2004年7月29日

SAS Forum ユーザ会

学術総会 2004

SAS/EMを用いた
組み合わせ分類変数の作成と効用

How to Make the Combination of Group Variables by SAS/EM ?

UFJ銀行

小 野 潔

(本報告は個人的見解です)

分析データ紹介

名前	役割	測定水準	タイ・	出力形式	入力形式	
A_GOAL	target	binary	char	$4.	$4.	目標
HOUSE_OCCU	rejected	nominal	char	$13.	$13.	居住&職種
TIMES_CAT	rejected	ordinal	char	$1.	$1.	回数ﾗﾝｸ
FLG_DEPT_OTHER	input	binary	num	BEST12.	12.	他社貸ﾌﾗｸﾞ
FLG_HIGH_SALARY	input	binary	num	BEST12.	12.	高年収ﾌﾗｸﾞ
FLG_SALARY	input	binary	num	BEST12.	12.	年収記入ﾌﾗｸﾞ
SEX	input	binary	num	BEST12.	12.	性別
INDUSTRY	input	nominal	char	$1.	$1.	業種
HOUSE	input	nominal	num	BEST12.	12.	住居形態
OCCUPATION	input	nominal	num	BEST12.	12.	職種
DEPT_REMAINDER_CAT	input	ordinal	char	$1.	$1.	他社借金ｶﾃｺﾞﾘ
DEPT_OTHER	input	ordinal	num	BEST12.	12.	他社借数
AGE_CAT	input	ordinal	char	$1.	$1.	年代
SUM_CAT	input	ordinal	char	$1.	$1.	契約金ｶﾃｺﾞﾘ
SALARY_CAT	input	ordinal	num	BEST12.	12.	年収

クラス変数

①職種のカテゴリー変数　　m種類

A.公務員	B.会社員	C.学生	D.主婦	E.自営業	・・・

②住居のカテゴリー変数　　n種類

a.自己所有	b.家族所有	c.社宅	d.借家	e.アパート	・・・

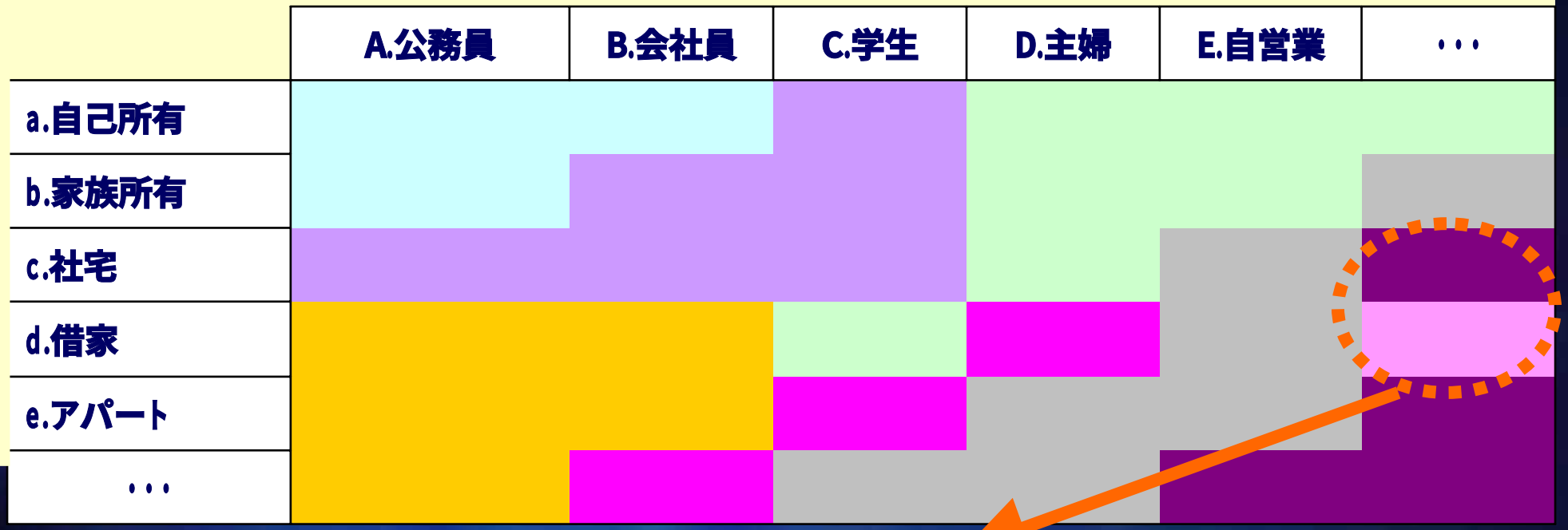
③職種＆住居のクラス変数　　m＊n種類

	A.公務員	B.会社員	C.学生	D.主婦	E.自営業	・・・
a.自己所有	A & a	B & a	C & a	D & a	E & a	・・・
b.家族所有	A & b	B & b	C & d	D & b	E & b	・・・
c.社宅	A & c	B & c	C & c	D & c	E & c	・・・
d.借家	A & d	B & d	C & d	D & d	E & d	・・・
e.アパート	A & e	B & e	C & e	D & e	E & e	・・・
・・・	・・・	・・・	・・・	・・・	・・・	・・・

分類（グループ）変数

SAS/EMは統計的手法により自動的に分類変数を作成

④職種＆住居の分類変数



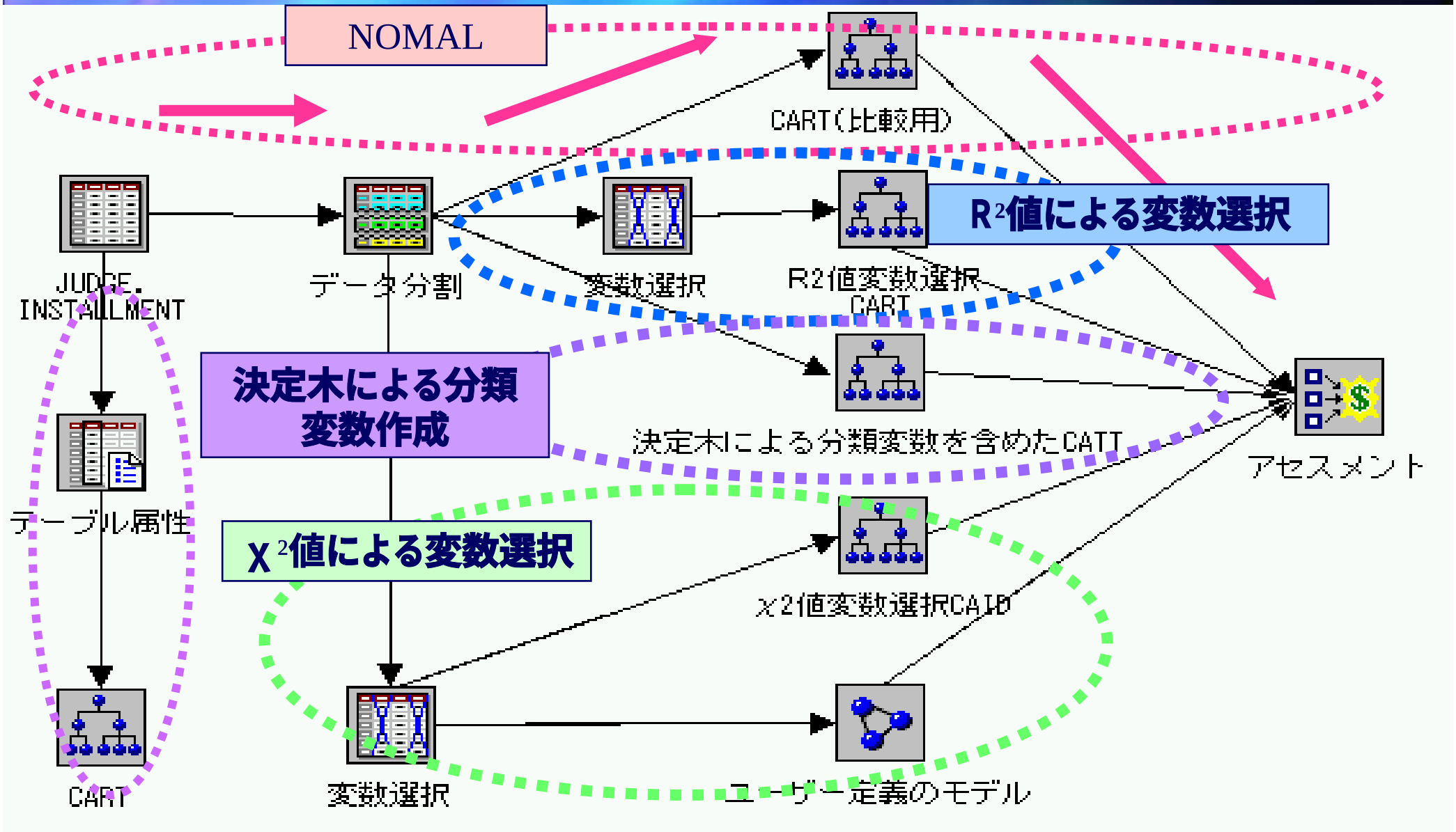
（問題）パレート順序がくずれているため、飛び地が発生.

（留意）上記の分類に意味はなく、単なる例示

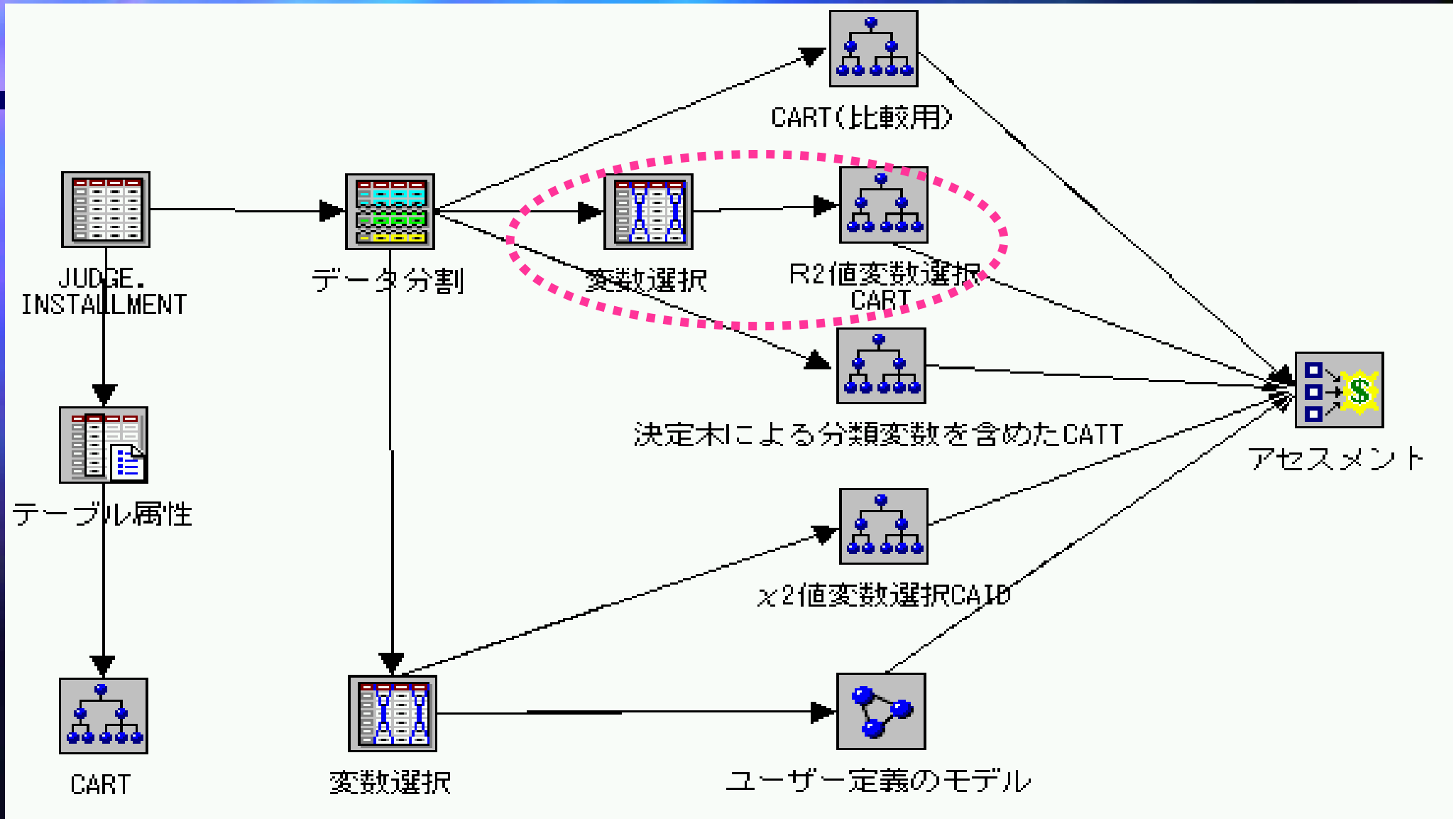
分類変数の利点

- 分類変数の特徴
  - 長所：モデルの精度や安定性の向上
  - クラス変数よりもオーバーフィッティングを抑制
  - 短所：カテゴリー変数の組み合わせは無数！！
  - 組み合わせは分析者の経験とノウハウから決定
  - 面倒なためあまり使われていない
  - 本報告：SAS/EMを用いて分類変数を半自動的に作成
- ポイント：選択基準による組み合わせ数の削減
  - R^2値による（SAS/EM標準装備）
  - χ^2値による（SAS/EM標準装備）
  - 決定木による（本採用の手法）

全体フロー


NOMAL
CART(比較用)
JUDGE.
INSTALLMENT
データ分割
変数選択
R2値変数選択
CART
R²値による変数選択
決定木による分類
変数作成
決定木による分類変数を含めたCART
アセスメント
テーブル属性
χ²値による変数選択
χ2値変数選択CART
CART
変数選択
ユーザー定義のモデル

R²値 (or F検定) による変数選択法



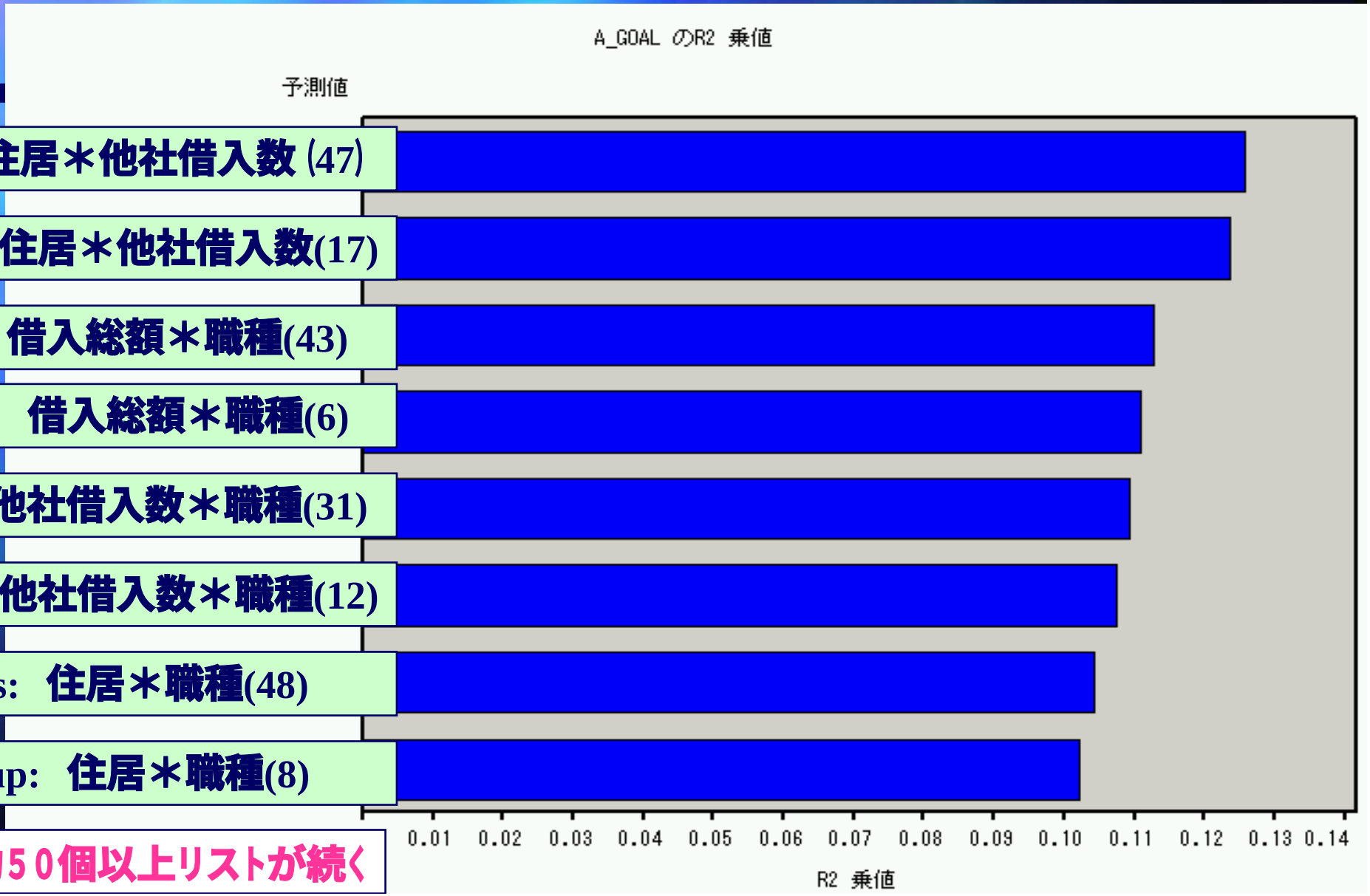
名前	役割	除外の理由	独立性
SEX	rejected	Grouped interaction variable GI_SEXHOUSE preferred	
SALARY_CAT	rejected	Grouped interaction variable GI_SALARY_CATAGE_CAT preferred	
FLG_SALARY	rejected	Grouped interaction variable GI_FLG_SALARYSUM_CAT preferred	
FLG_HIGH_SALARY	rejected	Grouped interaction variable GI_FLG_HIGH_SALARYSUM_CAT preferred	
FLG_DEPT_OTHER	rejected	Low R2 w/ target	
OCCUPATION	rejected	Grouped interaction variable GI_OCCUPATIONHOUSE preferred	
HOUSE	rejected	Grouped interaction variable GI_HOUSEDEPT_OTHER preferred	
SUM_CAT	rejected	Grouped interaction variable GI_SUM_CATINDUSTRY preferred	
AGE_CAT	rejected	Grouped interaction variable GI_SALARY_CATAGE_CAT preferred	
DEPT_OTHER	rejected	Grouped interaction variable GI_HOUSEDEPT_OTHER preferred	
DEPT_REMAINDER_	rejected	Grouped interaction variable GI_SALARY_CATDEPT_REMAINDER_CAT pref	
INDUSTRY	rejected	Grouped interaction variable GI_SUM_CATINDUSTRY preferred	
G_HOUSE	input		
GI_HOUSEDEPT_OT	input		
GI_SUM_CATINDUS	input		
GI_SALARY_CATAG	input		
GI_OCCUPATIONHO	input		
GI_HOUSEAGE_CAT	input		
GI_SALARY_CATSU	input		
GI_SALARY_CATHO	input		
GI_HOUSESUM_CAT	input		
GI_AGE_CATINDUS	input		
GI_OCCUPATIONAG	input		
GI_SALARY_CATIN	input		
GI_AGE_CATDEPT_	input		
GI_HOUSEINDUSTR	input		
GI_SALARY_CATOC	input		

SAS/EMが自動的に不採用と判定

R²値(or F検定) 選択基準による変数選択

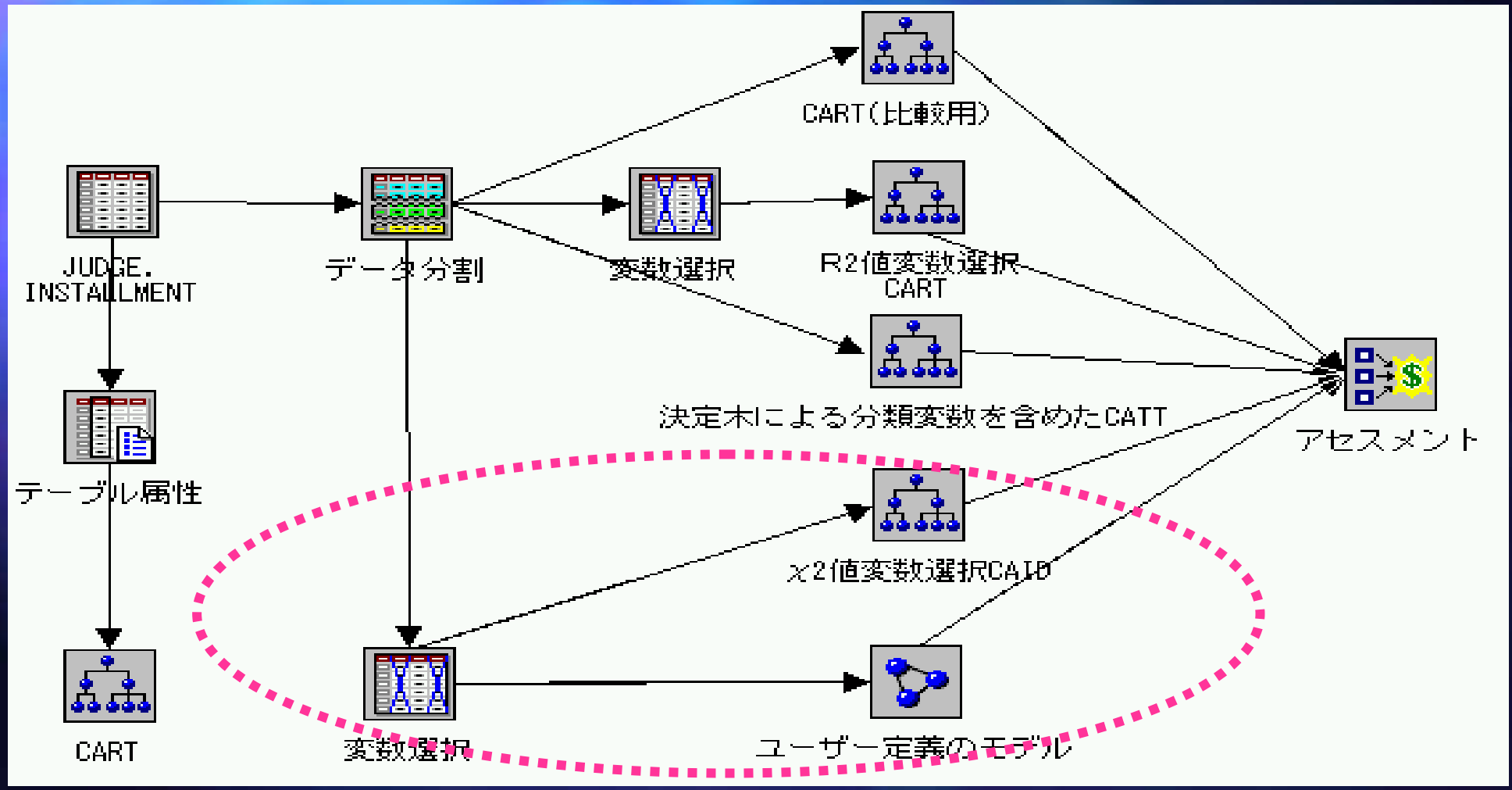
括弧内の数字は
Classの自由度＝(m-1)*(n-1)-1

R²値の影響度



実際は約５０個以上リストが続く

χ²値による変数選択法

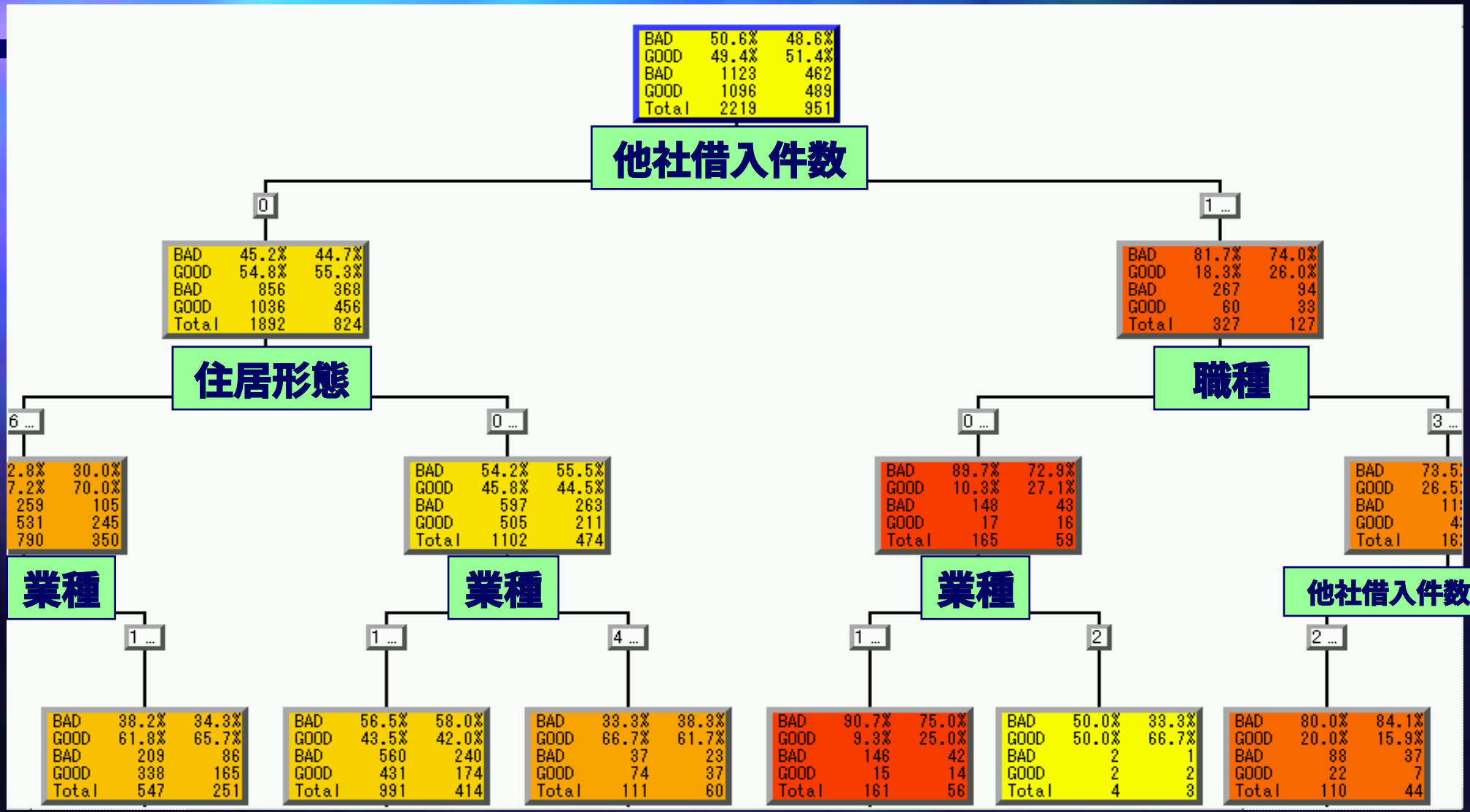


χ^2値による変数選択

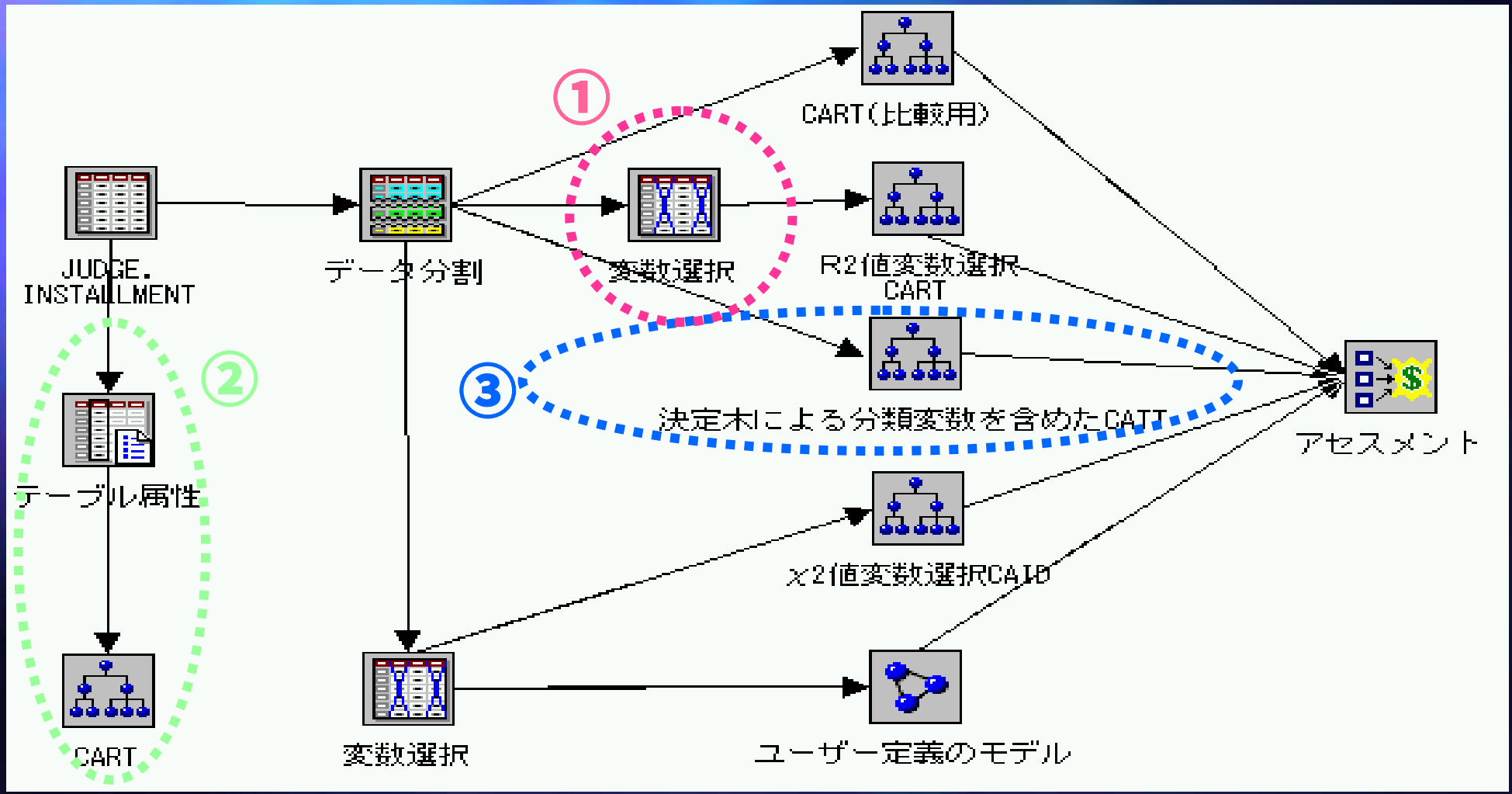
名前	役割	除外の理由	独立性	欠損値 (%)	水準の数	ラベル
SEX	rejected	Small chi-square		2%	2	性別
SALARY_CAT	rejected	欠損値 %		74%	8	年収
FLG_SALARY	rejected	Small chi-square		0%	2	年収記入フラグ
FLG_HIGH_SALARY	rejected	Small chi-square		0%	2	高年収フラグ
FLG_DEPT_OTHER	rejected	Small chi-square		0%	2	他社借フラグ
OCCUPATION	input			0%	7	職種
HOUSE	input			0%	8	住居形態
SUM_CAT	input			0%	8	契約金カテゴリ
AGE_CAT	input			0%	9	年代
DEPT_OTHER	input			0%	6	他社借数
DEPT_REMAINDER_	rejected	Small chi-square		0%	9	他社借金カテゴリ
INDUSTRY	input			0%	6	業種

SAS/EMが自動的に
採用変数を判定

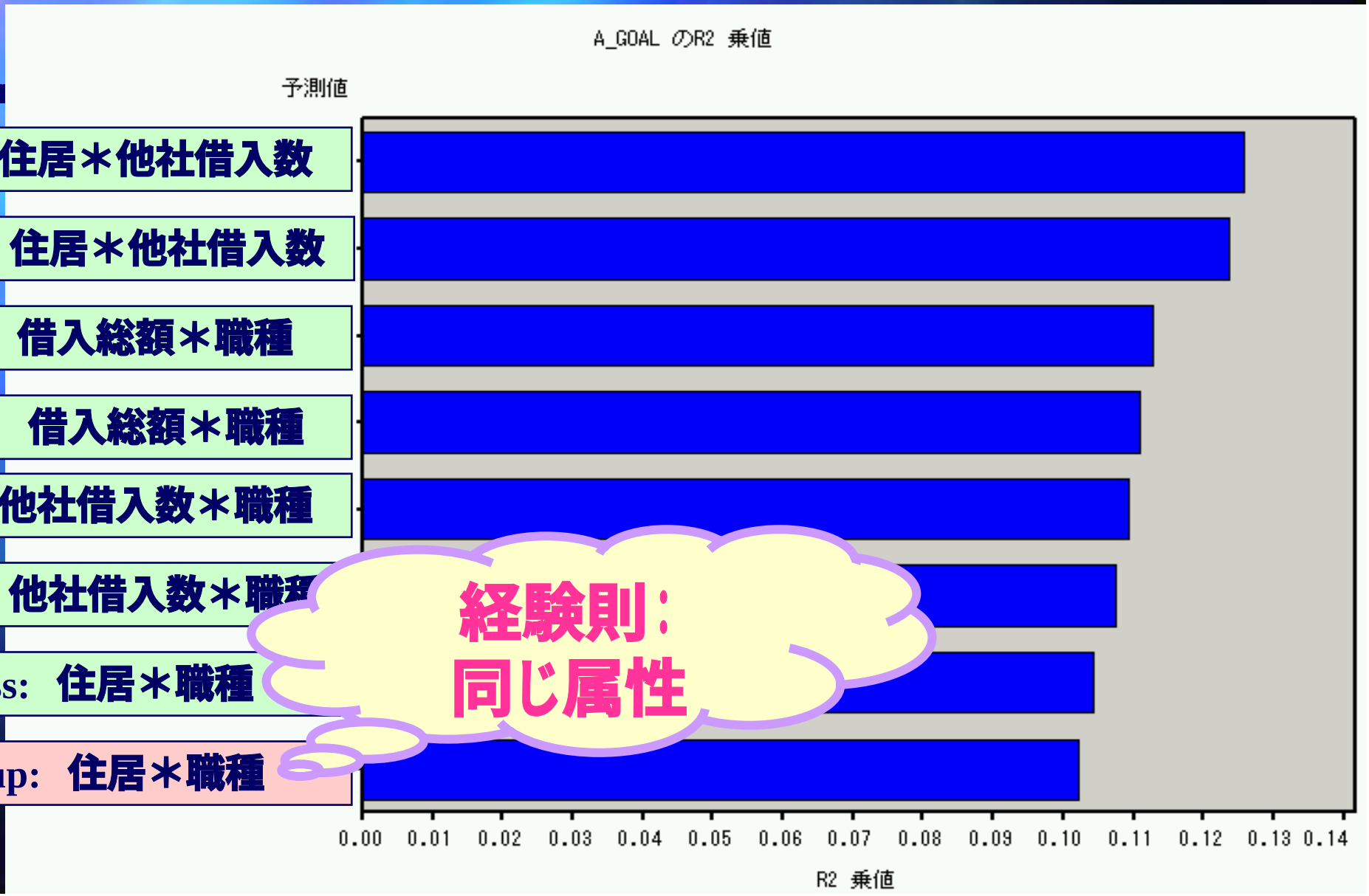
χ^2値による決定木



決定木による分類変数作成



① R²値の影響度

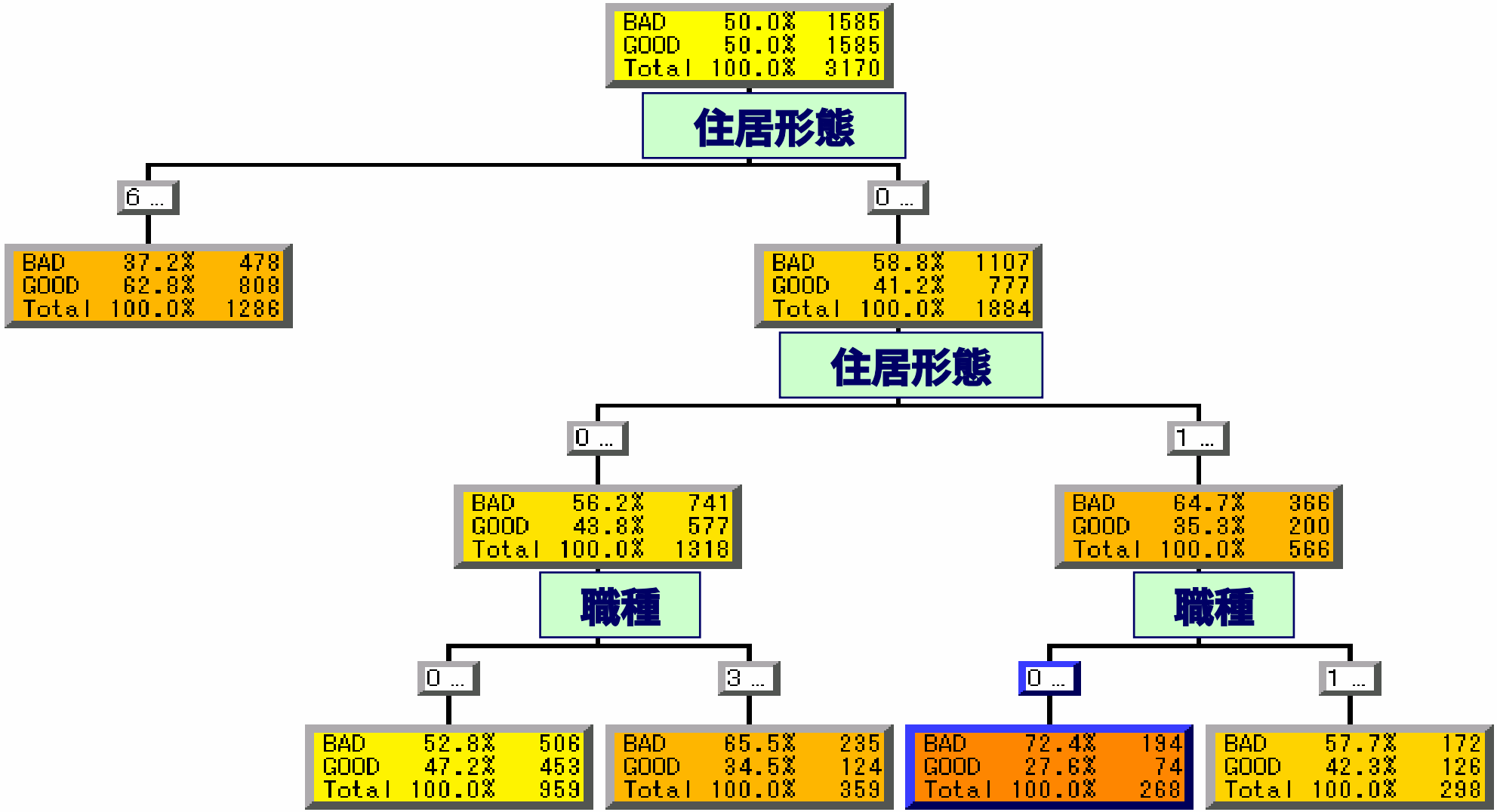


経験則：
同じ属性

② 決定木による分類変数作成時の入力変数

名前	キープ	役割	新しい役割	新しい測定水準	
A_GOAL	Yes	target	target	binary	
TIMES_CAT	Yes	rejecte	rejected	ordinal	回数ﾗﾝｸ
HOUSE_OCCU	Yes	rejecte	rejected	nominal	居住&職種
OCCUPATION	Yes	input	input	nominal	職種
HOUSE	Yes	input	input	nominal	住居形態
SALARY_CAT	Yes	input	rejected	ordinal	年収
SUM_CAT	Yes	input	rejected	ordinal	契約金ｶﾃｺﾞﾘ
AGE_CAT	Yes	input	rejected	ordinal	年代
DEPT_OTHER	Yes	input	rejected	ordinal	他社借数
DEPT_REMAINDER_CAT	Yes	input	rejected	ordinal	他社借金ｶﾃｺﾞﾘ
INDUSTRY	Yes	input	rejected	nominal	業種
SEX	Yes	input	rejected	binary	性別
FLG_SALARY	Yes	input	rejected	binary	年収記入ﾌﾗｸﾞ
FLG_HIGH_SALARY	Yes	input	rejected	binary	高年収ﾌﾗｸﾞ
FLG_DEPT_OTHER	Yes	input	rejected	binary	他社貸ﾌﾗｸﾞ

②_2　住居形態と職種の変数のみの決定木



SAS/EMによるプログラミング

- 決定木のノード結果のスコアを利用する法
  - if 職種 in (1,2,5) then node=3
  - and if 住居形態 in (0,4) then node=4
  - and if .....;
- 決定木のルール(if-then)を利用する法
  - If 職種 in(1,2,5) then node=3;
  - If 職種 in (1,2,5) and 住居形態 in(0,4) then node=4;
- (注) 例示のため、次頁とは連動していない

「決定木ノードのスコア」を利用する方法
(ネスト状のためわかりずらい)

```
IF _FNORVAL IN ('1' ,'2' ,'5' ) THEN DO;
      _FORMAT = PUT( OCCUPATION , BEST12.); %DMNORMCP( _FORMAT, _FNORVAL);
          IF _FNORVAL IN ('0' ,'4' ,'5' ) THEN DO;
             _NODE_  = 14;P_  GOALBAD  = 0.72388059701492;  P_A_GOALGOOD = 0.27611940298507;
                          I_A_GOAL  = 'BAD' ;U_A_GOAL  = 'BAD' ;_DECNUM = 1;
ELSE DO;
      _FORMAT = PUT( OCCUPATION , BEST12.); %DMNORMCP( _FORMAT
……(以下省略)……
```

分類変数のコード

「ルールの保存」を利用する方法
(プログラムの書き換え必要)

```
IF  住居形態 IS ONE OF: 6 7 AND 職種 IS ONE OF: 0 2 4 1   THEN  NODE:2   N:1286  BAD:37.2%  GOOD :62.8%
IF  職種 IS ONE OF: 3 5 6   AND 住居形態 IS ONE OF: 0 3 4 THEN  NODE:13  N:359   BAD:65.5%  GOOD:34.5%
IF  職種 IS ONE OF: 0 5 AND 住居形態 IS ONE OF: 1 2 5 THEN  NODE:14  N:268  BAD:72.4%  GOOD:27.6%
……(以下省略)……
```

上記をSASコードに直せば、入り子状でないルールが得られ、わかりやすい形式になる。

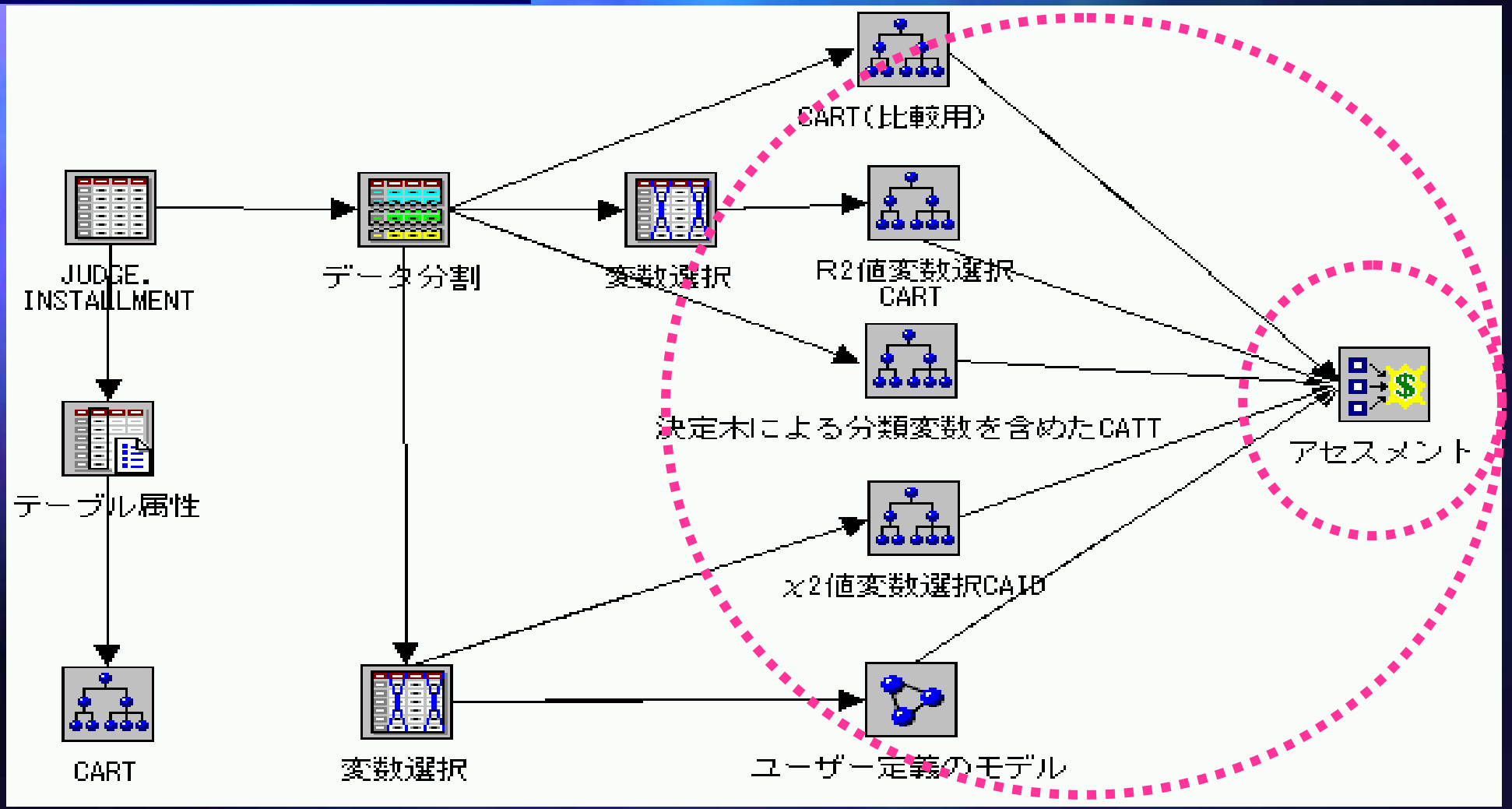
```
   IF HOUSE IN(0 4 1 3 2 5) AND OCCU IN(4 1)  THEN HOUSE_OCCU='AD041325_OC41';
ELSE IF HOUSE IN(6 7) AND OCCU IN(0 2 4 1)    THEN HOUSE_OCCU='AD67_OC0241';
ELSE IF HOUSE=7  AND OCCU IN(3 6 5)           THEN HOUSE_OCCU='AD7_OC365';
ELSE IF HOUSE=6  AND OCCU IN(3 6 5)           THEN HOUSE_OCCU='AD6_OC365';
ELSE IF HOUSE IN(0 4 3) AND OCCU IN(0 2)      THEN HOUSE_OCCU='AD043_OC02';
ELSE IF HOUSE IN(0 4 3) AND OCCU IN(3 6 5)    THEN HOUSE_OCCU='AD043_OC365';
ELSE IF HOUSE IN(1 2 5) AND OCCU IN(0 5)      THEN HOUSE_OCCU='AD125_OC05';
ELSE IF HOUSE IN(1 2 5) AND  OCCU IN(3 6 2)   THEN HOUSE_OCCU='AD125_OC362';
   LABEL HOUSE_OCCU='居住&職種';
```

決定木による分類変数を含めた入力ノード

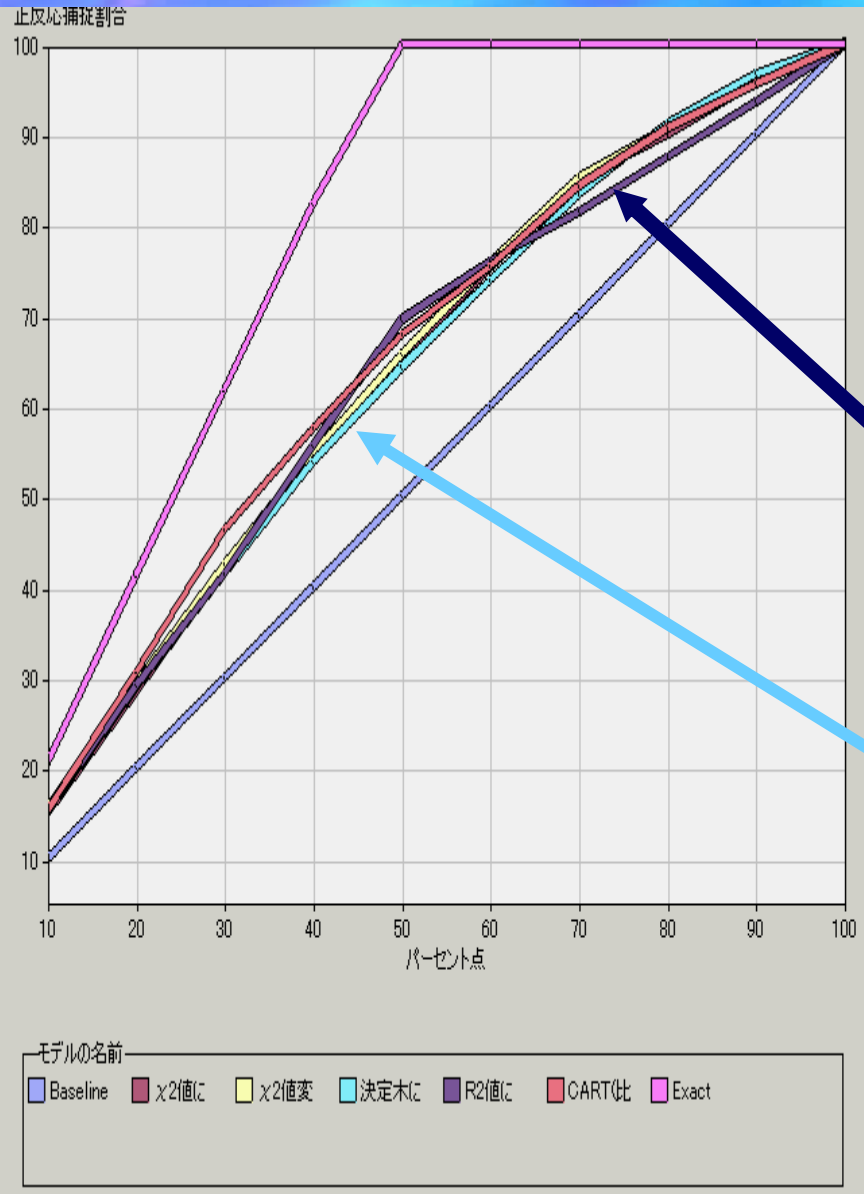
名前	状態	役割	測定水準		
A_GOAL	use	target	binary		
TIMES_CAT	don't use	rejected	ordinal		回数ﾗﾝｸ
HOUSE_OCCU	use	rejected	nominal		居住&職種
SUM_CAT	use	input	nominal		契約金ｶﾃｺﾞﾘ
AGE_CAT	use	input	nominal		年代
DEPT_REMAINDER_CAT	use	input	nominal		他社借金ｶﾃｺﾞﾘ
INDUSTRY	use	input	nominal		業種
SALARY_CAT	use	input	nominal		年収
DEPT_OTHER	use	input	nominal		他社借数
OCCUPATION	don't use	input	nominal		職種
HOUSE	don't use	input	nominal		住居形態
SEX	use	input	binary		性別
FLG_SALARY	use	input	binary		年収記入ﾌﾗｸﾞ
FLG_HIGH_SALARY	use	input	binary		高年収ﾌﾗｸﾞ
FLG_DEPT_OTHER	use	input	binary		他社貸ﾌﾗｸﾞ

モデル比較

**分類変数はどのような
モデルでも使用可能**



モデル精度の比較



ツール	名前	説明	Root ASE	Valid:Root ASE
User Defined	χ2値による分類変数	User Defined		
Tree	χ2値変数選択	Tree	0.4463611267	0.4714527138
Tree	R2値による分類変数	Tree	0.4481928424	0.4645239832
Tree	CART(比較用)	Tree	0.4539123068	0.4614434824
Tree	決定木による分類変数	Tree	0.4599742563	0.4717525476

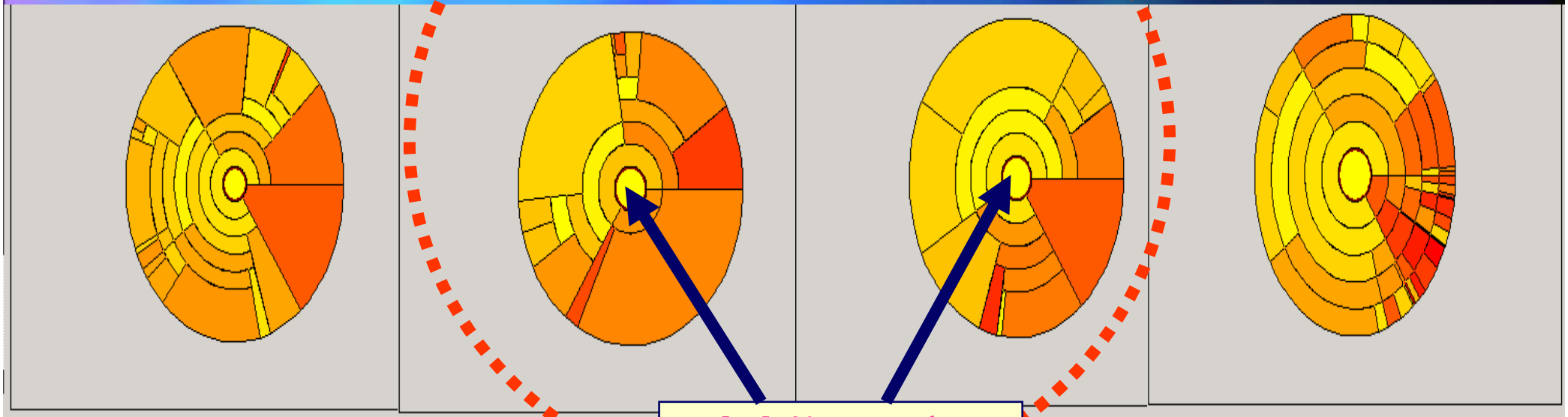
決定木の構造比較

NORMAL

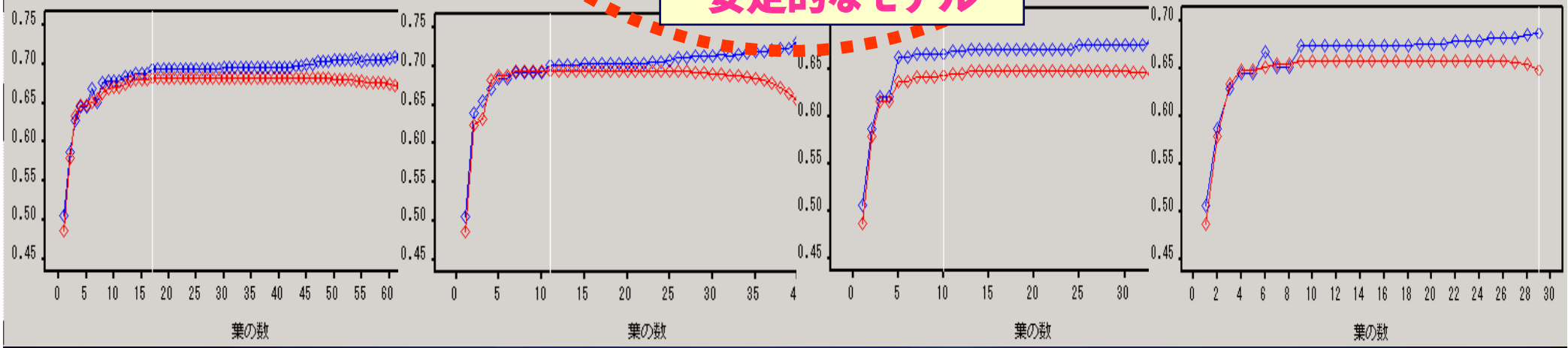
R^2値に基づく

決定木に基づく

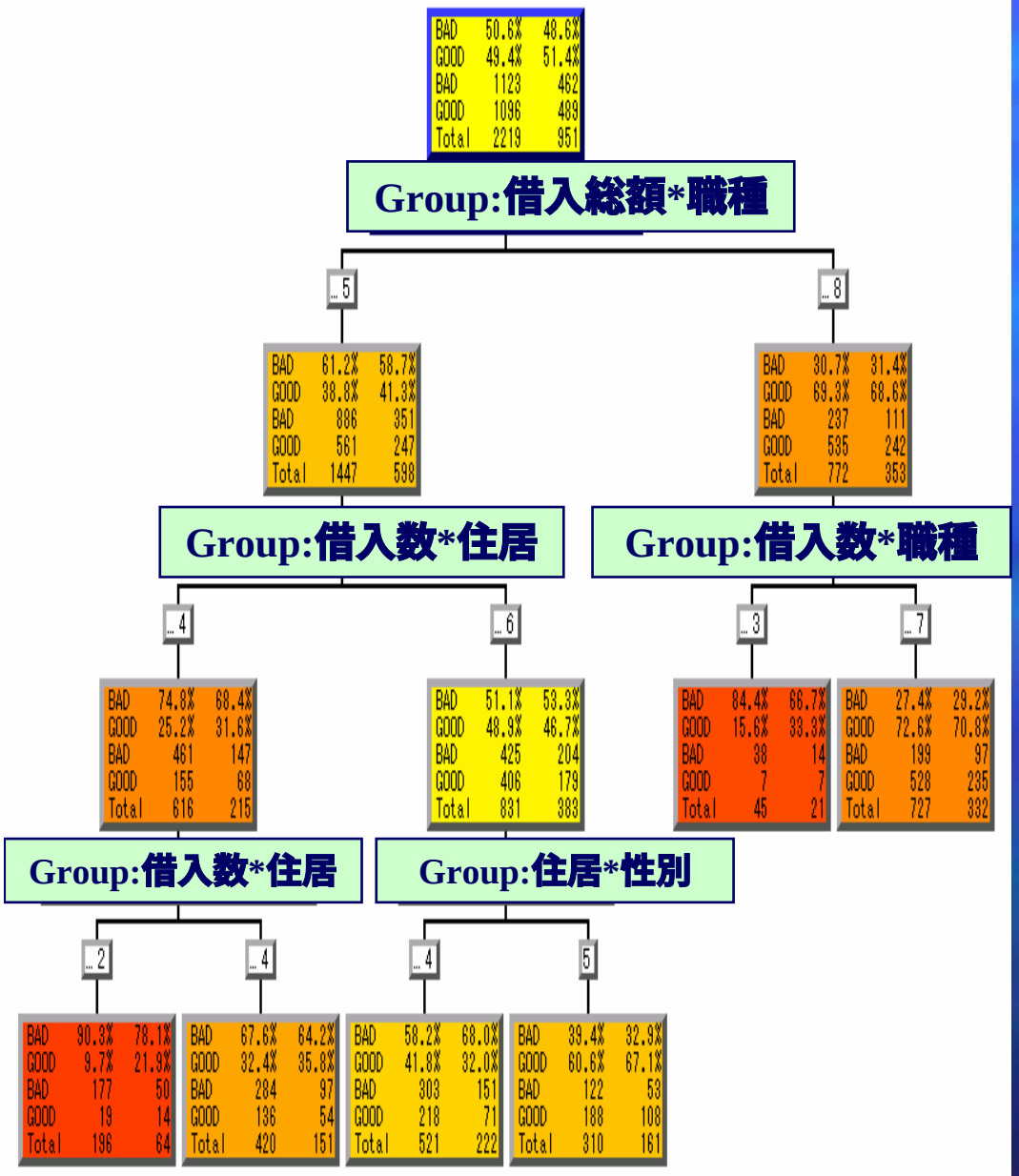
χ^2値に基づく



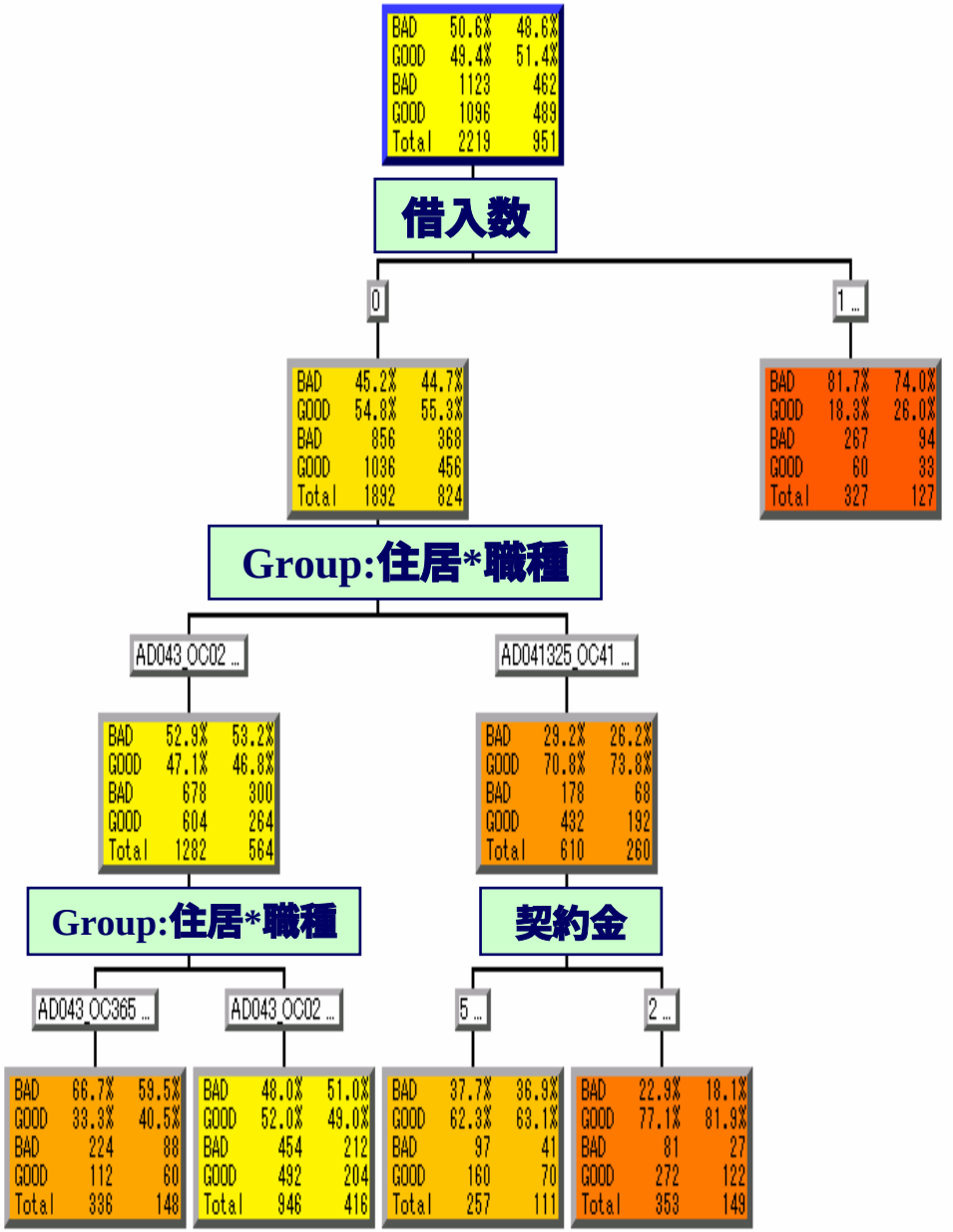
安定的なモデル



R²値による分類変数採用



決定木による分類変数採用



結　論

- 分類変数とモデルの安定性
- 組み合わせ分類変数
  - R^2値による分類変数作成（自動的）
  - X^2値による分類変数作成（自動的）
  - 決定木による分類変数作成（半自動的）
- SAS/EMによる簡易プログラム作成
- 将来：アドホックな知識発見への糸口


Question & Answer