§2.3.2　C&R Treeにより作成されたモデルの解釈

C&R Treeで作成された新規契約の予測モデルの解釈を行います。モデル情報は、モデルナゲットに含まれています。

操作手順

1. ストリームキャンバス内の**新規契約**モデルナゲットをダブルクリックします。

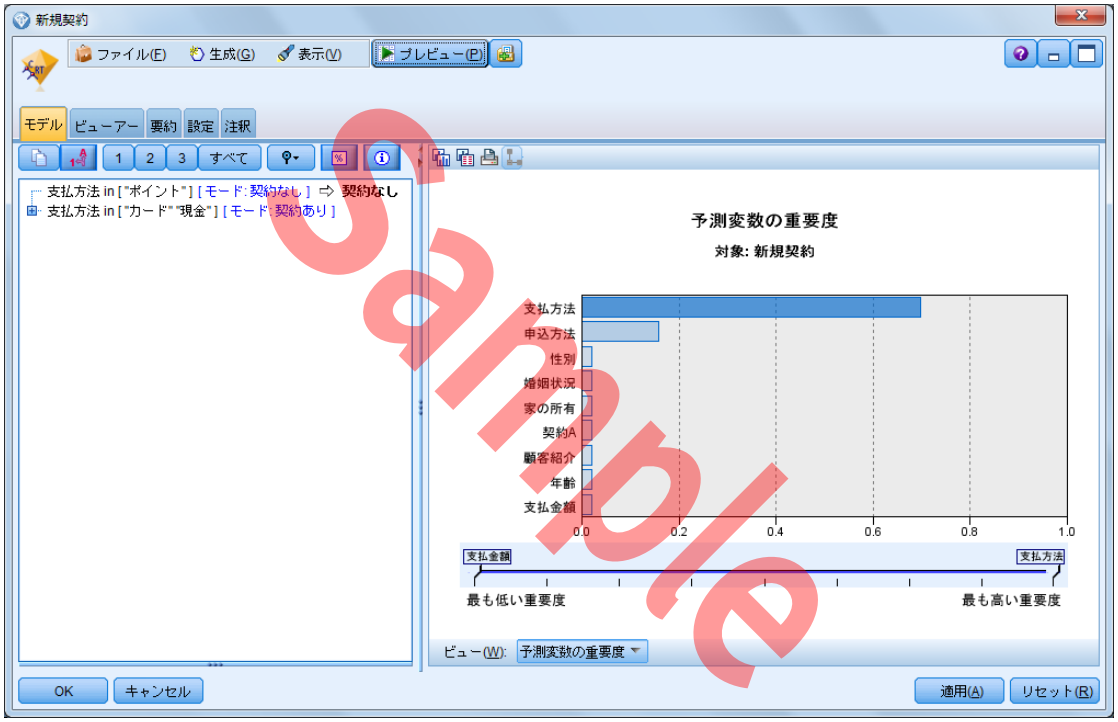

Figure2.3.11　新規契約モデルのモデルタブ

モデルタブでは、左側に予測ルールが表示され、右側に予測変数の重要度が表示されます。**予測変数の重要度**は、ディシジョンツリーの生成について、どの入力フィールドが寄与しているかを棒グラフによる視覚的な表現で確認できるものです。この例では、新規契約について支払方法が最も寄与していることが分かります。

操作手順

2. **ビューアー**タブをクリックします。

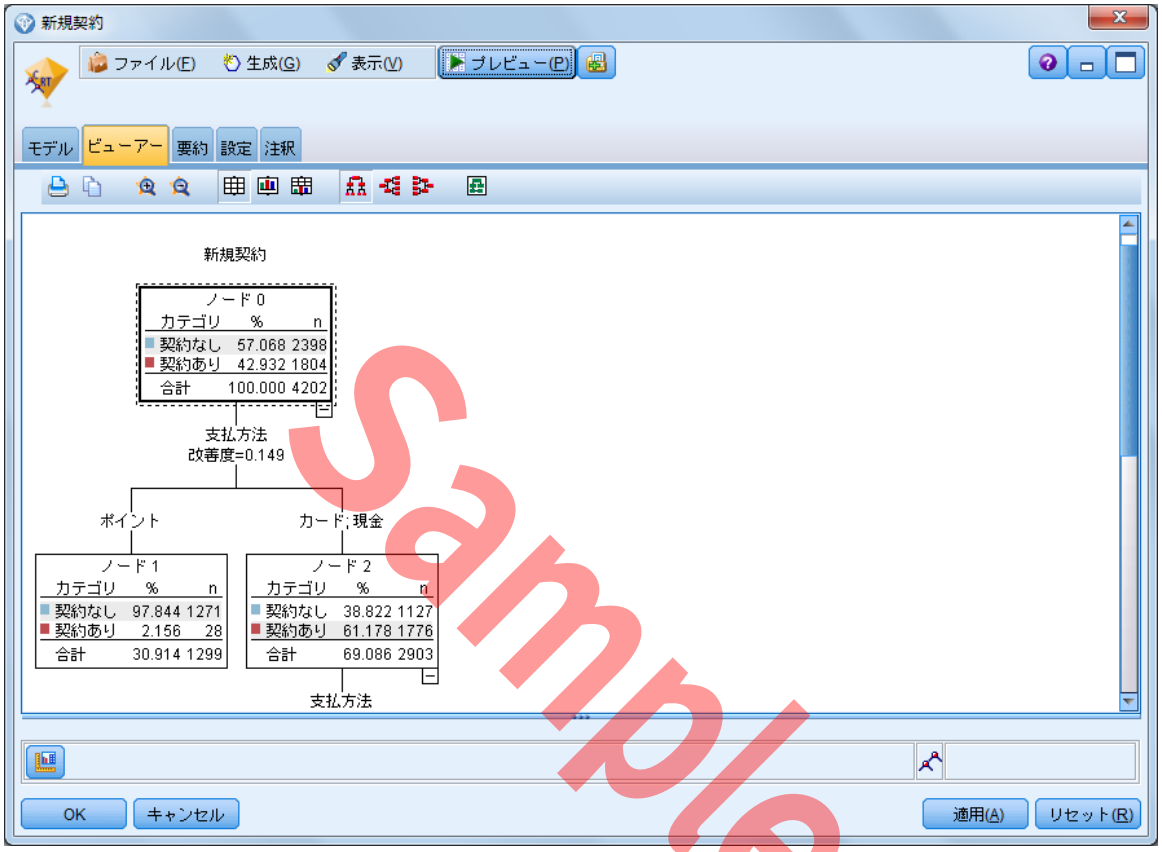

Figure2.3.12　新規契約モデルのビューアータブ

C&R Treeによって生成された**ツリー図**が表示されます。ツリー図の表示形式は、ツールバーのボタンを利用して、グラフ表示や横展開の表示などに切り替えることができます。

Figure2.3.13　ツリー図のツールバー

	印刷	ツリー図を印刷します。
	クリップボードにコピー	ツリー図をクリップボードにコピーします。
	拡大する	ツリー図を1段階拡大します。
	縮小する	ツリー図を1段階縮小します。
	度数情報を表で表示	ツリー図を表で表示します。
	度数情報をグラフで表示	ツリー図の表示をグラフ表示にします。
	度数情報をグラフと表で表示	ツリー図の表示をグラフと表の両方で表示します。
	上から下へ	ツリー図の表示を上から下へ表示させます。
	左から右へ	ツリー図の表示を左から右へ表示させます。
	左から右へ	ツリー図の表示を右から左へ表示させます。
	ツリーマップウィンドウを表示または隠す	ツリーマップウィンドウの表示を切り替えます。

Table2.3.1　ツリー図のツールバーのボタン

POINT

ツリー図の表示は、ツールバーのボタンを利用することで、拡大や縮小表示、グラフ表示やテーブル表示の切替やツリー図の表示形式の変更などを行うことができます。

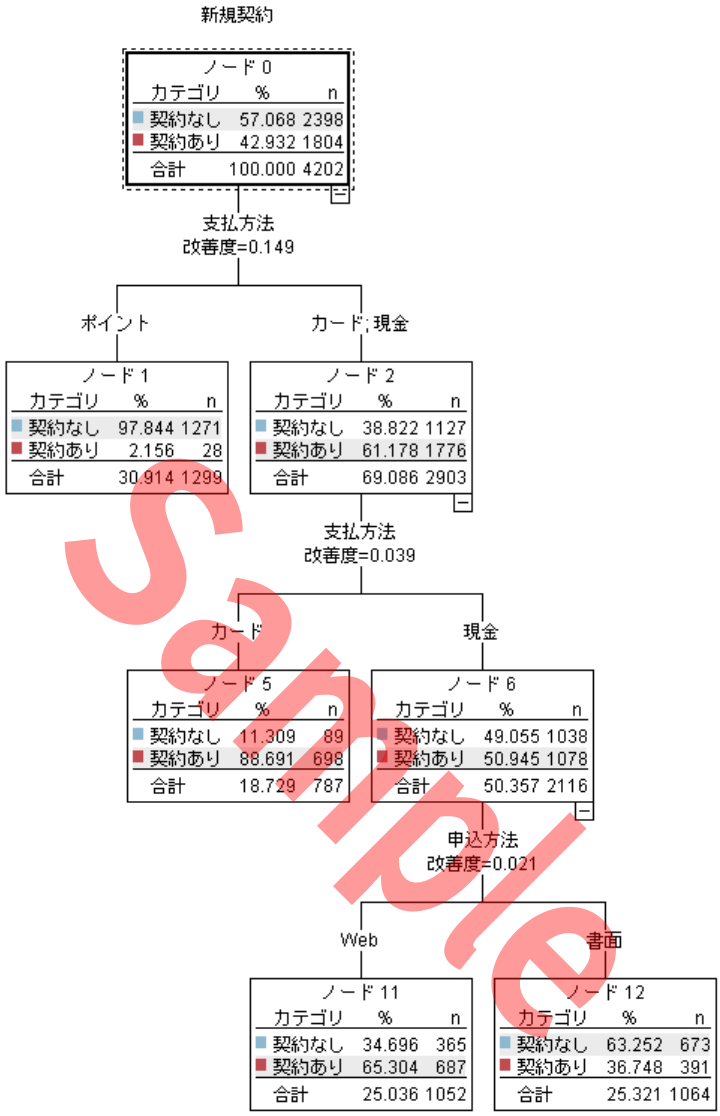

Figure2.3.13　ツリーの全体図

ツリー図のセグメントの1つ1つは**ノード**と呼ばれます。最初にあるのが対象フィールドを表しており**ルートノード**と呼ばれます。また、ツリーの分岐は親子に例えられ、上に位置するのが**親ノード**、分岐されたノードを**子ノード**と呼びます。子ノードを持たないノードは、**ターミナルノード**と呼ばれます。

まず、ルートノードに注目すると、新規**契約あり**は**42.932%**、**契約なし**は**57.068%**です。この例では、**支払方法**によって最初の分岐が行われていることが分かります。

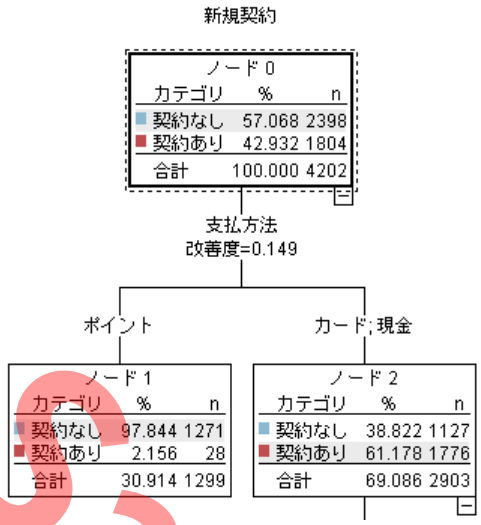

Figure2.3.14　新規契約を説明するツリー図の第1分岐

支払方法＝**ポイント**の場合、**契約なし**は**97.844%**です。ポイントによる支払を選択している顧客は、新規契約なしが非常に多いようです。

支払方法＝**カード**もしくは**現金**の場合、**契約あり**は**61.178%**です。カードもしくは現金による支払いを選択している顧客は、新規契約ありが多いようです。

このような違いを多数のフィールドの中から目で探すことは大変ですが、ディシジョンツリーを用いることで、比較的容易に発見することができます。

POINT

各ノードの最もパーセンテージの高いカテゴリは、網掛けで強調されます。

TIPS

改善度は、親ノードから子ノードに分岐した際の不純度の減少量です。改善度が大きいものから分岐のフィールドとして選択されます。

次に、カードもしくは現金による支払方法を選択しているノード(ノード2)に注目すると、再度、**支払方法**によって分岐されていることが分かります。

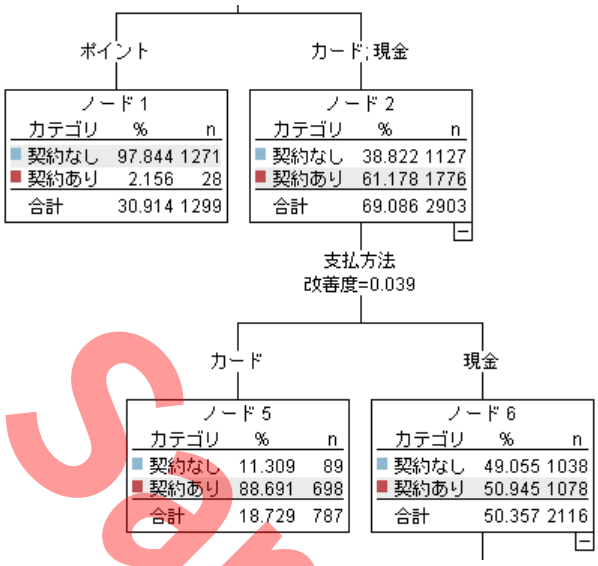

Figure2.3.15　新規契約を説明するツリー図の第2分岐

支払方法＝**カード**の場合、**契約あり**は**88.691%**です。カードによる支払を選択している顧客は、新規契約ありが非常に多いようです。

支払方法＝**現金**の場合、**契約あり**は**50.945%**です。現金による支払いを選択している顧客は、新規契約ありと新規契約なしでほとんど差がないようです。

POINT

支払方法が**ポイント**のノードは子ノードがないため**ターミナルノード**になります。予測ルールはターミナルノードから導かれるため、**支払方法＝ポイント**の顧客は新規**契約なし**と予測されることになります。その際の確信度は**97.844%**です。

次に、現金による支払方法を選択しているノード(ノード6)に注目すると、**申込方法**によって分岐されていることが分かります。

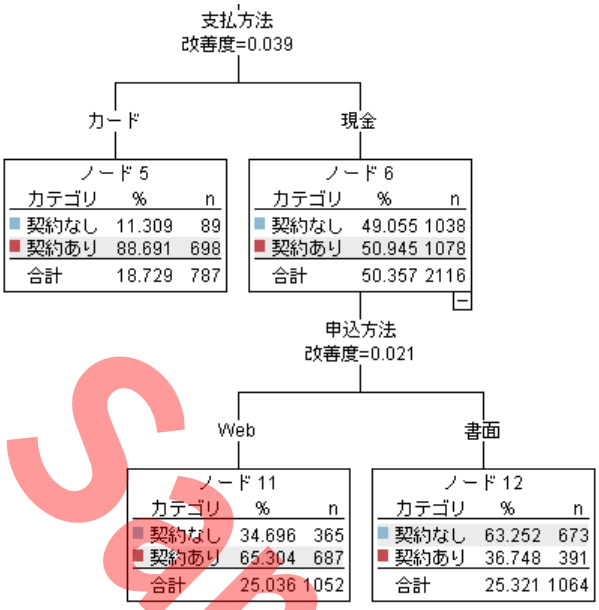

Figure2.3.16　新規契約を説明するツリー図の第3分岐

申込方法＝**Web**の場合、**契約あり**は**65.304%**です。支払方法が現金の場合でもWebによる申込みをした顧客は、新規契約ありが多いようです。

申込方法＝**書面**の場合、**契約なし**は**63.252%**です。書面による申込みをした顧客は、新規契約なしの方が多いようです。

TIPS

この例では含まれていませんが、分岐フィールドとして**連続型**フィールドが選ばれた場合は、2分岐のためのすべての分割点が探索され、最も差を認める値によって分割されます。

操作手順

3. **OK**ボタンをクリックして、モデルビューアーを閉じます。

§2.4.3　予測値の確認

モデルナゲットには、ツリー図とターミナルノードから生成されている予測ルールの情報が含まれています。ストリームでモデルナゲットを通過したデータには、予測値やその確信度が含まれることになります。ここでは、C&R Treeによる**予測値**と**確信度**をテーブルノードを用いて確認してみます。

操作手順

1. ストリームキャンバスの**新規契約**モデルナゲットを選択しておきます。
2. **出力**パレットの**テーブル**ノードをストリームキャンバスに挿入します。
3. **テーブル**ノードを**新規契約**モデルナゲットからリンクします。

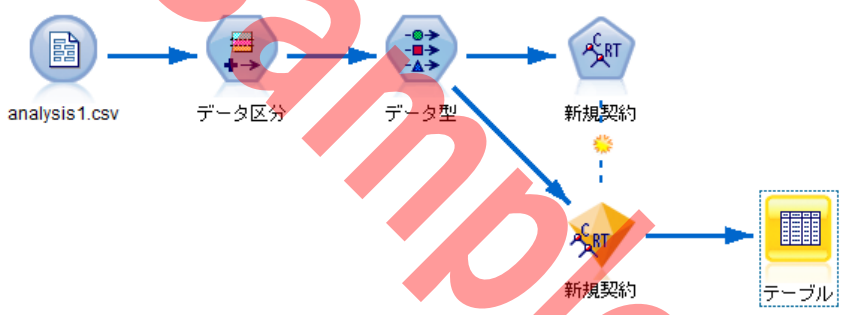

Figure2.3.17　新規契約モデルナゲットからリンクされたテーブルノード

次に、テーブルを実行します。テーブルはツールバーの実行ボタンを利用することで実行できます。

操作手順

4. ツールバーの**選択内容を実行**ボタンをクリックします。

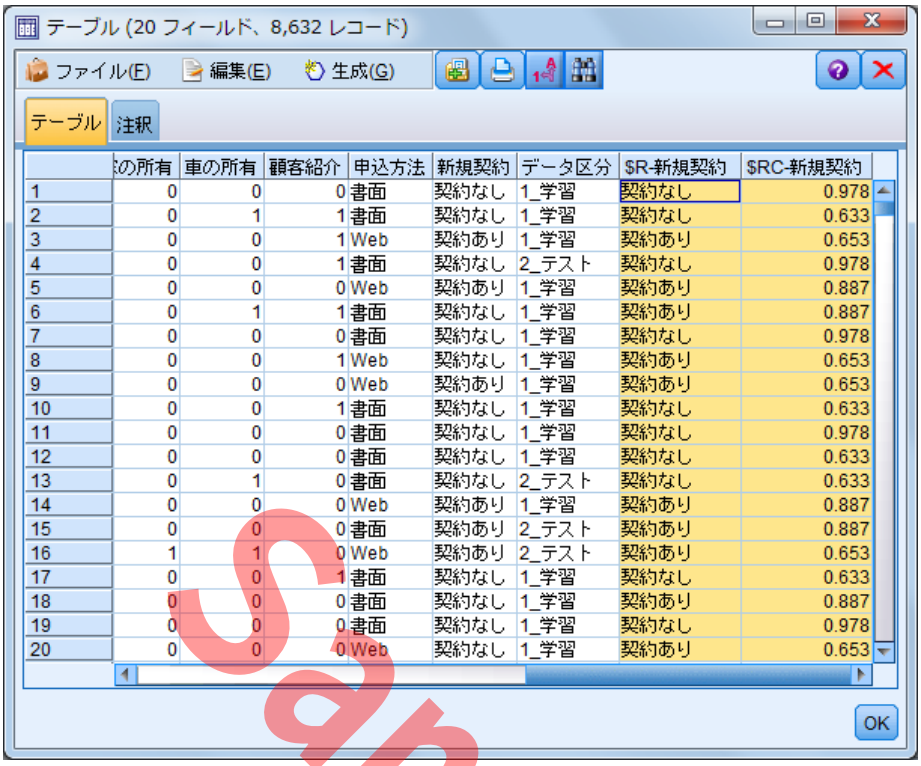

Figure2.3.18　テーブルに出力された予測値と確信度を含むデータ

データが読み込まれ、テーブルに出力されます。タイトルバーの表示から、20フィールド、8,632レコードが含まれていることが分かります。

新規契約は、元のデータに含まれる**実測値**であり、このフィールドには実際の新規契約ありと契約なしが記録されています。

$R-新規契約**は、C&R Treeのモデルによって予測された**予測値**であり、このフィールドには予測された新規契約ありと契約なしが記録されています。**$RC-新規契約は予測の確信度です。

新規契約と**$R-新規契約**の値が一致している場合、予測モデルが出した予測結果が的中していることを意味し、一致しない場合は予測が間違っていることを意味します。そこで、**クロス集計表**を用いることで、全体としてどの程度一致しているかを調べることができます。

操作手順

5. テーブルを閉じます。

§2.4.3　精度分析ノードによる予測精度の確認

精度分析ノードは、予測モデルの精度を評価するためのノードです。1 つ以上のモデルナゲットについて、**予測値**と**実際値**(対象フィールド)の的中率をさまざまな方法で評価することができ、学習データ区分とテストデータ区分の結果の比較も可能です。

この例では、新規契約の予測結果について、精度分析ノードによる精度の評価を行います。

操作手順

1. ストリームキャンバスの**新規契約**モデルナゲットを選択しておきます。
2. **出力**パレットの**精度分析**ノードをクリックします。

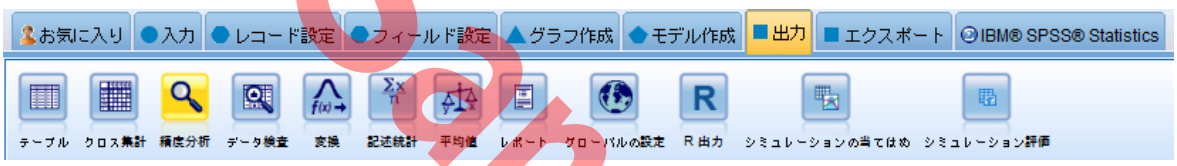

Figure2.4.5　出力パレットの精度分析ノード

操作手順

3. **精度分析**ノードを新規契約モデルナゲットからリンクします。

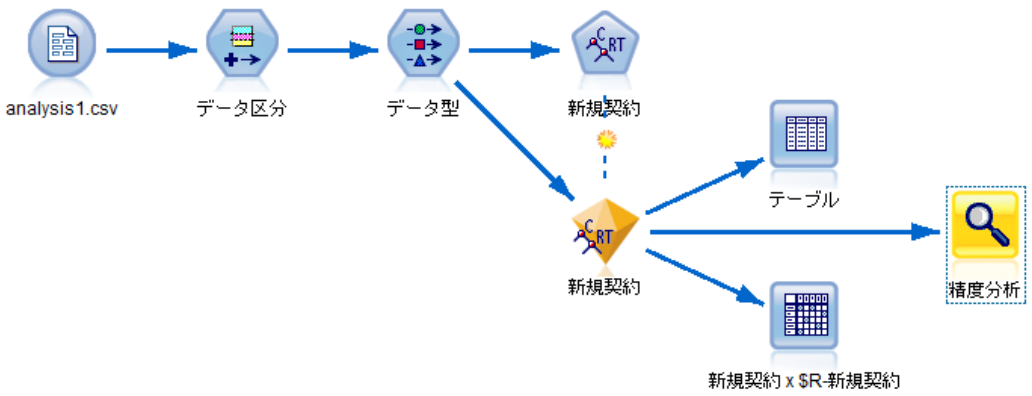

Figure2.4.6　ストリームキャンバスに配置した精度分析ノード

操作手順

4. **精度分析**ノードをダブルクリックして編集画面を表示します。
5. **評価メトリック(AUC & Gini、バイナリ-分類子のみ)**を選択します。

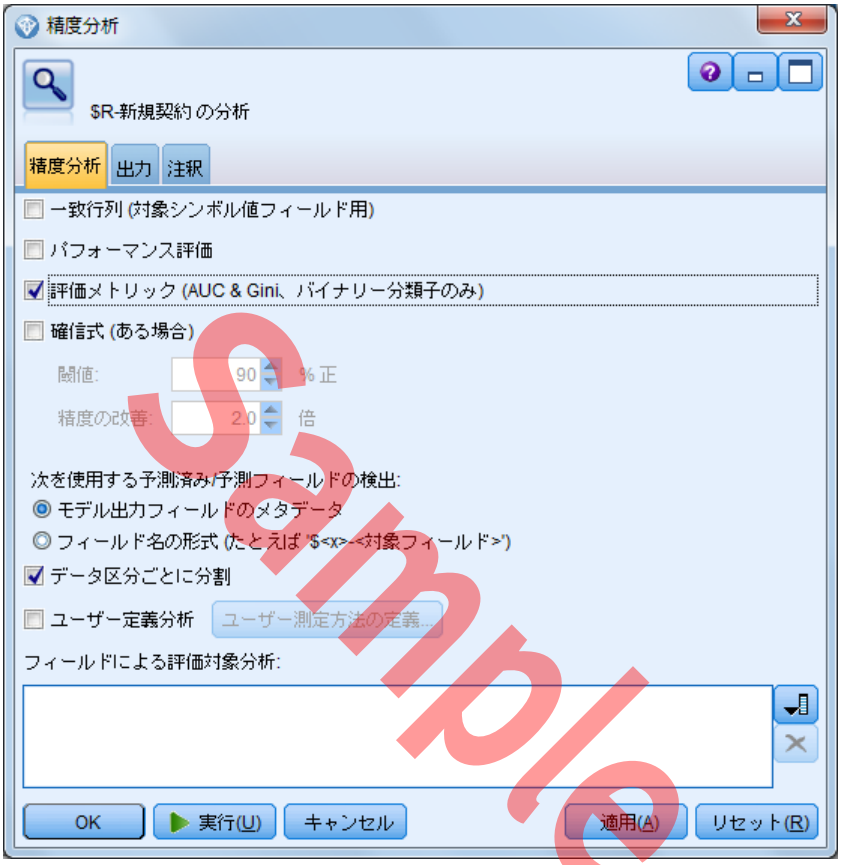

Figure2.4.7　精度分析ノードの精度分析タブ

一致行列(対象シンボル値フィールド用)を選択すると、実測値と予測値のクロス集計表を出力することができます。この設定はカテゴリごとの一致を調べる場合に使用します。このオプションはカテゴリ型の対象フィールドを用いている場合にのみ有効です。

TIPS

一致行列(対象シンボル値フィールド用)オプションで出力されるクロス集計表には、度数のみが表示されるため、一致率を調べる場合はクロス集計ノードを使用します。

評価メトリック（AUC & Gini、バイナリ-分類子のみ）を選択すると、モデルから計算されるAUCとGiniを出力することができます。AUCはモデルの精度の指標として知られ、値が1に近いほどモデルの予測精度が高いと評価します。このオプションはカテゴリ型の対象フィールドを用いている場合にのみ有効です。

その他、確信度に関するレポートを出力する設定や、任意のフィールドによる層別での比較を行うことができますが、この例では使用しません。

操作手順

6. **実行**ボタンをクリックします。

§2.4.4　精度分析の結果の解釈

精度分析の結果を確認します。実測値を含む**新規契約**と予測値を含む**$R-新規契約**が比較され、その一致率と不一致率が示されています。また、学習データ区分とテストデータ区分に分割されている場合は、それぞれの結果を表示します。

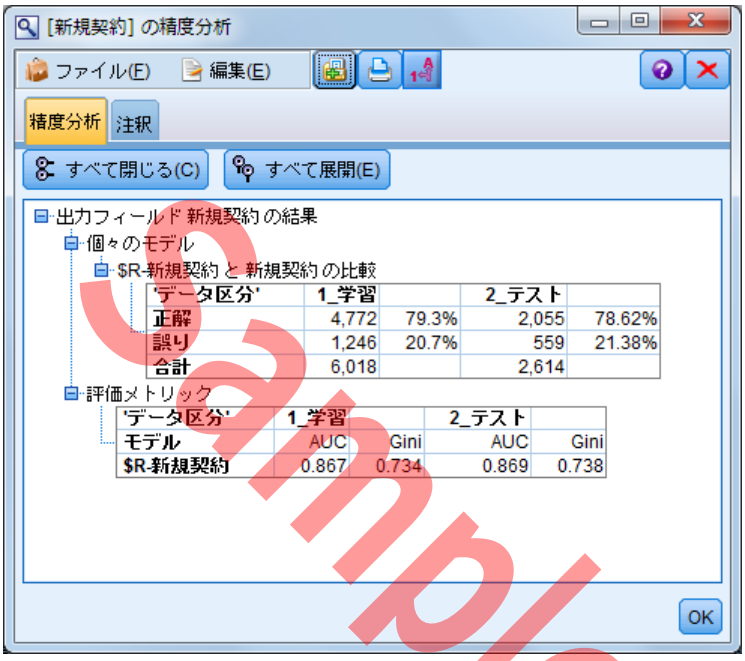

Figure2.4.8　精度分析の出力結果

学習データ区分では、実測値と予測値の一致率は**79.3%**です。およそ80%のレコードに関する予測が的中しています。

テストデータ区分では、実測値と予測値の一致率は**78.62%**です。学習データ区分における的中率とほぼ同じ的中であることが分かります。

POINT

学習データ区分とテストデータ区分の正解率が大きく異なる場合は、予測モデルの過剰な適合などが示唆されるため、基本的にモデルの見直しを検討します。

評価メトリックでは、**AUC**の値に注目します。この値が1に近いほどモデルの予測精度が高いことを意味します。この例では、**学習区分**データの**AUC=0.867**であり、**テスト区分**データの**AUC=0.869**で良好です。学習区分データとテスト区分データの値に大きな違いは見られません。

POINT

AUCはモデルの精度を表す指標の1つであり、値が**1**に近いほど精度が高いことを意味します。AUCが0.5の場合、ランダムに予測する場合と同じ精度を意味するため、AUCは0.5より大きくできるだけ1に近いほうが望ましい値です。一般に、**0.7**～**0.8**以上のAUCを良好の目安として用います。