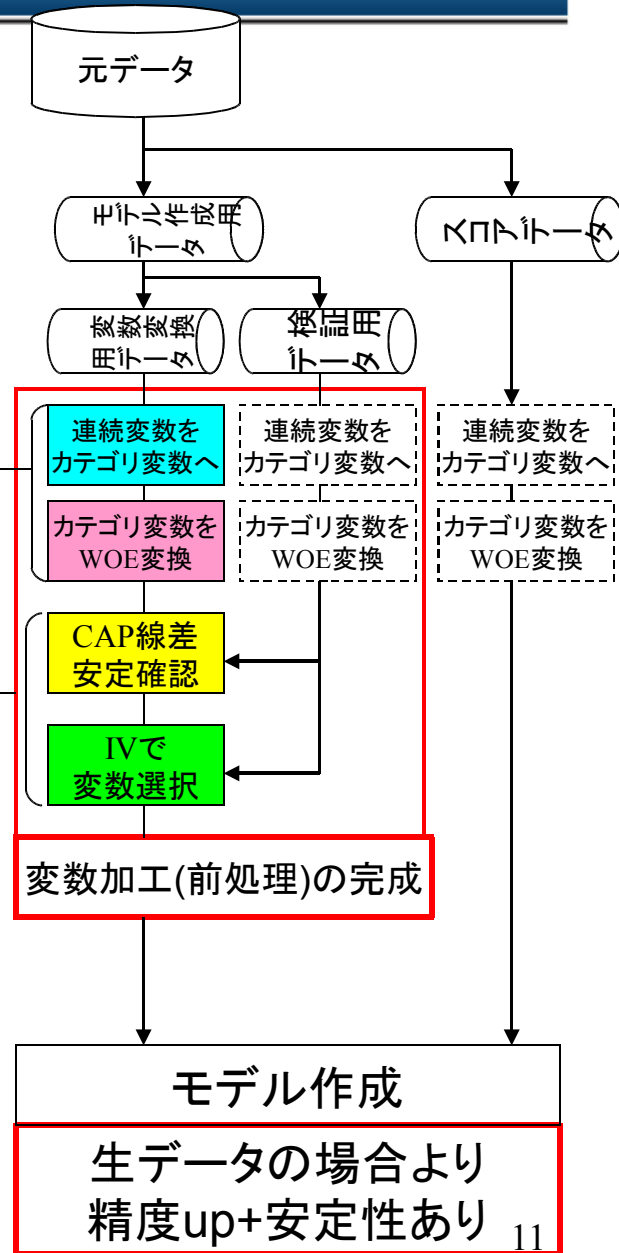


安定した変数を採用

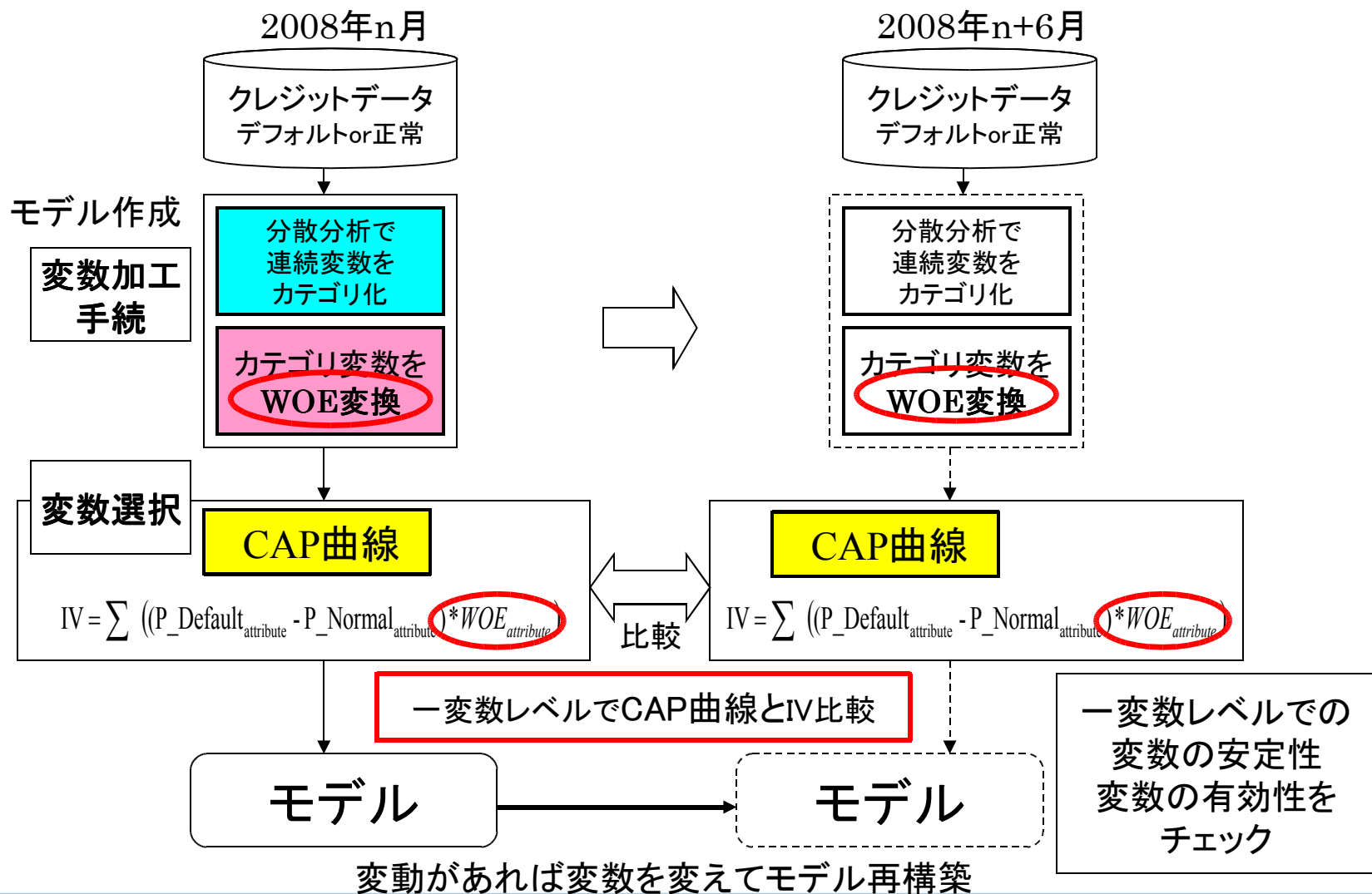
$$IV = \sum ((P_Default_{attribute} - P_Normal_{attribute}) * WOE_{attribute})$$

ex変数選択のための一覧表

	CAP曲線の最大差	Information Value	評価		モデル作成用変数として
			安定性	影響力	
年齢層	0.05	0.066	○	×	
変数A	0.02	0.6	○	○	採用
変数B	0.12	0.4	×	○	
変数C	0.4	0.4	○	○	採用
変数D	0.2	0.1	○	△	
...



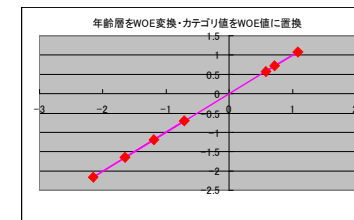
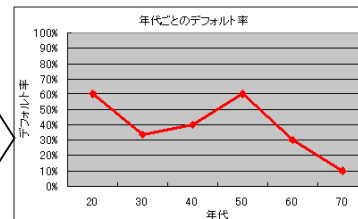
6 モデル・モニタリングでのWOEの利用方法



7 まとめ

WOE変換

①生データを線型に変換



②生データの場合より精度up+安定性のあるモデルの作成が可能

連続変数のカテゴリ化、変数の有効性、変数の安定性のチェックも必要

③モデル・モニタリングに利用可能