Vertica バージョン 8.0 の新機能：機械学習の拡張機能

原文はこちら

Vertica 7.2.2 で追加された予測分析のための機械学習の機能が、Vertica 8.0 において、全体的により直感的で使いやすくなっていることに気がつくでしょう。API は合理化され、より使い始めやすくなりました。また、データ準備のための機能も追加されています。
Vertica には定期的に機械学習に関する新機能が追加されています。将来のリリースにおいて、より拡張された機能をご期待ください！
機械学習の技術を初めてお使いの場合は、本ブログページの「Vertica での機械学習とは？それを使って何ができるのか？」のセクションをご覧ください。

拡張機能の概要

新機能	説明
余分なセットアップが不要	以前は、機械学習の関数は、オプションのパッケージに含まれていました。パッケージは Vertica の RPM に付随しておりましたが、クラスタ内の各ノード上で別々のインストールを実施する必要がありました。8.0 では、機械学習の関数は、Vertica のインストールに含まれており、自動的に利用できるようになりました。
一般に公開される	Vertica 8.0 では、機械学習の関数は、public スキーマに存在し、どのユーザーでも利用可能となりました。7.2.2 において機械学習の関数が存在していた v_ml スキーマは、8.0 では非推奨となります。
モデル用のテーブル	トレーニングされた機械学習のモデルのリストを格納するテーブルが、V_CATALOG スキーマにシステムテーブルとして保持されるようになりました。7.2.2 において同情報が格納されていた v_ml.models テーブルは、8.0 では非推奨となります。
新しい関数名	Vertica 8.0 では、機械学習の関数名は、よりシンプルで直感的になりました。
モデルのサマリー情報取得のための簡略化されたシンタックス	7.2.2 では、各アルゴリズムタイプ毎に別々のモデルのサマリー関数がありました。 8.0 では、すべてのモデルに対して、ひとつの SUMMARIZE_MODEL 関数が提供されています。新しい関数は、アルゴリズムを検出します。

Vertica バージョン 8.0 における予測分析のための機械学習

8.0 では、機械学習の関数の多くが新しい名前に変更されています。新しい名前は自然言語により近く、8.0 では public スキーマに保持されるため、スキーマ名で修飾されなくなりました。New Features Guide に新旧の関数名がリスト化されています。

サポートされている 3 つのアルゴリズムの各モデルの作成方法を示すサンプルコードについては、7.2.2 のブログ（Learn More From Your Data With Machine Learning Algorithms）を参照ください。7.2.2 での関数名を 8.0 での新しい関数名に置き換えるだけで、7.2.2 で動作したように 8.0 でも動作するサンプルとなります。

タスク	8.0 での関数名	説明
データの準備	**NORMALIZE** **BALANCE** **TABLESAMPLE**	数値列の値を縮小します。 分類モデルのためのトレーニングデータにおけるクラスの分布をバランスします。 データからランダムサンプルを取得します。
モデルの作成	**LOGISTIC_REG** **LINEAR_REG** **K_MEANS**	二項ロジスティック回帰、線形回帰、または k-Means クラスタリングを実行するモデルを作成します。
モデルの評価	**CONFUSION_MATRIX** **LIFT_TABLE** **ROC** **ERROR_RATE** **MSE** **RSQUARED**	混同行列、リフトテーブル、ROC 曲線を生成する、あるいは、エラーレートを見つけて分類モデルを評価します。 線形回帰モデルを評価するために、平均二乗誤差、または、R 二乗値を求めます。
モデルの要約	**SUMMARIZE_MODEL**	機械学習のモデルは非常にブラックボックスです。モデルの要約は、モデルに関する重要な情報を明らかにするためのメカニズムです。どのような計算が実行されていますか？ それはどのような前提ですか？ 要約は、トレーニングされたモデルのメトリスクを解釈し、それを新しいデータに適用する方法を学習するのに役立ちます。
モデルの適用	**PREDICT_LOGISTIC_REG** **PREDICT_LINEAR_REG** **APPLY_KMEANS**	ロジスティック回帰、線形回帰、または k-Means モデルを新しいデータに適用します。
モデル管理のための DDL	**DELETE_MODEL** **RENAME_MODEL**	モデルの削除または名前の変更をします。
モデルのサマリー情報取得のための簡略化されたシンタックス		7.2.2 では、各アルゴリズムタイプ毎に別々のモデルのサマリー関数がありました。 8.0 では、すべてのモデルに対して、ひとつの SUMMARIZE_MODEL 関数が提供されています。新しい関数は、アルゴリズムを検出します。

データ準備のための新しいオプション

BALANCE 関数と SELECT での TABLESAMPLE 句は、Vertica 8.0 での新しい機能となります。どちらも機械学習にとって重要なデータ準備の技術となります。

BALANCE

バランシングは、分類モデルの予測品質を向上させるために、トレーニングデータ内の分類の分布を変更する技術です。
672 の金融取引のデータセットがあるとします。2 件の取引が、不正とされていたとします。

```
/***View the imbalanced training data
=> SELECT fraud, COUNT(fraud) FROM financial_transactions GROUP BY fraud;
 fraud | COUNT
-------+-------
 t     |     2
 f     |   670
```

このデータセットを使用して、日々発生する数十万件の取引における不正を検出するための分類モデルをトレーニングしたいとします。しかしながら、トレーニングアルゴリズムは、不正であるものと不正でないものの比率が著しく不均衡であるために、予測能力を開発するためのトレーニングデータにおいて不正の十分な事例を検出することができない場合があるかもしれません。言い換えれば、新しい入力データにおける不正を正確に予測するための高品質モデルのトレーニングは、どのアルゴリズムでも非常に困難です。これは、トレーニングデータのバランスを取って新しいモデルを作成することが理にかなっているシナリオです。

```
/*** balance the training data
=> SELECT BALANCE('balance_fin_data', 'financial_transactions', 'fraud',
'weighted_sampling');
```

BALANCE 関数は、現在、データのバランスをとるための 1 つのメカニズムをサポートしています。優勢なクラスをアンダーサンプリングするために重み付きサンプリングを使用します。

```
/*** View the balanced training data
=> SELECT fraud, COUNT(fraud) FROM balance_fin_data GROUP BY fraud;
 fraud | COUNT
-------+-------
 t     |     2
 f     |   237
```

新しいモデルをトレーニングするために、バランスがとられたデータセットである balance_fin_data を使います。今回は、モデルに予測力がある可能性があります。混同行列は、おそらくいくつかの偽陰性と偽陽性を示すでしょう。リフトチャートと ROC 曲線を確認してください。モデルの品質に満足するまで、バランシングからトレーニング、テストまでの全プロセスを繰り返します。
BALANCE 関数の詳細につきましては、Vertica documentation をご覧ください。

ランダム サンプリング

サンプリングは、より大きな母集団から代表的なデータセットを導き出すための標準的な統計的手法です。Vertica 8.0 では、SELECT 文の FROM 句で TABLESAMPLE 句を指定することにより、ランダムなサンプルを取得することができます。

```
/** Return a random sample containing approximately 50% of the rows in a
table of customers
=> SELECT * FROM customers TABLESAMPLE(50);
```

TABLESAMPLE 句は、予測モデルのトレーニングとテストに必要なデータセットを取得するのに便利です。SELECT での TABLESAMPLE の使い方についての詳細は、Vertica documentation をご覧ください。

Vertica バージョン 7.2.2 からのアップグレード

Vertica バージョン 7.2.2 で機械学習のモデルを作成し、Vertica バージョン 8.0 へのアップグレード後も作成したものを使用したい場合、アップグレードスクリプトを実行する必要があります。以下の手順をご覧ください。
Upgrading the Advanced Analytics Package 7.2.x to Machine Learning 8.0.x

Vertica での機械学習とは？それを使って何ができるのか？

機械学習とは、データから学ぶために複雑な数学的アルゴリズムを使用するモデルを構築および展開するプロセスを指します。機械学習は、非常に大きなデータセットに適用すると最も効果的です。したがって、Big Data の高速処理用に設計された Vertica には当然適しています。最近の機械学習の導入により、Vertica は Big Data 分析のためのプラットフォームとしての価値を大幅に高めています。
機械学習には、教師ありと教師なしという大きく 2 つのタイプがあります。Vertica は、現在、2 つの教師ありのアルゴリズムと 1 つの教師なしのアルゴリズムをサポートしています。

Vertica での教師あり学習

教師ありのアルゴリズムは、各データポイントの期待されるアウトプットを含むトレーニングデータから学習します。トレーニングプロセスの結果であるモデルは、可能性のある結果を予測するために関心のある母集団に適用することができます。たとえば、過去の金融行動に基づくクレジットスコアや、過去の購買行動に基づくプロモーションの対象とする顧客層の中でもっとも優れた顧客を予測するために使われます。
Vertica は、教師あり学習の 2 つの主要な方式である分類と回帰のアルゴリズムをサポートしています。

- 分類は、可能性のあるカテゴリを予測します。例えば、分類分析を使用して、高いまたは低いテストスコア、あるいは、顧客離れする可能性があるか否かなどについて予測することができます。Vertica は、分類分析では、**二項ロジスティック回帰**をサポートします。

- 回帰は、連続した値を予測します。例えば、**回帰分析**を使用して、過去のデータに基づいてテストスコアまたは不動産の価値を予測することができます。Vertica は、回帰分析では、線形回帰をサポートしています。

Vertica での教師なし学習

教師なしのアルゴリズムは関心のある母集団から直接学習します。教師なし学習のアプリケーションの中には、顧客セグメンテーション、異常検出、市場バスケット分析などがあります。教師なし学習の結果は新しいデータに適用されることがありますが、予測は教師なし学習の主な目的ではありません。
Vertica は、教師なし学習のためのクラスタリングアルゴリズムである **k-Means** をサポートしています。クラスタリングでは、データ内の自然なグルーピングが検出されるため、クラスタ内のアイテムは、クラスタ外のアイテムよりも相互に共通項が多い状態となります。

機械学習関連情報

Vertica での機械学習に関する情報を得るためには、下記のドキュメントを参照してください。

概要説明：

- Machine Learning Functions in the *SQL Reference*
- Machine Learning for Predictive Analytics in *Analyzing Data*

個別のトピック：

- Data Preparation
- Clustering Data Using k-Means
- Building a Logistic Regression Model
- Building a Linear Regression Model