

Oracle Data Miner 4.0の使用

概要

目的

このチュートリアルでは、Oracle Database 12c Release 1上でデータマイニング・アクティビティを実行するための Oracle Data Miner 4.0の使用をカバーします。 Oracle Data Miner 4.0 は、 Oracle SQL Developer バージョン4.0のエクステンションとして提供されます。このレッスンでは、ビジネス上の課題を解決するために、分類モデルを作成することでData Minerを使い方を学びます

Oracle SQL Developerは、データベース開発者のためのフリーのグラフィカルツールです。SQL Developerでは、データベース・オブジェクトを参照し、SQL文やSQLスクリプトの実行、およびPL/SQL文の編集・デバッグができます。SQL Developer 4.0に含まれるData MinerはOracle Database 11g Release 2とOracle Database 12cに対応しています

所要時間

約45分間

導入

データマイニングは、データからパターンや傾向を抽出することにより、データのかたまりから有益な情報を抽出するプロセスです。データマイニングは以下のような多様なビジネス上の課題解決のために利用できます：

- 個人の行動の予測。たとえば、プロモーションの申し出に応答する可能性の高い顧客や特定の製品を購入する可能性のある顧客の抽出(分類)
- 対象となる人々やアイテムのプロファイルの検索(Decision Treeによる分類)
- 集合からセグメントまたはクラスタの発見(クラスタリング)
- より多くのターゲット属性に関連する要因の特定(属性重要度)
- 同時発生するイベントや購買の発見(相関、マーケットバスケット分析)
- 異常値やレアなイベントの検出(異常検出)

Oracle Data Miningを利用して、ビジネス上の課題を解決するフェーズは以下の通りです:

1. データマイニングおよびビジネス目標の観点での課題定義
2. データ収集および準備
3. モデルの構築と検証
4. 展開

課題の定義とビジネス目標

データマイニングの実施時に、ビジネス上の課題をデータマイニングの機能の観点で明確に定義する必要があります。たとえば小売業、電話会社、金融機関および他のエンタープライズ企業では、古くからの忠実な顧客のライバル社への切り替えという行為である顧客の「解約」に注視しています。「顧客の解約を解決するためにデータマイニングを使いたい」というのは、あまりにも漠然としています。ビジネス上の観点から、不満を持つ顧客の流出をくいとめることより、離れてしまった顧客を呼びもどすことのほうが、現実的に遙かに困難で費用がかかることがあります。さらに、企業にとっての価値が低い顧客には興味がないかもしれません。このようにデータマイニングによって、解約する可能性の高い顧客を予測し、潜在的に価値の高い顧客が解約するかどうかを予測することがビジネス上の課題となります

データ収集と準備

データマイニングにおける一般的な経験則は、個々のデータについてできるだけ多くの情報を収集し、有益である可能性のデータを任意にフィルタリングできるようにすることです。具体的には、いくつかの属性は重要で

はないかもしれませんと考へるかもしれません、容易に削除するべきではありません。ODMのアルゴリズムによって削除するかどうかを決定できます。目標は任意の個人に対して適用できる行動のプロファイルを構築することですので、あなたは、名前、住所、電話番号等の特定の識別子を削除するべきです（ただし、郵便番号のような特定の個人を識別することなく一般的な場所を示す属性は役に立つかもしれません）。一般的には、データ収集および準備のフェーズで、データマイニング・プロジェクトの時間と労力の50%以上を費やすと言われています

モデルの構築と検証

Oracle Data Minerでは、ワークフローの作成プロセスは、モデルの構築およびテスト中の困難なタスクの多くを自動化します。これは、ビジネス上の課題を解決するのに最も良いアルゴリズムがなんであるかを事前にすることはとても困難なので、通常、いくつかのモデルを作成しテストします。完全なモデルというものは存在せず、最良な予測モデルを検索するということは、必然的に最も制度の高いモデルを決定するということではなく、ビジネス上の目標という観点から許容されるエラーの種類を決定するということになります

展開

Oracle Data Miningは、実用的な結果を生成しますが、正しいものを素早く届けなければその結果は有用ではありません。Oracle Data Minerのユーザ・インターフェースは、結果を出力するためのオプションがいくつか用意されています

シナリオ

このレッスンは、分類モデルによって解決できるビジネス上の課題に焦点を当てています。このシナリオでは、ABC社は、保険を購入する可能性が最も高い顧客を識別したいと考えています

注: このチュートリアルでは、「データの準備および収集」というフェーズは既に完了しており、サンプルのデータセットにはすべての必要なデータフィールドが含まれています。よって、このレッスンでは「モデルの構築と評価」フェーズに主に焦点を当てています

ソフトウェア要件

次のソフトウェアが必要になります:

- 次のソフトウェアにアクセス可能もしくはインストール済み:
 - Oracle Database:
 - 必要最低バージョン: Oracle Database 11g Enterprise Edition, Release 2 (11.2.0.1) と Data Mining Option
 - 推奨バージョン: Oracle Database 12c Enterprise Edition, Release 12.1 と Advanced Analytics Option
 - SQL Developer 4.0

前提条件

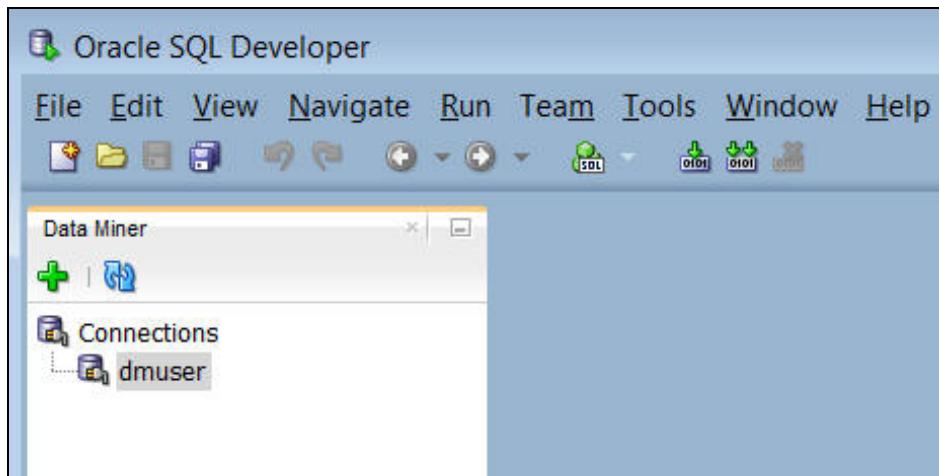
チュートリアルを開始する前に、Oracle Data Minerを含めたOracle SQL Developer 4.0をインストールしてください

注: もしまだOracle Data Minerのセットアップをしていない場合、次のレッスンを実施してください: **Oracle Data Miner 4.0のセットアップ**

Data Minerプロジェクトの作成

Data Minerプロジェクトを作成し、Data Minerワークフローを構築する前に、必要なData Minerの機能に簡単にアクセスするために、SQL Developer内のData Minerインターフェース・コンポーネントを整理しておくと便利です

開始するには、SQL Developerインターフェースのエレメント([接続]タブや[レポート]タブなどが含まれる)をすべて閉じ、以下のように、Data Minerタブのみを開きます:



上に示したように、Data Minerユーザ(dmuser)が作成されており、SQL Developerの接続が確立されています。

「**Oracle Data Miner 4.0のセットアップ**」チュートリアルで、DMUSERというデータベース・アカウントとSQL Developerの接続を作成する方法を学びます。このユーザは、マイニングに用いるサンプルデータへのアクセス権を持っています

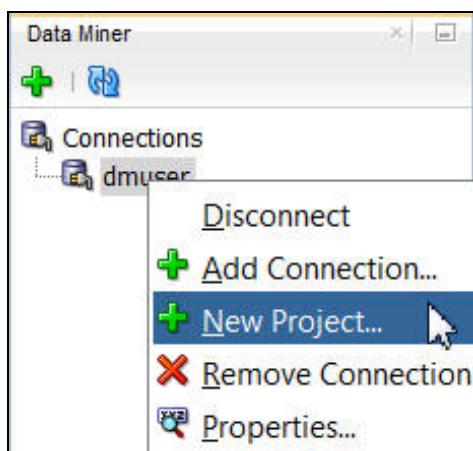
注: もし、Data Minerタブが開いていない場合、SQL Developerのメニューから、表示> Data Miner > Data Minerの接続を選択します

Data Minerプロジェクトの作成

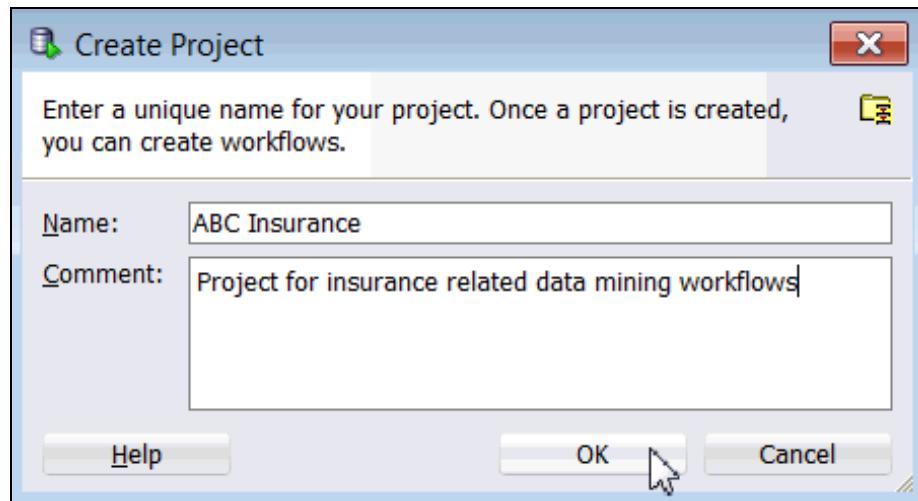
Data Miner ワークフローの作業を開始する前に、1つ以上のワークフローのコンテナとしてData Minerプロジェクトを作成する必要があります

Data Miner プロジェクトを作成するには、次の手順を実行します:

1. 以下のように、Data Miner タブで、dmuserを右クリックし新規プロジェクトを選択します:

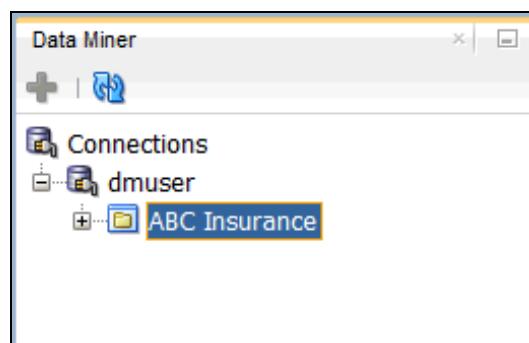


2. プロジェクトの作成ウィンドウで、プロジェクト名(この例ではABC Insurance)を入力し、OKをクリックします



注: オプションでこのプロジェクトの意図を説明するコメントを入力することができます。この説明は、いつでも変更できます

結果: 新規プロジェクトがデータマイニングのユーザ接続ノードの下に表示されます



Data Minerワークフローの構築

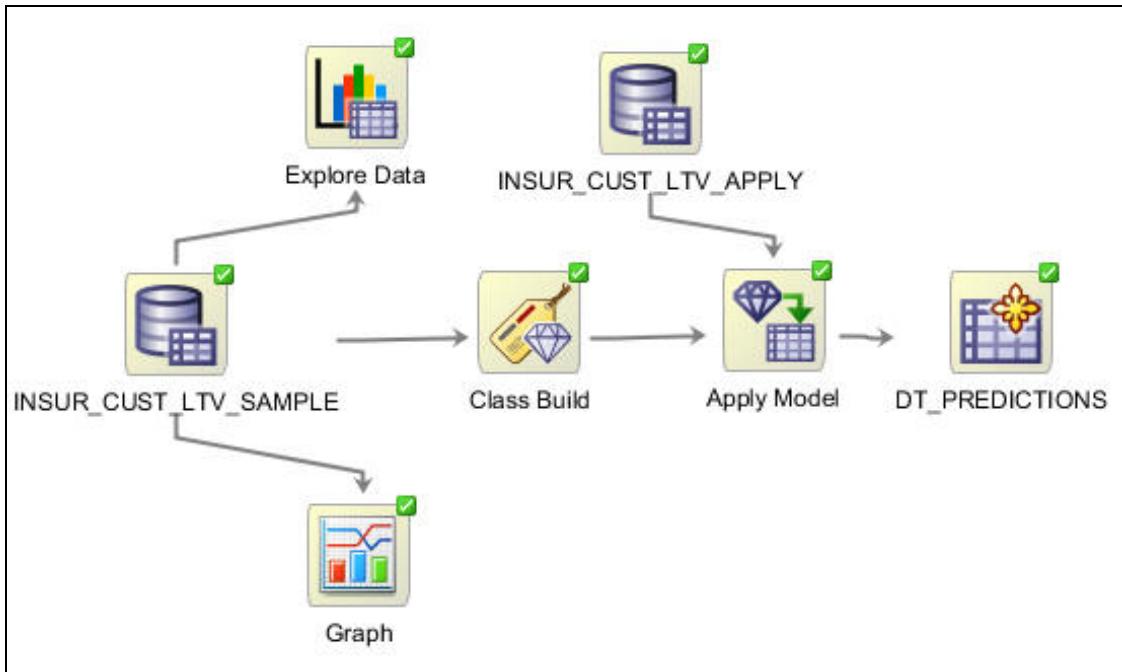
Data Minerワークフローは、データマイニング・プロセスをあらわす接続ノードの集合です

ワークフロー:

- データマイニングサーバのための指示を提供します。たとえば、「これらの特性を持つモデルを構築します」というワークフローを定義すると、ワークフローに返す結果とともにデータマイニングサーバでモデルが構築されます
- グラフィカル環境からデータマイニング・プロセスの作成、分析およびテストを対話的に実施できます
- より大きなプロセスの1サイクルのみをテストし、分析するために使う、もしくは特定のビジネス上の課題を解決するためにデザインされたプロセスのすべてのフェーズをカプセル化することができます

Data Minerワークフローには何が含まれる?

視覚的には、以下のようにワークフローウィンドウが表示され、作成使用としているデータマイニング・プロセス・フローのグラフィカルな表現を提供します:



注:

- プロセスの各要素は、ノードと呼ばれるグラフィカルなアイコンで表示されます
- 各ノードは、特定の指示を含む明確な目的を持ち、多くの方法で個々の定義を設定・修正します
- 一緒にリンクされる場合には、ワークフローノードは特定のデータマイニングの課題を解決されたことにより、モデリングプロセスを構築します

これから学ぶように、任意のノードをワークフローエリアに単にドラッグ&ドロップすることでワークフローに追加できます。各ノードには、デフォルトのプロパティが含まれています。必要に応じてプロパティを変更し、次のステップに進むための準備をします

データマイニングシナリオのサンプル

このトピックでは、保険を購入する可能性が最も高い既存顧客を予測するデータマイニング・プロセスを作成します。

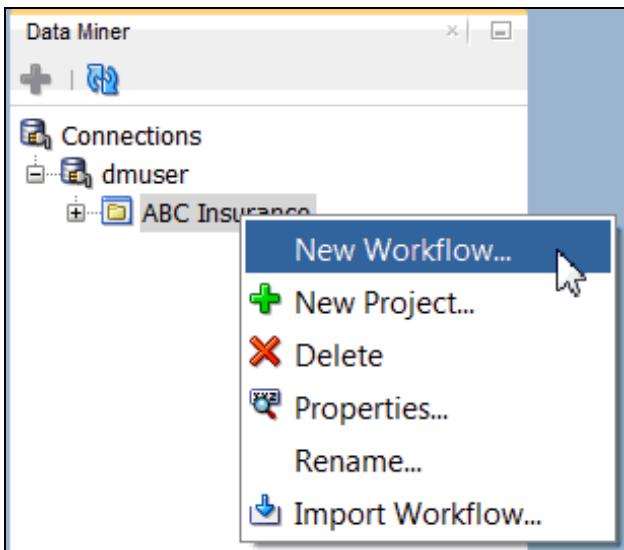
この目標を達成するために、以下を実施してワークフローを構築します:

- ソースデータを特定し、検討する
- いくつかの分類モデルを構築し、比較する
- 最も実用的な結果を生成するモデルを選択し、実行する

このプロセスのためのワークフローを作成するには、次の手順を実行します

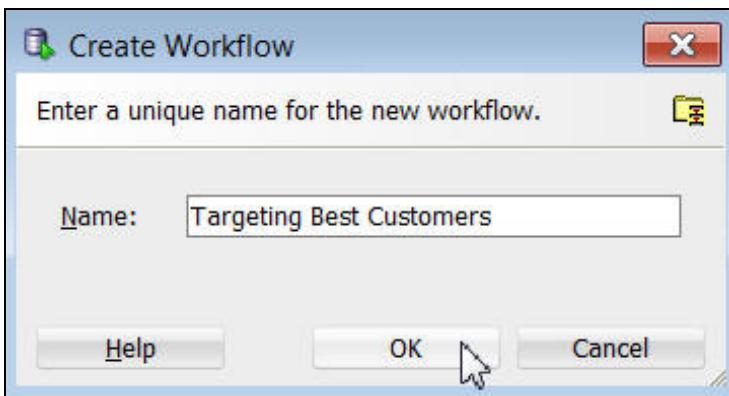
ワークフローの作成とデータソースの追加

1. プロジェクト(ABC Insurance)を右クリックし、メニューから、新規ワークフローを選択します



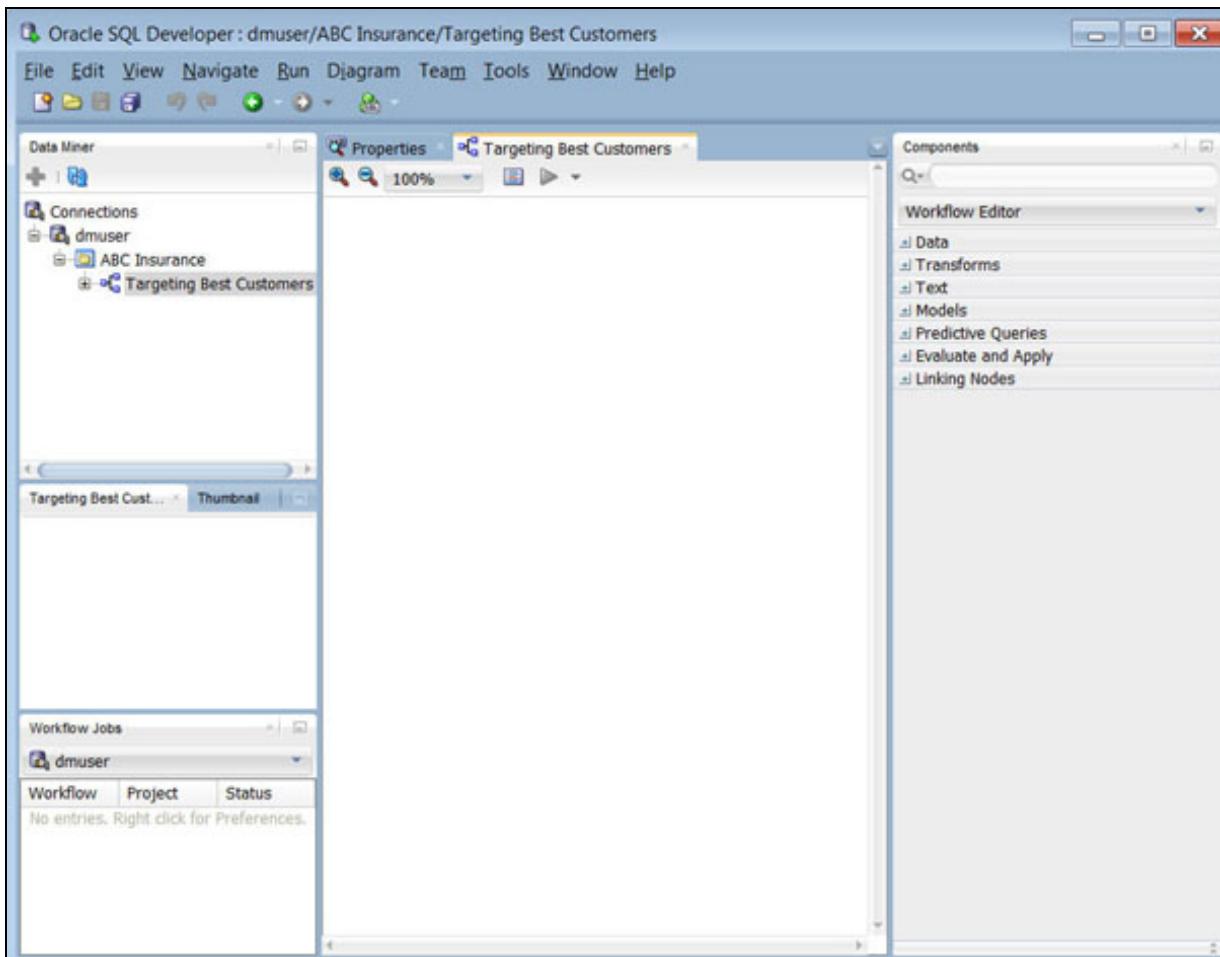
結果: ワークフローの作成ウィンドウが表示されます

2. ワークフローの作成ウィンドウで、名前に**Targeting Best Customers**を入力し、OKをクリックします



結果:

- SQL Developerウィンドウの中央:
 - 指定した名前のタブ名のついた空のワークフローウィンドウが表示されます
 - また、[プロパティ]タブが、初回は同じエリアにあります
- インタフェースの右側上部に、ワークフローエディタの[コンポーネント]タブが表示されます
- また、他の3つのOracle Data Minerインターフェース要素が左側下部に開いています:
 - ワークフローの名前を持つ[構造]タブ
 - [サムネイル]タブ
 - [ワークフロー・ジョブ]タブ

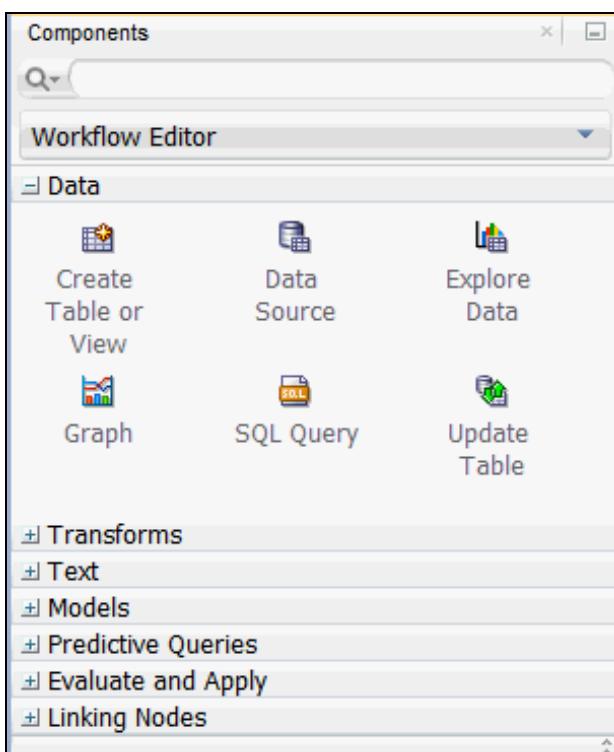


後で見やすいように、ニーズに合わせてSQL Developerウィンドウ内のData Minerタブペインを、開く、閉じる、サイズ変更、および移動することができます

3.

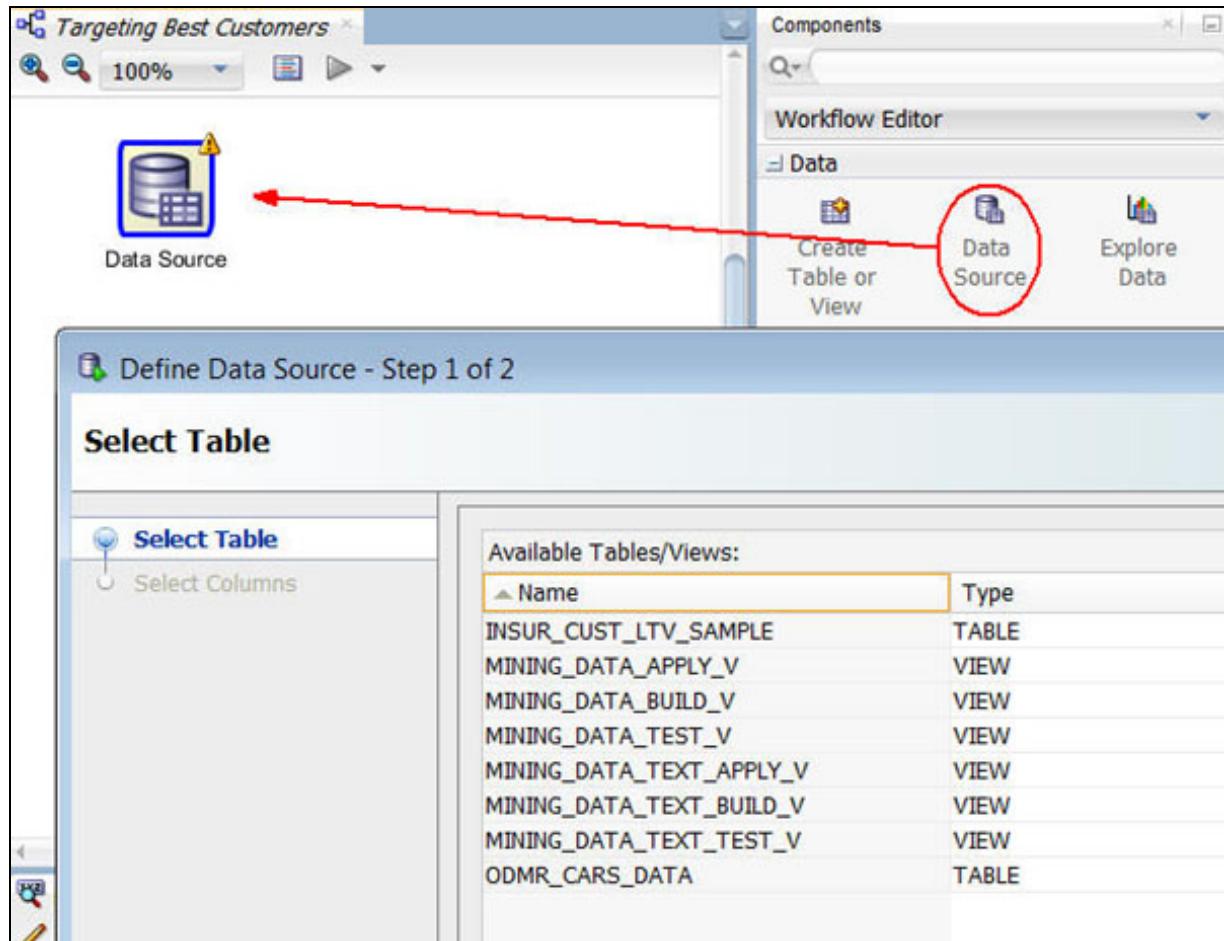
4. ワークフローの最初の要素はソースデータです。ここでは、ワークフローにデータソースノードを追加し、データソースとして INSUR_CUST_LTV_SAMPLE 表を選択します

A. [コンポーネント]タブで、**データ** カテゴリをドリルします。以下のように、6つのデータノードグループが表示されます：



B. ワークフローペインにデータソースノードをドラッグ&ドロップします

結果: ワークフローペインにデータソースノードが表示され、データソースの定義ウィザードが開きます

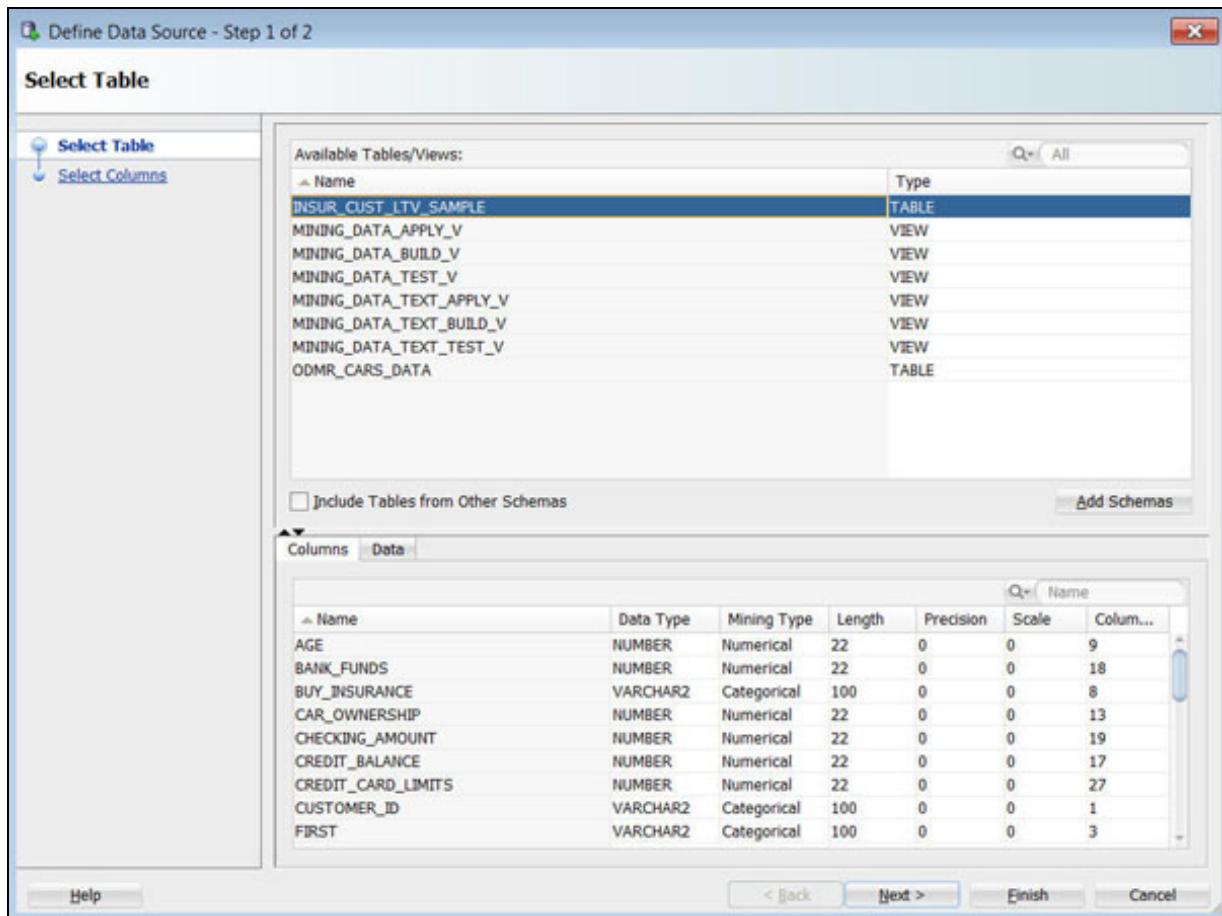


Notes:

- Oracle Data Minerによって、ワークスペースノード名とモデル名は自動で作成されます。この例では、「データソース」という名前が生成されました。このレッスンで示したものと全く同じノード、モデル名を取得できない場合があります
- プロパティ・インスペクタを使用して、ワークスペースとモデルを変更することができます

5. ウィザードのステップ1:

- A. 以下のように、使用可能な表/ビューリストから**INSUR_CUST_LTV_SAMPLE**を選択します:

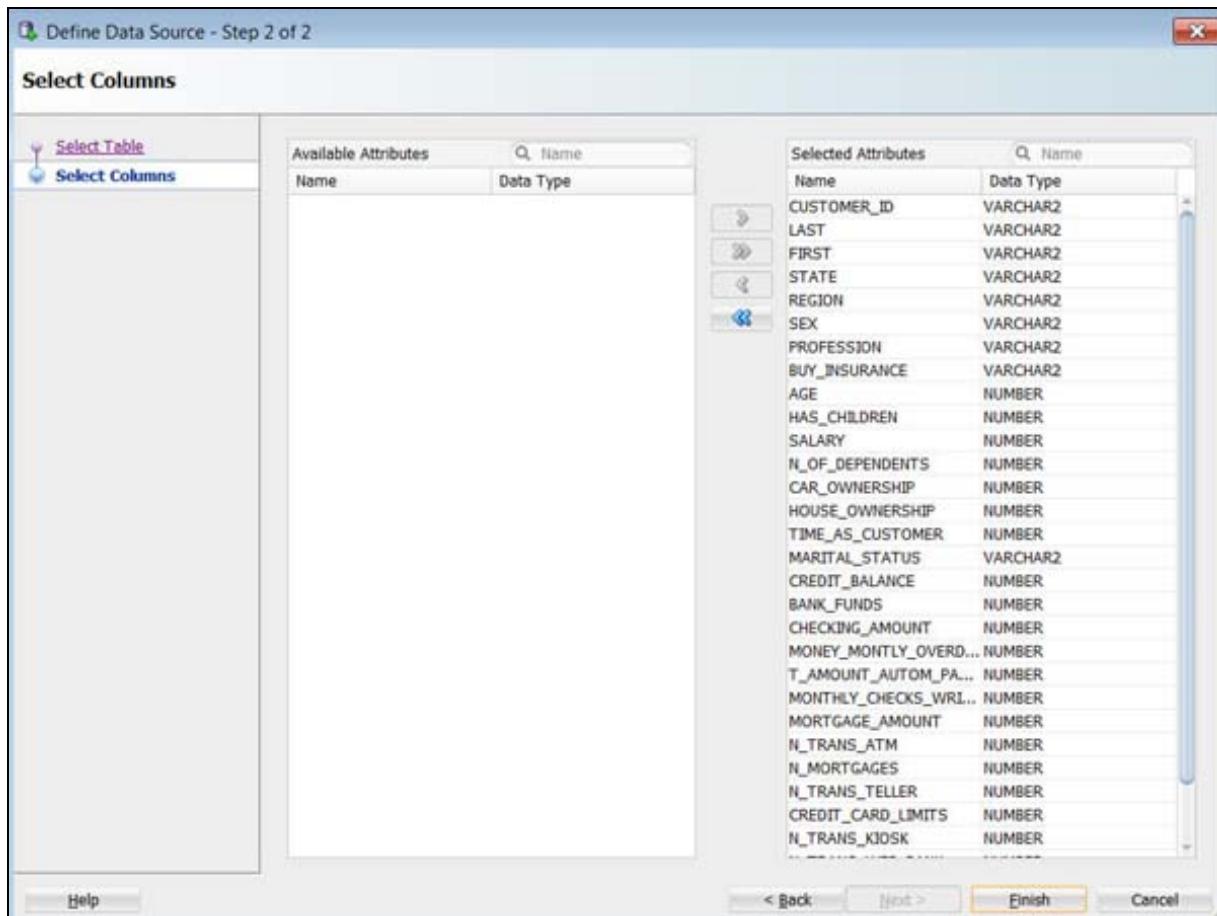


注: ウィザード内の下のペインで選択した表を表示し内容を確認できます。[列]タブには、表構造についての情報が表示され、[データ]タブには、選択した表もしくはビューからデータのサブセットが表示されます。

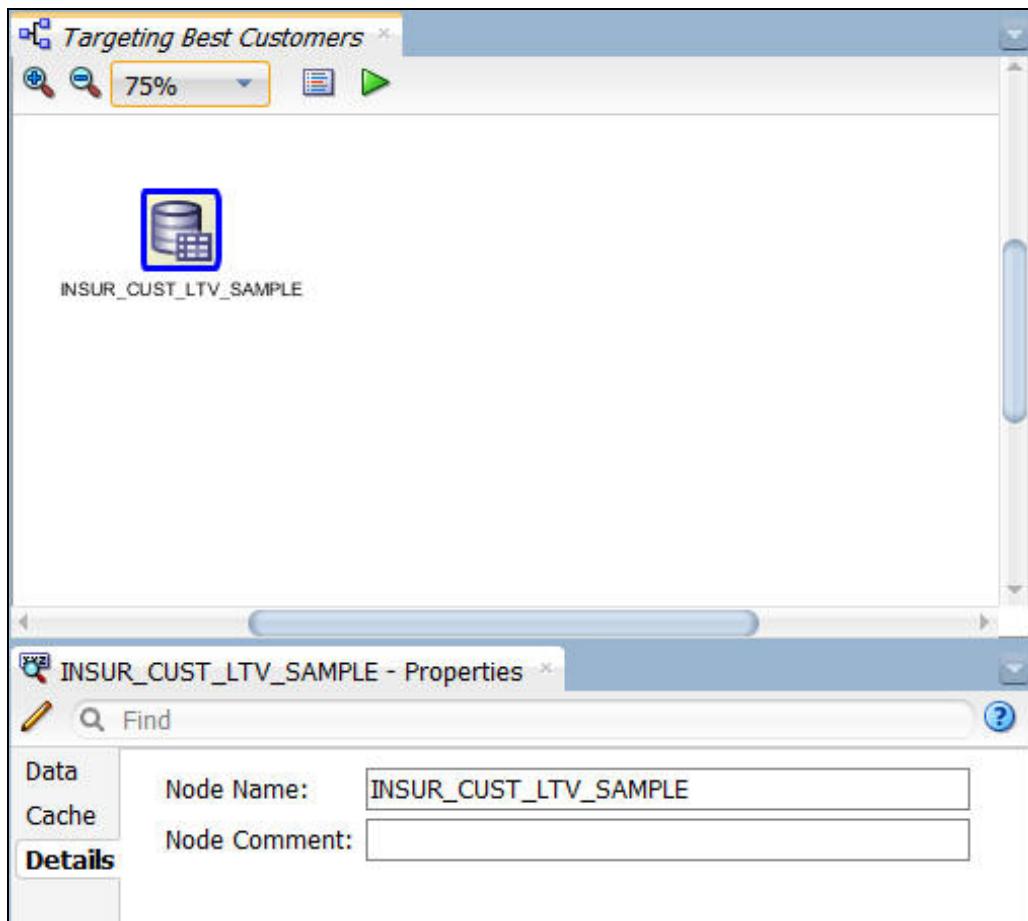
B. 次をクリックします

6. ウィザードのステップ2では、データソース内の不必要的列を削除できます。今回は、表に定義されたすべての属性を残しておきます

ウィザードの下部にある終了をクリックします



結果: 下に示すように、データソースノード名が選択した表名に更新され、ノードに関連づけられるプロパティが[プロパティ]タブに表示されます



Notes:

- ズームオプションで別の値を入力するか選択することで、ワークフローキャンバス内のノードの

サイズを変更できます。上図では既にズームプルダウンリストより**75%**が選択されています

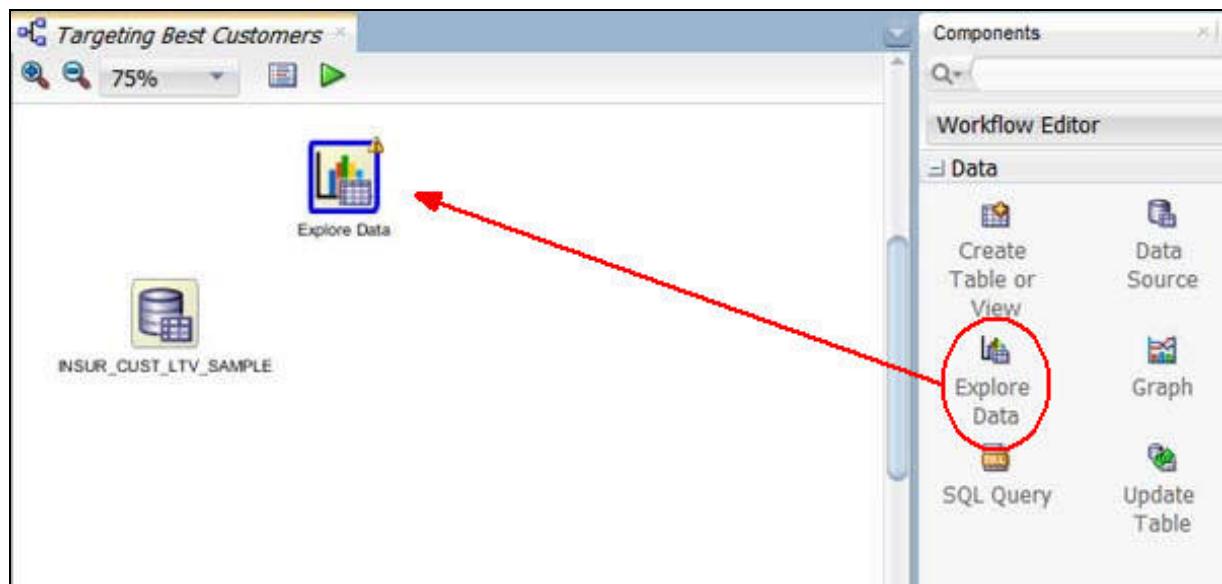
- [サムネイル]タブでは、より大きなワークフローウィンドウの小さなディスプレイが用意されています。ワークフローウィンドウの周囲のノードをドラッグすると、サムネイル表示が自動的に調整されます
- SQL Developer内の別の場所に任意のData Minerのタブを移動することができます。デフォルトでワークフローペインの右下に配置される[プロパティ]タブを、単に目的の場所にドラッグすることで移動しています
- [プロパティ]タブでは:
 - データセクションを使用して、表またはビューの列に関する情報を表示します
 - データキャッシュセクションを使用して、結果表示を最適化するために出力データのキャッシュを生成します
 - 上図のように詳細セクションを使用して、ノード名の変更や各ノードに関するコメントを追記します

ソースデータの分析

ソースデータの分析のためにデータの参照ノードを使います。グラフノードでもデータの可視化は可能です。これらはオプションのステップですが、Oracle Data Minerでは、このツールによって、選択したデータにより定義したビジネス上の課題を解決する基準を満たしているかどうかを確認できます

次の手順に従ってください:

1. 以下のようにデータグループからデータの参照をワークフローにドラッグ&ドロップします:



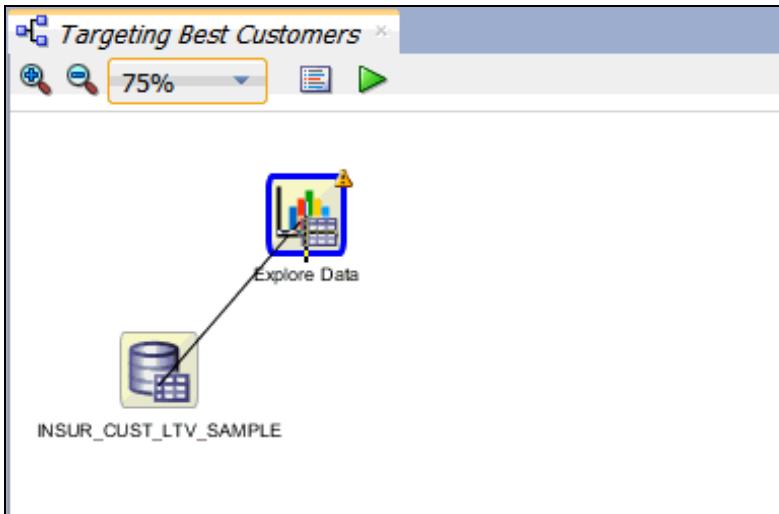
結果: 新たにワークフローペインにデータの参照ノードが表示されます

注:

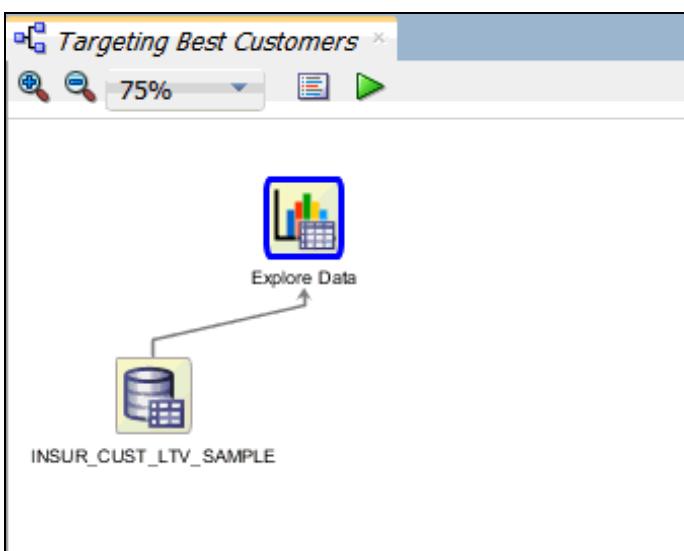
- ノードの周囲の境界線にある黄色い情報アイコン(!)は、ノードが完全ではないことを示しています。よって、データ参照ノードを使う前に、少なくとも1つの追加ステップが必要です
- この場合、ソースデータを参照するためにデータ参照ノードにデータソースノードから接続する必要があります

2. データソースとデータ参照ノードを接続するために以下の手順を行います:

- A. データソースノード(INSUR_CUST_LTV_SAMPLE)を右クリックし、ポップアップメニューから接続を選択し、ポインタをデータの参照ノードにドラッグします:

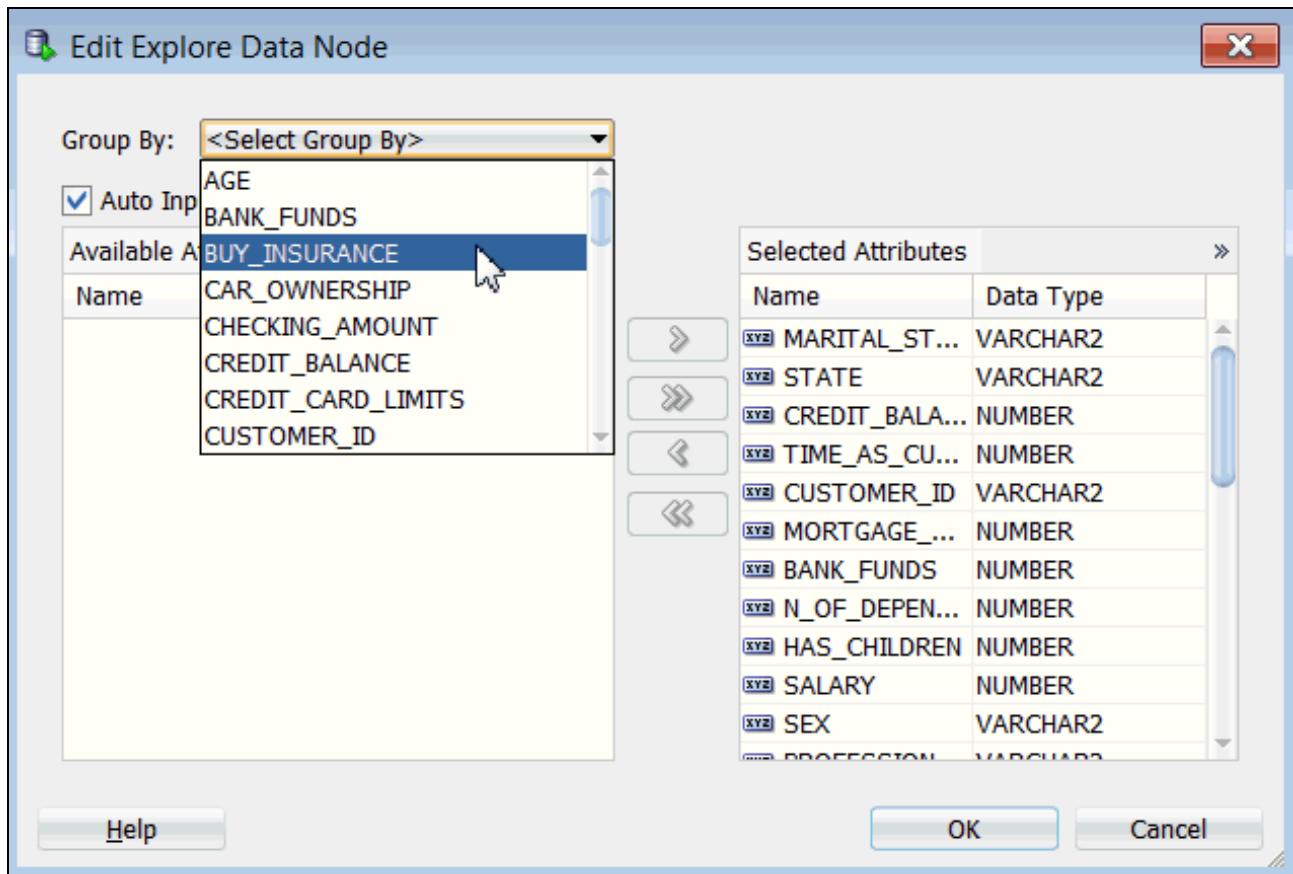


B. 次に、データの参照ノードをクリックし2つのノードを接続します。結果、表示はこのようになります:



3. 次に、データソースの「グループ化基準」を選択します

- データの参照ノードをダブルクリックし、データの参照ノードの編集ウィンドウを表示します
- グループ化基準リストから、以下のように**BUY_INSURANCE** 属性を選択します:

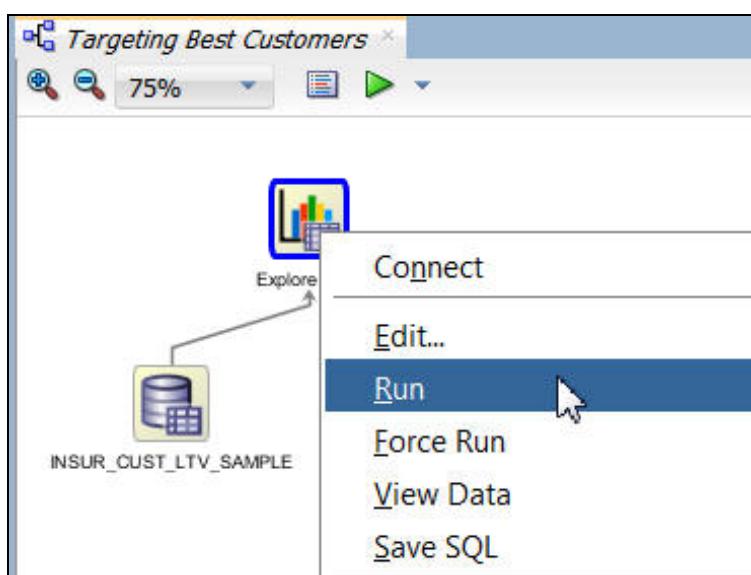


C. 次に**OK**をクリックします

注: 選択した属性ウィンドウでは、ソースデータから任意の属性を削除（または再追加）することができます

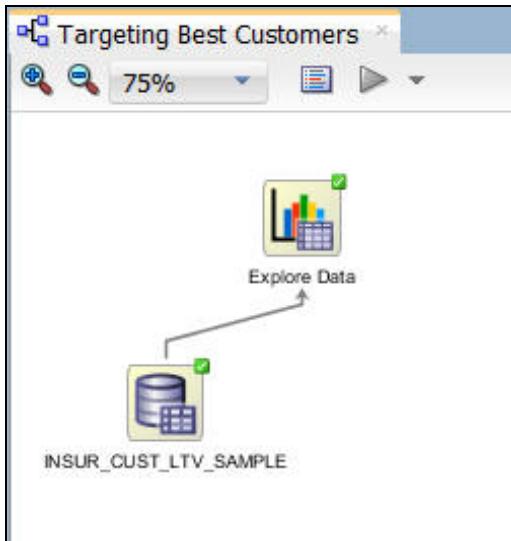
4.

5. 次にデータ参照ノードを右クリックし**[実行]**を選択します



結果:

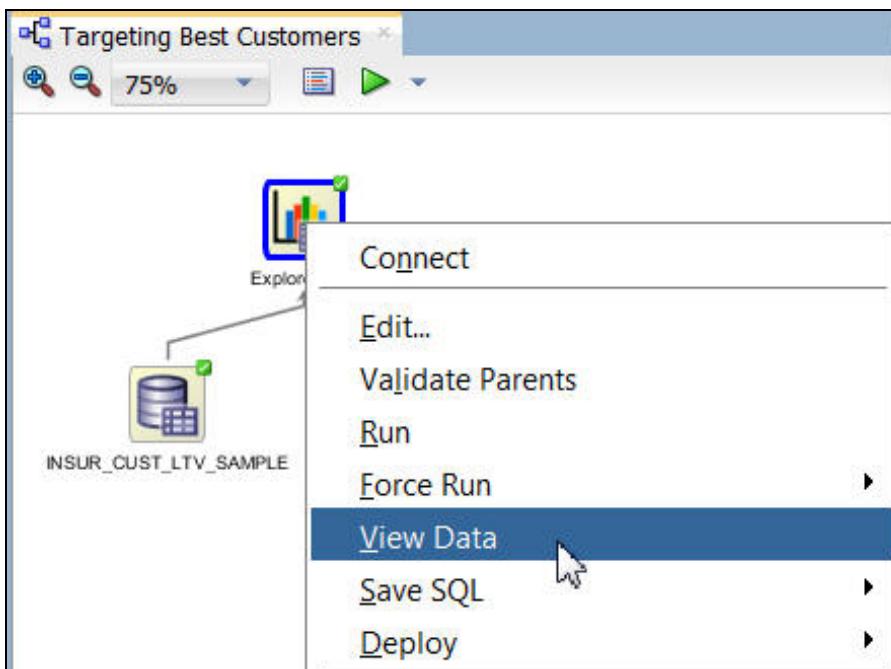
- Data Minerはワークフロー文書を保存し、ステータス情報をノードの処理中にワークフローペイントの上部にあるステータスバーに表示します
- 各ノードの処理中には、ノードの境界線上に緑の歯車アイコンが表示されます
- 更新が完了すると、データソースおよびデータの参照ノードの境界線には、以下のように緑のチェックマークが表示されます:



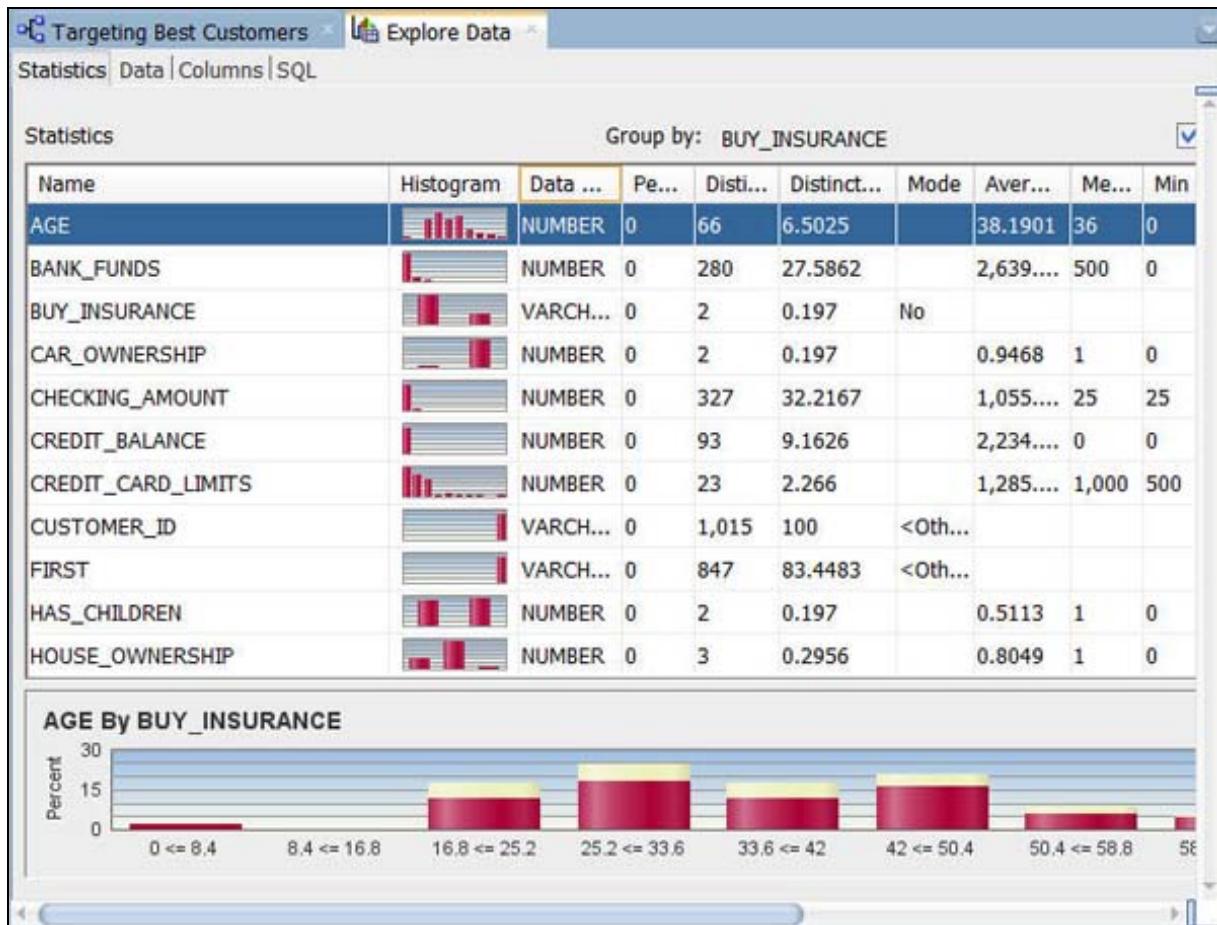
注: ワークフローキャンバスから任意の処理を実行すると、指定した手順はOracle Data Minerサーバによって実行されています

6. データの参照ノードの結果を確認するには、次の手順を実行します:

A.データの参照ノードを右クリックし、メニューから**データの表示**を選択します



結果: 以下のようにデータの参照ノードのための新たなタブが開きます



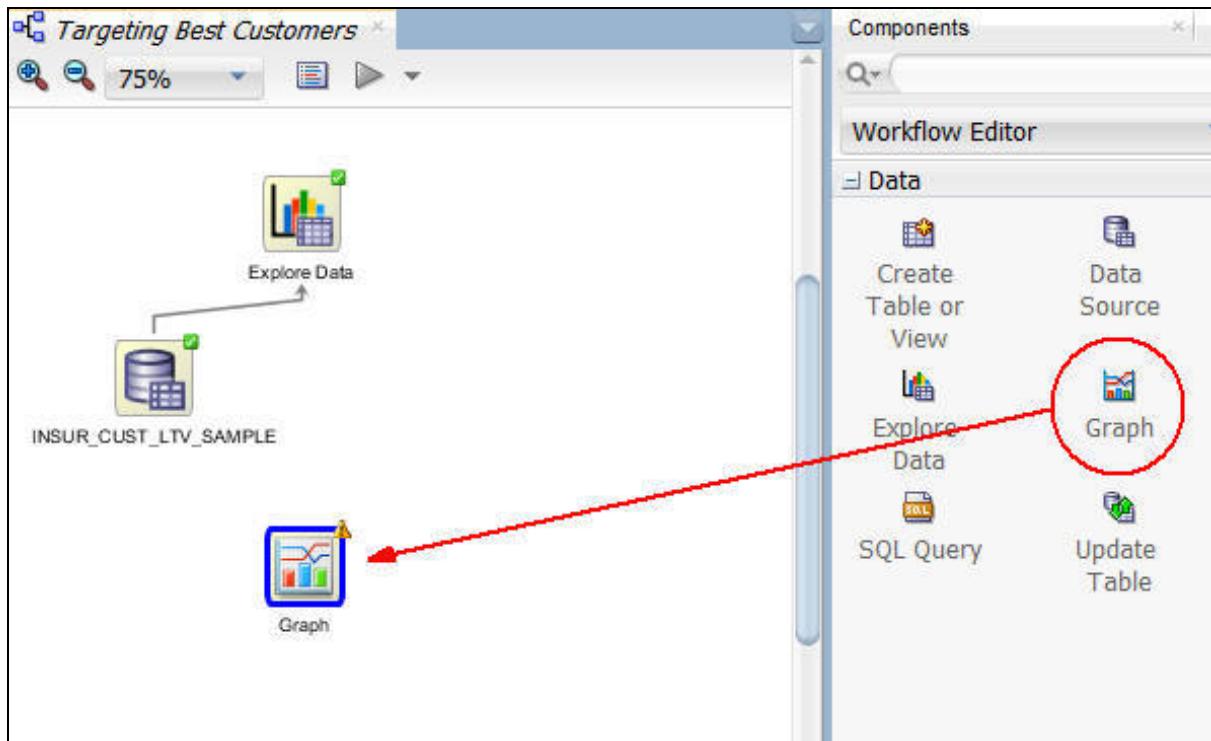
注:

- Data Minerは、定義したデータセット内の各属性に関するさまざまな統計を、「グループ化基準」属性に定義した属性(BUY_INSURANCE)について算出します。次の列が表示されます:ヒストグラムのサムネイル、データ型、個別値、個別値パーセント、モード値、平均値、中間値、最小値、最大値、標準偏差、分散
- この表示によって可視化され、データを検証でき、また手動でデータのパターンや構造を調査できます

- B. [名前]リストから属性を選択すると、下のウィンドウに関連するヒストグラムが表示されます
- C. ソースデータの分析を実施したら、クローズアイコン(X)をクリックしてデータの参照タブを閉じます
- 次に、グラフノードを使用してさらにデータを可視化します

7.

8. 以下のように、データグループからグラフノードをワークフローにドラッグ&ドロップします:

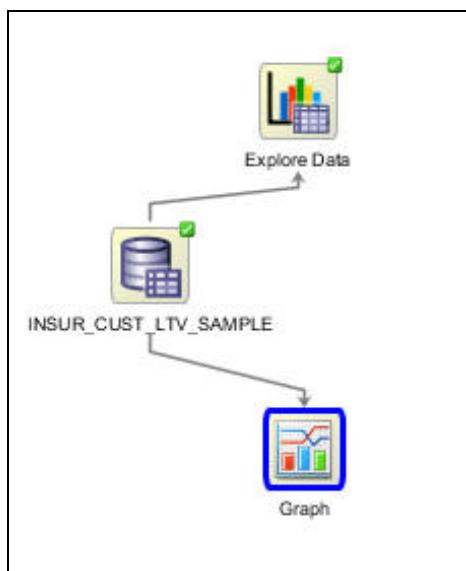


結果: ワークフローペインに新たにグラフノードが表示されます。前に見たように、黄色い情報アイコン(!)が表示され設定が完全ではないことを示しています

9.

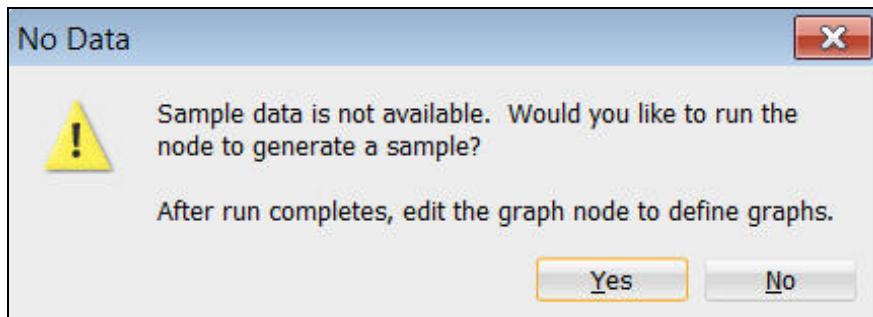
10. データソースノードにグラフノードを接続するには、以下の手順を使用します:

- データソースノード(INSUR_CUST_LTV_SAMPLE)を右クリックし、ポップアップメニューから接続を選択し、グラフノードにポインタをドラッグします
- 次に、2つのノードを接続するためにグラフノードをクリックします。以下のように結果が表示されます:



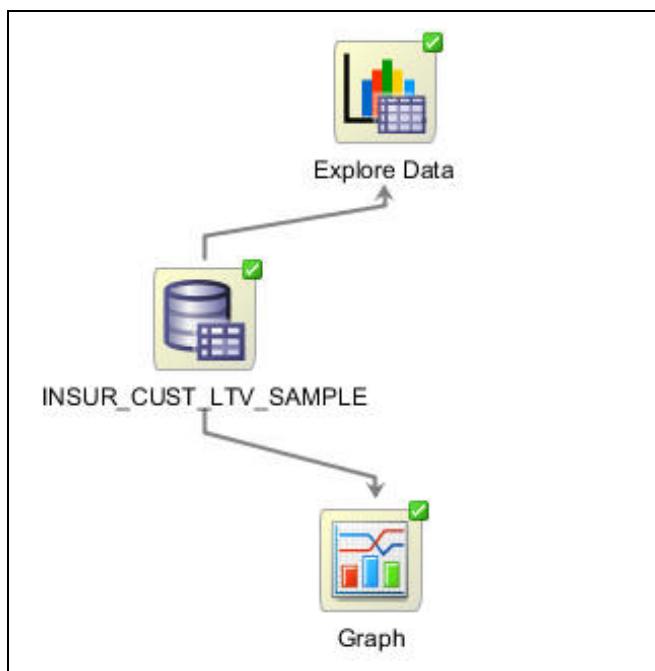
11. 次に、グラフノードを右クリックし、メニューから編集をクリックします

結果: 次の情報ダイアログが表示されます



はい をクリックし、ノードを実行しサンプルを生成します

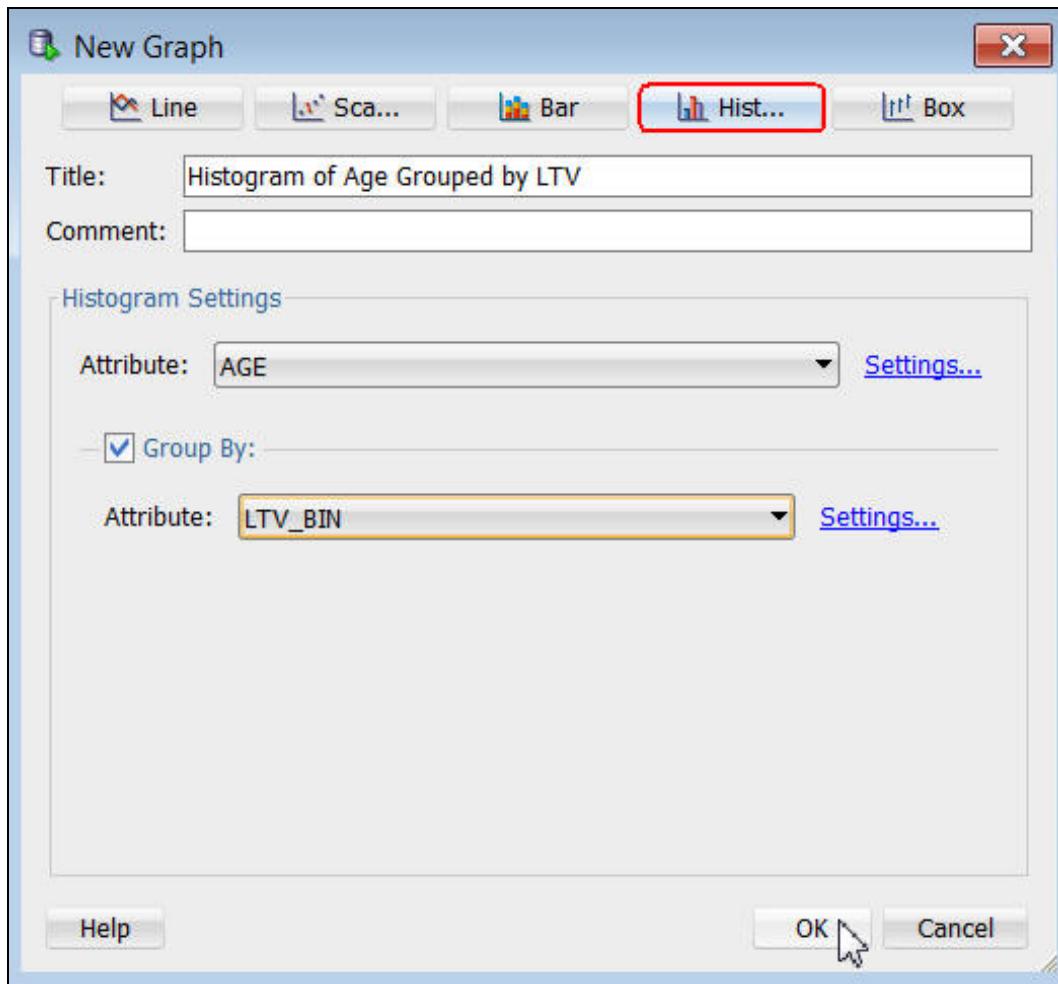
結果:処理が終了すると、ワークフローは以下のようになります:



12. ここで、新規グラフウィンドウを表示するためにグラフノードをダブルクリックします。以下の属性を設定します:

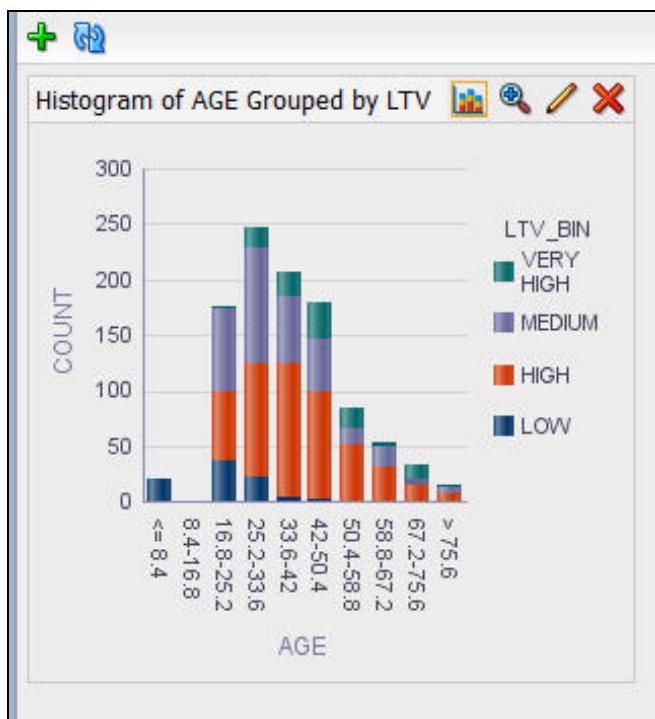
- 上部のHistogram ボタンをクリックし、グラフタイプを選択します
- Title ポックスで、 **Histogram of AGE Grouped by LTV**と入力します
- ヒストグラムの設定エリアで属性の値に**AGE** を選択します
- 次に、 **グループ化基準オプション**を有効にします
- グループ化基準オプションの属性に、 **LTV_BIN**を選択します

新たなグラフウィンドウは以下のように設定します:

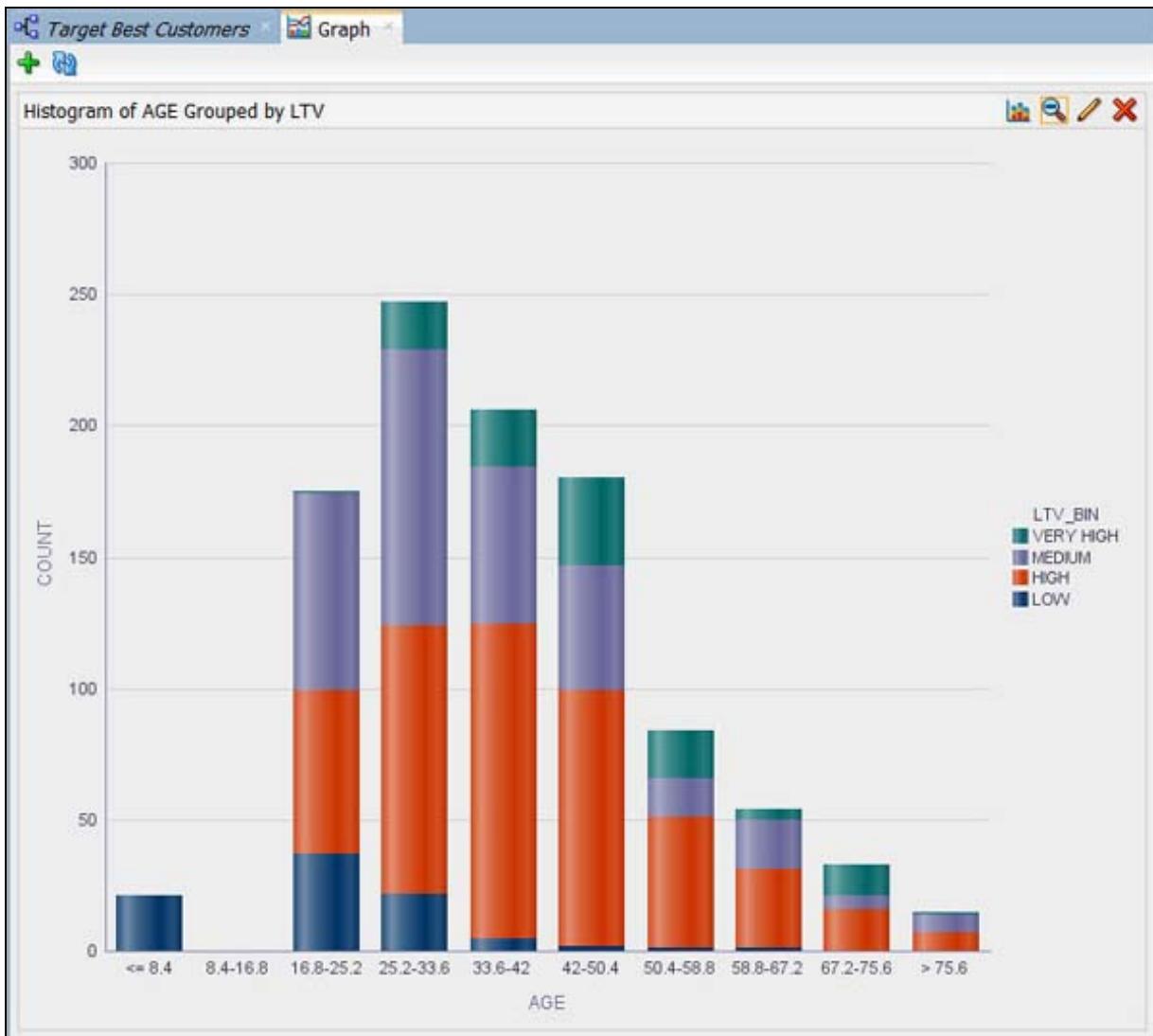


F. OKをクリックします

結果: 以下のようなグラフが表示されます:



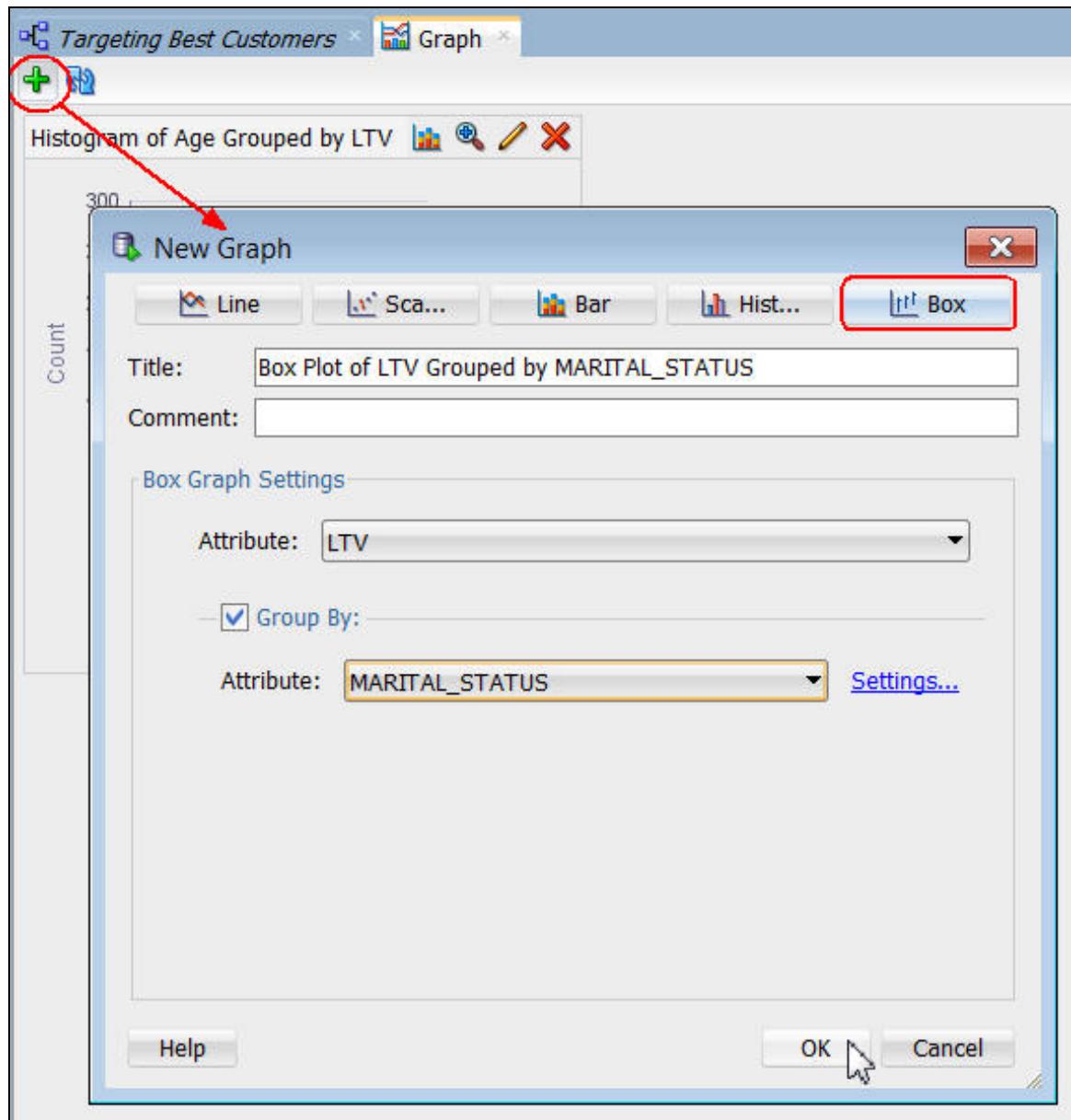
G. オプションで、以下のようにグラフをフルウィンドウで表示するにはズームインツールを選択します:



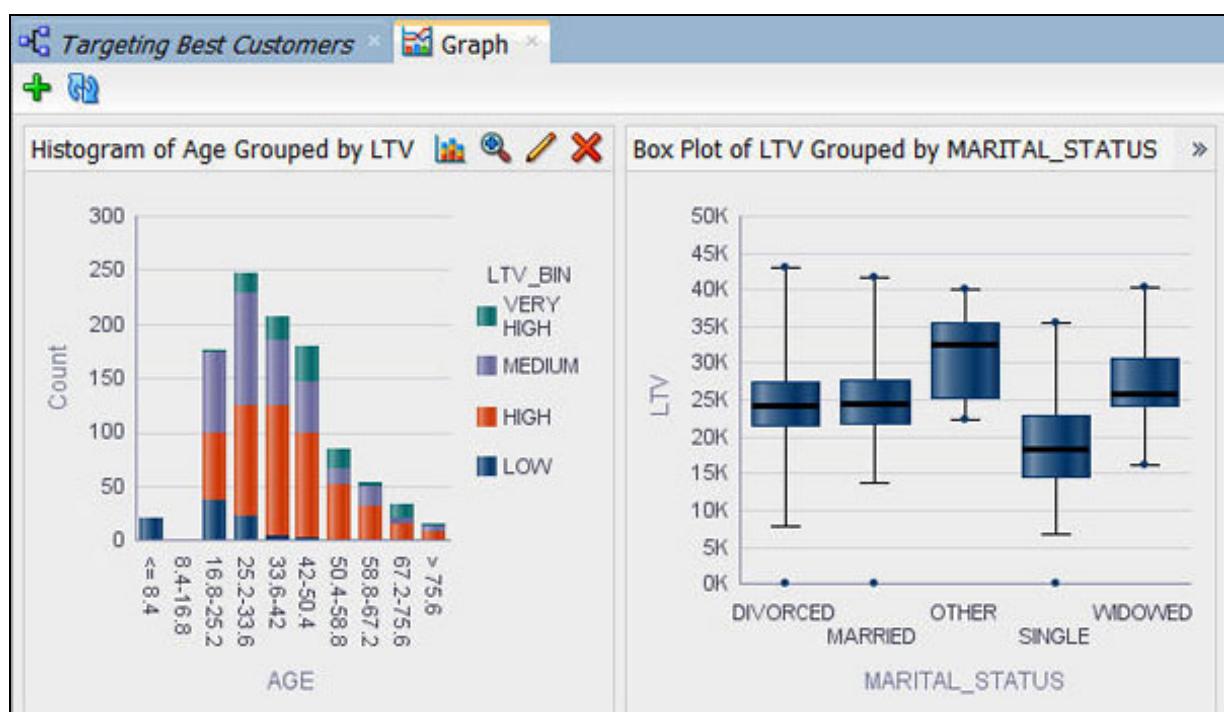
注: 単にズームアウトツール（ズームインツールを切替）をクリックするとオリジナルサイズに戻ります

13.

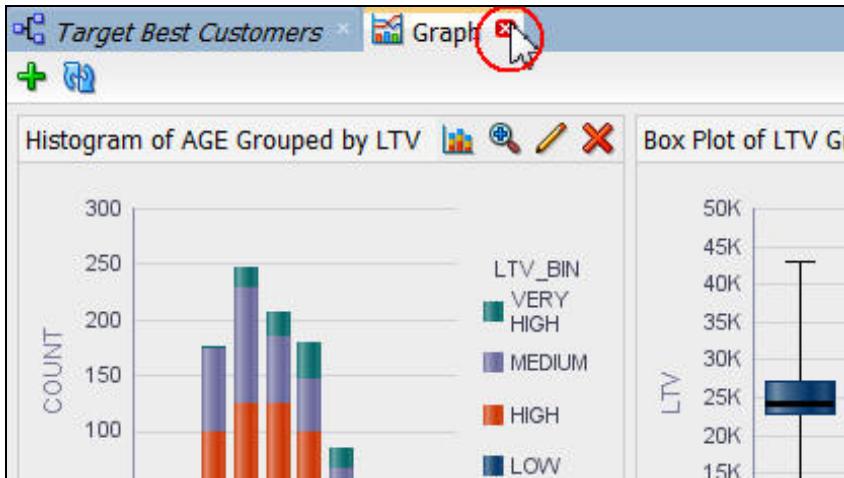
14. 以下のように、新規グラフツール(緑の "+" アイコン)をクリックするだけで、1つのノード内に追加のグラフを作成できます:



注: MARITAL_STATUSでグループ化したLTV(ライフタイムバリュー)のボックスグラフを作成します。以下のような結果が表示されます:



15. グラフノードで分析を実施したら、以下のようにクローズアイコン(X)をクリックしてグラフタブを閉じます:



次に、データベースのデータマイニングのパワーを使用して、高レベル分析の演習を実施します

分類モデルの作成

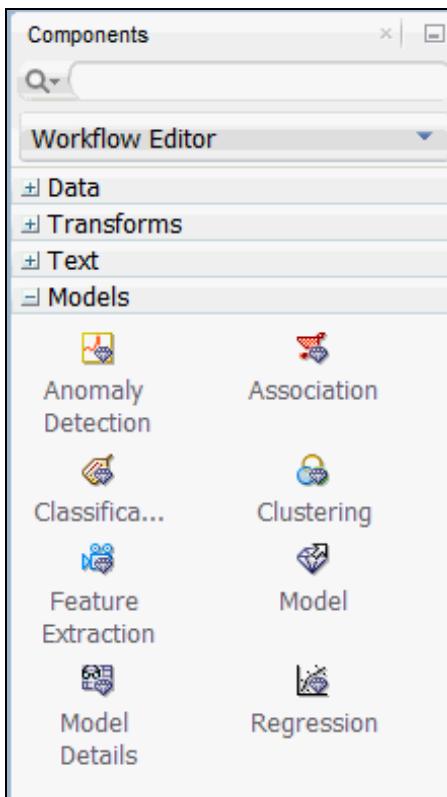
このチュートリアルの概要セクションで、個人の行動の予測には分類モデルを使うことを紹介しています。このシナリオでは、あなたは保険を購入してくれそうな顧客を予測したいとします。したがって、今回は分類モデルを用います。

Oracle Data Minerでは、分類モデルを作成するとアルゴリズムの異なる4つのモデルが作成されます。分類ノード内のもてるはすべて同じターゲットとケースIDを持ちます。このデフォルトの構成は、最良の予測をするアルゴリズムの発見が容易にできます。ここでは、すべてのアルゴリズムを使用して分類ノードを定義します。

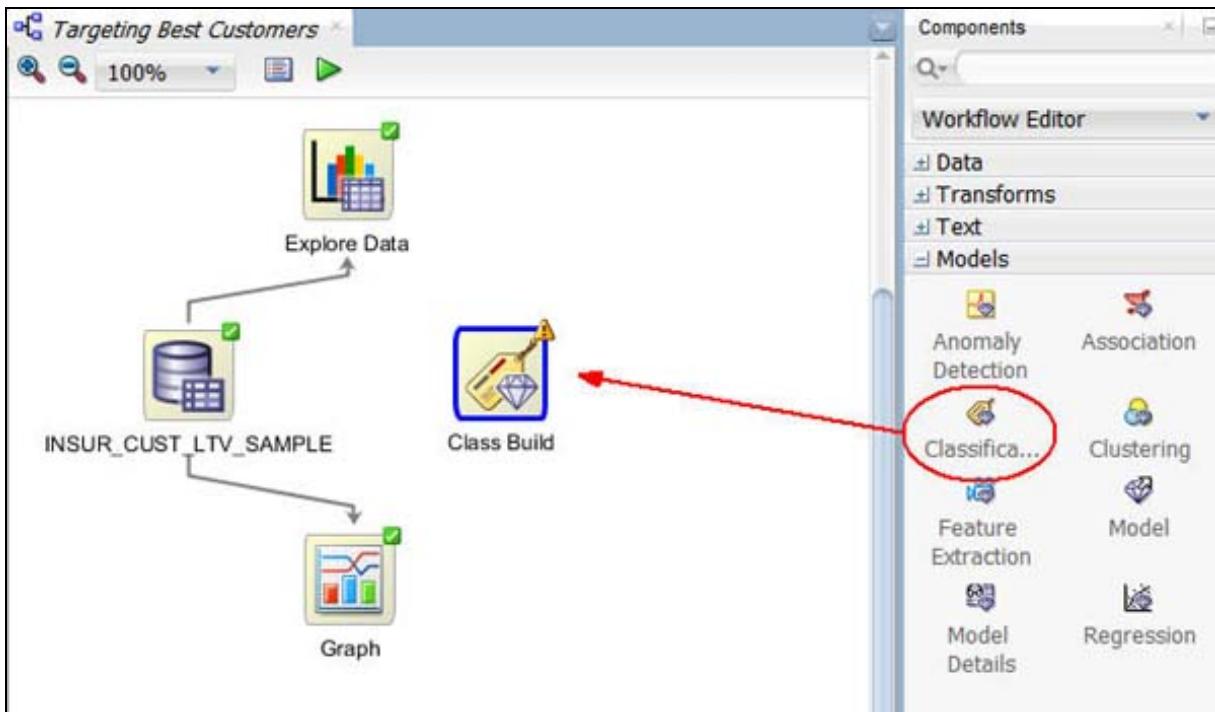
では、次のトピックでは各モデルを実行し検証します。

デフォルトの分類モデルを作成するには、次の手順を実行します:

1. A. はじめに[コンポーネント]で、データカテゴリを折りたたみ、**モデル**カテゴリを展開します:



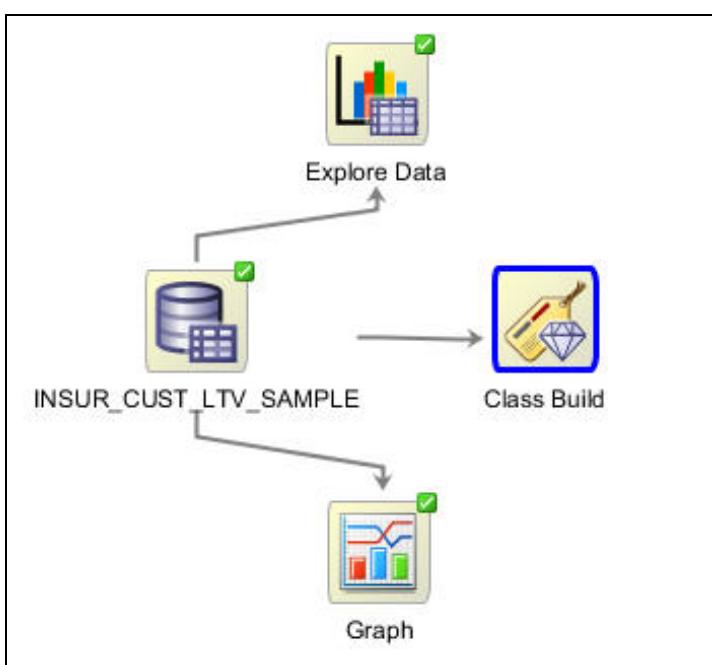
- B. 次に、[コンポーネントタブから]分類ノードをワークフローペインにドラッグ&ドロップします:



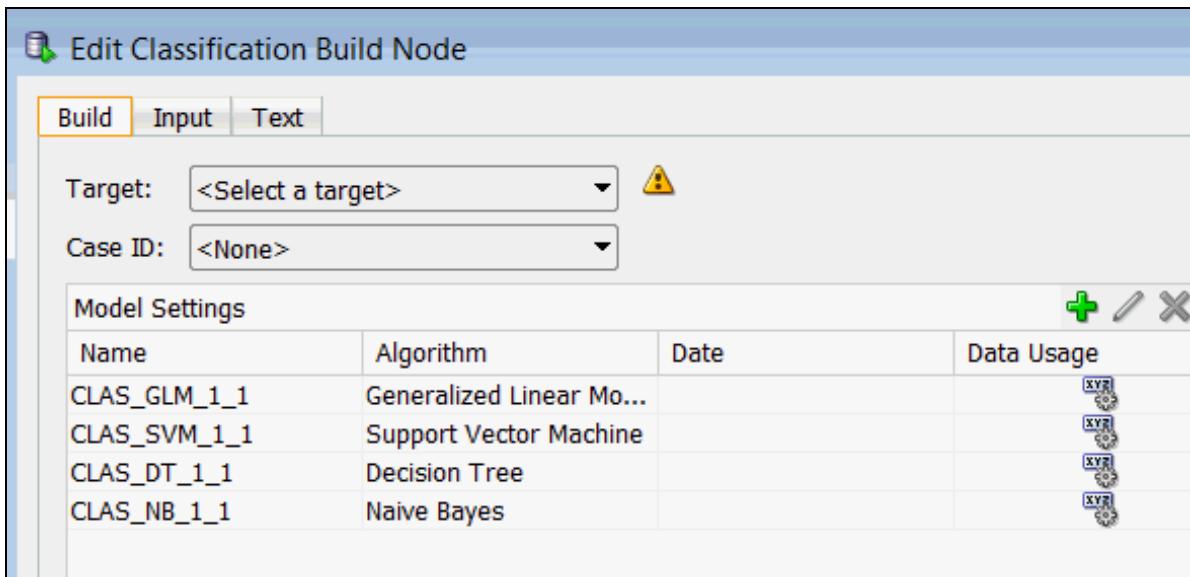
結果: 「分類構築」という名前のノードがワークフローに表示されます:

注:

- 先に述べたように、境界線上の黄色い感嘆符は、ノードが完全になる前により多くの情報の設定が必要なことを示しています
 - この場合、2つのアクションが必要です:
 1. ソースデータノードと分類構築ノードの間に接続を作成する必要があります
 2. 2つの属性を分類構築プロセス用に指定する必要があります
2. まず、先に説明したのと同じように、分類構築ノードにデータソースノードを接続します



結果: 分類ビルド・ノードの編集ウィンドウが表示されます

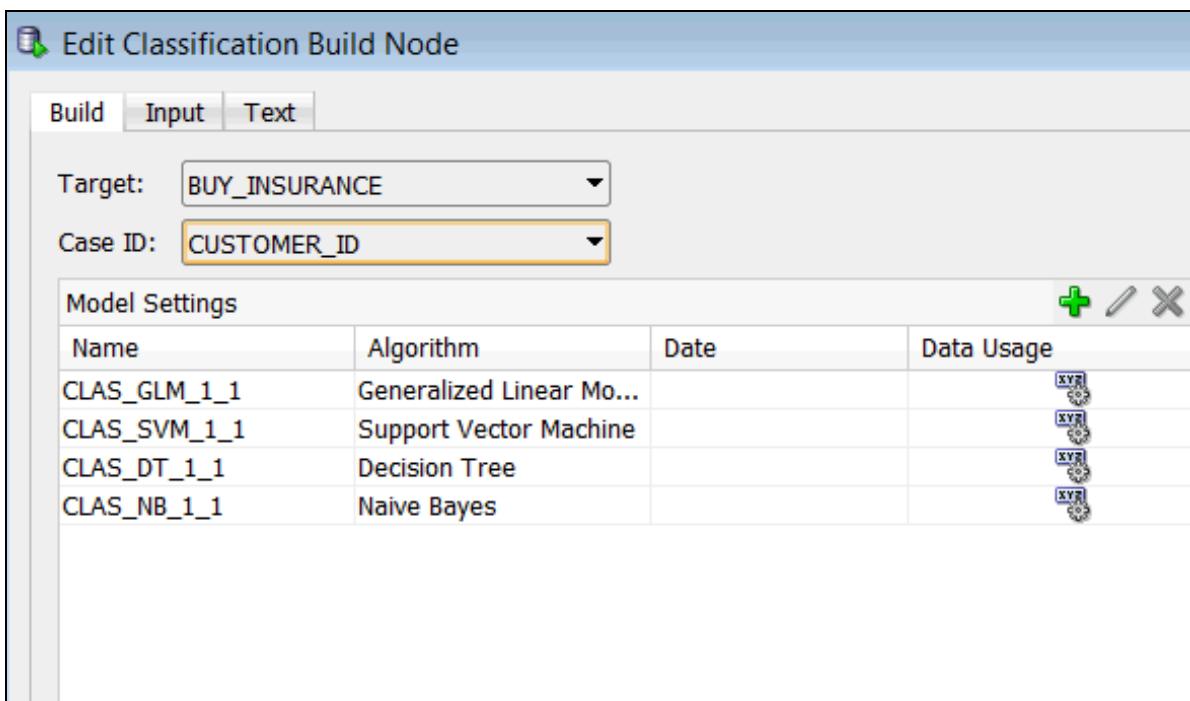


注:

- 黄色い!"!"マークが表示されているターゲットフィールドに注意してください。これは、この項目のために属性を選択しなければならないことを意味します
- 各モデルの名前が自動で生成されますが、この例とは異なる場合があります

3. 分類ビルド・ノードの編集ウィンドウでは:

- ターゲット属性として**BUY_INSURANCE**を選択します
- ケースID属性として**CUSTOMER_ID**を選択します



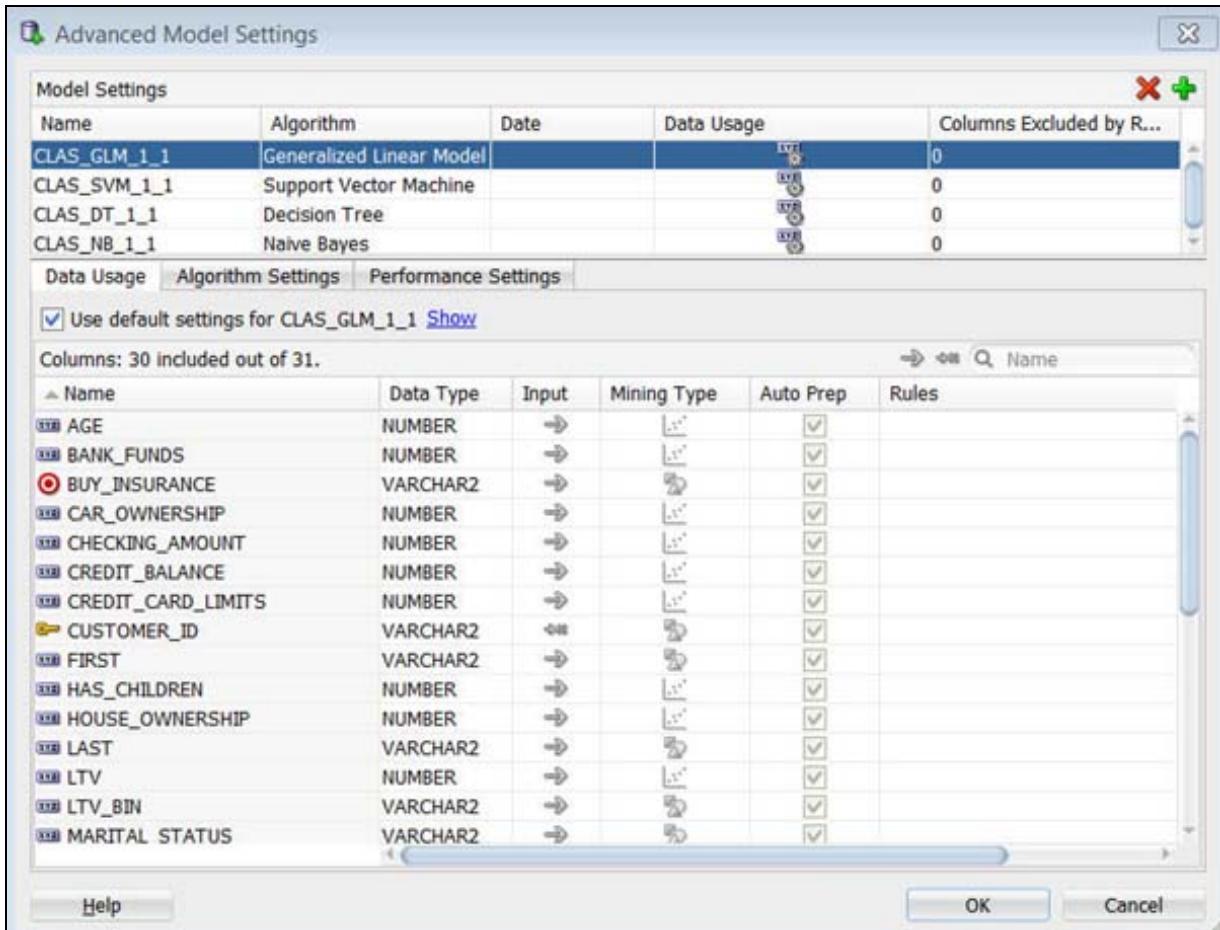
注:

- 必須ではありませんが、各レコードを一意にするためにケースIDを定義することをおすすめします。これはモデルの再現性を支援し、優れたデータマイニングの取り組みと一致するものです
- 先の述べたように、分類モデルのためのすべての4つのアルゴリズムがデフォルトで選択されています。特に指定しない限り、これらは自動的に実行されます。

4. オプションで、任意のアルゴリズムをダブルクリックして、リストされているアルゴリズムの特定の設定を変更できます

- 例えれば以下のように、最初のアルゴリズムをダブルクリックして、詳細モデル設定ウィンドウを表示

します:

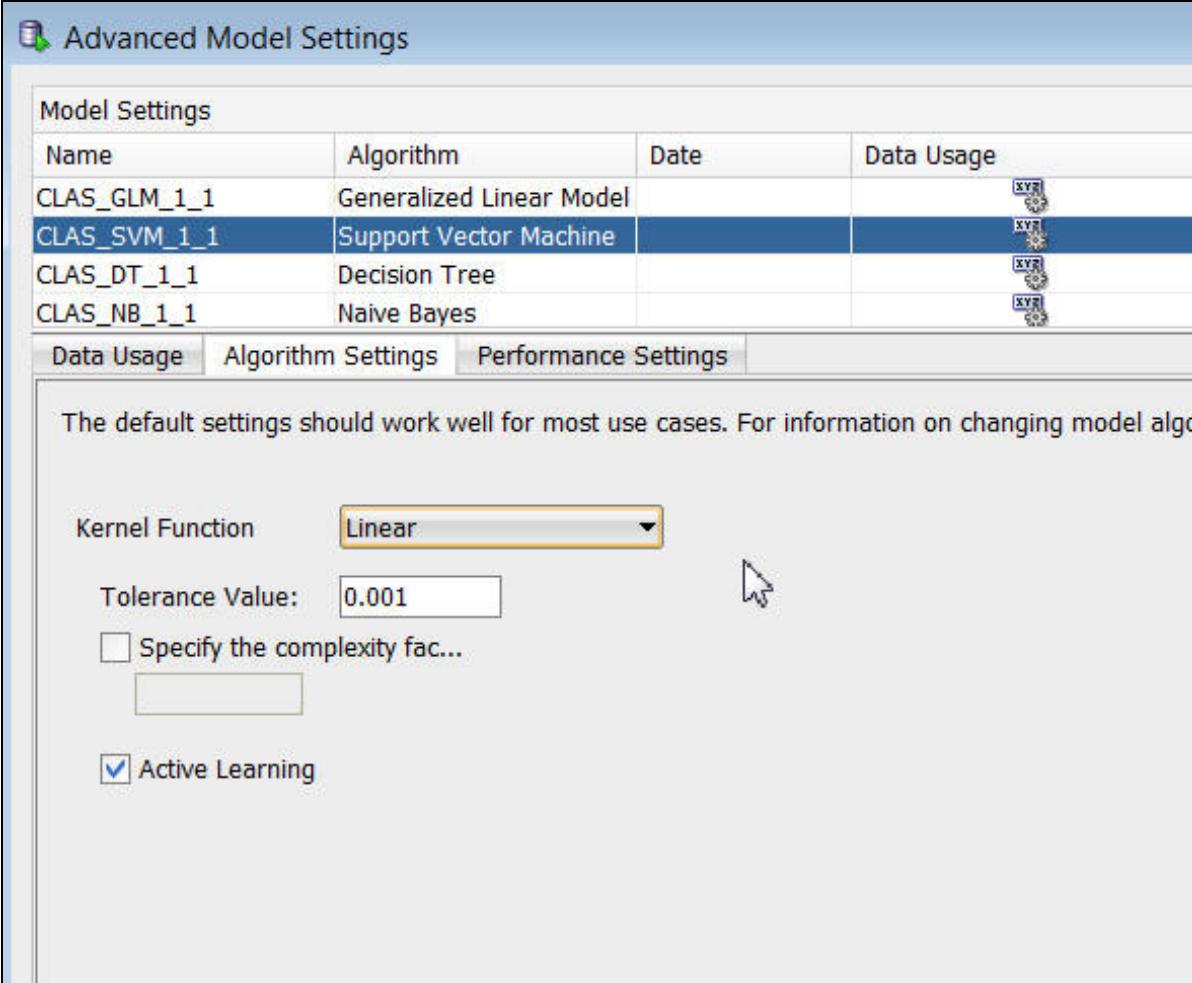


注:

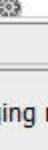
- 詳細モデル設定ウィンドウでは、4つの分類アルゴリズムのそれぞれについて、データの使用方法、アルゴリズム設定、パフォーマンス設定について変更できます
- このウィンドウから赤い"x"もしくは緑の "+" アイコンを用いて、任意のアルゴリズムを選択解除(再選択)できます

B. **Support Vector Machine** アルゴリズムを選択し、**アルゴリズム設定**タブをクリックします

C. 次に、以下のように、カーネル・ファンクションオプションで**線形**を選択します:

Advanced Model Settings

Model Settings

Name	Algorithm	Date	Data Usage
CLAS_GLM_1_1	Generalized Linear Model		
CLAS_SVM_1_1	Support Vector Machine		
CLAS_DT_1_1	Decision Tree		
CLAS_NB_1_1	Naive Bayes		

Data Usage Algorithm Settings Performance Settings

The default settings should work well for most use cases. For information on changing model algo...

Kernel Function: **Linear**

Tolerance Value: **0.001**

Specify the complexity fac...

Active Learning

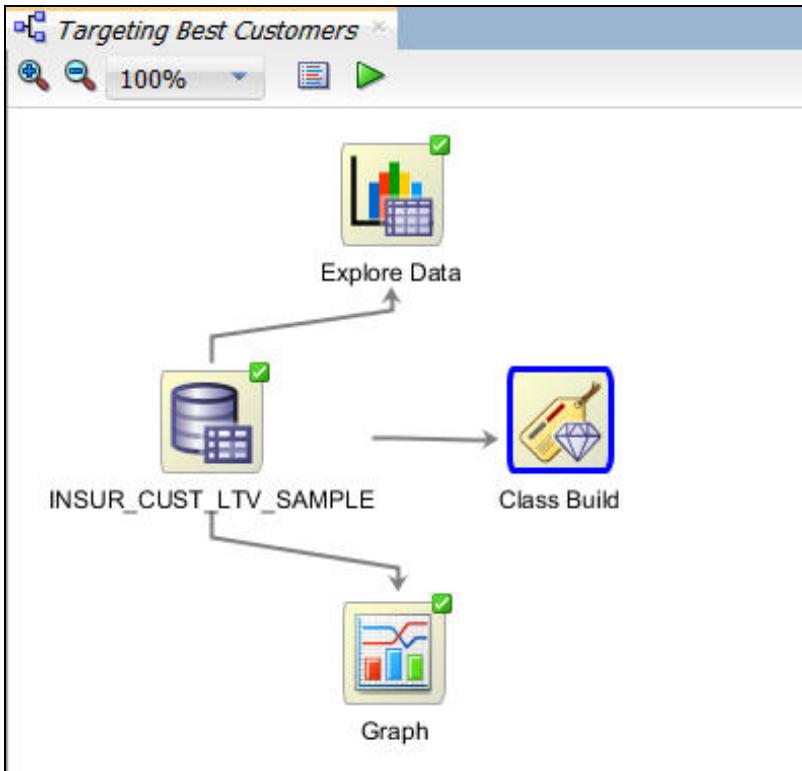
注: モデルの結果をわかりやすく解釈するために、Support Vector Machine (SVM)アルゴリズムのデフォルトの設定であるシステム決定から線形に値を変更しています

D. 各アルゴリズムのほかのタブも気軽に確認してください。しかし、ほかの設定はデフォルトから変更しないでください

E. 確認し終わったら、**OK**をクリックしSVMアルゴリズムの設定を保存し、詳細モデル設定ウィンドウを閉じます

5. 最後に、分類ビルド・ノードの編集ウィンドウで**OK**をクリックし、変更を保存します

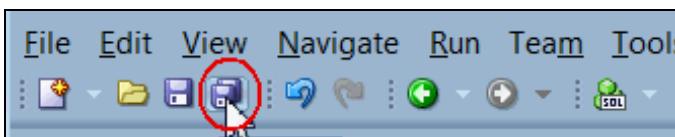
結果: 分類構築ノードを実行する準備が整いました



注: プロパティタブの**モデル**セクションでは、以下のように各線括したアルゴリズムの現在のステータスを参照できます:

Name	Output	Build	Test	Tune	Algorithm	C...
CLAS_GLM_1_1	Not...	Not...	Not...	Auto...	Generaliz...	
CLAS_SVM_1_1	Not...	Not...	Not...	Auto...	Support ...	
CLAS_DT_1_1	Not...	Not...	Not...	Auto...	Decision ...	
CLAS_NB_1_1	Not...	Not...	Not...	Auto...	Naive Bayes	

6. メインツールバーですべて保存をクリックしワークフローを保存します



モデルの構築

このトピックでは、ソースデータを元に選択したモデルを構築します。この操作は「トレーニング」と呼ばれ、このモデルはトレーニングデータから実行するときには「学習」と呼ばれます。

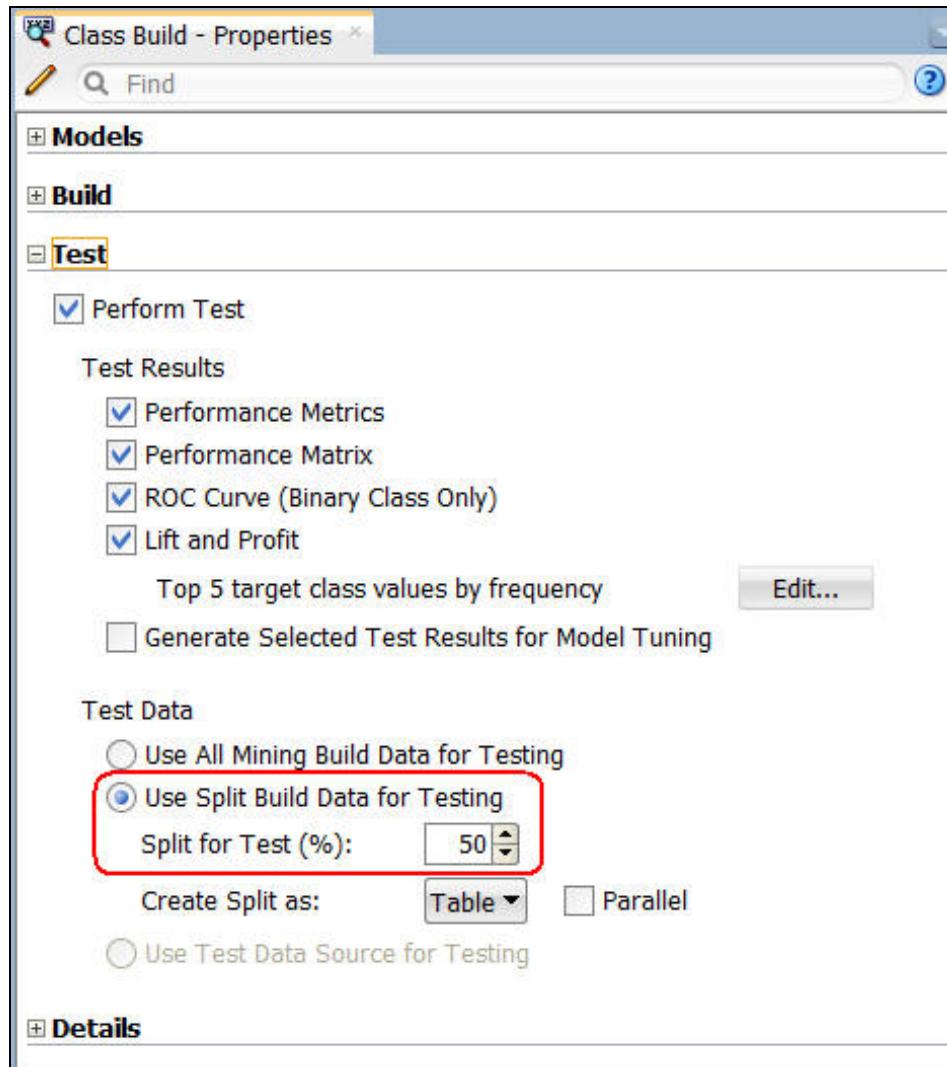
一般的なデータマイニングの実践構築(トレーニング)はソースデータの一部に対して行われ、その後、データの残りの部分に対してモデルをテストします。デフォルトでOracle Data Minerは、40/60に分割したデータを用いたアプローチを用います

モデルを構築する前に、分類構築ノードを選択し、プロパティタブからテストセクションを選択します。テストセクションでは以下を指定することができます:

- 構築プロセス中にテストを実行するかしないか
- どのテスト結果を生成するか

- テストデータの管理をどうするか

テストデータエリアでは以下のように値を**50**に変更してテストデータの分割を50/50に指定します



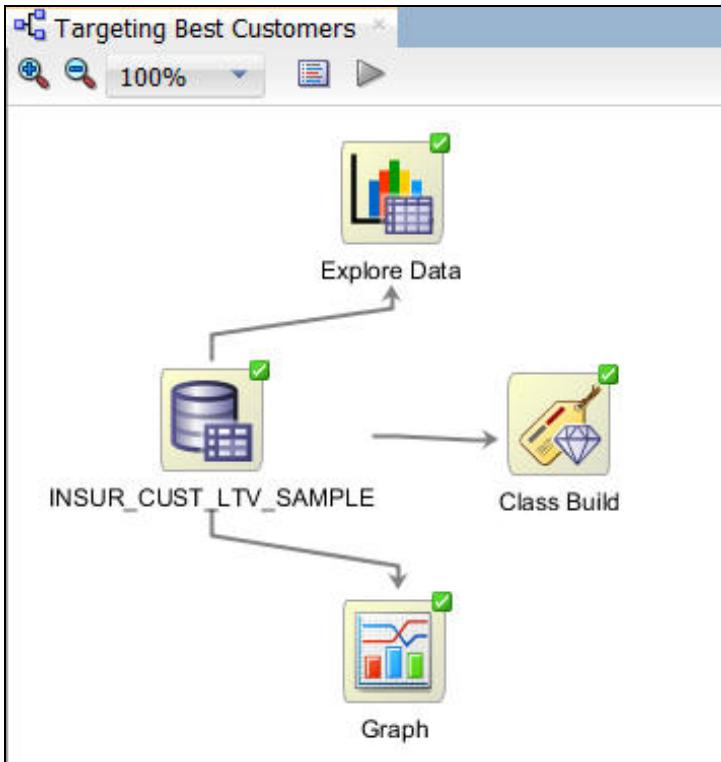
次に、モデルを構築します

1. 分類構築ノードを右クリックし、ポップアップメニューから実行を選択します

注:

- ノードを実行すると、ノードに定義されたすべてのモデルが構築、テストされます
- 前と同様、サーバプロセスが実行中はノードの境界線上に緑のギアアイコンが表示され、ワークフローウィンドウの上部にステータスが表示されます

構築が完了すると、すべてのノードの境界線に緑のチェックが表示されます



また、プロパティ・インスペクタを使用して構築についての情報を確認することができます

2. ワークフローで分類構築ノードを選択し、プロパティタブで**モデルセクション**を選択します

Name	Output	Build	Test	Tune	Algorithm
CLAS_GLM_1_1	→	✓ 7/31...	✓ 7/3...	Auto...	Generaliz...
CLAS_SVM_1_1	→	✓ 7/31...	✓ 7/3...	Auto...	Support V...
CLAS_DT_1_1	→	✓ 7/31...	✓ 7/3...	Auto...	Decision T...
CLAS_NB_1_1	→	✓ 7/31...	✓ 7/3...	Auto...	Naive Bayes

注:

- 4つのモデルすべてが構築に成功している
- モデルはすべて同じターゲット(BUY_INSURANCE)を持つが異なるアルゴリズムを使っている
- ソースデータが自動でテストデータと構築データに分割される

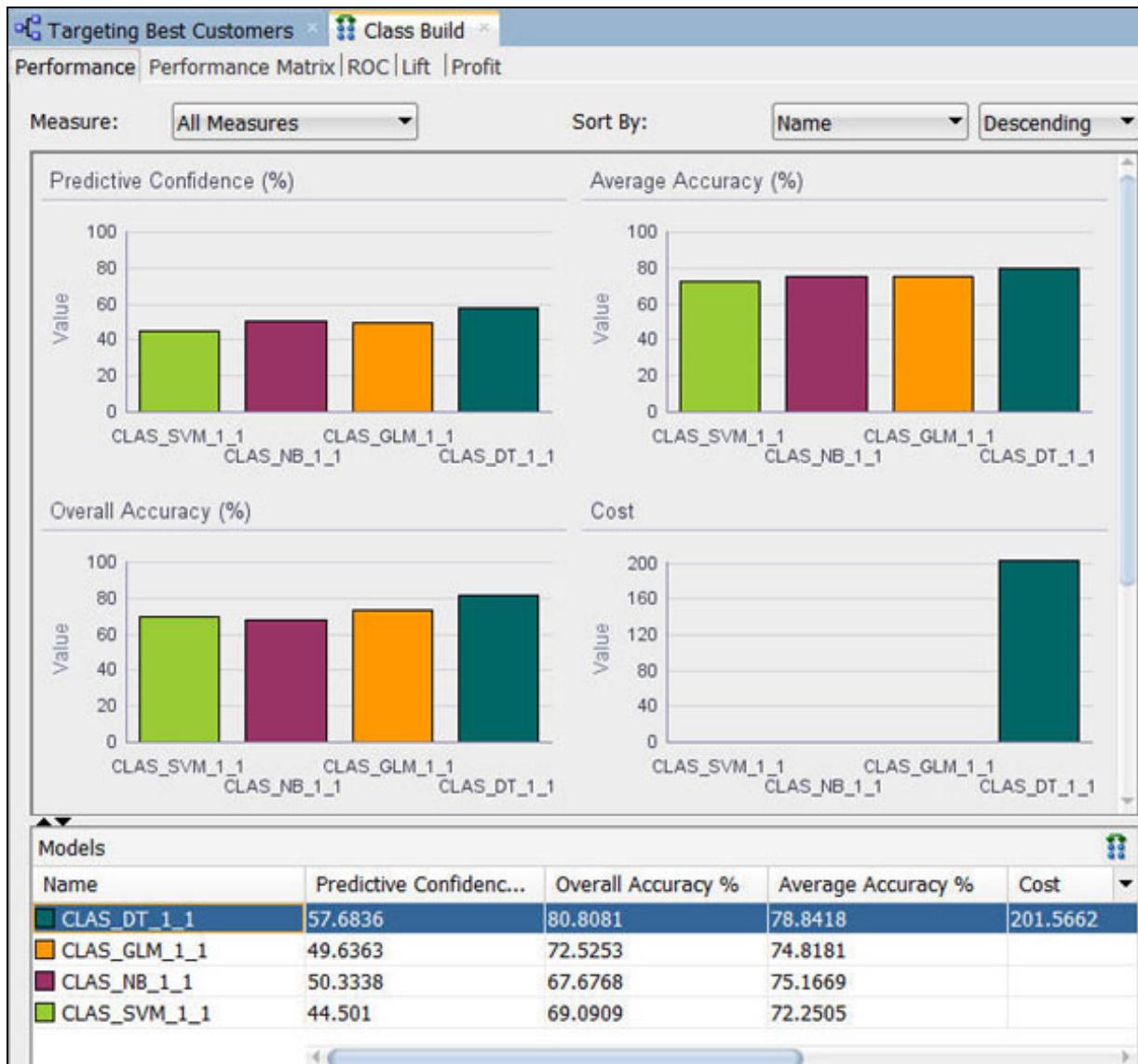
モデルの比較

選択したモデルを構築・トレーニングした後、比較フォーマットで表示してすべてのモデル結果を評価できます。ここでは、4つのすべての分類モデル結果を相対的に比較します。

次の手順に従います:

1. 分類構築ノードを右クリックし、メニューから**テスト結果の比較**を選択します

結果: 分類構築タブが新たに開き、パフォーマンスタブでは、以下のように4つのモデルの比較情報が表示されます:

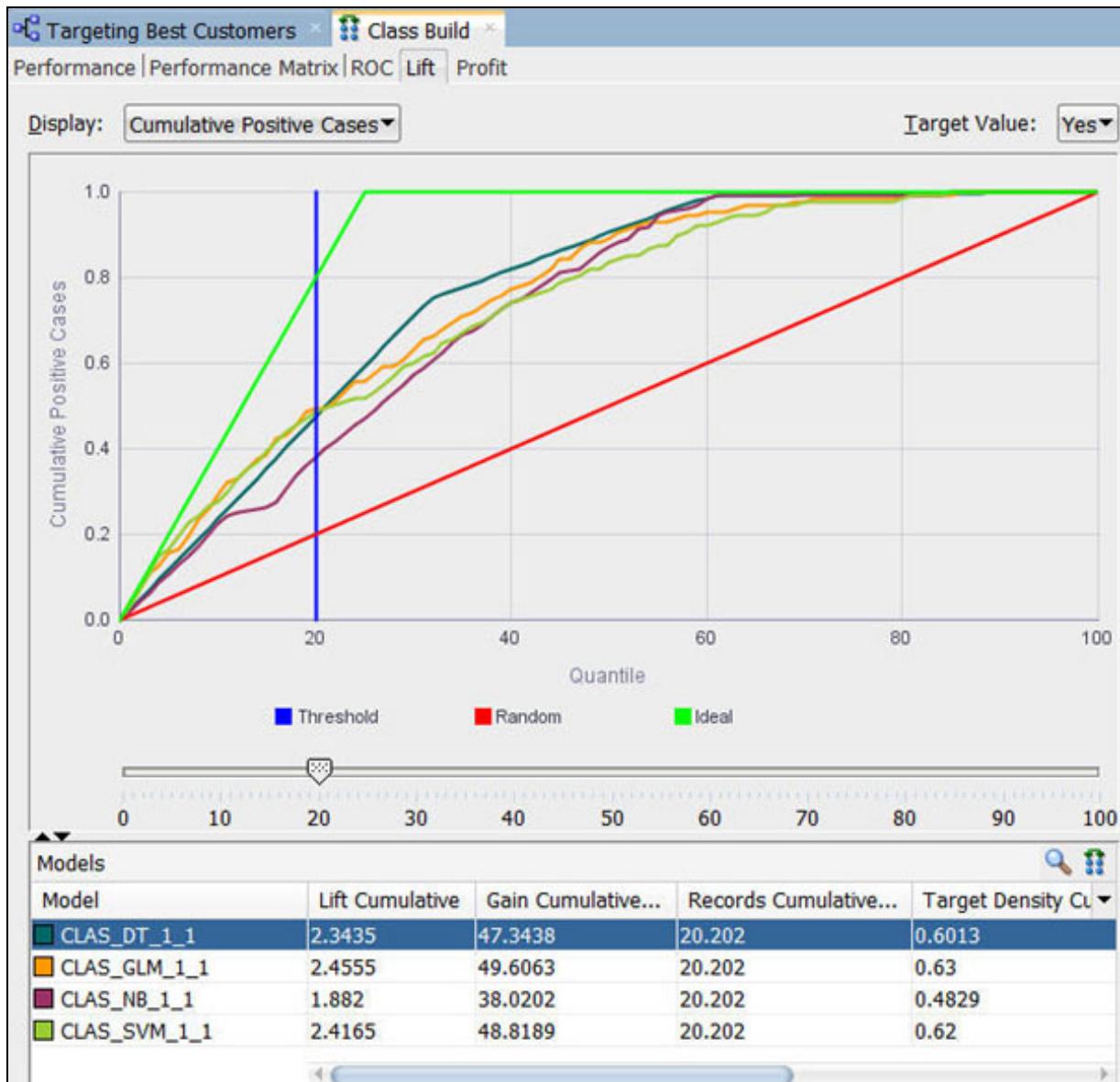


注:

サンプルデータが非常に小さいので、ここで得られる数値はチュートリアルで示すものとは多少異なる場合があります。また、ヒストグラムの色はこの例に示したものと異なっていてもかまいません

- 比較結果には5つのタブが含まれます: パフォーマンス、パフォーマンス・マトリックス、ROC、リフト、利益
- パフォーマンスタブでは、各モデルについての、予測信頼度、平均精度および全体精度という情報を数値およびグラフィカルな情報として提供します
- パフォーマンスタブによると、ディシジョンツリー(DT)モデルが最も高い予測信頼度、全体精度、および平均精度を出しているように見えます。ほかのモデルの結果はまちまちです

2. リフトタブを選択します。そして、グラフ上部右のターゲット値をYesに変更します



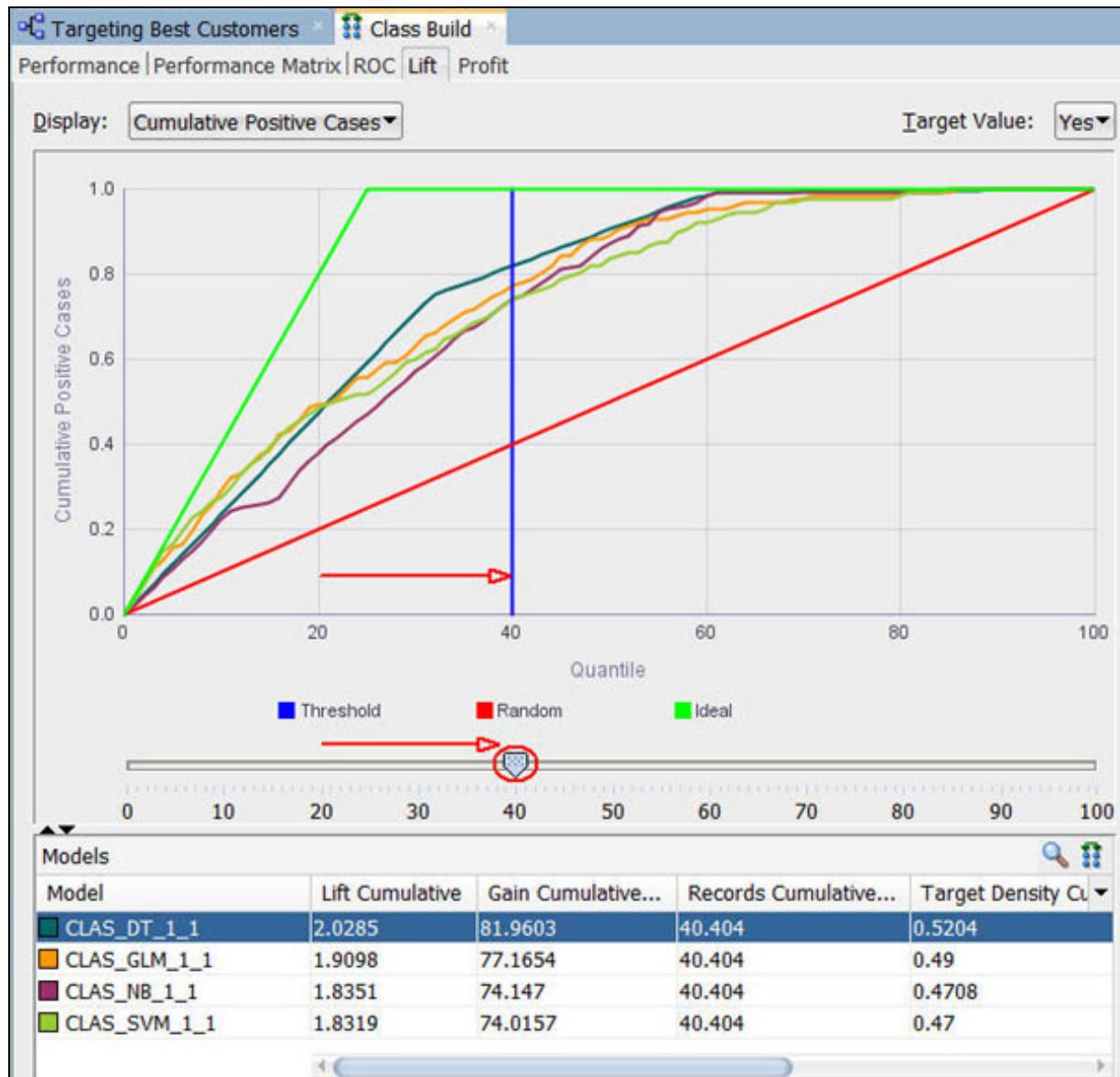
注:

- リフトタブでは、各モデルのリフト値をグラフィカルに表示します。ランダムケースの場合の赤いラインおよびしきい値のための縦の青いラインがあります
- リフトはモデルテストの一種です。実際に正のターゲット値になるものを「高速に」見つけるための評価基準です
- リフトビューアは、各モデル内の指定されたターゲット値のリフト値を比較します
- リフトビューアでは、累積リフトと累積リフトの値を表示します

上記の例では、20分位でDT、一般化線形モデル(GLM)、およびSupport Vector Machine (SVM)のモデルは、累積リフトと累積利益%の値はとても近いものを示しています

リフトタブで、以下のようにスライダツールを用いて、グラフのX軸に沿って分位の測定ラインを移動させることができます。左右に移動させるとモデル内のデータは自動的に更新されます。以下のイメージを参考してください

- 分位単位で移動させると、DTモデルの累積リフトおよび累積利益%は以下のように40分位で他のモデルを追い越します
- 50分位以上になると、SVMモデルが停滞しているのに比べ、NBモデルのリフトと利益が増加しているように見えます。しかし、DTモデルは継続的に高いリフトと利益の値を出してきています



3. 次に、パフォーマンス・マトリックスタブを選択します

The screenshot shows the Oracle Data Miner 4.0 interface with the 'Class Build' tab selected. A performance matrix is displayed, comparing four models based on correct predictions. The models and their values are:

Models	Correct Predictions %	Correct Predictions C...	Total C...
CLAS_GLM_1_1	72.5253	359	495
CLAS_NB_1_1	67.6768	335	495
CLAS_DT_1_1	80.8081	400	495
CLAS_SVM_1_1	69.0909	342	495

Below the matrix, a 'Target Value Details' section is shown, set to 'Correct Predictions'.

注: パフォーマンス・マトリックスでは、DTモデルの正しい予測%の値がほぼ81%に達していて、最も高い値をだしています。次点のGLMは72.5%です

4. GLMとDTモデルの詳細を比較しましょう

まず、モデルのターゲット値の詳細を表示するためにGLMモデルを選択します。各モデルの「ターゲット値」はBUY_INSURANCE属性であることを思い出してください

Targeting Best Customers Class Build

Performance Performance Matrix ROC Lift Profit

Display: Compare Models

Models	Correct Predictions %	Correct Predictions Count	Total
CLAS_GLM_1_1	72.5253	359	495
CLAS_NB_1_1	67.6768	335	495
CLAS_DT_1_1	80.8081	400	495
CLAS_SVM_1_1	69.0909	342	495

Target Value Details

Measure: Correct Predictions

Target Value	CLAS_GLM_1_1
No	70.1087
Yes	79.5276

注: GLMモデルは、保険を購入するという顧客に関して70%の正しい予測を返し、買わないという顧客については79.5%の正しい予測をしています

次に、DTモデルを選択します

Targeting Best Customers Class Build

Performance Performance Matrix ROC Lift Profit

Display: Compare Models

Models	Correct Predictions %	Correct Predictions Count	Total
CLAS_GLM_1_1	72.5253	359	495
CLAS_NB_1_1	67.6768	335	495
CLAS_DT_1_1	80.8081	400	495
CLAS_SVM_1_1	69.0909	342	495

Target Value Details

Measure: Correct Predictions

Target Value	CLAS_DT_1_1
No	82.8804
Yes	74.8031

注: DTモデルでは、保険を購入する顧客についての予測は82.9%、購入しない顧客に関しては74.8%の正しい予測をしています

- 初回の分析をした結果、より深くディシジョンツリー・モデルを検証することにします。分類構築タブ ウィンドウを閉じます

特定のモデルの選択と検証

前回トピックで実行された分析により、ディシジョンツリー・モデルを以降の分析で用いることにします

ディシジョンツリー・モデルを検証するために次の手順を実行します

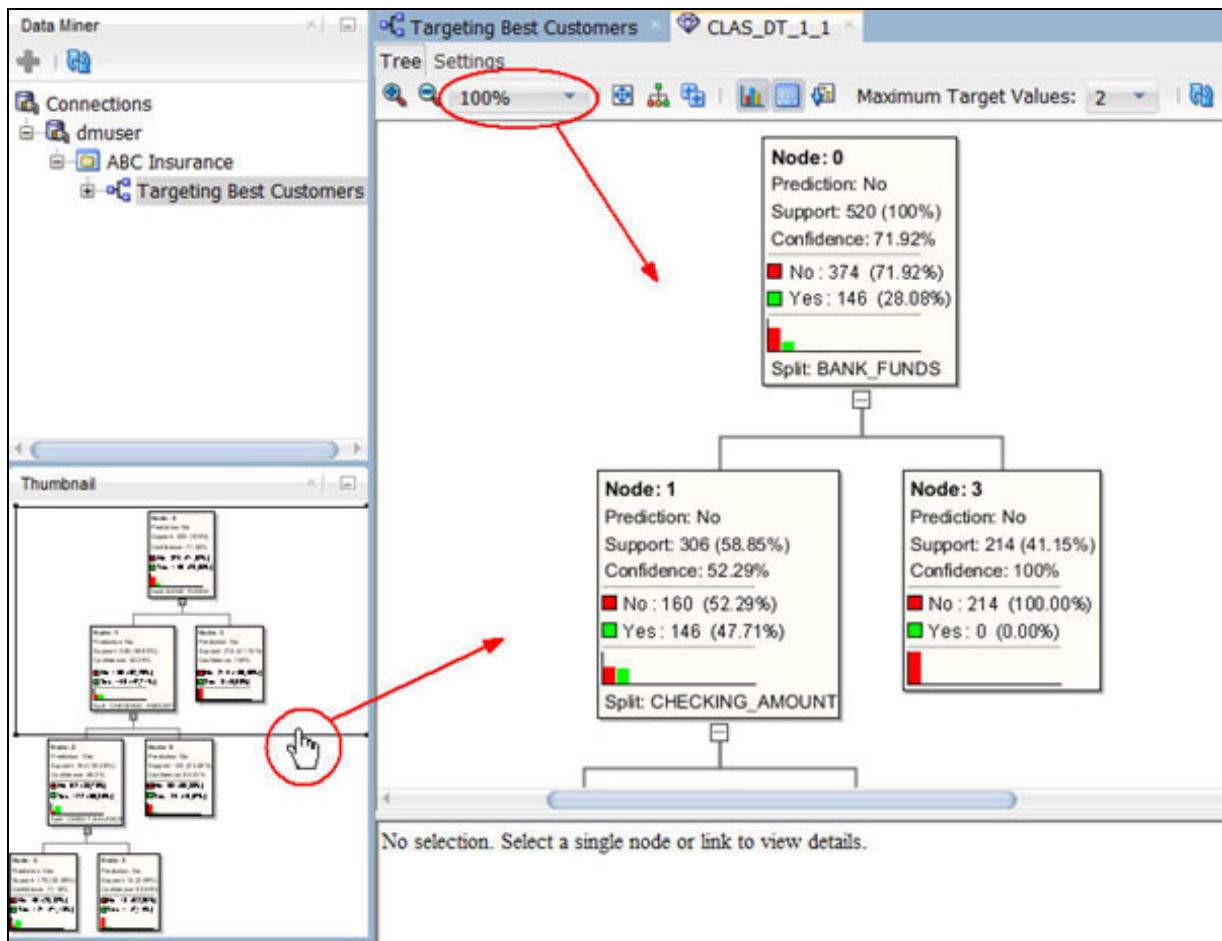
- ワークフローペインに戻り、再び分類構築ノードを右クリックし、**モデルの表示> CLAS_DT_1_1** を選択します。(注: ディシジョンツリー・モデルの名前が異なることがあります)

結果: ウィンドウが開き、ディシジョンツリーがグラフィカルに表示されます

- このインターフェースは、いくつかのナビゲーション表示機能が提供されています:

- サムネイルタブでは、ツリー全体の高レベルなビューを提供しています。たとえば、プライマリ表示ウィンドウ内ではノードをいくつかのみ表示していますが、サムネイルタブでは、このツリーには5つのレベルが含まれることを確認できます
- サムネイルタブのボックスを動かすことでプライマリウィンドウ内のビューを動的に動かすことができます。また、ディシジョンツリー表示内の別の場所を表示するには、プライマリ表示ウィンドウ内のスクロールバーを使用することができます
- 最後に、表示可能なコンテンツのサイズを増加もしくは減少させるためにプライマリ表示ウィンドウのズーム率を変更できます

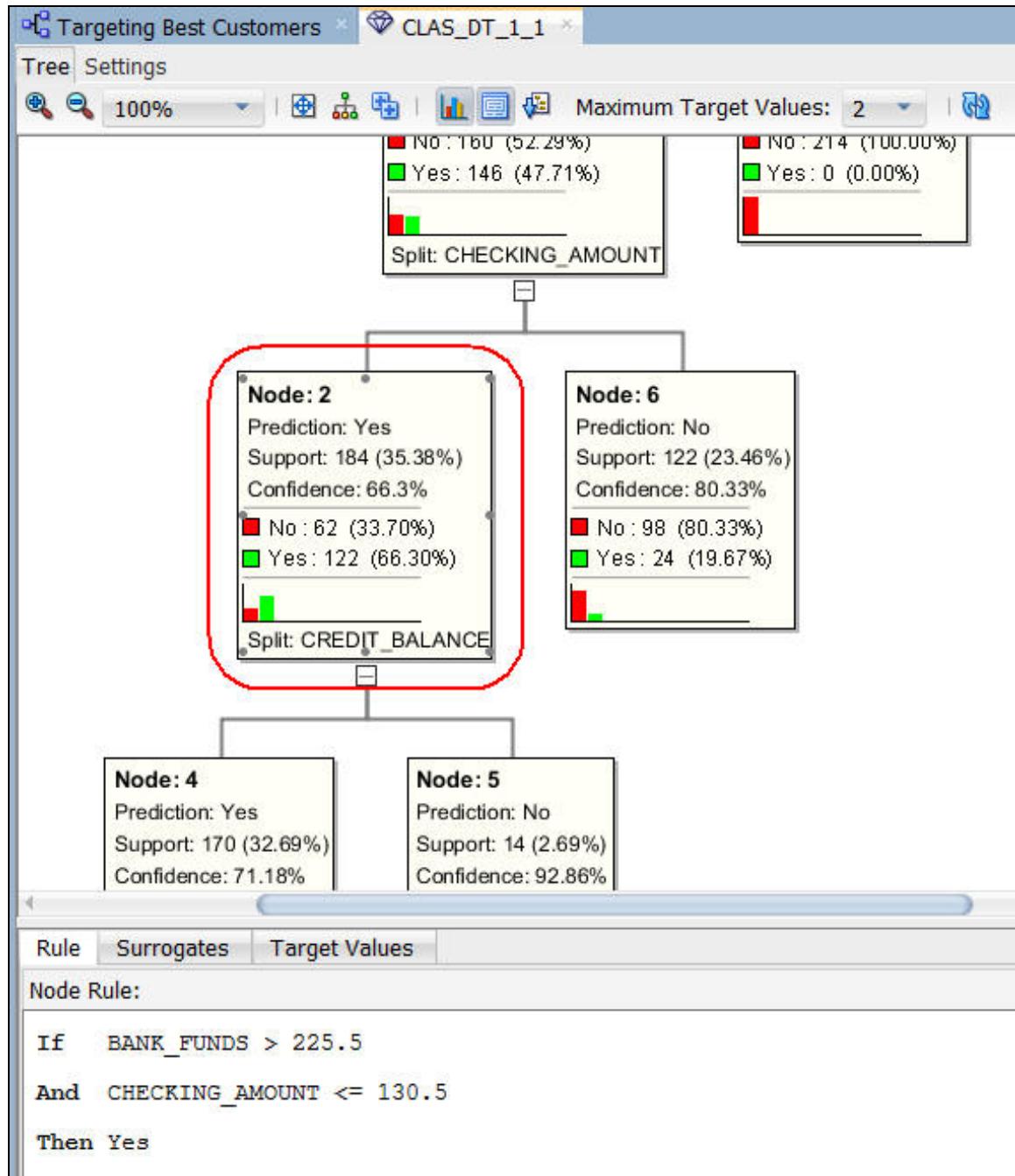
たとえば、ディシジョンツリー表示ウィンドウで100%ズームに設定します



- まず、移動してノード 2を選択します

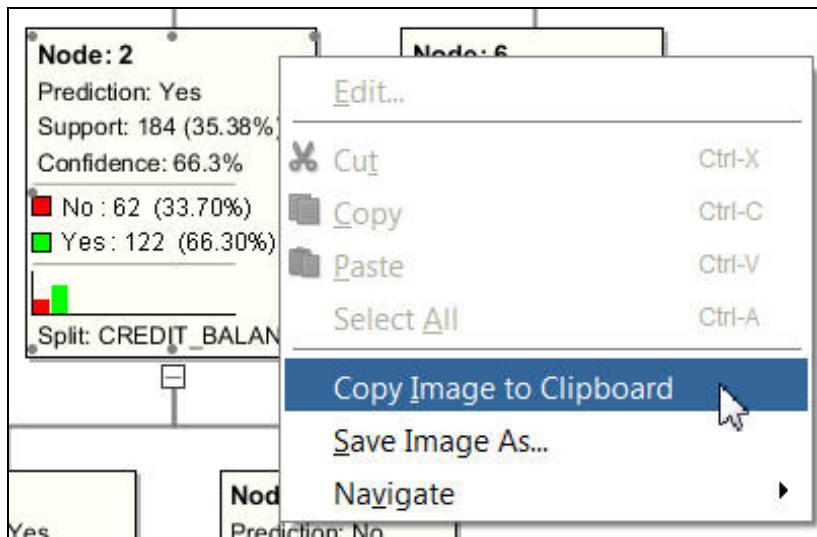
注:

- ディシジョンツリーの各レベルには、IF/THEN文で定義されるルールが表示されます。ツリーにレベルが追加されるには、新たな条件がIF/THEN文で追加されることです
- ツリーの各ノードに対して、ボックスに個々のノードについての要約情報が表示されます
- また、以下のように個々のノードを選択すると、IF/THEN文ルールがルールタブに表示されます
- 一般的に、ディシジョンツリー・モデルは非常に大きなレベルのセットを表示し、また、各レベルのノードにさらにツリーが含まれています。しかし、このレッスンのデータセットは通常のデータマイニングのセットよりもとても小さいのでこのディシジョンツリーもとても小さいです

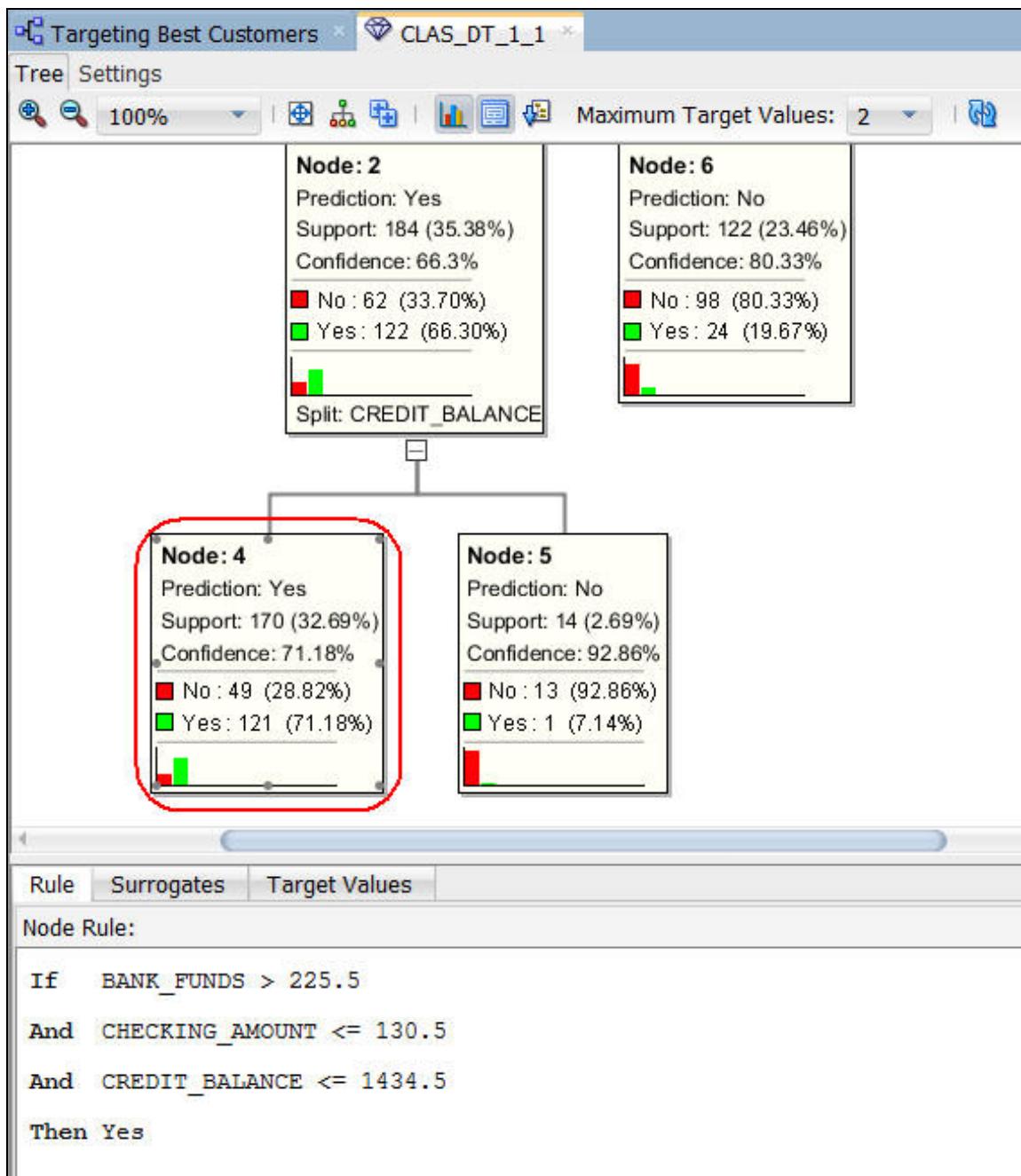


注:

- このレベルは、まずBANK_FUNDS属性で分割し、2番目にCHECKING_AMOUNT属性で分割しています
- ノード2は、BANK_FUNDSの値が225.5より大きく、CHECKING_AMOUNTの値が130.5より小さい場合に、66.3%の確率でこの条件の顧客が保険を購入するであろうと予測しています
- Data Miner 4.0を用いると、UIの画像からチャートイメージをコピーして貼り付けることができます。そして、この画像を別の文書にも貼り付けることができます。たとえば、今回はディシジョンツリーのノード2を選択し、クリップボードにコピーします。



4. 次に、ツリーのレベルの下方のノード4を選択します

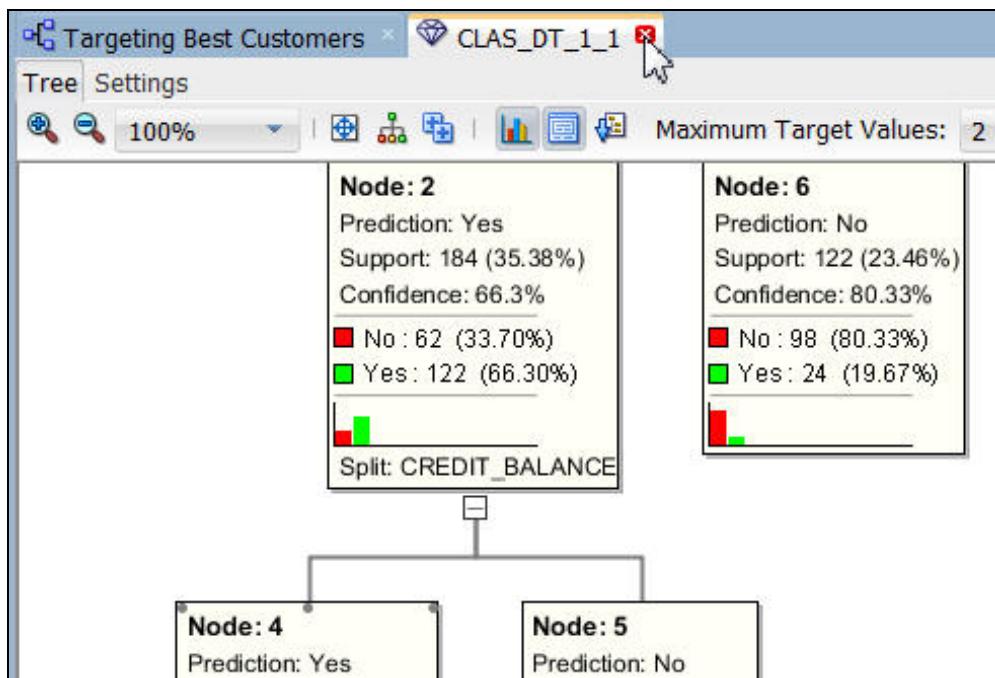


注:

- このツリーの下位レベルでは、最終的にはCREDIT_BALANCE属性で分割されます

- このノードは、BANK_FUNDSの値が225.5より大きく、CHECKING_AMOUNTの値が130.5に等しく、CREDIT_BALANCEの値が1434.5に等しい場合、71%の確率で顧客が保険を購入することを予測しています

5. 以下にあるように、ディシジョンツリーの表示タブを閉じます:



モデルの適用

このトピックでは、ディシジョンツリー・モデルを適用し、結果表示用の表を作成します。モデルを「適用」し、保険を購入する可能性がある顧客を予測します。.

モデルを適用するには、次の手順を実行します:

- まず、分類構築ノード内から必要なモデル(複数でも可能)を指定します
- 第二に、ワークフローに新規データソースノードを追加します。(このノードは「適用」するためのデータとします)
- 第三に、ワークフローに「適用」ノードを追加します
- 次に、分類構築ノードと新規データソースノードをそれぞれ適用ノードに接続します
- 最後に、モデルから予測結果を得るために適用ノードを実行します

モデルを適用し、結果を表示するには、以下の手順を実行します:

- ワークフロー上で、分類構築ノードを選択します。次に、[プロパティ]タブからモデルセクションを表示し、DTモデル以外のモデルの選択を解除します。

モデルの選択を解除するには、各モデルの出力列にある大きな緑色の矢印をクリックします。この動作により列に小さな赤い"x"が追加され、次に構築では使用されないことを示します

作業が終了したら、プロパティタブのモデルタブは以下のようになります:

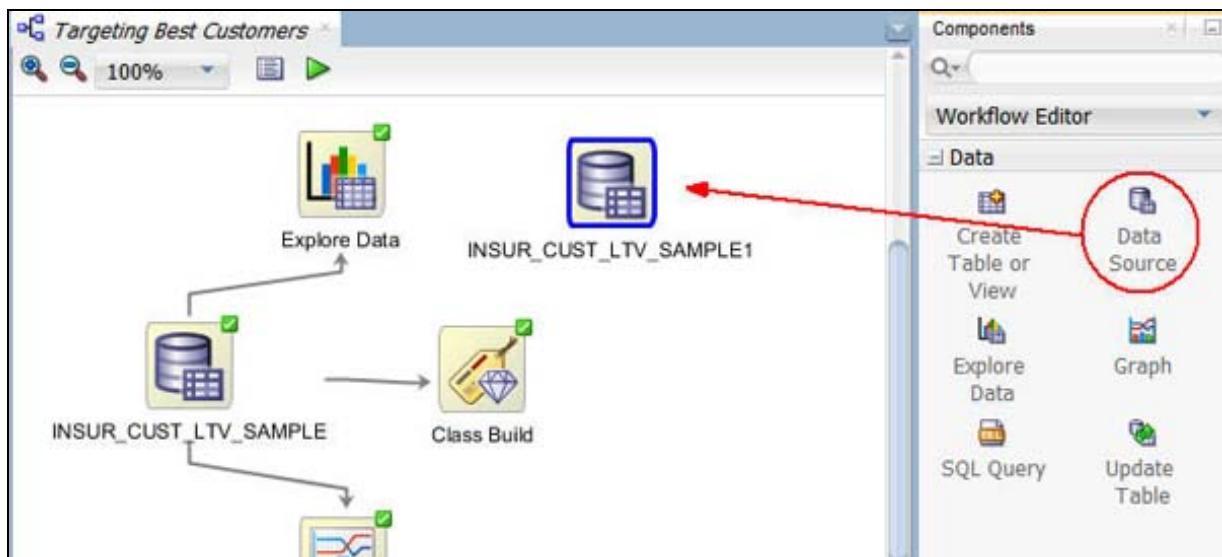
Name	Output	Build	Test	T...	Algorithm	Co...
CLAS_GLM_1_1	xx	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	7/3...	Aut...	Generaliz...
CLAS_SVM_1_1	xx	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	7/3...	Aut...	Support ...
CLAS_DT_1_1	→	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	7/3...	Aut...	Decision ...
CLAS_NB_1_1	→x	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	7/3...	Aut...	Naive Ba...

注: この場合、DTモデルだけが後続のノードに処理が渡されます

2. 次に、ワークフローに新たなデータソースノードを追加します。注: 「適用」用のデータソースとして同じ表を使う場合であっても、ワークフローに2番目のデータソースノードを追加する必要があります

A. 以下のように、[コンポーネント]タブの[データ]カテゴリからワークフローキャンバスに、データソースノードをドラッグ&ドロップします。自動的にデータ・ソースの定義ウィザードが開きます

B. データ・ソースの定義ウィザードで、**NSUR_CUST_LTV_SAMPLE** 表を選択し終了をクリックします



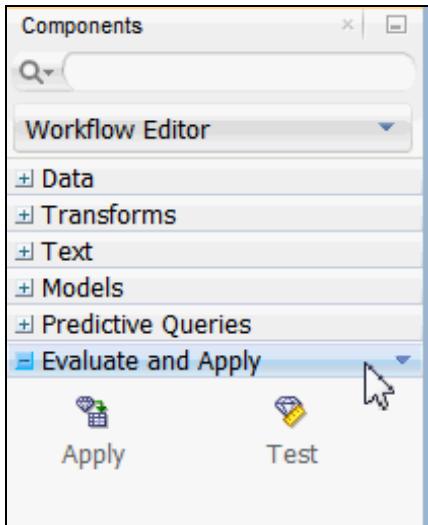
結果: **INSUR_CUST_LTV_SAMPLE1**という名前の新規データソースノードが、ワークフローキャンバスに表示されます

3. 新しいデータソースノードを選択し、プロパティタブの詳細セクションを使って、以下のようにノード名を**INSUR_CUST_LTV_APPLY**に変更します:

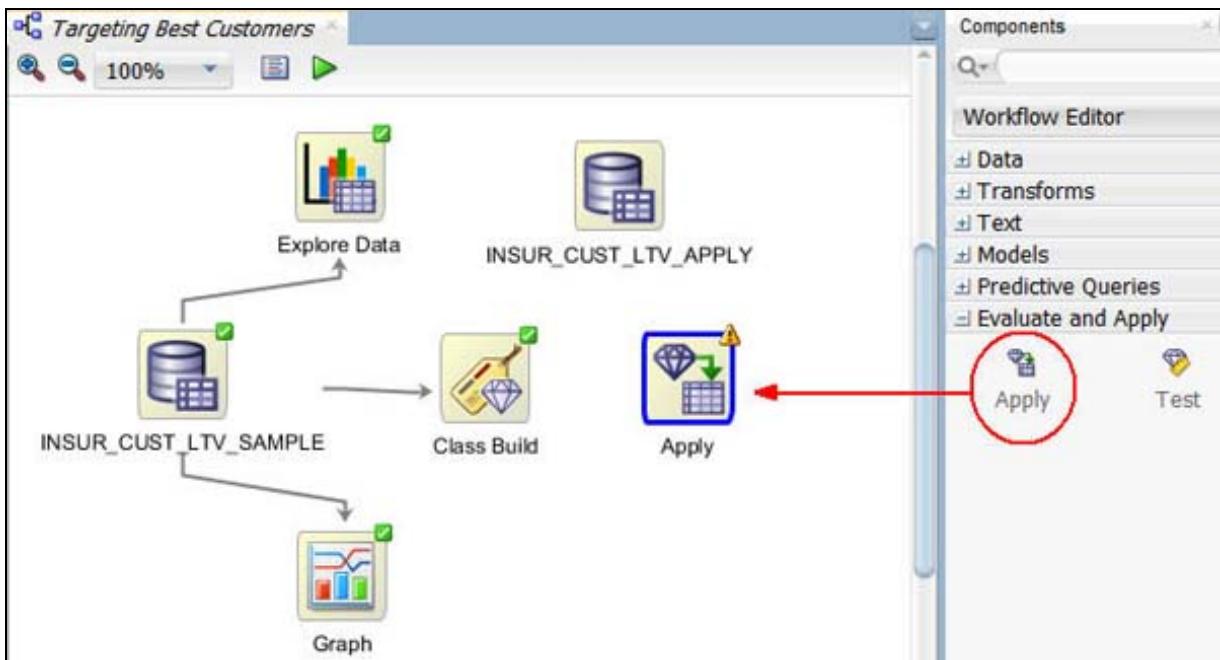
Node Name:	INSUR_CUST_LTV_APPLY
Node Comment:	

結果: 新規表名がワークフロー上にも反映されます

4. 次に、コンポーネントタブ内の評価と適用カテゴリを展開します



5. 以下のように、適用ノードをワークフローキャンバスにドラッグ&ドロップします:

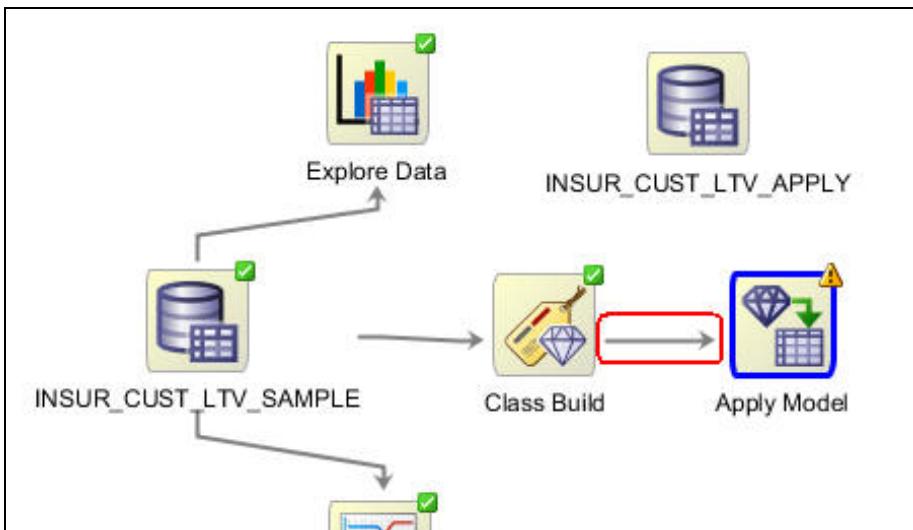


注: ノードの境界線上の感嘆符は、適用ノードが実行される前に多くの情報が必要であることを示します

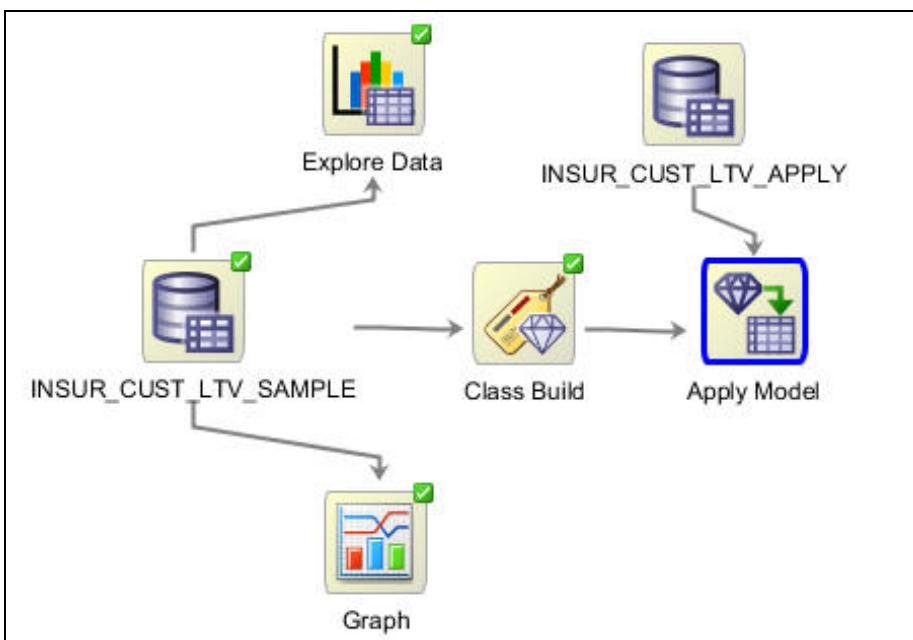
6. プロパティ・インスペクタの詳細タブを用いて、適用ノードの名前を**Apply Model**に変更します

7.

8. 以下に記載されている方法で、分類構築ノードをApply Modelノードに接続します



9. 次に、INSUR_CUST_LTV_APPLY ノードをApply Modelノードに接続します:



注:

- 2つめのリンクが完了すると、Apply Modelノードの境界線上の黄色い感嘆符の表示が消えます
- これはノードが実行する準備が出てきていることを示しています

10. 適用するモデルノードを実行する前に、結果のアウトプットについて検討します。デフォルトでは、各顧客のための情報のための2つの列を作詞します:

- 予測値(YesまたはNo)
- 予測の確率

しかし、本当に特定の顧客と予測された情報を関連づけることができるよう、各顧客の情報を理解してください

この情報を取得するには、適用されたアウトプットに3つめの列: CUSTOMER_ID を追加する必要があります。アウトプットに顧客IDを追加するには次の手順に従います:

A. Apply Modelノードを右クリックし、編集をクリックします

結果: 適用ノードの編集ウィンドウが表示されます。予測、予測確率および予測コストの列が予測タブに自動的に追加定義されています。

Edit Apply Node

Predictions Additional Output

Automatic Settings

Output Apply Columns

Column	Function	Parameter(s)	Model
CLAS_DT_1_1_PRED	Prediction		CLAS_DT_1_1
CLAS_DT_1_1_PROB	Prediction Probability	Prediction: <Most Likely>	CLAS_DT_1_1
CLAS_DT_1_1_PCST	Prediction Cost	Prediction: <Most Likely>	CLAS_DT_1_1

B. 追加出力タブを選択し、以下のように緑の"+"ボタンをクリックします:

Edit Apply Node

Predictions Additional Output

Output Data Columns

Column	Alias	Data Type
--------	-------	-----------

+ (Green plus sign button)

C. 出力データ列の編集ダイアログが表示されます:

- 使用可能な属性リストから**CUSTOMER_ID**を選択します
- シャトルコントロールを使用して、選択した属性リストに移動します
- 最後に**OK**をクリックします

Edit Output Data Column Dialog

Available Attributes

Name	Data Type
MARITAL_STATUS	VARCHAR2
STATE	VARCHAR2
CREDIT_BALANCE	NUMBER
TIME_AS_CUST...	NUMBER
MORTGAGE_AM...	NUMBER
BANK_FUNDS	NUMBER
N_OF_DEPENDE...	NUMBER
HAS_CHILDREN	NUMBER
SALARY	NUMBER
SEX	VARCHAR2

Selected Attributes

Name	Data Type
CUSTOMER_ID	VARCHAR2

Buttons: Help, OK (circled), Cancel

結果: 以下に表示されているように、CUSTOMER_ID列が追加出力タブに追加されています:

The screenshot shows the 'Edit Apply Node' dialog box. At the top, there are two tabs: 'Predictions' (which is selected) and 'Additional Output'. Below the tabs is a section titled 'Output Data Columns' containing a table. The table has three columns: 'Column', 'Alias', and 'Data Type'. There is one row in the table with the value 'CUSTOMER_ID' in the 'Column' column, an empty 'Alias' column, and 'VARCHAR2' in the 'Data Type' column. At the bottom of the dialog, there is a 'Default Column Order:' dropdown menu set to 'Data Columns First'.

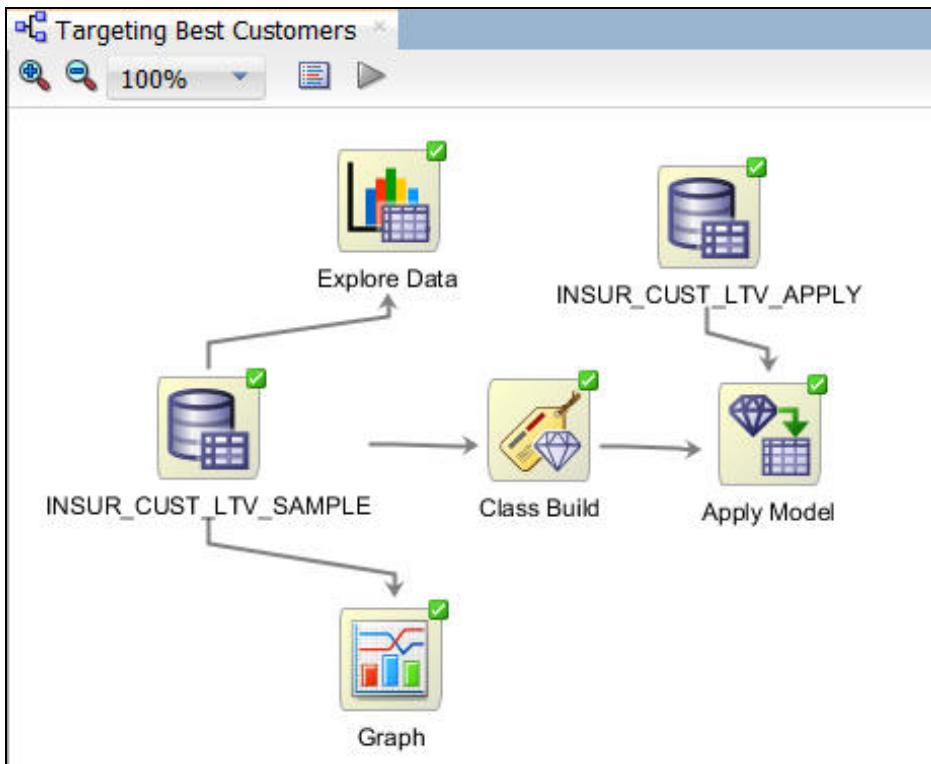
また、デフォルト列順序がデータ列が先に設定されており、適用列が後に配置されます。希望する場合、この順序を切り替えることができます

D. 最後に、変更を適用するために適用ノードの編集ウィンドウで**OK**をクリックします

11. これで、モデルを適用する準備ができました。Apply Modelノードを右クリックしメニューから**実行**を選択します

結果:前と同様、ワークフロードキュメントが自動的に保存され、実行中は小さな歯車アイコンがそれぞれのノードに表示されます。また、実行ステータスはワークフローペインの上部に表示されています

処理が完了すると、サーバプロセスが正常に完了したことを示すために、すべてのワークフローノードに緑のチェックマークアイコンが表示されます

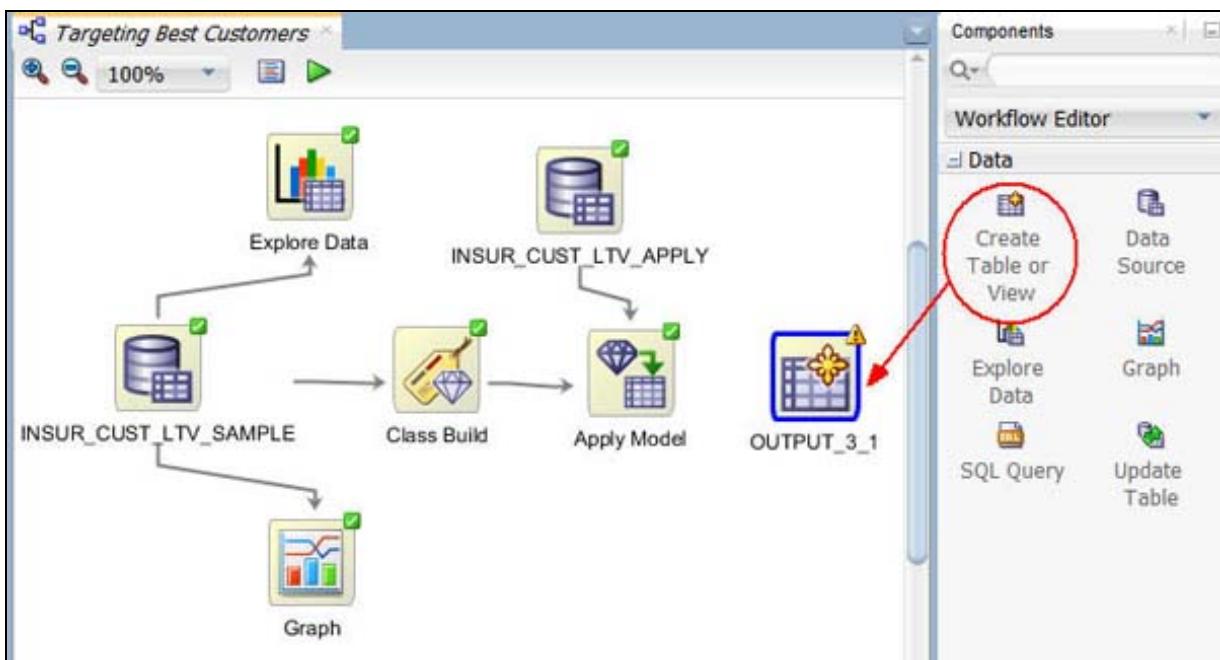


12. 必要に応じて、モデルの予測結果（「適用」の結果）を格納するデータベース表に作成することができます

この表は任意の目的に利用できます。たとえば、アプリケーションにこの表から予測値を組み込むことができ、顧客への割引レター等やその他の適切なアクションのための示唆を与えてくれます

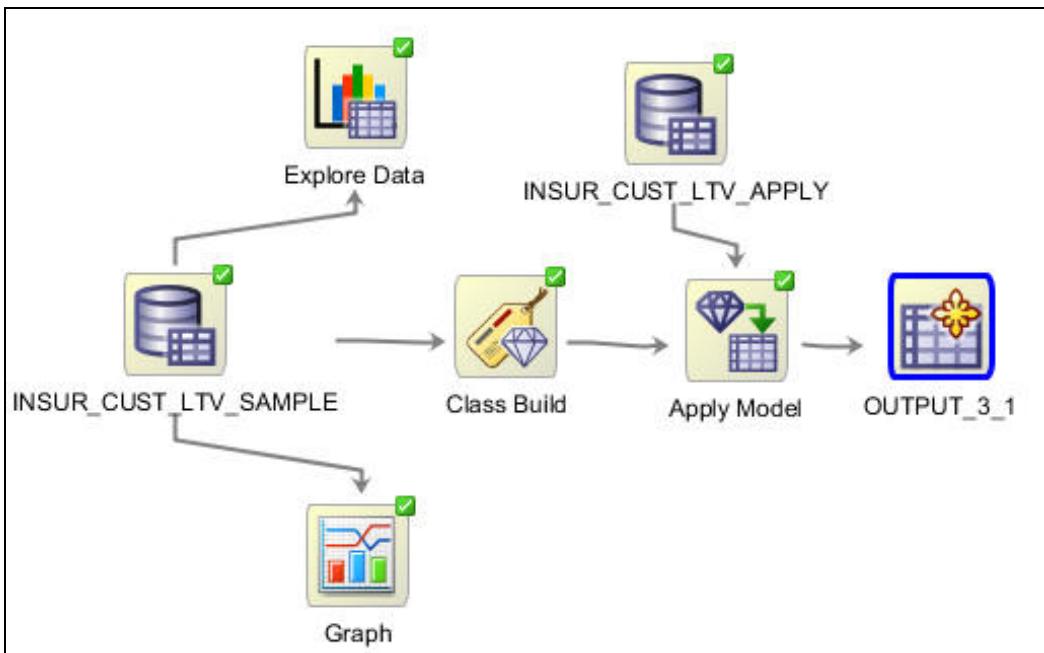
モデルの予測結果の表を作成するには、次の手順を実行視します:

- A. 以下のようにコンポーネントペインのデータカテゴリから、**表またはビューの作成**をワークフローウィンドウにドロップします:



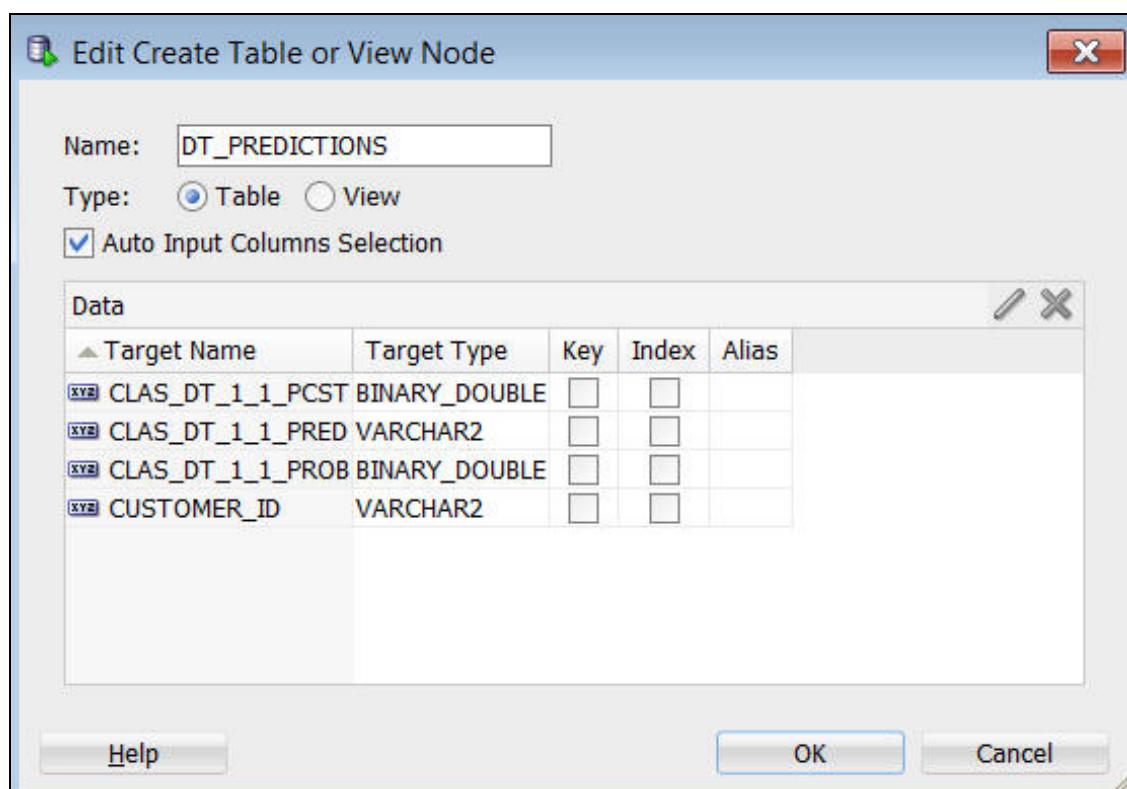
結果: OUTPUTノードが作成されます (OUTPUTノードの名前は以下の例とは異なる場合があります)

- B. **Apply Model** ノードを**OUTPUT**ノードに接続します



C. 作成される表の名前を指定するには以下の手順を実行します(そうしないと、Data Minerはデフォルトの名前で表を作成します):

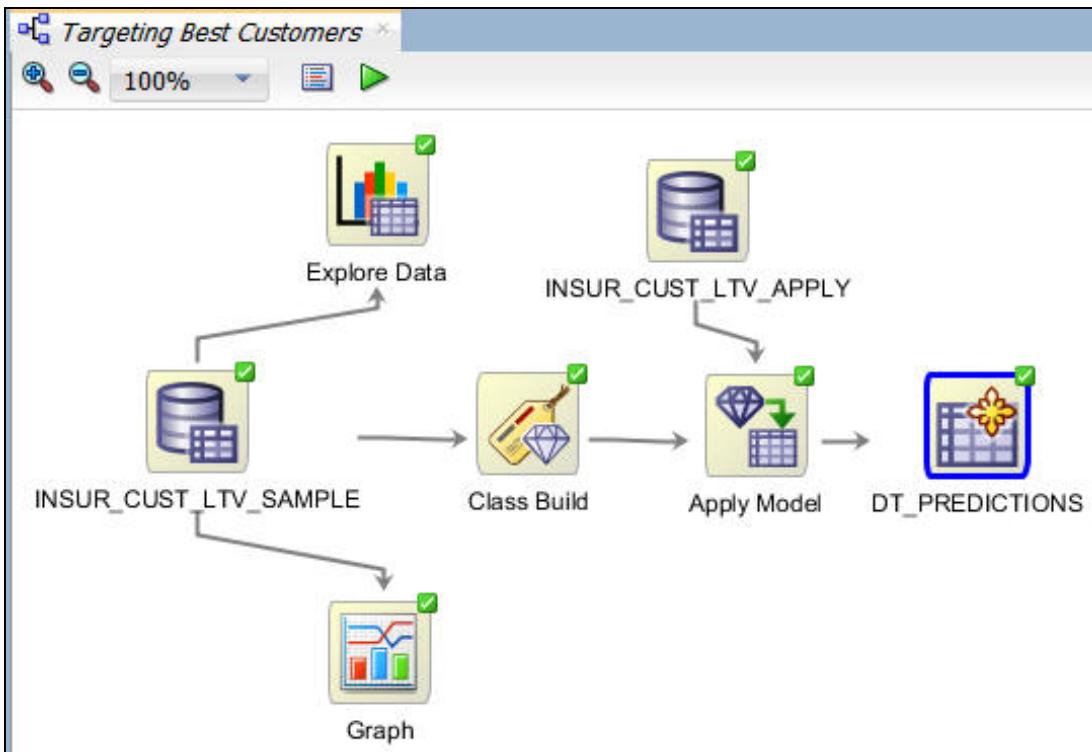
1. OUTPUTノードを右クリックし、メニューから**編集**を選択します
2. 表またはビュー作成ノードの編集ウィンドウで、以下のようにデフォルトの表名から**DT_PREDICTIONS**に名前を変更します:



3. 次に**OK**をクリックします

D. 最後に、DT_PREDICTIONSノードを右クリックし、メニューから**実行**を選択します

結果: プロセスが実行されると、ワークフロードキュメントが自動的に保存されます。完了すると、すべてのノードに以下のように緑のチェックマークがつきます:



注: アウトプットノード(DT_PREDICTIONS)を実行すると、スキーマ内に表が作成されます

13.

14. 結果を表示するには:

A. DT_PREDICTIONS表ノードを右クリックし、メニューから**データの表示**を選択します

結果: 新しいタブに表の内容が表示されます:

- 表には、3つの予測データと顧客ID列の4つの列が含まれます
- 次に示すように、ソートボタンを使用して任意の列を元に表を並べ替えることができます
- ここでは、以下のようにソートされます:
 - まず、予測結果値(CLAS_DT_1_1_PRED)を降順で選択します (保険を購入するという予測結果が"Yes"の列が最初にくることを意味します)
 - 次に、予測確率(CLAS_DT_1_1_PROB)を降順で選択します(表表示の一番上は予測確率の高いものになることを意味します)

The screenshot shows the Oracle Data Miner interface with a 'Sort...' dialog box overlaid. The dialog box is titled 'Select columns to sort by'. It contains two lists: 'Available Columns' on the left and 'Selected Columns' on the right. In the 'Available Columns' list, 'CLAS_DT_1_1_PCST (Asc)' and 'CUSTOMER_ID (Asc)' are shown. In the 'Selected Columns' list, 'CLAS_DT_1_1_PRED (D)' and 'CLAS_DT_1_1_PROB (D)' are selected. To the right of the lists are sorting options: 'Ascending' (radio button), 'Descending' (radio button, highlighted with a red box), and 'Nulls First' (checkbox). Below the lists are four arrow buttons: a top-right arrow (highlighted with a red box), a double-right arrow, a left-right arrow, and a bottom-right arrow. At the bottom of the dialog are 'Apply Sort' and 'Cancel' buttons. A red arrow points from the 'Sort...' button in the main interface to this dialog box.

	CLAS_DT_1_1_PRED	CLAS_DT_1_1_PROB	CLAS_DT_1_1_PCST	CUSTOMER_ID
1 No	1.0	0.0	CU14507	
2 No				
3 No				
4 No				
5 No				
6 No				
7 No				
8 No				
9 No				
10 No				
11 No				
12 No				
13 No				
14 No				
15 Yes				
16 No				
17 No				
18 No				
19 No				
20 No				
21 No	1.0	0.0	CU15231	
22 No	1.0	0.0	CU6993	
23 No	0.8032786885245902	0.7006512463507748	CU259	

B. ソートの適用をクリックし、結果を表示します:

The screenshot shows the Oracle Data Miner interface with the results of the sorting applied. The table is now ordered by 'CLAS_DT_1_1_PRED' and 'CLAS_DT_1_1_PROB' in descending order. The first few rows show 'Yes' values for 'CLAS_DT_1_1_PRED' and non-zero values for 'CLAS_DT_1_1_PROB'.

	CLAS_DT_1_1_PRED	CLAS_DT_1_1_PROB	CLAS_DT_1_1_PCST	CUSTOMER_ID
1 Yes	Yes	0.711764705882353	0.4007549543881723	CU6691
2 Yes	Yes	0.711764705882353	0.4007549543881723	CU7296
3 Yes	Yes	0.711764705882353	0.4007549543881723	CU1547
4 Yes	Yes	0.711764705882353	0.4007549543881723	CU1722
5 Yes	Yes	0.711764705882353	0.4007549543881723	CU13184
6 Yes	Yes	0.711764705882353	0.4007549543881723	CU13664
7 Yes	Yes	0.711764705882353	0.4007549543881723	CU14750
8 Yes	Yes	0.711764705882353	0.4007549543881723	CU7922
9 Yes	Yes	0.711764705882353	0.4007549543881723	CU3239
10 Yes	Yes	0.711764705882353	0.4007549543881723	CU7412
11 Yes	Yes	0.711764705882353	0.4007549543881723	CU3147
12 Yes	Yes	0.711764705882353	0.4007549543881723	CU14708
13 Yes	Yes	0.711764705882353	0.4007549543881723	CU14746
14 Yes	Yes	0.711764705882353	0.4007549543881723	CU8606

注:

- 適用ノードを実行するたびに、Oracle Data Minerは異なるサンプルをとります。データおよび表の並びは実行のたびに変わることがあります。よって、表のサンプルはここで表示しているものと異なる場合があります。データ量が少ないのでこのレッスンのスキーマの場合、特にこれは明らかです
- フィルタボックスにWHERE句を入力することでデータをフィルタリングできます
- 表の内容は、Oracle Application ExpressやOracle BI Answers、Oracle BI DashboardsなどのOracleの提供するアプリケーションを使用して表示できます

C. 結果表示を確認し、DT_PREDICTIONS表のタブを閉じ、すべて保存をクリックします

まとめ

このレッスンでは、SQL Developer 4.0に含まれるグラフィカルユーザインターフェースOracle Data Minerを使って、「分類」予測データマイニングを作成しビジネス課題を検証・解決しました

このチュートリアルでは、以下のことを学びました:

- Data Minerインターフェースコンポーネントについて
- Data Minerプロジェクトの作成
- 顧客の行動を予測するために分類モデルを使ったワークフロードキュメントの構築

リソース

Oracle Data Miningについて詳しくは:

- OTNのOracle Data MiningおよびOracle Advanced Analyticsのページ
- Oracle Learning Libraryにある他のOBE
- Data Mining概要マニュアル:
 - Oracle Database 12c Release 1 (12.1)
 - Oracle Database 11g Release 2 (11.2)

謝辞

主なカリキュラム開発者: Brian Pottle

他の貢献者: Charlie Berger, Mark Kelly, Margaret Taft, Kathy Talyor