Chapter 1	モデリングの概要	1-1
1.1.1	モデリングの概要	
1.1.2	2 教師あり学習の概要	
1.1.3	アソシエーションの概要	
1.1.4	セグメンテーションの概要	1-10
1.1.5	う すべてのモデルの概要	
1.2.1	モデルの検証	1-14

# 

2.1.1	ディシジョンツリーの概要	2-2
2.1.2	ディシジョンツリーの手法	2-3
2.1.3	CHAIDの概要	2-5
2.1.4	C&R Treeの概要	2-6
2.1.5	QUESTの概要	2-7
2.1.6	C5.0の概要	2-8
2.2.1	ストリームの確認	2-9
2.3.1	C&R Tree によるモデル作成の実行	2-13
2.3.2	C&R Tree により生成されたモデルの解釈	2-23
2.3.3	予測値の確認	2-30
2.4.1	クロス集計表による予測精度の確認	2-32
2.4.2	クロス 集計表の 結果の 解釈 	2-35
2.4.3	精度分析ノードによる予測精度の確認	2-36
2.4.4	精度分析の結果の解釈	2-39
2.4.5	評価ノードによる予測精度の確認	2-41
2.4.6	評価ノードの結果の解釈	2-46

Chapter 3 オ-	ートメーション(自動数値)	3-1
3.1.1	自動数値モデルの概要	
3.1.2	ストリームの確認	3-3
3.2.1	オートメーションによる自動数値モデルの作成	3-7
3.3.1	オートメーションによる自動数値モデルの結果の解釈	3-16
3.4.1	モデルのアンサンブル	3-24
3.4.2	予測値の確認	

アソシエーションの概要......4-2 4.1.14.1.2 4.1.3 4.2.1 4.2.2 Aprioriによるアソシエーションモデルの実行(1)......4-13 4.3.1 4.3.2 Aprioriよるアソシエーションルールの結果の解釈(1)......4-18 4.4.1 Aprioriによるアソシエーションモデルの実行(2)......4-27 442 Aprioriよるアソシエーションルールの結果の解釈(2)......4-29 アソシエーションルールのフィルタリング......4-32 4.4.3 4.5.1 アソシエーションルールによる予測......4-33

5.1.15.1.2 K-Meansの概要..... ...... 5.1.3 5.1.4 5.2.15.2.2 5.3.1 5.3.2 5.4.1レコードの所属クラスターの確認......5-27 属性フィールドと所属クラスターの分析......5-29 5.4.2 5.4.3 5.5.1 5.5.2

 スコアリングとエクスポート	Chapter 6 ス
 スコアリングの概要とストリーム	6.1.1
 新規データセットの設定	6.1.2
 新規データへのスコアリング	6.2.1
 エクスポートの概要	6.3.1
 Microsoft Excelへのエクスポート	6.3.2





本書では、IBM SPSS Modeler 18.2Jを使用しています。 IBM およびSPSSは、International Business Machines Corp.の登録商標です。

本書を無断で複写複製(コピー)することは、著作権法上の例外を除き、禁じられています。

# 3 オートメーション(自動数値)

IBM SPSS Modelerでは、モデル作成のためのオートメーションノードが含まれています。オート メーションでは、目的や用途に応じた複数のモデリング手法を同時に実行させて、精度の良いもの から順番に表示して評価することができます。オートメーションの予測モデルとして、カテゴリ型 フィールドを対象とする自動分類ノードと、連続型フィールドを対象とする自動数値ノードがあり ます。

## Contents



# §3.1.1 自動数値モデルの概要

予測のためのモデリングでは、カテゴリ型フィールドを対象とする場合と、連続型フィー ルドを対象とする場合があります。連続型フィールドを対象とするモデリング手法は、個別 にそれぞれのノードを用いるか、オートメーションの自動数値モデルで実行することができ ます。自動数値モデルによって指定できるのは、以下の手法です。

	線型回帰	線型回帰分析を実行するノードです。対象フィールド は連続型です。
	1次	線型モデルを作成するためのノードです。
	一般化線型	リンク関数(接続関数)を使用することで、さまざまな 分布のデータを扱う一般化線型モデルを作成します。
(PP)	Random Trees	入力フィールドのランダムサンプリングを行って多 数のツリーを構築し、予測精度を向上させます。
	KNN	類似性に基づいてレコードを分類して予測を行う Nearest Neighbor Modelsを作成します。
CHALD	CHAID	統計的仮説検定に基づくディシジョンツリーの手法 ズす。多分岐のツリーが構成されます。
A CRT	C&R Tree	不純度に基づくディシジョンツリーの手法です。常に 2分岐のツリーが構成されます。
	ニューラルネット	入力層と出力層の間に隠れ層を持つ多層型ニューラ ルネットワークモデルを構築するノードです。

Table3.1.1 連続型フィールドを対象とする主要なモデリングノード

#### POINT

**自動数値**ノードを使用すると、連続型フィールドを対象とするモデリング手法を、さまざ まなパラメーター設定で複数実行して、結果の精度の良いモデルを採用することができま す。

## §3.1.2 ストリームの確認

オートメーションの自動数値によるモデリングの例として、**Chapter3.str**をIBM SPSS Modelerに読み込みます。このストリームには、データソースとしてカンマ区切りの analysis1.csvがセットされています。顧客の属性や利用サービスに関するデータが記録され ています。フィールド(変数)は16個、8,632行のデータが含まれています。

ここでは、顧客の支払金額を予測するモデル作成の例で、実行手順を確認します。

操作手順

- 1. ファイルメニュー > ストリームを開くを選択します。
- 2. / ModelerC / Chpter3.strを選択します。



Figure3.1.1 Chapter3.strファイルのストリーム

3. ストリームのデータ区分ノードをダブルクリックして編集画面を表示します。

	×
◆ 生成(G)     ● ブレビュー(P)	ଡ □ □
設定 注釈	
データ区分フィールド: データ区分	
データ区分: 💿 学習とテスト(I) 🔵 学習、テスト、検証	$(\underline{V})$
学習データ区分のサイズ: 50 🗢 ラベル: 学習	値 =  "1_学習"
テストデータ区分のサイズ: 50 🗢 ラベル: テスト	値 = 『2_テスト"
検証データ区分のサイズ: 0 🗢 ラベル: 検証	<b>値 = </b> "3_検証"
合計サイズ: 100%	$\mathbf{O}$
値: 🔵 システム定義の値 ("1"、"2" および "3") を[	使用
● ラベルをシステム定義の値の後に結合する	
○ ラベルを値として使用	
✓ ランダム シードの設定	
シード: 1234567 🗘 生成	
□ 一意のフィールドを使用してデータ区分を割り当てる:	-
OK キャンセル	適用( <u>A</u> ) リセット( <u>R</u> )

Figure3.1.2 データ区分ノード

**データ区分**ノードは、レコードをランダムに**学習**データ区分と**テスト**データ区分に分割する 場合に使用します。デフォルトでは、学習データ区分に**50%**、テストデータ区分に**50%**確保さ れます。また、ランダムシードの設定が有効にすることで、乱数の値が固定されるため、スト リームを実行するごとに異なるレコードが割り当てられることを防ぐことができます。

#### TIPS

特にレコード数が少ない場合、学習データにはテストデータより多めのサンプルを割り当てます。学習用に70%、テスト用に30%の割合は比較的よく利用されます。

- 4. **OK**ボタンをクリックして、データ区分ノードの編集画面を閉じます。
- 5. ストリームのデータ型ノードをダブルクリックして編集画面を表示します。

🌶 データ型					>
デーク型 つい	ビュー(P)				0 - 0
<b>↓ ○</b> Ø	● 値の読み込み	値の	消去	すべての値	の消去
フィールドー	尺度	値	欠損値	検査	ロール
◇ 顧客番号	☞ 連続型	[47,84042		なし	⊘なし
◇ 年齢	☞ 連続型	[18,63]		なし	▲ 入力
A 性別	🖁 フラグ型	M/F		なし	<b>入</b> 力
▲ 婚姻状況	🕈 フラグ型	未婚/既婚	-	なし	<u>کر ک</u>
A 支払方法	♣ 名義型	カード,ポ		なし	◎ なし
◇ 支払金額	☞ 連続型	[214,30363]		なし	
◇ 契約A	🕈 フラグ型	1/0		なし	🔉 እ. ታ
契約B	🕈 フラグ型	1/0		なし	🔪 እ.
◇ 契約C	🕈 フラグ型	1/0		なし	🔪 ኢታ
◇ サービスA	🕈 フラグ型	1/0		なし	🔪 ኢታ
◇ サービスB	🖁 フラグ型 💋	1/0		なし	🔪 ኢታ
Ø サービスC	🖁 フラグ型	1/0		なし	🔪 ኢታ
◇ 家の所有	🖁 フラグ型	1/0		なし	🔪 ኢታ
◇ 車の所有	🖁 フラグ型	1/0		なし	🔪 ኢታ
🚫 顧客紹介	🖁 フラグ型	1/0		なし	🔪 ኢታ
A 申込方法	🧕 フラグ型	書面/Web		なし	🔪 ኢታ
A 新規契約	8 フラグ型	契約あり <mark>/</mark> …		なし	⊘ なし
🔺 データ区分 📃	▲ 名義型	"1_学習","		なし	📲 データ区分
● 現在のフィーノ	レザを表示 〇 未使用の	のフィールド副	設定を表示		
ОК	キャンセル			適用( <u>A</u> )	リセット( <u>R</u> )

Figure3.1.3 データ型ノード

**データ型**ノードで、フィールドの**尺度**と**ロール**を設定します。この例では、**支払金額**を予測 の目的フィールドとして**対象**に割り当てています。

**顧客番号と支払方法、新規契約**の3つはモデリングには使用しません。

また、**データ区分**のフィールドによって、学習データ区分でモデル作成、テストデータ区分 でモデルのテストが自動的に行えるようになっています。

#### POINT

予測に寄与しないフィールドや、使用することに意味のないフィールドなどはそのロール を**なし**に設定しておきます。

#### TIPS

入力フィールドや対象フィールドの指定は、各モデリングノードでも設定することが可能 ですが、データ型ノードのロールとして設定しておくと、モデリング手法を変える度にフ ィールドの指定を行う必要がなくなるため効率的です。

#### 操作手順



# §3.2.1 オートメーションによる自動数値モデル作成の実行

この例では、顧客の**支払金額**の予測を行う例として、オートメーションの**自動数値**ノード を利用したモデル作成の手順を確認します。

操作手順

1. モデル作成パレットのすべてサブパレットをクリックします。



Figure3.2.2 ストリームキャンバスに配置した自動数値ノード

ストリームキャンバスに挿入したノードは、自動的に対象フィールドの名称になります。こ の例では、**支払金額**になっています。

入力フィールドと対象フィールドの設定は、データ型ノードで完了しているため、このまま でもモデリングを実行することができますが、実行前に重要なパラメーターを確認しておきま す。

- 3. 自動数値ノードをダブルクリックして編集画面を開きます。
- 4. モデルタブを開きます。



Figure3.2.3 自動数値ノードのモデルタブ

**モデル**タブでは、実行したモデルの選択の基準を設定します。**データ区分データを使用**が有効になっており、上流のデータ区分ノードの結果を受けて、学習用データとテストデータに区分した処理が行われます。

**モデルのランク付け**は、複数のモデルの精度を評価する基準の選択です。デフォルトでは、 相関(実測値と予測値の相関)が選択されています。その他には、使用フィールド数、相対誤差 が選択可能です。使用モデル数は、上位いくつまでのモデルを採用するかの設定であり、デフ ォルトでは3個です。

この例では、デフォルトのまま進めます。

- 5. **エキスパート**タブを開きます。
- 6. モデルの選択ドロップダウンリストからModelerで実行(目的を個別に指定)を選択し ます。

🛞 支払金額	X
実行されるモデルの推定数:13 フィールド モデル エキスパート ≣	<b>⑦ □ □</b> 股定 注釈
モデルの選択:	Modeler で実行 (目的を個別に指定) 🖌
使用? モデルタイプ	モデル バラメータ モデル数
🗹 🐹 線型回帰	デフォルト 1
☑ 🐹 一般化線型	デフォルト 1
□ 🌋 一般化線型エンジン	デフォルト 1
<ul> <li>KNN アルゴリズム</li> </ul>	デフォル
Linear-AS	デンォルト 1
LSVM	デフォルト 1
Random Trees	デフォルト 1
<ul> <li>□ 単一モデルの構築に費やされる最大時間の制</li> <li>停止規則</li> </ul>	限 15 😴 分
OK ▶ 実行( <u>U</u> ) + +	- ンセル 適用( <u>A</u> ) リセット( <u>R</u> )

Figure3.2.4 自動数値ノードのエキスパートタブ

**エキスパート**タブでは、実行するモデリング手法の選択と各手法のパラメーター設定を行う ことができます。デフォルトでは、線型回帰、一般化線型、C&R Tree、CHAID、線型、ニュ ーラルネットが選択されており、それぞれ1つのモデルがデフォルトのパラメーター設定によ って作成されます。数パターンのパラメーターを設定することによって、同じモデリング手法 でも複数のモデル作成を行えるようになっています。

ここでは、線型回帰を例にパラメーター設定の変更手順を確認してみます。

- 7. 線型回帰のモデルパラメータのセルをクリックします。
- 8. 指定を選択します。

🛞 支払金額			×
です。 フィールド	されるモデルの推定数: 13 • モデル エキスパート	設定 注釈	❷ □ □
モデルの選択	l.	Modeler で実行 (目的を個別	に指定) 👻
使用?	モデルタイプ	モデル パラメータ モデル養	t
✓	線型回帰	デフォルト 👻 1	
<b>~</b>	😥 一般化線型	デフォルト 指定 1	
	🌿 一般化線型エンジン	デフォルト 1	
	KNN アルゴリズム	デフォルト 1	
	Linear-AS	デフォルト 1	
<b>V</b>	LSVM	デフォルト 1	
	Random Treès	デフォルト 1	
<ul> <li>□ 単一モデルの構築に費やされる最大時間の制限</li> <li>15 ♀ 分</li> <li>停止規則</li> </ul>			
ОК	▶ 実行(世) キ・	▶ンセル 適用( <u>A</u> )	リセット( <u>R</u> )

Figure3.2.5 線型回帰のモデルパラメータの指定

#### POINT

モデルタイプは、**モデルパラメータ**を個別に指定することができ、さまざまな設定による モデル作成を同時に試行することができます。



クラスタリングは、入力フィールドの類似したパターンを持つレコードをクラスター(セグメント)に分割するための分析手法です。入力フィールドのみに関心があるため、クラシフィケーションの手法と異なり対象フィールドの概念を持ちません。クラスタリングによって、例えば、既存の顧客のレコードを顧客セグメントに分類することができます。

## Contents



## Keyword

クラスタリング / クラスター分析 / K-Means / Two-Step / Kohonen / 異常値検査 / 自動クラスタリング / クラスターのプロフィール /

# §5.1.1 クラスタリングの概要

クラスタリングは、入力フィールドの類似したパターンを持つレコードをクラスター(セグ メント)に分割するための分析手法です。入力フィールドのみに関心があるため、クラシフィ ケーションの手法と異なり対象フィールドの概念を持ちません。クラスタリングによって、 既存の顧客のレコードを潜在的な顧客セグメントに分類することができます。

<b>P</b>	自動 クラスタリング	複数のクラスタリングモデリングを同時に試行して、 精度の高い結果を採用する場合に使用します。
K	K-Means	K個のクラスター数を指定することで実行される代表 的なクラスタリングのモデリングノードです。
	Two-Step	Two-Stepでは、2段階のクラスター化手法が用いられ、 最適なクラスター数を自動的に決定されます。
	Kohonen	ニューラルネットワークの手法で、ネットワークの学 習によってレコードを類似セグメントに分類します。
	異常値検査	Two-Stepの方法によりセグメント化を行い、正常なパ ターンに合致しないレコードを特定します。
	Table5.1.1 ク	ラスタリング(セグメンテーション)のノード

IBM SPSS Modelerでは、クラスタリングのためのノードとして、自動クラスタリング、 K-Means、Two-Step、Kohonen、異常値検査の5つがあり、モデル作成パレットに含まれて います。

#### POINT

**クラスタリング**は、多数の入力ノードに基づいて類似グループ(クラスター/セグメント) を識別することができる分析手法です。

# §5.3.2 K-Meansによるクラスタリングの結果の解釈

クラスタリングの実行が完了すると、ストリームにモデルナゲットが作成されます。この モデルに、クラスタリングの結果が含まれています。

#### 操作手順

1. K-Meansのモデルナゲットをダブルクリックして編集します。



Figure 5.3.7 生成されたK-Meansモデル

**K-Means**モデルビューアーが表示されます。左側のメインビュー領域にモデル要約が表示 されます。この例では、7個のフィールドに基づいて3個のクラスターに分類されています。ま た、クラスターの品質はクラスタリングの精度の評価の参考となります。 **クラスターの品質**は、レコードとクラスター中心の距離を計算しており、値が1に近いほど 所属するクラスターと最近隣クラスターとの差があることを意味し、精度が高いと評価します。 このスナップショットを使用して、クラスターの精度が悪いかどうかをすばやく確認できます。

#### TIPS

すべてのレコードに対するシルエット平均は、(B-A)/max(A,B)となります。Aはクラスタ ー中心へのレコードの距離、Bはレコードが属さない最近隣クラスター中心へのレコード の距離です。シルエット係数1は、すべてのケースはクラスター中心に直接配置されてい るということを意味します。値-1は、すべてのケースが他のクラスターのクラスター中心 にあることを意味します。平均の0の値は、ケースが自身のクラスター中心と、その他の 最近隣クラスターとの間で等距離にあることを意味します。

#### TIPS



2. 左側のメインビュー領域の下部にあるビューをクラスターに変更します。

クラスター	クラスター-3	クラスター-2	クラスター-1
ラベル			
説明			
サイズ	62.3%	19.4%	18.2%
	(5382)	(1678)	(1572)
入力	家の所有	2所本	家の所有
	0 (100.0%)	1 (101.0%)	0 (93.4%)
	<b>a</b>	契約A T (93.8%)	契約A 0 (100.0%)
	<b>契約</b> 時	契約B	契約B
	(100.0%)	1 (99.9%)	0 (50.2%)
	契約C	契約C	契約C
	1 (74.8%)	1 (89.7%)	0(79.4%)
	支払方法	支払方法	支払方法
	現金 (52.6%)	ポイント (48.7%)	現金 (50.5%)
	支払金額	支払金額	支払金額
	9,031.64	8,062.89	12,204.76
	車の所有	車の所有	車の所有
	0(87.6%)	0(83.1%)	0(79.7%)

#### クラスター

入力値 (予測値)の重要度 ■1.0■0.8■0.6■0.4■0.2■0.0

Figure5.3.8 クラスタービューのクラスター中心による表示

クラスタービューでは、各クラスターサイズとしてnと%、各入力フィールドの要約情報を 確認することができます。第1クラスターは1,572レコードで全体の18.2%、第2クラスターは 1,678レコードで全体の19.4%、第3クラスターは5,382レコードで全体の62.3%です。

#### TIPS

クラスターの表示順番の設定は、下部のツールバーに含まれる**クラスターをサイズでソート、クラスター名でソート、クラスターをラベルでソート**の3種類選択することができます。デフォルトではサイズでソートされています。

入力フィールドとして**家の所有**の特徴に注目すると、第3クラスターでは**家の所有=0**が **100%**であり、このクラスターで家の所有がない特徴を持ちます。一方、第2クラスターで は**家の所有=1**が**100%**であり、このクラスターは家の所有がある特徴を持ちます。家の所 有は、クラスターの構成に寄与しているようです。

また、**車の所有**の特徴に注目すると、3つのクラスターのいずれも**車の所有=0**が80%ほどであり、車の所有によるクラスターの違いはあまりなさそうです。このようなフィールドは、モデリングから除外しても良いかもしれません。

#### TIPS

入力フィールドの表示順番の設定は、下部のツールバーに含まれる入力値をクラスター内 の重要度でソート、入力値を名前でソート、入力値をデータでソートの3種類選択するこ とができます。デフォルトではクラスター内の重要度でソートされています。

3. 左側の領域の下部にあるツールバーからセルには絶対分布を表示をクリックします。



Figure 5.3.9 クラスタービューのツールバー

#%	セルには クラスター中心を表示	連続型フィールドでは平均値、カテゴリ型フィールド の最頻カテゴリの%が表示されます。	
	セルには 絶対分布を表示	フィールドの絶対分布を表示します。濃い赤はクラス ター分布、淡い赤は全体のデータを表示します。	
	セルには 相対分布を表示	フィールドの相対分布を表示します。濃い赤はクラス ター分布、淡い赤は全体のテータを表示します。	
	セルには 基本情報を表示	フィールド名のみを表示します。よりコンパクトな表示です。	
Table5.3.1 セルの内容ツールバーのボタン			

クラスター

入力値 (予測値)の重要度



Figure 5.3.10 クラスタービューの絶対分布による表示

絶対分布では、全体の分布とクラスターの分布を度数に基づいて表示します。視覚的に確認 することができるため、平均値やパーセンテージの表示とあわせて各クラスターの特徴を評価 します。

4. 左側の領域で、クラスターの名前をCTRLキーを使用してすべて選択します。



Figure5.3.12 カテゴリ型フィールドの比較の例

#### POINT

クラスターの比較では、カテゴリ型フィールドは最頻カテゴリが表示されます。

連続型フィールドの場合は箱ひげ図で表示され、全体の中央値と4分位範囲を示します。四 角形のポイントマーカーと水平線は、それぞれ各クラスターの中央値と4分位範囲を示します。

**支払金額**フィールドに注目すると、クラスター1は全体の中央値より大きく、クラスター2 とクラスター3では全体の中央値より低いことが分かります。



#### StatsGuild Inc.

これらの出力を参考にしながら、クラスターの解釈とプロフィール作成を行います。

#### 操作手順

5. **OK**ボタンをクリックして、モデルビューアーを閉じます。

#### POINT

クラスタリングは、多数の入力フィールドに基づいて、レコードを類似するグループに分類するための手法であり、分類されたクラスターが説明のつく結果になっているかどうか を評価するのは分析者です。この作業はプロフィール作成と呼ばれます。

