Spectrum™ Technology Platform

バージョン 12.0 SP1

Machine Learning ガイド

目次

10 - Java Model Scoring

11 - Machine Learning モデル管理

1 - はじめに

このセクションの構成

Machine Learning モジュール

Spectrum™ Technology Platform Machine Learning モジュールを使用すると、数値データをビニングし、教師ありと教師なしの機械学習モデルを適合し、これらのモデルでデータをスコアリングできます。

> **注：** Machine Learning モジュールは、Windows と Linux の各オペレーティング システムでのみサポートされています。

Binning

Binning は、目標情報を考慮に入れずに、連続変数のレコードをグループ (ビン) に分類します。均等幅ビンと均等個数ビンという 2 つのいずれかの方法で、教師なしビニングを実行できます。

Binning Lookup

Binning Lookup は、Binning ステージを使用してデータフローで作成された既存のビンを使用して、以前に定義されたビニングを新しいデータに適用します。

K-Means Clustering

K-Means Clustering は、分析クラスタリングに基づくモデルを作成します。このクラスタリングでは、一連のレコードをデータ値に基づく類似レコードのクラスタに分割します。

Linear Regression

Linear Regression は、持続的目標と入力変数を使用するデータセットからモデルを作成します。

Logistic Regression

Logistic Regression は、バイナリ目標と入力変数を使用するデータセットからモデルを作成します。

主成分分析

主成分分析 (PCA) は、相関のある可能性がある変数群の観測データの集合を、主成分と呼ばれる線型相関のない変数の値の集合に変換する統計的な処理です。

Random Forest Classification

Random Forest Classification は、バイナリまたは多項目標と入力変数を使用するデータセットからモデルを作成します。

Random Forest Regression

Random Forest Regression は、持続的目標と入力変数を使用するデータセットからモデルを作成します。

Java Model Scoring

この機能は、機械学習モデルの適合を行った時に作成された式を使用して、新しいデータをスコアリングします。

Machine Learning モデル管理

Machine Learning モデル管理では、Spectrum™ Technology Platform サーバー上のすべての機械学習モデルが管理できます。モデルのエクスポーズ、アンエクスポーズ、削除が可能です。また、各モデルの詳細情報を表示して、同じタイプの任意の 2 つのモデルを比較できます。

> **注：** Machine Learning モジュールは、Java Model Scoring のモデリング アルゴリズムに、基盤となる H2O.ai ライブラリを使用します。

Machine Learning ワークフロー

標準的な機械学習ワークフローは、1 つ以上のデータフローで行われる以下のステップで構成されます。

1. Data Integration など、Spectrum の他のモジュールを使用して、データにアクセスします。
2. Data Integration、Data Quality、および各種の Core モジュールなど、Spectrum の他のモジュールのステージを使用して、データを準備します。
3. 機械学習モデルを適合し、データフローを実行してから、モデル ステージの [モデル出力] タブを確認します。必要に応じてモデルに微調整を加え、データフローを再実行します。その後、Machine Learning モデル管理ツールのモデル評価出力全体を確認する必要があります。モデルを 1 度に 1 つずつ確認するか、2 つのモデルを比較することができます。
4. (オプション) モデルをデータのスコアリングに使用する場合は、モデルを Machine Learning モデル管理ツールでエクスポーズします。これにより、そのモデルは Java Model Scoring ステージで使用可能になります。

    a. 上のステップ 1 ～ 2 によって Spectrum™ Technology Platform データフローを作成し、ステップ 3 を Java Model Scoring ステージに置き換えます。このデータフローをバッチ モードで実行するように設定し、更新されたデータに適用されたモデル スコアを、ファイルに設定します (自然な処理の流れとして、X または入力として使用されたフィールドがステップ 1 ～ 2 で更新されます)。

b. あるいは、Spectrum™ Technology Platform の Web サービスを使用してオンデマンドでデータをスコアリングします。例えば、Web サイトにアクセスして顧客 ID とモデル入力を取得し、それらをスコアリングして、顧客向けに Web コンテンツをカスタマイズするプロセスにそのスコアを返します。

5. (オプション) モデル スコアは、Data Hub グラフ データベースにエンティティ プロパティとして展開するか、マップ上に展開するか、または CES アプリケーションに展開することもできます。

2 - Binning

このセクションの構成

Binning の概要

Binning ステージは、目標情報を考慮に入れずに、連続変数をグループ (ビン) に分類する、教師なしビニングとして知られる処理を実行します。取得されるデータには、レンジ、個数、各レンジ内の値の割合などがあります。

ビニングの実行には、次のような利点があります。

- データが欠落しているレコードをモデルに含めることができる。
- 外れ値がモデルに与える影響を制御または緩和することができる。
- 最終モデルの係数の重みを同等にすることによって、特性によって尺度が異なる問題を解決する。

Spectrum™ Technology Platform の教師なしビニングでは、データを同じサイズのビンに分割する均等幅ビン、または、データをほぼ同数のレコードを含むグループに分割する均等個数ビンが使用できます。Binning ステージでは、均等幅ビンは、[Equal Ranges] ビン、均等個数ビンは、[Equal Count] ビンと呼ばれます。

コマンド ラインの命令を使用して、ビニングの一覧を表示したり、ビニングを削除したりできます。『管理ガイド』の「**管理ユーティリティ**」セクションの「Machine Learning モジュール」を参照してください。

基本オプションの設定

1. レンジ幅均等とレコード数均等のどちらの **[ビニング スタイル]** を実行するかを選択します。
2. **[NULL 値ビン]** で、空のビン フィールド (データが欠落しているために値が不明であることを表します) の処理方法を選択します。NULL 値を最高ビンに割り当てる場合は **[Highest]**、最低ビンに割り当てる場合は **[Lowest]** を選択します。最低ビンは、必ずビン 1 です。
3. **[ターゲット内部ビン]** をクリックして、両端のビンの間のビン数を入力します。レンジ幅均等ビニングを実行する場合は、内部ビン処理を選択しても、**[ビン幅]** を選択してもよいですが、両方を選択することはできません。レコード数均等ビニングを実行する場合は、内部ビン処理しか実行できません。
4. レンジ幅均等ビニングを実行し、内部ビン処理ではなくビン幅を選択する場合は、**[ビン幅]** をクリックして、各ビンに含める個数を入力します。
5. データをビニングに含める各フィールドに対し、**[含める]** をクリックします。このリストには、数値フィールドしか表示されません。

6. **[OK]** をクリックして、設定を保存します。

ビニング出力

Binning ステージには 2 つの出力ポートがあります。1 つめのポートは、すべての入力フィールドに加えて、選択された各入力フィールドに対するビニング済みフィールドを出力します。例えば、入力フィールドに Name、Age、Income のフィールドがあり、Age と Income に対してビニングを実行する場合、1 つめのポートからは、以下のフィールドが出力されます。

- Name
- Age
- Binned_Age
- Income
- Binned_Income

2 つめのポートは、選択された各入力フィールドに対する 4 種類の情報を出力します。例えば、Age に対してビニングを実行する場合、2 つめのポートからは、以下のフィールドが出力されます。

- Age_Bins
- Age_BinValue
- Age_Count
- Age_Percentage

3 - Binning Lookup

このセクションの構成

Binning Lookup の概要

Binning Lookup は、**Binning** ステージを使用してデータフローで作成された既存のビンを使用して、以前に定義されたビニングを新しいデータに適用します。

ビニングのプロパティの定義

1. **[プライマリ ステージ] / [展開済みステージ] / [Machine Learning]** の下で、**[Binning Lookup]** ステージをクリックしてキャンバス上にドラッグし、データフロー内の所望の位置に配置して、他のステージに接続します。入力ステージにはビニングされるデータが含まれている必要があることに注意してください。1 つの出力ステージがビニングされた出力のために必要です。オプションで、ビニング サマリを取得するために第 2 の出力ステージを接続できます。
2. **[ビニング名]** ドロップダウンから適切なものを選択します。これらは、Binning ステージを使用するデータフローによって作成された既存のビンの名前です。
3. **[ビニング タイプ]** および **[説明]** フィールドは、手順 2 で選択したビニング名によってインポートされるため、編集できません。
4. **[入力]** グリッドには、Binning ステージでのビニングのために含められた各フィールドがデータ タイプとともに表示されます。
5. **[OK]** をクリックして、設定を保存します。

ビニング出力

このタブには、Binning Lookup ステージによってビニングされるフィールドとデータ タイプが表示されます。ビニング ステージを使用して生成される出力の詳細については、ビニング出力（10 ページ） を参照してください。現在、ビニングされたフィールドをビニング出力タブで編集するオプションが用意されています。ユーザは "ビニングされた Spectrum フィールド" を使用して、ビニングされたフィールドに新しい名前を付けることができます。ビニングされたフィールドを含めたり除外したりするオプションも用意されています。

4 - K-Means Clustering

このセクションの構成

K-Means Clustering の概要

K-Means Clustering は、分析クラスタリングに基づくモデルを作成します。このクラスタリングでは、一連のレコードをデータ値に基づく類似レコードのクラスタに分割します。

モデルを作成するにはまず、[モデルのプロパティ] タブを設定する必要があります。[基本オプション] タブと [高度なオプション] タブには、ジョブを完了するために十分なデフォルト設定が指定されていますが、ニーズに合わせてその設定を変更することができます。続いてジョブを実行すると、結果として得られたモデル出力詳細情報の限定版が [モデル出力] タブに表示されます。モデルは Spectrum™ Technology Platform サーバーに格納され、出力全体は、Machine Learning モデル管理ツールで確認できます。

モデルのプロパティの定義

1. **[プライマリ ステージ] / [展開済みステージ] / [Machine Learning]** の下で、**[K-Means Clustering]** ステージをクリックしてキャンバス上にドラッグし、データフロー内の所望の位置に配置して、他のステージに接続します。入力ステージは、モデルの入力変数フィールドを含むデータ ソースでなければならないことに注意してください。出力ステージは、[基本オプション] タブで [スコア] 入力データ オプションを選択しない限り、不要です。Machine Learning モデル管理ツールとは独立して出力を取得する場合は、出力ステージを接続することもできます。
2. K-Means ステージをダブルクリックして、**[K-Means Clustering オプション]** ダイアログ ボックスを表示します。
3. デフォルトのモデル名を使用しない場合は、**[モデル名]** を入力します。
4. 必要に応じて、**[上書き]** チェックボックスをオンにして、既存モデルを新しいデータで上書きします。
5. モデルのクラスタ数をデフォルト値 (5) 以外にする場合は、**[クラスタの数]** を入力します。
6. 必要に応じて、モデルの **[説明]** を入力します。
7. モデルにデータを追加するフィールドの **[含める]** をクリックします。
8. **[モデル データ タイプ]** ドロップダウンを使って、入力フィールドを数値、カテゴリ値、日付と時刻のいずれのフィールドとして使うかを指定します。
9. **[OK]** をクリックして、モデルと設定を保存するか、次のタブで操作を続行します。

基本オプションの設定

1. 数値列を標準化し、平均が 0 で偏差が 1 になるようにするには、**[標準化]** をオンのままにします。

   標準化を使わない場合、実際の寄与の大きさではなく尺度の違いから、他の属性と比べて偏差が大きくなる変数によって左右される要素が結果に含まれる場合があります。
2. モデル予測 (スコア) を表す列を入力データに追加するには、**[入力データを記録]** をオンにします。
3. **[クラスタ数を見積もる]** をオンにすると、K-Means アルゴリズムによって、モデルに含めるクラスタ数の判定が試みられます。[モデルのプロパティ] タブで所望のクラスタ数を指定した場合でも、データから判断して異なるクラスタ数の方が適切であることが、この処理によって検出される可能性があります。
4. 入力データがトレーニングおよびテストのデータ サンプルにランダムに分割される場合は、1 ～ 100 の値を **[トレーニング データの比率]** に指定します。
5. ステップ 5 で指定した値を 100 から引いた値を **[テスト データの比率]** に入力します。
6. データフローを何度実行しても、データが必ず同じ方法でテスト データとトレーニング データに分割されるようにするには、**[テスト データ用シード]** に数値を入力します。フローを実行するたびにランダムな分割を行う場合は、このフィールドをオフにします。
7. **[OK]** をクリックして、モデルと設定を保存するか、次のタブで操作を続行します。

高度なオプションの設定

1. **[定数フィールドを無視]** をオンにすると、各レコードで値が同じフィールドがスキップされます。
2. **[アルゴリズムのシード]** をオンにしてシード数を入力すると、データフローを何度実行しても、データが必ず同じ方法でテスト データとトレーニング データに分割されます。フローを実行するたびにランダムな分割を行う場合は、このフィールドをオフにします。
3. **[初期化]** ドロップダウンで、正しい初期化モードを選択します。

   Furthest 最初の中心点はランダムに初期化しますが、2つめの中心点はそれから最も遠いデータ ポイントになるように初期化します。互いに大きく分散するように、中心点を初期化します。

Plus-Plus (++)	標準の *k*-means の再帰的最適化を行う前に、クラスタの中心を初期化します。*k*-means++ の初期化を行うと、アルゴリズムによって、最適な *k*-means ソリューションに O(log *k*) 近似のソリューションが検出されることが保証されます。
Random	こちらがデフォルトです。N 個のオブザベーション集合から K 個のクラスタを、各オブザベーションの選択確率が等しくなるようにランダムに選択します。

4. **[アルゴリズムのシード]** をオンにしてシード数を入力すると、データフローを何度実行しても、データが必ず同じ方法でテスト データとトレーニング データに分割されます。フローを実行するたびにランダムな分割を行う場合は、このフィールドをオフにします。
5. 相互検証を実行する場合は **[N フォールド]** をオンにし、フォールドの数を入力します。
6. 相互検証を実行する場合は、**[フォールド割り当て]** をオンにしてドロップダウン リストから選択します。このフィールドは、**[N フォールド]** に値が入力済みの場合のみ適用可能です。

Auto	こちらがデフォルトです。オプションの自動選択をアルゴリズムに任せます。現在、[ランダム] が選択されます。
Modulo	データセットをフォールドに等分し、シードを基準としません。
Random	データを n フォールドのサブセットにランダムに分割します。大きなデータセットに最適です。

7. **[最大反復回数]** をオンにし、実行する必要があるトレーニング反復回数を入力します。
8. **[OK]** をクリックして、モデルと設定を保存するか、次のタブで操作を続行します。

モデル出力

このタブには、適合されたモデルの評価に使用するメトリクスが表示されます。これらのフィールドは編集できません。[トレーニング] 列には、必ずデータが含まれます。[基本オプション] タブでテストとトレーニングの分割を選択した場合は、[テスト] 列にもデータが設定されます。ただし、[高度なオプション] タブで N フォールド検証を選択した場合を除きます。その場合は、[N フォールド] 列にデータが設定されます。出力を再生成するには **[出力]** ボタンをクリックし、出力全体を Machine Learning モデル管理ツールに表示するには **[モデルの詳細]** をクリックします。

5 - Linear Regression

このセクションの構成

Linear Regression の概要

Linear Regression では、持続的目標と入力変数を使用するデータセットからモデルを作成して、機械学習を実行することができます。

モデルを作成するにはまず、[モデルのプロパティ] タブを設定する必要があります。[基本オプション] タブと [高度なオプション] タブには、ジョブを完了するために十分なデフォルト設定が指定されていますが、ニーズに合わせてその設定を変更することができます。続いてジョブを実行すると、結果として得られたモデルの限定版が [モデル出力] タブに表示されます。出力全体は、Machine Learning モデル管理ツールで確認できます。

モデルのプロパティの定義

1. **[プライマリ ステージ] / [展開済みステージ] / [Machine Learning]** の下で、**[Linear Regression]** ステージをクリックしてキャンバス上にドラッグし、データフロー内の所望の位置に配置して、他のステージに接続します。入力ステージは、モデルの目標フィールドと入力変数フィールドの両方を含むデータ ソースでなければならないことに注意してください。出力ステージは、[基本オプション] タブで [スコア] 入力データ オプションを選択しない限り、不要です。Machine Learning モデル管理ツールとは独立して出力を取得する場合は、出力ステージを接続することもできます。
2. [Linear Regression] ステージをダブルクリックして、**[Logistic Regression オプション]** ダイアログ ボックスを表示します。
3. デフォルトのモデル名を使用しない場合は、**[モデル名]** を入力します。
4. **[上書き]** チェックボックスをオンにして、既存モデルを新しいデータで上書きします。
5. **[目標フィールド]** ドロップダウンをクリックし、数値フィールドを選択します。
6. モデルの **[説明]** を入力します。
7. データをモデルに追加したいそれぞれのフィールドで **[含める]** をクリックします。目標フィールドとして選択したフィールドは必ず含めてください。
8. **[モデル データ タイプ]** ドロップダウンを使用して、各入力フィールドを数値、カテゴリ値、日付と時刻のいずれのフィールドとして使うかを指定します。
9. **[OK]** をクリックして、モデルと設定を保存するか、次のタブで操作を続行します。

基本オプションの設定

1. 数値列を標準化し、平均が 0 で偏差が 1 になるようにするには、**[標準化]** をオンのままにします。

   標準化を使わない場合、実際の寄与の大きさではなく尺度の違いから、他の属性と比べて偏差が大きくなる変数によって左右される要素が結果に含まれる場合があります。
2. モデル予測 (スコア) を表す列を入力データに追加するには、**[入力データを記録]** をオンにします。
3. ドロップダウン リストから **[リンク機能]** を選択します。これは、ランダムなコンポーネントと体系的なコンポーネントとの間のリンクを指定するものです。応答で期待される値を説明変数の線型予測因子にどのように関連付けるかを示します。

特定	0 未満または 1 を超えるような無意味な「確率」を予測します。線型確率モデルを得るために二項データで使用されることがあります。 *g(p) = p*
逆変換	実際の見積もり値のためのリンク関数の逆関数を計算します。 *g(μi)=1μi*
ログ	定められた時間および空間の範囲内での発生回数をカウントします。 *g(μi)=log(μi)*

4. 欠落データの処理方法を指定するには、**[スキップ]** または **[平均値を補完]** をオンにします。後者のオプションを選択すると、欠落データの代わりに平均値が追加されます。
5. 入力データがトレーニングおよびテストのデータ サンプルにランダムに分割される場合は、1 ～ 100 の値を **[トレーニング データの比率]** に指定します。
6. ステップ 5 で指定した値を 100 から引いた値を **[テスト データの比率]** に入力します。
7. データフローを何度実行しても、データが必ず同じ方法でテスト データとトレーニング データに分割されるようにするには、**[テスト データ用シード]** に数値を入力します。フローを実行するたびにランダムな分割を行う場合は、このフィールドをオフにします。
8. **[OK]** をクリックして、モデルと設定を保存するか、次のタブで操作を続行します。

高度なオプションの設定

1. **[定数フィールドを無視]** をオンにすると、各レコードで値が同じフィールドがスキップされます。
2. **[p 値を計算]** をオンにすると、パラメータを予測するための p 値が計算されます。
3. モデルの作成時に共線列を自動的に削除するには、**[共線列を削除]** をオンにします。これにより、返されるモデルでは係数が 0 になります。

   このオプションは、**[p 値を計算]** がオンになっている場合は常にオンにする必要があります。
4. 定数項 (切片) をモデルに含めるには、**[定数項 (切片) を含める]** をオンにします。

   [共線列を削除] がオンの場合は、このオプションを必ずオンにする必要があります。
5. ドロップダウン リストから **[ソルバー]** を選択します。CoordinateDescent および CoordinateDescentNaive は現時点で実験用であることに注意してください。

Auto	入力データとパラメータに基づいてソルバーが決定されます。
CoordinateDescent	最も内側のループにおける循環座標降下のバージョンを更新する共分散を用いた IRLSM。
CoordinateDescentNaive	最も内側のループにおける循環座標降下のバージョンを単純に更新する共分散を用いた IRLSM。
IRLSM	予測因子が少数のときの問題や、L1 ペナルティによるラムダ検索の問題に最適です。
LBFGS	多数の列が含まれるデータセットに最適です。

6. **[アルゴリズムのシード]** をオンにしてシード数を入力すると、データフローを何度実行しても、データが必ず同じ方法でテスト データとトレーニング データに分割されます。フローを実行するたびにランダムな分割を行う場合は、このフィールドをオフにします。
7. 相互検証を実行する場合は **[N フォールド]** をオンにし、フォールドの数を入力します。
8. 相互検証を実行する場合は、**[フォールド割り当て]** をクリックしてドロップダウン リストから選択します。このフィールドは、**[N フォールド]** に値が入力済みで、**[フォールド フィールド]** が指定されていない場合にのみ適用可能です。

Auto	オプションの自動選択をアルゴリズムに任せます。現在、[ランダム] が選択されます。
Modulo	データセットをフォールドに等分し、シードを基準としません。

Random データを n フォールドのサブセットにランダムに分割します。大きなデータセットに最適です。

9. 相互検証を実行する場合は、**[フォールド フィールド]** をオンにして、相互検証フォールド インデックス割り当てを含むフィールドをドロップダウン リストから選択します。

   このフィールドは、**[N フォールド]** と **[フォールド割り当て]** に値が入力されていない場合のみ適用可能です。

10. **[最大反復回数]** をオンにし、実行する必要があるトレーニング反復回数を入力します。
11. **[目標イプシロン]** をオンにして、収束のしきい値を入力します。この値は 0 ～ 1 の間でなければなりません。目標値がこのしきい値に満たない場合、モデルは収束します。
12. **[ベータ イプシロン]** をオンにして、収束のしきい値を入力します。この値は 0 ～ 1 の間でなければなりません。目標値がこのしきい値に満たない場合、モデルは収束します。現在のベータ変化の L1 正則化がこのしきい値に満たない場合、収束の使用を検討してください。
13. **[OK]** をクリックして、モデルと設定を保存するか、次のタブで操作を続行します。

モデル出力

このタブには、適合されたモデルの評価に使用するメトリクスが表示されます。これらのフィールドは編集できません。[トレーニング] 列には、必ずデータが含まれます。[基本オプション] タブでテストとトレーニングの分割を選択した場合は、[テスト] 列にもデータが設定されます。ただし、[高度なオプション] タブで N フォールド検証を選択した場合を除きます。その場合は、[N フォールド] 列にデータが設定されます。

ジョブを実行すると、結果として得られたモデルが Spectrum™ Technology Platform サーバーに格納されます。出力を再生成するには **[出力]** ボタンをクリックし、出力全体を Machine Learning モデル管理ツールに表示するには **[モデルの詳細]** をクリックします。

6 - Logistic Regression

このセクションの構成

Logistic Regression の概要

Logistic Regression では、バイナリ目標と入力変数を使用するデータセットからモデルを作成して、機械学習を実行することができます。

モデルを作成するにはまず、[モデルのプロパティ] タブを設定する必要があります。[基本オプション] タブと [高度なオプション] タブには、ジョブを完了するために十分なデフォルト設定が指定されていますが、ニーズに合わせてその設定を変更することができます。続いてジョブを実行すると、結果として得られたモデルの限定版が [モデル出力] タブに表示されます。出力全体は、Machine Learning モデル管理ツールで確認できます。

モデルのプロパティの定義

1. **[プライマリ ステージ] / [展開済みステージ] / [Machine Learning]** の下で、**[Logistic Regression]** ステージをクリックしてキャンバス上にドラッグし、データフロー内の所望の位置に配置して、他のステージに接続します。入力ステージは、モデルの目標フィールドと入力変数フィールドの両方を含むデータ ソースでなければならないことに注意してください。出力ステージは、[基本オプション] タブで [スコア] 入力データ オプションを選択しない限り、不要です。Machine Learning モデル管理ツールとは独立して出力を取得する場合は、出力ステージを接続することもできます。
2. Logistic Regression ステージをダブルクリックして、**[Logistic Regression オプション]** ダイアログ ボックスを表示します。
3. デフォルトのモデル名を使用しない場合は、**[モデル名]** を入力します。
4. **[上書き]** チェックボックスをオンにして、既存モデルを新しいデータで上書きします。
5. **[目標フィールド]** ドロップダウンをクリックして "カテゴリ" を選択します。
6. モデルの **[説明]** を入力します。
7. データをモデルに追加したいそれぞれのフィールドで **[含める]** をクリックします。目標フィールドとして選択したフィールドは必ず含めてください。
8. **[モデル データ タイプ]** ドロップダウンを使用して、各入力フィールドを数値、カテゴリ値、日付と時刻のいずれのフィールドとして使うかを指定します。
9. **[OK]** をクリックして、モデルと設定を保存するか、次のタブで操作を続行します。

基本オプションの設定

1. 数値列を標準化し、平均が 0 で偏差が 1 になるようにするには、**[標準化]** をオンのままにします。

   標準化を使わない場合、実際の寄与の大きさではなく尺度の違いから、他の属性と比べて偏差が大きくなる変数によって左右される要素が結果に含まれる場合があります。
2. モデル予測 (スコア) を表す列を入力データに追加するには、**[入力データを記録]** をオンにします。
3. データがサンプル済みで、応答の平均が実態を反映していない場合は、**[プライア]** をオンにし、p(y==1) の事前確率をテキスト フィールドに入力します。
4. 欠落データの処理方法を指定するには、**[スキップ]** または **[平均値を補完]** をオンにします。後者のオプションを選択すると、欠落データの代わりに平均値が追加されます。
5. 入力データがトレーニングおよびテストのデータ サンプルにランダムに分割される場合は、1 ～ 100 の値を **[トレーニング データの比率]** に指定します。
6. ステップ 5 で指定した値を 100 から引いた値を **[テスト データの比率]** に入力します。
7. データフローを何度実行しても、データが必ず同じ方法でテスト データとトレーニング データに分割されるようにするには、**[テスト データ用シード]** に数値を入力します。フローを実行するたびにランダムな分割を行う場合は、このフィールドをオフにします。
8. **[OK]** をクリックして、モデルと設定を保存するか、次のタブで操作を続行します。

高度なオプションの設定

1. **[定数フィールドを無視]** をオンにすると、各レコードで値が同じフィールドがスキップされます。
2. **[p 値を計算]** をオンにすると、パラメータを予測するための p 値が計算されます。
3. モデルの作成時に共線列を自動的に削除するには、**[共線列を削除]** をオンのままにします。これにより、返されるモデルでは係数が 0 になります。

   このオプションは、**[p 値を計算]** がオンになっている場合は常にオンにする必要があります。
4. 定数項 (切片) をモデルに含めるには、**[定数項 (切片) を含める]** をオンにします。

   [共線列を削除] がオンの場合は、このオプションを必ずオンにする必要があります。

5. ドロップダウン リストから **[ソルバー]** を選択します。CoordinateDescentNaive および CoordinateDescentNaive は現時点で実験用であることに注意してください。

Auto	入力データとパラメータに基づいてソルバーが決定されます。
CoordinateDescentNaive	最も内側のループにおける循環座標降下のバージョンを更新する共分散を用いた IRLSM。
CoordinateDescentNaive	最も内側のループにおける循環座標降下のバージョンを単純に更新する共分散を用いた IRLSM。
IRLSM	予測因子が少数のときの問題や、L1 ペナルティによるラムダ検索の問題に最適です。
L_BFGS	多数の列が含まれるデータセットに最適です。

6. **[アルゴリズムのシード]** をオンにしてシード数を入力すると、データフローを何度実行しても、データが必ず同じ方法でテスト データとトレーニング データに分割されます。フローを実行するたびにランダムな分割を行う場合は、このフィールドをオフにします。
7. 相互検証を実行する場合は **[N フォールド]** をオンにし、フォールドの数を入力します。
8. 相互検証を実行する場合は、**[フォールド割り当て]** をオンにしてドロップダウン リストから選択します。このフィールドは、**[N フォールド]** に値が入力済みで、**[フォールド フィールド]** が指定されていない場合にのみ適用可能です。

Auto	オプションの自動選択をアルゴリズムに任せます。現在、[ランダム]が選択されます。
Modulo	データセットをフォールドに等分し、シードを基準としません。
Random	データを n フォールドのサブセットにランダムに分割します。大きなデータセットに最適です。
Stratified	分類問題の応答変数に基づいて、フォールドを層化します。データセットをトレーニング データとテスト データに分割する際に、観測値を複数のクラスからすべてのセットに均等に分散します。これは、クラスの数が多く、データセットが比較的小さい場合に便利です。

9. 相互検証を実行する場合は、**[フォールド フィールド]** をオンにして、相互検証フォールド インデックス割り当てを含むフィールドをドロップダウン リストから選択します。

   このフィールドは、**[N フォールド]** と **[フォールド割り当て]** に値が入力されていない場合のみ適用可能です。
10. **[最大反復回数]** をオンにし、実行する必要があるトレーニング反復回数を入力します。
11. **[目標イプシロン]** をオンにして、収束のしきい値を入力します。この値は 0 ～ 1 の間でなければなりません。目標値がこのしきい値に満たない場合、モデルは収束します。

12. **[ベータ イプシロン]** をオンにして、収束のしきい値を入力します。この値は 0 ～ 1 の間でなければなりません。目標値がこのしきい値に満たない場合、モデルは収束します。現在のベータ変化の L1 正則化がこのしきい値に満たない場合、収束の使用を検討してください。
13. **[OK]** をクリックして、モデルと設定を保存するか、次のタブで操作を続行します。

モデル出力

このタブには、適合されたモデルの評価に使用するメトリクスが表示されます。これらのフィールドは編集できません。[トレーニング] 列には、必ずデータが含まれます。[基本オプション] タブでテストとトレーニングの分割を選択した場合は、[テスト] 列にもデータが設定されます。ただし、[高度なオプション] タブで N フォールド検証を選択した場合を除きます。その場合は、[N フォールド] 列にデータが設定されます。

ジョブを実行すると、結果として得られたモデルが Spectrum™ Technology Platform サーバーに格納されます。出力を再生成するには **[出力]** ボタンをクリックし、出力全体を Machine Learning モデル管理ツールに表示するには **[モデルの詳細]** をクリックします。

7 - 主成分分析

このセクションの構成

主成分分析の概要

主成分分析 (PCA) は、相関のある可能性がある変数群の観測データの集合を、主成分と呼ばれる線型相関のない変数の値の集合に変換する統計的な処理です。

モデルを作成するにはまず、[モデルのプロパティ] タブを設定する必要があります。[基本オプション] タブと [高度なオプション] タブには、ジョブを完了するために十分なデフォルト設定が指定されていますが、ニーズに合わせてその設定を変更することができます。続いてジョブを実行すると、結果として得られたモデルの限定版が [モデル出力] タブに表示されます。出力全体は、Machine Learning モデル管理ツールで確認できます。モデルの出力に問題がなければ、そのモデルを公開してスコアリング データフローで使用することができます。

モデルのプロパティの定義

1. **[プライマリ ステージ] / [展開済みステージ] / [Machine Learning]** の下で、**[PCA Options]** ステージをクリックしてキャンバス上にドラッグし、データフロー内の所望の位置に配置して、他のステージに接続します。入力ステージは、モデルの主成分が含まれているデータ ソースでなければならないことに注意してください。出力ステージは必要ありませんが、Machine Learning モデル管理ツールとは独立して出力を取得する場合は、出力ステージを接続することもできます。
2. [PCA Options] ステージをダブルクリックして、**[PCA オプション]** ダイアログ ボックスを表示します。
3. デフォルトのモデル名を使用しない場合は、**[モデル名]** を入力します。
4. 必要に応じて、**[上書き]** チェックボックスをオンにして、既存モデルを新しいデータで上書きします。
5. モデルに含める **[主要コンポーネント]** の数を入力します。
6. 必要に応じて、モデルの **[説明]** を入力します。
7. **[入力]** テーブルで、モデルにデータを追加するフィールドの [含める] をクリックします。
8. **[モデル データ タイプ]** ドロップダウンを使って、入力フィールドをカテゴリ値、日付と時刻、数値、文字列、ユニーク ID のいずれのフィールドとして使うかを指定します。
9. **[OK]** をクリックして、モデルと設定を保存するか、次のタブで操作を続行します。

基本オプションの設定

1. 第1主成分をスキップするには、**[すべての因子レベルを使用]** をオフのままにしておきます。この場合、データ内の分散が最大になります。このチェックボックスをオンにすると、第1主成分が保持されます。
2. モデル予測 (スコア) を表す列を入力データに追加するには、**[入力データを記録]** をオンにします。
3. トレーニング データに対する適切な **[変換]** を選択します。

平均除去	各列の平均値を減算します。
スケール除去	各列の標準偏差による除算を行います。
なし	
標準化	各列に対して、平均の減算とその範囲 (最大値と最小値の差) による除算を行います。
正規化	ゼロ平均と単位分散を使用します。こちらがデフォルトです。

4. **[欠落データ]** の処理方法を指定するには、**[スキップ]** または **[平均値を補完]** をオンにします。後者では、欠落データの代わりに平均値が追加されます。
5. **[OK]** をクリックして、モデルと設定を保存するか、次のタブで操作を続行します。

高度なオプションの設定

1. **[定数フィールドを無視]** をオンにすると、各レコードで値が同じフィールドがスキップされます。
2. ドロップダウン リストから **[PCA メソッド]** を選択します。GLRM および Power は現時点で実験用であることに注意してください。

GLRM	一般化された低ランク モデルのフィッティングを L2 損失関数によって正則化なしに行います。局所行列代数を使用して SVD を求めます。このオプションは、[基本オプション] タブで **[すべての因子レベルを使用]** をオンにしている場合にのみ有効になります。

GramSVD	Gram 行列の分散型計算を使用した後、JAMA パッケージを使用した局所 SVD を実行します。
Power	べき乗による反復法を使用して SVD を計算します。
Randomized	ランダム化された部分空間反復法を使用します。

3. トレーニングの反復回数を制限しない場合は、**[最大反復回数]** はオフのままにしておきます (デフォルトの状態)。トレーニングの反復回数を制限するには、このボックスをオンにして数値を入力します。
4. **[OK]** をクリックして、モデルと設定を保存するか、次のタブで操作を続行します。

モデル出力

このタブには、適合されたモデルの評価に使用するメトリクスが表示されます。これらのフィールドは編集できません。

ジョブを実行すると、結果として得られたモデルが Spectrum™ Technology Platform サーバーに格納されます。出力を再生成するには **[出力]** ボタンをクリックし、出力全体を Machine Learning モデル管理ツールに表示するには **[モデルの詳細]** をクリックします。

8 - Random Forest Classification

このセクションの構成

Random Forest Classification の概要

Random Forest Classification では、バイナリまたは多項目標と入力変数を使用するデータセットからモデルを作成して、機械学習を実行することができます。

モデルを作成するにはまず、[モデルのプロパティ] タブを設定する必要があります。[基本オプション] タブと [高度なオプション] タブには、ジョブを完了するために十分なデフォルト設定が指定されていますが、ニーズに合わせてその設定を変更することができます。続いてジョブを実行すると、結果として得られたモデルの限定版が [モデル出力] タブに表示されます。出力全体は、Machine Learning モデル管理ツールで確認できます。

> **注：** Random Forest Classification とそのオプションに関する詳細情報を参照するには、ここをクリックします。

モデルのプロパティの定義

1. **[プライマリ ステージ] / [展開済みステージ] / [Machine Learning]** の下で、**[Random Forest Classification]** ステージをクリックしてキャンバス上にドラッグし、データフロー内の所望の位置に配置して、他のステージに接続します。入力ステージは、モデルの目標フィールドと入力変数フィールドの両方を含むデータ ソースでなければならないことに注意してください。出力ステージは、[基本オプション] タブで [スコア] 入力データ オプションを選択しない限り、不要です。Machine Learning モデル管理ツールとは独立して出力を取得する場合は、出力ステージを接続することもできます。
2. [Random Forest Classification] ステージをダブルクリックして、**[Random Forest Classification オプション]** ダイアログ ボックスを表示します。
3. デフォルトのモデル名を使用しない場合は、**[モデル名]** を入力します。
4. **[上書き]** チェックボックスをオンにして、既存モデルを新しいデータで上書きします。
5. **[目標フィールド]** ドロップダウンをクリックし、数値フィールドを選択します。
6. **[多項レベル]** をクリックし、目標フィールドに存在する 3 つ以上のカテゴリを入力します。このフィールドを有効にすると、**[入力データを記録]** フィールドが無効になることに注意してください。
7. モデルの **[説明]** を入力します。
8. データをモデルに追加したいそれぞれのフィールドで **[含める]** をクリックします。目標フィールドとして選択したフィールドは必ず含めてください。

9. **[モデル データ タイプ]** ドロップダウンを使用して、各入力フィールドを数値、カテゴリ値、日付と時刻のいずれのフィールドとして使うかを指定します。
10. **[OK]** をクリックして、モデルと設定を保存するか、次のタブで操作を続行します。

基本オプションの設定

1. **[ツリーの数]** に、お使いのモデルでのツリー数の最大値を入力します。
2. **[最大深度]** に、モデルに含めるレベル数の最大値を入力します。
3. **[最小行数]** に、モデルに含める行数 (またはレコード数) の最小値を入力します。
4. **[ビンの数 (数値)]** に、ヒストグラムを構築したうえで最良のポイントで分割するビンの数を入力します。
5. **[ビンの数 (最上位レベル)]** に、ルート レベルで必要なビンの数の最小値を入力します。
6. **[ビンの数 (カテゴリ別)]** に、ヒストグラムを構築したうえで最良のポイントで分割するビンの数の最大値を入力します。
7. **[サンプル レート]** をオンにし、各ツリーでサンプルとして使用される行の比率を入力します。0.0 ～ 999 の値を使用できます。
8. **[各ツリーの列サンプル レート]** をオンにし、各ツリーの列に対するサンプリング率を入力します。0.0 ～ 1.0 の値を使用できます。
9. **[各レベルの列数]** をオンにし、すべてのレベルでの列のサンプリングに対する相対変化量を入力します。有効な値の範囲は、1.0 から、選択した入力予測因子の数値までです。デフォルトは 1.0 です。
10. モデル予測 (スコア) を表す列を入力データに追加するには、**[入力データを記録]** をオンにします。
11. 入力データがトレーニングおよびテストのデータ サンプルにランダムに分割される場合は、1 ～ 100 の値を **[トレーニング データの比率]** に指定します。
12. ステップ 5 で指定した値を 100 から引いた値を **[テスト データの比率]** に入力します。
13. **[テスト データ用シード]** により、データフローを何度実行してもデータが必ず同じ方法でテスト データとトレーニング データに分割されるようになります。フローを実行するたびにランダムな分割を行う場合は、このフィールドをオフにします。
14. **[OK]** をクリックして、モデルと設定を保存するか、次のタブで操作を続行します。

高度なオプションの設定

1. **[定数フィールドを無視]** をオンにすると、各レコードで値が同じフィールドがスキップされます。
2. **[クラスのバランスをとる]** をオンにすると、クラス分布のバランスを取るために大多数のクラスでアンダーサンプリングが行われるか、少数のクラスでオーバーサンプリングが行われます。
3. **[ヒストグラム タイプ]** を選択します。

   Auto バケットが最小値から最大値まで (最大値 - 最小値)/N の刻み幅でビニングされます。このオプションで、最適な分割ポイントを見つけるために使用するヒストグラムのタイプを指定します。

   QuantilesGlobal 各バケットに含める個体数を均等にします。個々の数値列 (二値以外) の `nbins` 個の分位を計算した後、2つの分位に挟まれた各バケットに含める内容を均等に (残余はランダムに) 取捨選択して合計 `nbins_top_level` 個のビンを生成します。

   Random 最小値から最大値までの N-1 個のポイントをサンプリングし、それらのポイントをソートしたリストから最適な分割ポイントを見つけます。

   RoundRobin すべてのヒストグラム タイプを (ツリーごとに 1 つずつ) 順に繰り返し使用します。

   UniformAdaptive 個々のフィーチャーをビニングして刻み幅 (個体数ではない) が均等のバケットを生成します。これは最速の方法ですが、分布に大きな偏りがあると分割が正確でなくなる可能性があります。

4. **[カテゴリ別エンコーディング]** を選択します。

   Auto 自動的に 列挙型 エンコーディングを実行します。

   Binary カテゴリを整数に変換してから 2 進数に変換し、その各桁を別々の列に割り当てます。次元数を減らしてデータをエンコードします (距離に歪みが生じます)。

   > **注 :** カテゴリ別のフィーチャーの列の数は 32 以下でなければなりません。

   Eigen カテゴリ別のフィーチャーの *k* 個の列についてのみ、ワンホット (one-hot) エンコーディング マトリックスを *k* 次元固有空間に投影し続けます。

列挙 すべてのヒストグラム タイプを (ツリーごとに 1 つずつ) 順に繰り返し使用します。

OneHotExplicit カテゴリごとに 1 つの列を生成し、列の各セルの値 "1" または "0" でその列のカテゴリが行に含まれているかどうかを表します。

5. **[アルゴリズムと N フォールドのシード]** をオンにしてシード数を入力すると、データフローを何度実行しても、データが必ず同じ方法でテスト データとトレーニング データに分割されます。フローを実行するたびにランダムな分割を行う場合は、このフィールドをオフにします。
6. 相互検証を実行する場合は **[N フォールド]** をオンにし、フォールドの数を入力します。
7. 相互検証を実行する場合は、**[フォールド割り当て]** をオンにしてドロップダウン リストから選択します。このフィールドは、**[N フォールド]** に値が入力済みで、**[フォールド フィールド]** が指定されていない場合にのみ適用可能です。

   Auto オプションの自動選択をアルゴリズムに任せます。現在、[ランダム]が選択されます。

   Modulo データセットをフォールドに等分し、シードを基準としません。

   Random データを n フォールドのサブセットにランダムに分割します。大きなデータセットに最適です。

   Stratified 分類問題の応答変数に基づいて、フォールドを層化します。データセットをトレーニング データとテスト データに分割する際に、観測値を複数のクラスからすべてのセットに均等に分散します。これは、クラスの数が多く、データセットが比較的小さい場合に便利です。

8. 相互検証を実行する場合は、**[フォールド フィールド]** をオンにして、相互検証フォールド インデックス割り当てを含むフィールドをドロップダウン リストから選択します。

   このフィールドは、**[N フォールド]** と **[フォールド割り当て]** に値が入力されていない場合のみ適用可能です。

9. **[停止の基準回数]** をオンにすると、指定した回数のトレーニングで Stopping_metric オプションの改善が見られないとき、トレーニングの停止前に失敗したトレーニングの回数が入力されます。この機能を無効にするには、0 を指定します。この指標は Validation データに基づいて計算されます (提供されている場合)。そうでなければ、トレーニング データが使われます。
10. **[停止指標]** を選択して、新しいツリーの生成を終了するタイミングを決定します。

    AUC ROC 曲線下面積。

    注 : 二項モデルにのみ適用できます。

Auto	デフォルトは `deviance` です。
Lifttopgroup	上位 1%。
Logloss	対数損失
Meanperclasserror	平均誤分類率。
Misclassification	(1 - (正しい予測数/合計予測数)) * 100 の値。
MSE	平均 2 乗誤差。予測変数の分散とバイアスを包含する誤差です。
RMSE	2 乗平均平方根誤差。モデルや評価関数によって予測された値 (サンプルや母集団の値) と実際に観測した値との差異を表します。MSE の平方根でもあります。

11. **[停止の基準許容値]** をオンにし、指標に基づく停止の相対許容誤差を指定する値を入力すると、改善がこの値未満の場合にトレーニングが終了します。このフィールドは、**[停止の基準回数]** をオンにしている場合にのみ有効になります。
12. **[最小分割改善]** をオンにし、2 乗誤差が低減したときに分割が行われるように最小の相対的な改善を指定する値を入力します。このオプションは、適切に実行すれば、過剰適合を減らす効果があります。最適な値は 1e-10...1e-3 の範囲でしょう。このフィールドは、**[停止の基準回数]** をオンにしている場合にのみ有効になります。
13. **[OK]** をクリックして、モデルと設定を保存するか、次のタブで操作を続行します。

モデル出力

このタブには、適合されたモデルの評価に使用するメトリクスが表示されます。これらのフィールドは編集できません。[トレーニング] 列には、必ずデータが含まれます。[基本オプション] タブでテストとトレーニングの分割を選択した場合は、[テスト] 列にもデータが設定されます。ただし、[高度なオプション] タブで N フォールド検証を選択した場合を除きます。その場合は、[N フォールド] 列にデータが設定されます。

ジョブを実行すると、結果として得られたモデルが Spectrum™ Technology Platform サーバーに格納されます。出力を再生成するには **[出力]** ボタンをクリックし、出力全体を Machine Learning モデル管理ツールに表示するには **[モデルの詳細]** をクリックします。

9 - Random Forest Regression

このセクションの構成

Random Forest Regression の概要

Random Forest Regression では、持続的目標と入力変数を使用するデータセットからモデルを作成して、機械学習を実行することができます。

モデルを作成するにはまず、[モデルのプロパティ] タブを設定する必要があります。[基本オプション] タブと [高度なオプション] タブには、ジョブを完了するために十分なデフォルト設定が指定されていますが、ニーズに合わせてその設定を変更することができます。続いてジョブを実行すると、結果として得られたモデルの限定版が [モデル出力] タブに表示されます。出力全体は、Machine Learning モデル管理ツールで確認できます。

> **注：** Random Forest Regression とそのオプションに関する詳細情報を参照するには、ここをクリックします。

モデルのプロパティの定義

1. **[プライマリ ステージ] / [展開済みステージ] / [Machine Learning]** の下で、**[Random Forest Regression]** ステージをクリックしてキャンバス上にドラッグし、データフロー内の所望の位置に配置して、他のステージに接続します。入力ステージは、モデルの目標フィールドと入力変数フィールドの両方を含むデータ ソースでなければならないことに注意してください。出力ステージは、[基本オプション] タブで [スコア] 入力データ オプションを選択しない限り、不要です。Machine Learning モデル管理ツールとは独立して出力を取得する場合は、出力ステージを接続することもできます。
2. [Random Forest Regression] ステージをダブルクリックして、**[ランダム フォレスト回帰オプション]** ダイアログ ボックスを表示します。
3. デフォルトのモデル名を使用しない場合は、**[モデル名]** を入力します。
4. 必要に応じて、**[上書き]** チェックボックスをオンにして、既存モデルを新しいデータで上書きします。
5. **[目標フィールド]** ドロップダウンをクリックし、数値フィールドを選択します。
6. 必要に応じて、モデルの **[説明]** を入力します。
7. データをモデルに追加したいそれぞれのフィールドで **[含める]** をクリックします。目標フィールドとして選択したフィールドは必ず含めてください。
8. **[モデル データ タイプ]** ドロップダウンを使用して、各入力フィールドを数値、カテゴリ値、日付と時刻のいずれのフィールドとして使うかを指定します。

9. **[OK]** をクリックして、モデルと設定を保存するか、次のタブで操作を続行します。

基本オプションの設定

1. **[ツリーの数]** に、お使いのモデルでのツリー数の最大値を入力します。デフォルトは 50 です。
2. **[最大深度]** に、モデルに含めるレベル数の最大値を入力します。デフォルトは 5 です。
3. **[最小行数]** に、モデルに含める行数 (またはレコード数) の最小値を入力します。デフォルト値は 10 です。
4. **[ビンの数 (数値)]** に、ヒストグラムを構築したうえで最良のポイントで分割するビンの数を入力します。デフォルト値は 20 です。
5. **[ビンの数 (最上位レベル)]** に、ルート レベルで必要なビンの数の最小値を入力します。デフォルトは 1024 です。
6. **[ビンの数 (カテゴリ別)]** に、ヒストグラムを構築したうえで最良のポイントで分割するビンの数の最大値を入力します。デフォルトは 1024 です。
7. **[サンプル レート]** をオンにし、各ツリーでサンプルとして使用される行の比率を入力します。0.0 ～ 1.0 の値を使用できます。
8. **[各ツリーの列サンプル レート]** をオンにし、各ツリーの列に対するサンプリング率を入力します。0.0 ～ 1.0 の値を使用できます。
9. **[各レベルの列数]** では、各レベルでランダムに選択する列の数を指定します。このオプションがオフになっている場合、デフォルト値の -1 が使用され、変数の数は、分類の場合は列数の平方根、回帰の場合は p/3 (p は予測因子の数) となります。このオプションをオンにすると、1 以上の値を指定できます。予測因子の数より大きい値は指定できません。
10. モデル予測 (スコア) を表す列を入力データに追加するには、**[入力データを記録]** をオンにします。
11. 入力データがトレーニングおよびテストのデータ サンプルにランダムに分割される場合は、1 ～ 100 の値を **[トレーニング データの比率]** に指定します。
12. ステップ 5 で指定した値を 100 から引いた値を **[テスト データの比率]** に入力します。
13. **[テスト データ用シード]** により、データフローを何度実行してもデータが必ず同じ方法でテスト データとトレーニング データに分割されるようになります。フローを実行するたびにランダムな分割を行う場合は、このフィールドをオフにします。
14. **[OK]** をクリックして、モデルと設定を保存するか、次のタブで操作を続行します。

高度なオプションの設定

1. **[定数フィールドを無視]** をオンにすると、各レコードで値が同じフィールドがスキップされます。
2. **[ヒストグラム タイプ]** を選択します。

   Auto バケットが最小値から最大値まで (最大値 - 最小値)/N の刻み幅でビニングされます。このオプションで、最適な分割ポイントを見つけるために使用するヒストグラムのタイプを指定します。

   QuantilesGlobal 各バケットに含める個体数を均等にします。個々の数値列 (二値以外) の `nbins` 個の分位を計算した後、2つの分位に挟まれた各バケットに含める内容を均等に (残余はランダムに) 取捨選択して合計 `nbins_top_level` 個のビンを生成します。

   Random 最小値から最大値までの N-1 個のポイントをサンプリングし、それらのポイントをソートしたリストから最適な分割ポイントを見つけます。

   RoundRobin すべてのヒストグラム タイプを (ツリーごとに 1 つずつ) 順に繰り返し使用します。

   UniformAdaptive 個々のフィーチャーをビニングして刻み幅 (個体数ではない) が均等のバケットを生成します。これは最速の方法ですが、分布に大きな偏りがあると分割が正確でなくなる可能性があります。

3. **[カテゴリ別エンコーディング]** を選択します。

   Auto 自動的に 列挙型 エンコーディングを実行します。

   Binary カテゴリを整数に変換してから 2 進数に変換し、その各桁を別々の列に割り当てます。次元数を減らしてデータをエンコードします (距離に歪みが生じます)。

   > **注 :** カテゴリ別のフィーチャーの列の数は 32 以下でなければなりません。

   Eigen カテゴリ別のフィーチャーの *k* 個の列についてのみ、ワンホット (one-hot) エンコーディング マトリックスを *k* 次元固有空間に投影し続けます。

   列挙 すべてのヒストグラム タイプを (ツリーごとに 1 つずつ) 順に繰り返し使用します。

OneHotExplicit カテゴリごとに 1 つの列を生成し、列の各セルの値 "1" または "0" でその列のカテゴリが行に含まれているかどうかを表します。

4. **[アルゴリズムと N フォールドのシード]** をオンにしてシード数を入力すると、データフローを何度実行しても、データが必ず同じ方法でテスト データとトレーニング データに分割されます。フローを実行するたびにランダムな分割を行う場合は、このフィールドをオフにします。
5. 相互検証を実行する場合は **[N フォールド]** をオンにし、フォールドの数を入力します。
6. 相互検証を実行する場合は、**[フォールド割り当て]** をオンにしてドロップダウン リストから選択します。このフィールドは、**[N フォールド]** に値が入力済みで、**[フォールド フィールド]** が指定されていない場合にのみ適用可能です。

   Auto オプションの自動選択をアルゴリズムに任せます。現在、[ランダム] が選択されます。

   Modulo データセットをフォールドに等分し、シードを基準としません。

   Random データを n フォールドのサブセットにランダムに分割します。大きなデータセットに最適です。

7. 相互検証を実行する場合は、**[フォールド フィールド]** をオンにして、相互検証フォールド インデックス割り当てを含むフィールドをドロップダウン リストから選択します。

   このフィールドは、**[N フォールド]** と **[フォールド割り当て]** に値が入力されていない場合のみ適用可能です。

8. **[停止の基準回数]** をオンにすると、指定した回数のトレーニングで Stopping_metric オプションの改善が見られないとき、トレーニングの停止前に失敗したトレーニングの回数が入力されます。この機能を無効にするには、0 を指定します。この指標は Validation データに基づいて計算されます (提供されている場合)。そうでなければ、トレーニング データが使われます。
9. **[停止指標]** を選択して、新しいツリーの生成を終了するタイミングを決定します。

   Auto デフォルトは `deviance` です。

   deviance 平均残差逸脱度 (MSE)。

   MAE 平均絶対誤差。2 つの連続変数の間の差異です。

   MSE 平均 2 乗誤差。予測変数の分散とバイアスを包含する誤差です。

   RMSE 2 乗平均平方根誤差。モデルや評価関数によって予測された値 (サンプルや母集団の値) と実際に観測した値との差異を表します。MSE の平方根でもあります。

   RMSLE 2 乗対数平均平方根誤差。予測値と実測値の比率を表します。

10. **[停止の基準許容値]** をオンにし、指標に基づく停止の相対許容誤差を指定する値を入力すると、改善がこの値未満の場合にトレーニングが終了します。
11. **[最小分割改善]** をオンにし、2 乗誤差が低減したときに分割が行われるように最小の相対的な改善を指定する値を入力します。このオプションは、適切に実行すれば、過剰適合を減らす効果があります。最適な値は 1e-10...1e-3 の範囲でしょう。このフィールドは、**[停止の基準回数]** をオンにしている場合にのみ有効になります。
12. **[OK]** をクリックして、モデルと設定を保存するか、次のタブで操作を続行します。

モデル出力

このタブには、適合されたモデルの評価に使用するメトリクスが表示されます。これらのフィールドは編集できません。[トレーニング] 列には、必ずデータが含まれます。[基本オプション] タブでテストとトレーニングの分割を選択した場合は、[テスト] 列にもデータが設定されます。ただし、[高度なオプション] タブで N フォールド検証を選択した場合を除きます。その場合は、[N フォールド] 列にデータが設定されます。

ジョブを実行すると、結果として得られたモデルが Spectrum™ Technology Platform サーバーに格納されます。出力を再生成するには **[出力]** ボタンをクリックし、出力全体を Machine Learning モデル管理ツールに表示するには **[モデルの詳細]** をクリックします。

10 - Java Model Scoring

このセクションの構成

Java Model Scoring の概要

Java Model Scoring では、機械学習モデルの適合を行った時に作成された式を使用して、新しいデータをスコアリングすることができます。

> **注：** モデルを Java Model Scoring ステージで使用できるようにするにはまず、Machine Learning モデル管理を介してそれを公開する必要があります。詳細については、**Machine Learning モデル管理の概要**（47ページ）を参照してください。

データをスコアリングするには、**[Java Model Scoring オプション]** ダイアログで 2 つのタブの設定を終える必要があります。まずモデルとそのタイプを指定してから、モデルのフィールドが Spectrum™ Technology Platform のフィールドに正しくマッピングされていることを確認します。続いて、出力を設定します。ジョブに含めるフィールドを選択し、ジョブを実行します。**[モデル出力]** タブには、Spectrum™ Technology Platform とモデルのデータ タイプのマッピングが含まれます。

ジョブに、ファイルまたはテーブルに出力を取得するステージが含まれている場合は、後続のデータフローまたは Web サービスでその出力を使用できます。

モデルのプロパティの定義

1. **[プライマリ ステージ] / [展開済みステージ] / [Advanced Analytics]** の下で、**[Java Model Scoring]** ステージをクリックしてキャンバス上にドラッグし、データフロー内の所望の位置に配置して、入出力ステージに接続します。入力ステージは、モデルの目標フィールドと入力変数フィールドの両方を含むデータ ソースでなければならないことに注意してください。ジョブをバッチ モードで実行する場合は、モデル スコアを取得するための出力ステージも必要です。あるいは、Spectrum™ Technology Platform の Web サービスを使用して、リアルタイムでデータをスコアリングします。
2. Java Model Scoring ステージをダブルクリックして、**[Java Model Scoring オプション]** ダイアログ ボックスを表示します。
3. 必要に応じて、スコアリング対象のモデルのタイプを **[タイプ フィルタ]** ドロップダウンから選択します。
4. モデルのスコアリングに使用する **[タイプ フィルタ]** を選択します。
5. **[モデル名]** をドロップダウンから選択します。
6. スコアリング対象のモデルのタイプを、**[モデル タイプ]** フィールドに入力します。

7. 必要に応じて、モデルの **[説明]** を入力します。
8. モデルの適合に使用したデータに含まれないカテゴリ レベルの行について、[Predicted_Value] 列にデータを返すには、**[不明なカテゴリ レベルを無視]** をオンにします。このボックスをオフのままにすると、これらの行の [Predicted_Value] 列は「Null/NA」を返します。
9. **[入力]** テーブルには、モデルの入力フィールドに関する情報が表示されます。これらのフィールドとそのデータ タイプは、Spectrum のフィールドとデータ タイプに自動的にマッピングされます。
10. **[OK]** をクリックしてこれらのオプションを保存するか、次のタブで操作を続行します。

モデル出力

[出力] テーブルには、モデルの出力フィールドに関する情報が表示されます。これらのフィールドとそのデータ タイプは、Spectrum のフィールドとデータ タイプに自動的にマッピングされます。

1. モデルの出力に含めるデータの各フィールドに対し、**[含める]** をクリックします。
2. **[OK]** をクリックしてモデルを保存します。

11 - Machine Learning モデル管理

このセクションの構成

Machine Learning モデル管理の概要

Machine Learning モデル管理の [モデル分析] タブには、Spectrum™ Technology Platform サーバー上にある機械学習モデルの全一覧が表示されます。テキスト ボックスに文字列を入力することによって、この一覧にフィルタを適用することができます。その文字列によって、テーブルのすべてのフィールドが検索されます。

これらのモデルに対して、複数の操作が実行できます。モデルのエクスポーズ、アンエクスポーズ、削除が可能です。エクスポーズされたモデルは Java Model Scoring ステージで、機械学習モデルの適合を行った時に作成された式を使用して、新しいデータをスコアリングするために使用されます。また、各モデルの詳細情報を表示できます。詳細情報は、データを表示するモデルのタイプによって異なります。最後に、同じタイプの任意の 2 つのモデルを比較できます。比較を実行すると、比較する各モデルに対して [モデルの詳細] タブに表示されるのと同じ情報が、左右に並んで表示されます。

Machine Learning モデル管理のモデル分析へのアクセス

Machine Learning モデル管理には、次の 3 つの方法でアクセスできます。

- Spectrum™ Technology Platform ようこそページを使用します。
  - Web ブラウザを起動し、次の Spectrum™ Technology Platform の Welcome ページを開きます。

    http://<サーバー名>:<ポート>

    例えば、Spectrum™ Technology Platform が "myspectrumplatform" という名前のコンピュータにインストールされており、デフォルトの HTTP ポート 8080 を使用している場合は、次のアドレスに移動します。

    http://myspectrumplatform:8080

  - **[Spectrum Machine Learning]** をクリックします。
  - **[Machine Learning リポジトリを開く]** をクリックします。
- いずれかのモデル構築ステージから **[モデルの詳細についてはここをクリック]** をクリックします。
- Web ブラウザを使用します。
  - Web ブラウザを起動し、以下の Spectrum™ Technology Platform の Machine Learning モデル管理ページを開きます。

http://<サーバー名>:<ポート>/machinelearning

例えば、Spectrum™ Technology Platform が "myspectrumplatform" という名前のコンピュータにインストールされており、デフォルトの HTTP ポート 8080 を使用している場合は、次のアドレスに移動します。

http://myspectrumplatform:8080/machinelearning

- Spectrum™ Technology Platform の有効なユーザ名とパスワードを入力します。
- ツールが起動したら、**[モデル分析]** タブをクリックします。

モデル管理のモデル分析操作

モデルを選択して該当するボタンをクリックすることによって、以下の操作を実行します。

	モデルをエクスポーズして、Java Model Scoring ステージで使用できるようにします。エクスポーズされていないモデルを、スコアリングに使用することはできません。
	モデルをアンエクスポーズします。
	モデルを削除します。 **注：** エクスポーズされているモデルを削除することはできません。ただし現時点では、他のユーザのモデルを削除できないようにするためのセキュリティ機能は装備されていません。
	モデル出力の詳細を表示します。K-Means Clustering ステージと Logistic Regression ステージからのこの情報は、[モデル出力] タブの [モデルの詳細についてはここをクリック] をクリックすることによっても参照できます。
	モデルを比較します。

[モデルの詳細] タブ

[モデルの詳細] 画面には、すべてのモデルに関する以下の情報が表示されます。

- **[モデル名]** — モデルの名前
- **[モデル タイプ]** — 機械学習モデルのタイプ

- **[ユーザ]** — モデルを作成したユーザのユーザ名
- **[説明]** — モデルの説明 (作成時に記述された場合)
- **[ステータス]** — モデルがエクスポーズされているかどうか
- **[データフロー名]** — モデルを生成したデータフローの名前
- **[作成時間]** — モデルが作成された日時

モデル タイプに応じて、その他の詳細情報が表示されます。

K-Means Clustering の詳細情報

[モデルの詳細] 画面には、K-Means Clustering モデルに関する以下の情報が表示されます。

モデル サマリ

- 行数
- クラスタ数
- カテゴリ列数
- 反復回数
- クラスタ内平方和
- 総平方和
- クラスタ間平方和

メトリクス

以下の項目に対するトレーニング、テスト、N フォールド データを提供します。

- 総クラスタ内平方和
- 総平方和
- クラスタ間平方和

セントロイド統計

各中心点 (セントロイド) に対する以下のトレーニング、テスト、N フォールド データを提供します。

- サイズ
- クラスタ内平方和

クラスタ平均

各中心点の詳細情報を提供します。内容は入力データによって異なります。クラスタとは、特定のクラスタリング アルゴリズムに基づいて類似と識別された、データ セットからのオブザベーションのグループです。

標準化されたクラスタ平均

各中心点の正規化情報を提供します。内容は入力データによって異なります。

Logistic Regression の詳細情報

[モデルの詳細] 画面には、Logistic Regression モデルに関する以下の情報が表示されます。

メトリクス

以下の項目に対するトレーニング、テスト、N フォールド データを提供します。

- 平均二乗誤差 (MSE)
- 平方根平均二乗誤差 (RMSE)
- オブザベーション数
- R の二乗 (R2)
- 対数損失 (Logloss)
- 曲線下面積 (AUC)
- ジニ係数
- クラスあたり平均誤差
- 赤池情報量基準 (AIC)
- 残差逸脱度
- Null 逸脱度
- Null 自由度
- 残差自由度

最大メトリクスしきい値

以下のメトリクスを使用するトレーニング、テスト、N フォールド データに対する、トレーニング最大メトリクスしきい値を提供します。

- f1 最大値
- f2 最大値
- f0point5 最大値
- 最大正確度
- 最大適合率
- 最大再現率
- 最大特異度
- absolute_mcc 最大値
- min_per_class_accuracy 最大値
- mean_per_class_accuracy 最大値

混同行列

真 (true) の値が既知の一連のトレーニング、テスト、N フォールド データに対するモデルのパフォーマンスを表します。

標準化係数チャート

入力がどれだけ変化すると目標が変化するかを表す、係数の相対値を提供することによって、最も重要な予測因子を示します。

GLM 係数

指数分布に従う結果の回帰モデルを推定する、一般化線形モデル (GLM: Generalized Linear Model) の係数を示します。

AUC 曲線

曲線下面積 (AUC)。使用モデルの中で、トレーニング、テスト、N フォールド データを使用して最も正確にクラスを予測するものを判定します。

リフト/ゲイン曲線

トレーニング、テスト、N フォールド データを使用してバイナリ分類モデルの予測能力を評価します。

Logistic Regression の詳細情報

[モデルの詳細] 画面には、Linear Regression モデルに関する以下の情報が表示されます。

メトリクス

以下の項目に対するトレーニング、テスト、N フォールド データを提供します。

- 平均二乗誤差 (MSE)
- 平方根平均二乗誤差 (RMSE)
- オブザベーション数
- R の二乗 (R2)
- 平均残差逸脱度
- 平均絶対誤差 (MAE)
- 二乗対数平均平方根誤差 (RMSLE)
- 赤池情報量基準 (AIC)
- 残差逸脱度
- Null 逸脱度
- Null 自由度
- 残差自由度

標準化係数チャート

特定の予測係数値の変化により目標値がどれだけ変化 (正または負の変化) するかを表す係数の相対値を提供することによって、最も重要な予測因子を示します。さらに、モデルの上位 25 の係数をグラフで示します。

GLM 係数

指数分布に従う結果の回帰モデルを推定する、一般化線形モデル (GLM: Generalized Linear Model) の係数を示します。

Random Forest Regression の詳細情報

[モデルの詳細] 画面には、Random Forest Regression モデルに関する以下の情報が表示されます。

メトリクス

以下の項目に対するトレーニング、テスト、N フォールド データを提供します。

- 平均二乗誤差 (MSE)
- 平方根平均二乗誤差 (RMSE)
- オブザベーション数
- R の二乗 (R2)
- 平均残差逸脱度
- 平均絶対誤差 (MAE)
- 二乗対数平均平方根誤差 (RMSLE)

変数重要度

以下のメトリクスを使用して、各変数の重要度の値を提供します。

- 相対的重要度
- 小数点以下桁数を含む重要度
- 比率

さらに、モデルの上位 25 の変数をグラフで示します。

Random Forest Classification の詳細情報 — 二項

[モデルの詳細] 画面には、Random Forest Classification の二項モデルに関する以下の情報が表示されます。

メトリクス

以下の項目に対するトレーニング、テスト、N フォールド データを提供します。

- 平均二乗誤差 (MSE)
- 平方根平均二乗誤差 (RMSE)
- オブザベーション数
- R の二乗 (R2)
- Logloss
- 曲線下面積 (AUC)
- ジニ
- クラスあたり平均誤差

最大メトリクスしきい値

以下のメトリクスを使用するトレーニング、テスト、N フォールド データに対する、トレーニング最大メトリクスしきい値を提供します。

- f1 最大値
- f2 最大値
- f0point5 最大値
- 最大正確度
- 最大適合率
- 最大再現率
- 最大特異度
- absolute_mcc 最大値
- min_per_class_accuracy 最大値
- mean_per_class_accuracy 最大値

混同行列

真 (true) の値が既知の一連のトレーニング、テスト、N フォールド データに対するモデルのパフォーマンスを表します。

変数重要度

以下のメトリクスを使用して、各変数の重要度の値を提供します。

- 相対的重要度
- 小数点以下桁数を含む重要度
- 比率

さらに、モデルの上位 25 の変数をグラフで示します。

AUC 曲線

曲線下面積 (AUC)。使用モデルの中で、トレーニング、テスト、N フォールド データを使用して最も正確にクラスを予測するものを判定します。

リフト/ゲイン曲線

トレーニング、テスト、N フォールド データを使用してバイナリ分類モデルの予測能力を評価します。

Random Forest Classification の詳細情報 — 多項

[モデルの詳細] 画面には、Random Forest Classification の**多項**モデルに関する以下の情報が表示されます。

メトリクス

以下の項目に対するトレーニング、テスト、N フォールド データを提供します。

- 平均二乗誤差 (MSE)
- 平方根平均二乗誤差 (RMSE)
- オブザベーション数
- R の二乗 (R2)
- Logloss
- クラスあたり平均誤差

混同行列

真 (true) の値が既知の一連のトレーニング、テスト、N フォールド データに対するモデルのパフォーマンスを表します。

変数重要度

以下のメトリクスを使用して、各変数の重要度の値を提供します。

- 相対的重要度
- 小数点以下桁数を含む重要度
- 比率

さらに、モデルの上位 25 の変数をグラフで示します。

主成分分析の詳細情報

[モデルの詳細] 画面には、主成分分析 (PCA) モデルに関する以下の情報が表示されます。

コンポーネントの重要度

主要コンポーネントを、以下のメトリクスに基づき重要度順に表示します。

- 標準偏差
- 寄与率
- 累積寄与率

回転

変数負荷量の行列をグラフで示します。コンポーネントのスコアを算出するには、正規化された元の各変数にこの重みを掛ける必要があります。

著作権に関する通知

USPS® 情報

Pitney Bowes Inc. は、ZIP + 4® データベースを光学および磁気媒体に発行および販売する非独占的ライセンスを所有しています。CASS、CASS 認定、DPV、eLOT、FASTforward、First-Class Mail、Intelligent Mail、LACS^{Link}、NCOA^{Link}、PAVE、PLANET Code、Postal Service、POSTNET、Post Office、RDI、Suite^{Link}、United States Postal Service、Standard Mail、United States Post Office、USPS、ZIP Code、および ZIP + 4 の各商標は United States Postal Service が所有します。United States Postal Service に帰属する商標はこれに限りません。

Pitney Bowes Inc. は、NCOA^{Link}® 処理に対する USPS® の非独占的ライセンスを所有しています。

Pitney Bowes Software の製品、オプション、およびサービスの価格は、USPS® または米国政府によって規定、制御、または承認されるものではありません。RDI™ データを利用して郵便送料を判定する場合に、使用する郵便配送業者の選定に関するビジネス上の意思決定が USPS® または米国政府によって行われることはありません。

データ プロバイダおよび関連情報

このメディアに含まれて、Pitney Bowes Software アプリケーション内で使用されるデータ製品は、各種商標によって、および次の 1 つ以上の著作権によって保護されています。

© Copyright United States Postal Service.All rights reserved.

© 2014 TomTom.All rights reserved.TomTom および TomTom ロゴは TomTom N.V の登録商標です。

© 2016 HERE

Fuente: INEGI (Instituto Nacional de Estadística y Geografía)

電子データに基づいています。© National Land Survey Sweden.

© Copyright United States Census Bureau

© Copyright Nova Marketing Group, Inc.

このプログラムの一部は著作権で保護されています。© Copyright 1993-2007 by Nova Marketing Group Inc. All Rights Reserved

© Copyright Second Decimal, LLC

© Copyright Canada Post Corporation

この CD-ROM には、Canada Post Corporation が著作権を所有している編集物からのデータが収録されています。

© 2007 Claritas, Inc.

Geocode Address World データ セットには、**http://creativecommons.org/licenses/by/3.0/legalcode** に存在するクリエイティブ コモンズ アトリビューション ライセンス (「アトリビューション ライセンス」) の下に提供されている GeoNames Project (**www.geonames.org**) からライセンス供与されたデータが含まれています。お客様による GeoNames データ (Spectrum™ Technology Platformユーザ マニュアルに記載) の使用は、アトリビューション ライセンスの条件に従う必要があり、お客様と Pitney Bowes Software, Inc. との契約と、アトリビューション ライセンスの間に矛盾が生じる場合は、アトリビューション ライセンスのみに基づいてそれを解決する必要があります。お客様による GeoNames データの使用に関しては、アトリビューション ライセンスが適用されるためです。

3001 Summer Street
Stamford CT 06926-0700
USA

www.pitneybowes.com