実践! ビジネス課題へのアナリティクス活用基礎講座

THE POWER TO KNOW.

18PC0

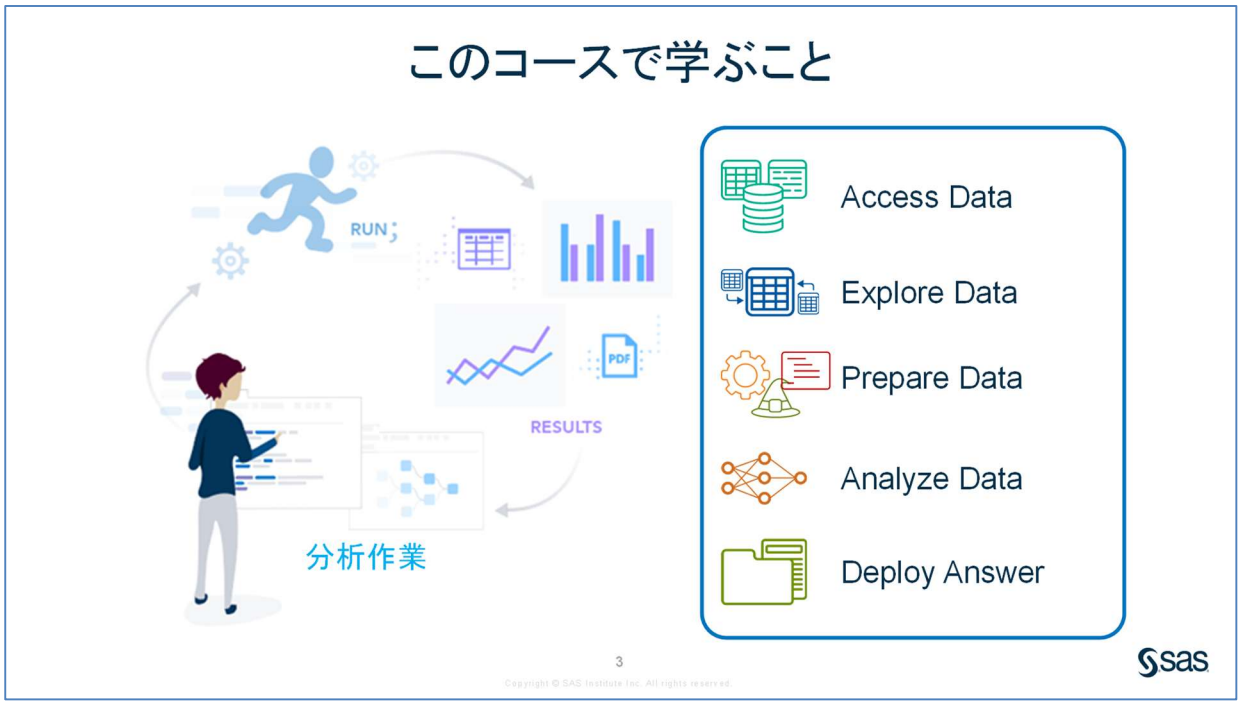
実践! ビジネス課題へのアナリティクス活用基礎講座

コースノート

目次

Lesson 1　　コースイントロダクション

1.1 本コースの学習目標



分析を依頼されたときに、まず、なにから始めますか？次に、なにを行いますか？

もし、自身で分析ができないと感じているならば、おそらくそれは、やり方を知らないだけなのかもしれません。しかし、逆にやり方さえわかっていれば、それは誰にでもできることなのかもしれません。

分析作業がいくつかのプロセスに分けることができることはご存知のことでしょう。それぞれのプロセスを実現するためには、その場面ごとに必要な知識を身に付けていなければなりません。その場面ごとに特化した、専門的な知識を身に付けている人のことを、「スペシャリスト」と呼ぶことがあります。

分析を始めようと思ったときに重要なことは、それぞれのプロセスにスペシャリスト並みの専門的な知識を身に付けることでしょうか。専門的な知識を身に付けるためには、多くの時間が必要になります。その知識を身に付ける時間を、ビジネス課題を抱える組織は、おそらく待ってくれません。

実は多くの場合に組織が求めているのは、内製化をするための即戦力、そのために広くビジネス全体を網羅して分析プロジェクトに参画し、自身で手を動かして活躍できる人材であるのではないでしょうか。その人物像に求められるのは、ひとつのプロセスに対して専門的な知識を持つスペシャリストではなく、全体のプロセスに最低限の知識を持つ「ジェネラリスト」としての能力です。

専門性を持つ知識が必要な場合には、それぞれのプロセスの専門家へアウトソースすることができるでしょう。一方で、幅広くそれぞれの専門家と話をするこのできる、ジェネラルな知識を持つ人材が必要であるとも言えます。

意思決定の迅速化、売上・利益の最大化を目標とする時に、そこに携わる人々は共通の認識を持ち、共通の用語で会話して、お互いのなすべき業務を理解して遂行しなければなりません。

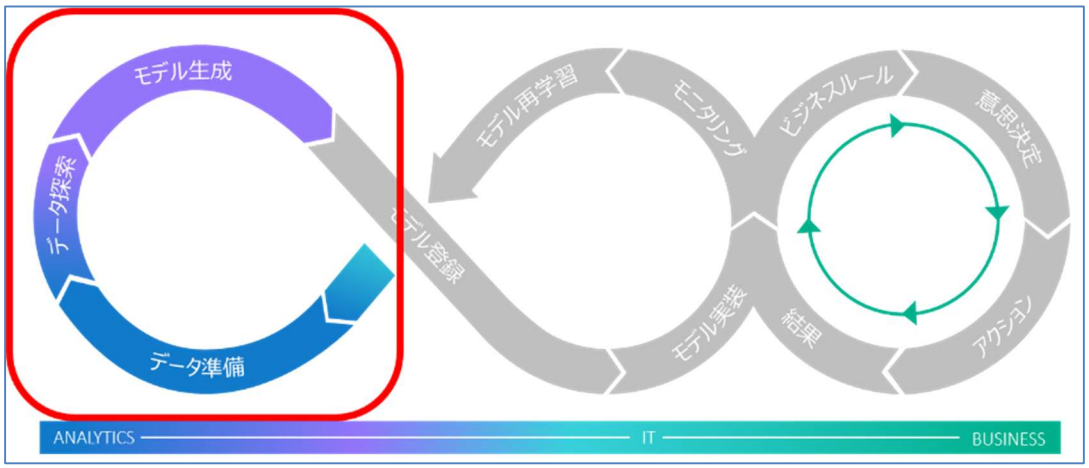
すべてのエリアに専門的な知識を持つことは一つの理想ではありますが、第一段階では広く知識を習得してジェネラリストを目指し、次の段階で専門的な内容を選択してスペシャリストを目指しましょう。

1.2 分析に必要なプロセス

本コースでは、アナリティクス・ライフサイクルの考え方を基にした分析のプロセスを、ハンズオンを中心として体験します。与えられた実践的な分析シナリオを解決するために、実際の業務を意識した流れでコースを進め、それぞれの場面で行う作業を理解します。

このコースを受講することにより、分析プロセス全体を理解し、自身で一通りの作業ができるようになることを目指します。

アナリティクス・ライフサイクルは「中心」を担う重要な考え方であり、機械学習の真の価値はアナリティクス・ライフサイクル全体を通した実用的な洞察から導き出されます。



その中で分析モデルの構築にフォーカスした、アナリティクス・ライフサイクルの 3 つのフェーズは、Data、Discovery、および Deployment です。

アナリティクス・ライフサイクルは、ローデータから価値を抽出することを目的とした一連のアクティビティです。価値の定義は、組織の目標と目的によって固有のものとなります。

Data フェーズ は、すべての分析活動の基盤となります。Discovery フェーズ は、あなたがこれまで知らなかった何かを発見する工程であり、Deployment フェーズは、あなたが分析から価値を導き出す場面です。

この 3 つのフェーズすべてを完全に理解して実施することが、データを価値に変え、影響力のある結果を生み出すことにつながります。どのフェーズからでも始めることはできますが、次のステップを知り、そこに到達する方法を検討することが重要です。

重要な Data のタスク

- データの分割
- まれなイベントの対処
- 欠損値の管理
- 非構造化データの追加
- 特徴抽出
- 極端な値や異常な値の処理
- 有用な入力の選択

7

効果的な機械学習モデルは、十分に準備・整備されたデータに基づいて構築されます。一般的に、成功する機械学習アプリケーションの考案に費やされる時間の 80%はデータの準備に費やされていると一般的に言われます。データの準備は、既存のデータを厳密に正しく変換し、クリーニングすることだけではありません。また、検討する必要がある特徴を十分に理解して、手元のデータが開始点として適切であることを確認することも含まれます。「garbage in、garbage out」と良く言われるように、データ準備の簡略化から価値あるモデルは得られません。時間をかけてデータを整理し、スライドにあるような一般的な問題に対応してください。

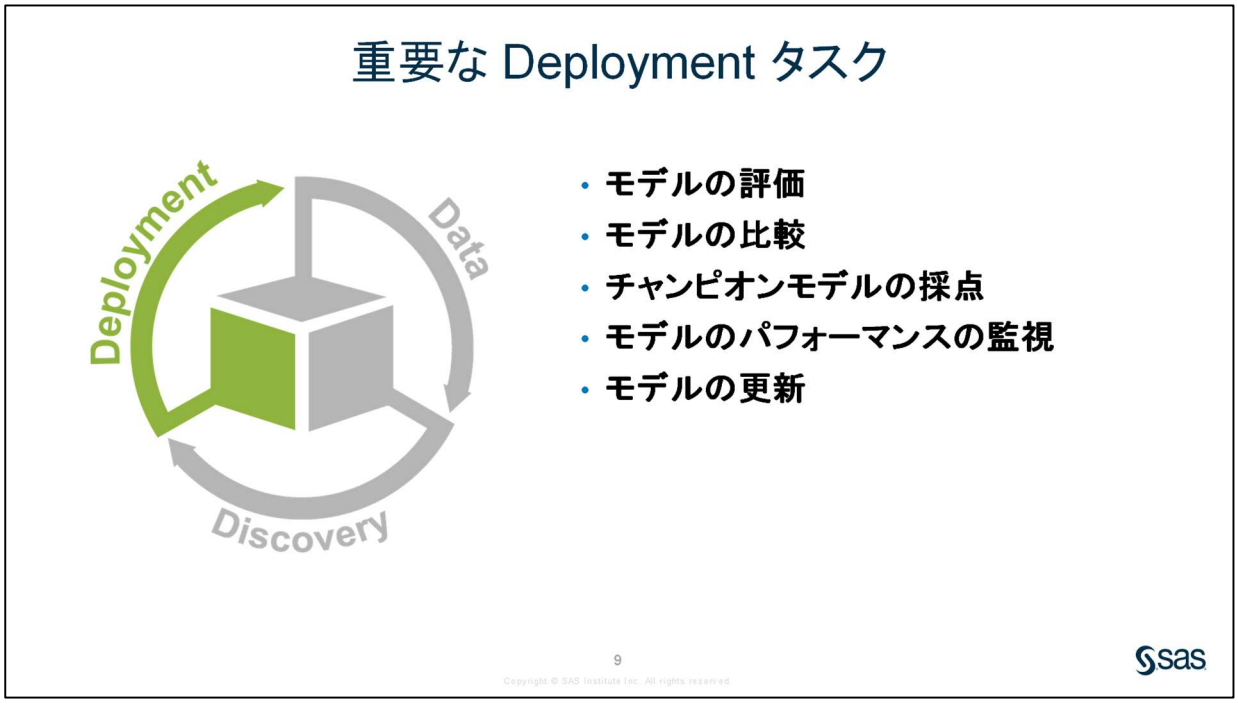
重要な Discovery タスク

- アルゴリズムの選択
- モデルの改善
- モデルの複雑さの最適化
- 正則化やモデルのハイパーパラメータのチューニング
- アンサンブルモデルの構築

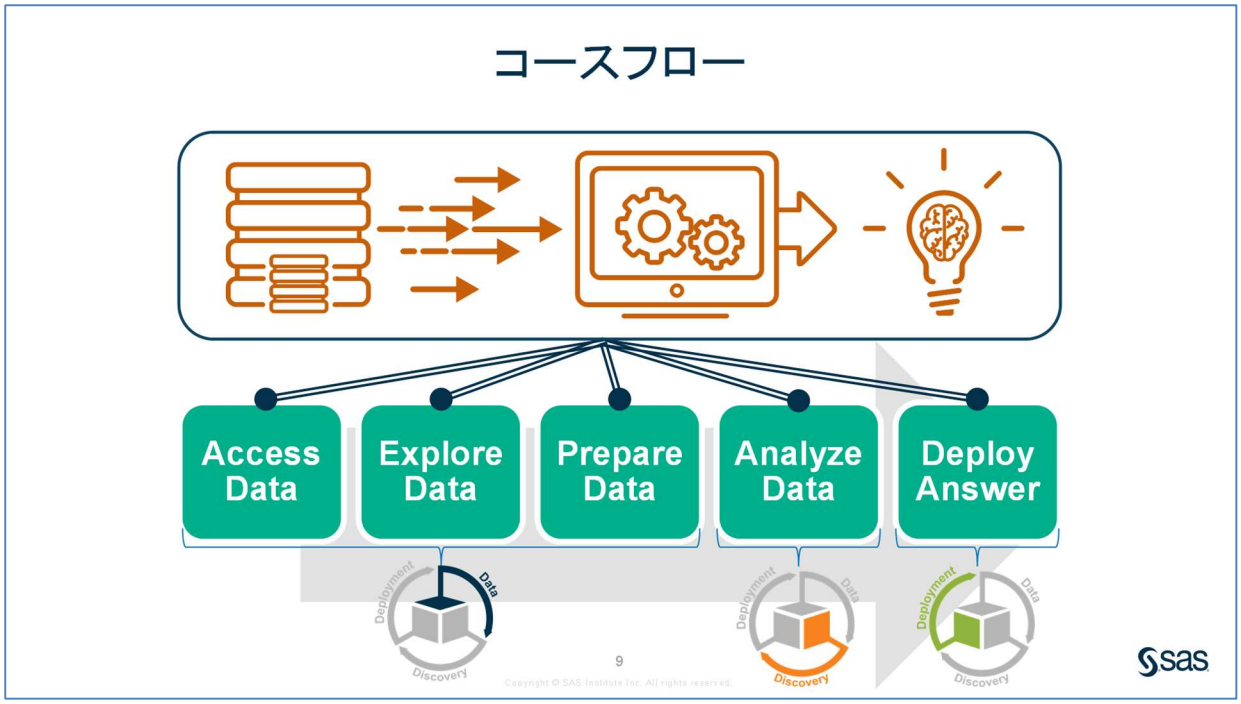
8

適切なデータを十分に用意し、モデリングに適した形式にデータをマッサージして、モデルに含める主要な機能を特定し、モデルの使用方法を確立することによって、強力な機械学習アルゴリズムを使用して予測モデルを構築したり、データのパターンを発見したりすることができ

ます。このフェーズは、特定のアプリケーションに最適なモデルを生成するアルゴリズム（およびそれらのアルゴリズムのオプションの構成）を特定するためのさまざまなアプローチを自由に実験できるようにする場面です。



通常、いくつかのモデルを構築します。したがって、最初に個々のモデルを評価してから、それらの複数のモデルを比較して、チャンピオンモデルと呼ばれる最適なモデルを決定することが重要です。その後、チャンピオンモデルが本番環境に展開されます（スコアリングと呼ばれるプロセス）。モデルを展開した後も、要件ごとにモデルを監視して更新する必要があります。



このスライドは、アナリティクス・ライフサイクルの工程を分割し、本コースで行う作業工程に変換したものです。

Access Data：様々なファイル形式のデータを、分析ツールで利用できるようにします。

Explore Data：データの構造や関係性、保存されている値について把握します。

Prepare Data：分析に適したデータの形に加工します。また必要な項目の作成や選択を行います。

Analyze Data：分析モデルを生成して評価を行います。

Deploy Answer：分析結果を考察し、ビジネス課題を解決するヒントを見つけ、意思決定をサポートするシナリオを作成します。

このコースでは、分析におけるプロセスを実際の操作を中心に体験し理解して、自身の環境で利用しているデータや、未知のデータに遭遇した際に、分析シナリオから同様のプロセスを実現していく力を身に付けることを目標としています。

本コースでは、機械学習から教師あり学習を行う際に必要な、データ分析の流れをハンズオンと共にご紹介いたします。

1.3 本コースのシナリオ

本コースの分析シナリオ

ポルトガル銀行では、顧客に対して定期預金の開設キャンペーンを実施しています。過去、現場の判断で多数の顧客に対して営業を実施していたため、販促費が増加しています。そこで、定期預金の開設見込みのある顧客に対してのみ重点的に営業を実施したいと考えています。

そこで、過去実績データを用いて、分析を活用することで効率的にアプローチが可能かを検証してほしいと、データサイエンス部に依頼がありました。また、結果を用いることで期待できる増収益の推定も期待されています。

※1回のコンタクトにかかる費用はおおよそ 3 EURと見積もっています。

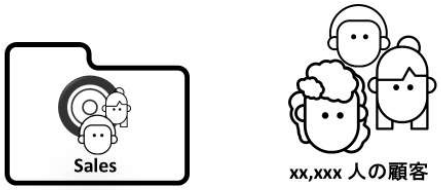
※定期預金を1件成約した際の粗利はおよそ 45 EURと見積もっています。

11

sas

シナリオから導き出した分析方法

- 過去のキャンペーン実施履歴と定期預金口座開設履歴のデータから、定期預金口座を開設するポテンシャルがある顧客か否かを分類する。
- 目的変数は定期預金口座を開設したか、していないかの2値
- 営業部にモデルの妥当性を説明するため、構築したモデルの判断根拠や重要変数を可視化
- 基礎集計情報と特徴量の重要度等の提供



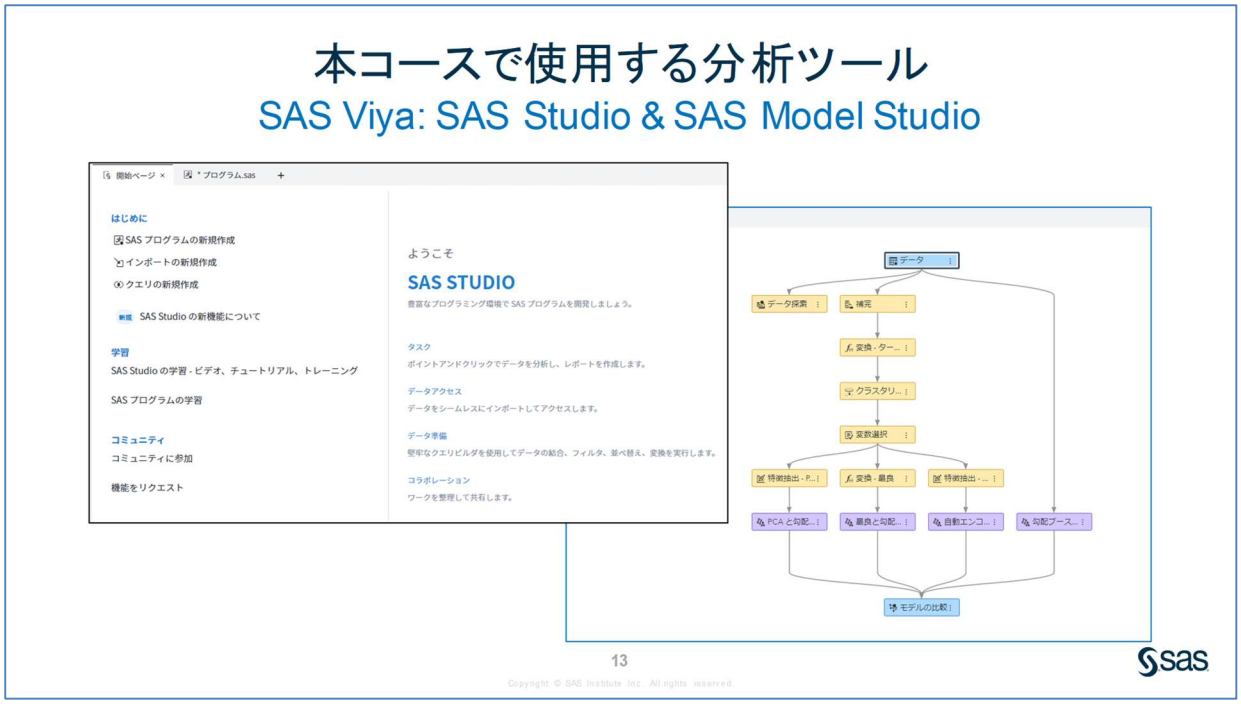
12

sas

【本コースで使用するデータについて】

K. Bache and M. Lichman (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

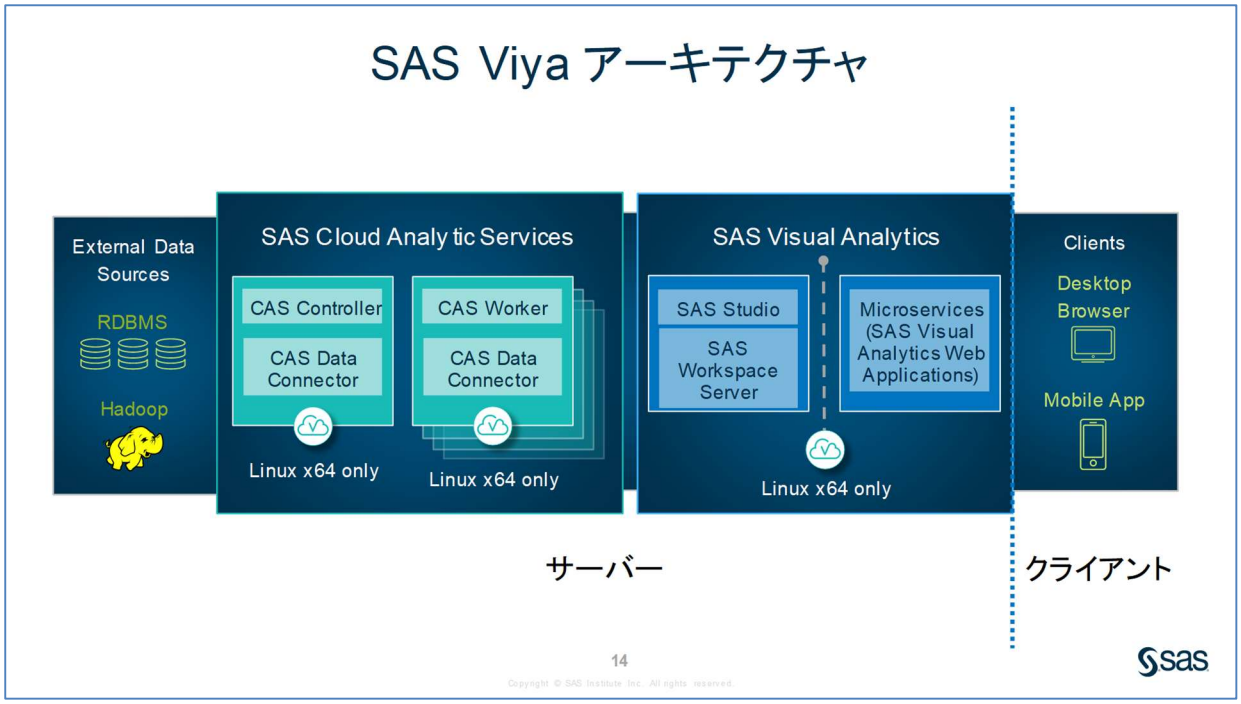
S. Moro, R. Laureano and P. Cortez (2011). Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, October, 2011. EUROSIS.



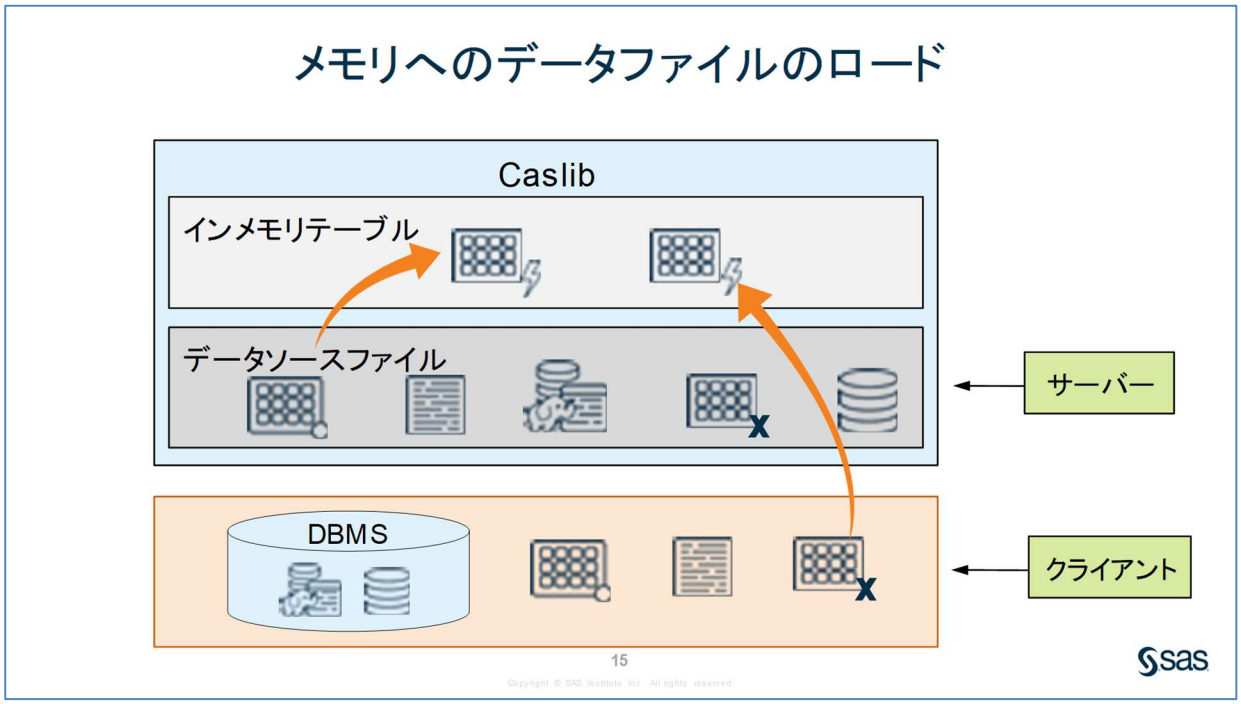
このコースでは、分析ツールとして SAS を使用します。様々な SAS アプリケーションの中から以下の二つを使用します：

・SAS Viya: SAS Studio：
SAS Studio は、Web ブラウザを介して SAS コードを記述および実行するツールです。SAS Studio を使用すると、既存のデータファイルへのアクセスや、新しいプログラムの作成ができます。事前定義済みプログラムであるスニペットを使用して SAS コードを生成したり、ステップを使用してフローを作成することで、プログラムを使用しないで作業を実施することも可能です。

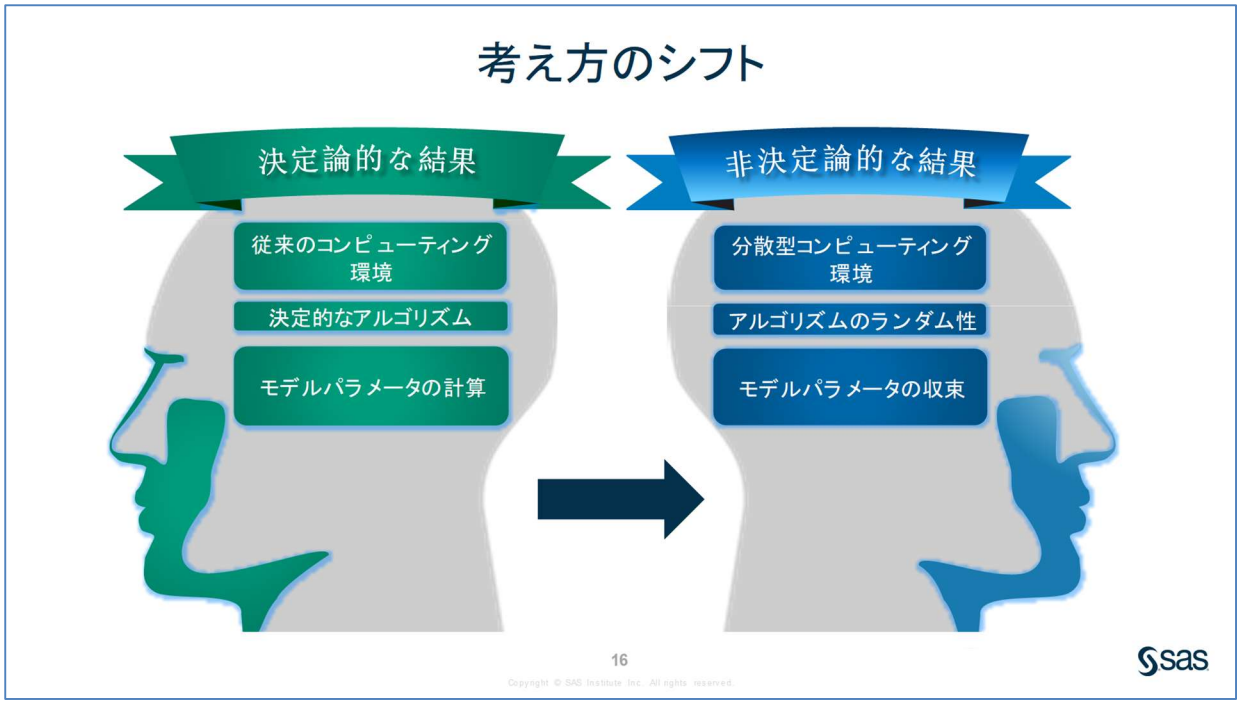
・SAS Viya: SAS Model Studio：
SAS Model Studio は、機械学習のモデル、時系列予測のモデル、テキストマイニングのモデルを GUI ベースの簡単なマウス操作で作成することができます。モデル生成プロセスをグラフィカルなフロー図として描き、実行するだけです。このフロー図のことを「パイプライン」と呼びます。



SAS Viya の中心には、SAS Cloud Analytic Services というインメモリの分析エンジンがあります。分散型のアーキテクチャを採用し、スケーラブルで高性能なマルチスレッドアルゴリズムを使用して、サイズを問わずインメモリデータに対して分析処理を迅速に実行します。サーバー側のインメモリデータに対して、Web ブラウザベースの目的に応じた各種クライアントアプリケーションで、ファイルアクセス、データ加工、分析、レポーティングといった処理を行うことができます。



クライアント側とサーバー側の両方のファイルをメモリ内のテーブルに直接ロードできます。



SAS Viya では、非決定論的な結果が出たり、再現性のある結果が得られなかったりすることがありますが、これは基本的に 2 つの理由によるものです。

- 分散コンピューティング環境
- 非決定論的アルゴリズム

分散コンピューティングでは、ケースが計算ノードに分割されるため、結果にばらつきが生じる可能性があります。コントローラやワーカーの管理が行き届いている同じサーバーでも、微妙に違う結果になるかもしれません。異なるサーバーでは、さらにその可能性が高くなります。CAS サーバーは、プールされたメモリを表し、コードをマルチスレッドで実行します。マルチスレッドでは、同じ命令を他の利用可能なスレッドに分配して実行する傾向があり、別々のデータの割り当てやサブセットを使用して、多くの異なるコアに多くの異なるキューを作成します。ほとんどの場合、複数のスレッドは、互いに独立しているが、より大きなテーブルの一部である、孤立したデータの集まりに対して実行します。そのため、あるスレッドで動作しているカウンタ（例えば n+1;）が、別のスレッドで動作しているカウンタとは異なる結果を生成することがありますが、これは各スレッドがデータの異なるサブセットを処理しているためです。そのため、複数のスレッドから得られた個々の結果をまとめない限り、スレッドごとに結果が異なる可能性があります。しかし、これはそれほど複雑なことではありません。それは、SAS Viya が処理結果のほとんどの照合と再組み立てを自動的に行うからです。

非決定論的アルゴリズムとは、決定論的アルゴリズムとは異なり、同じ入力であっても、実行するたびに異なる挙動を示す可能性があるアルゴリズムです。並列アルゴリズムでは、競合状態のため実行ごとに異なる動作をすることがあります。確率的アルゴリズムの動作は、乱数ジェネレータに依存します。非決定論的アルゴリズムは、決定論的アルゴリズムを使って正確な解を得るにはコストがかかりすぎる場合に、解の近似値を求めるためによく使われます（Wikipedia）。SAS Visual Data Mining and Machine Learning のモデルの中には、非決定論的なプロセスで作成されているものがあります。つまり、モデルを実行し、そのモデルを保存し、モデルを閉じて、後でレポートを再度開いたり、レポートを印刷したりすると、表示される結果が異なる場合があります。

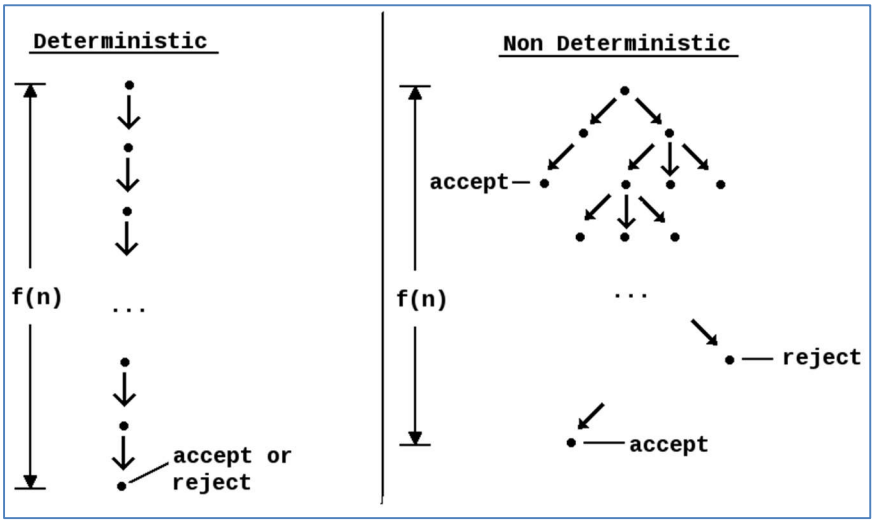


Image source:By Eleschinski2000 - With a paint program., CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=43528132

f(n) ステップを実行する決定論的アルゴリズムは、常に f(n) ステップで終了し、常に同じ結果を返します。f(n) ステップを行う非決定論的アルゴリズムは、異なる実行で同じ結果を返さないかもしれません。非決定論的アルゴリズムは、固定の高さのツリーサイズが無限になる可能性があるため終了しない場合があるかもしれません。

考え方が全く違うのです！

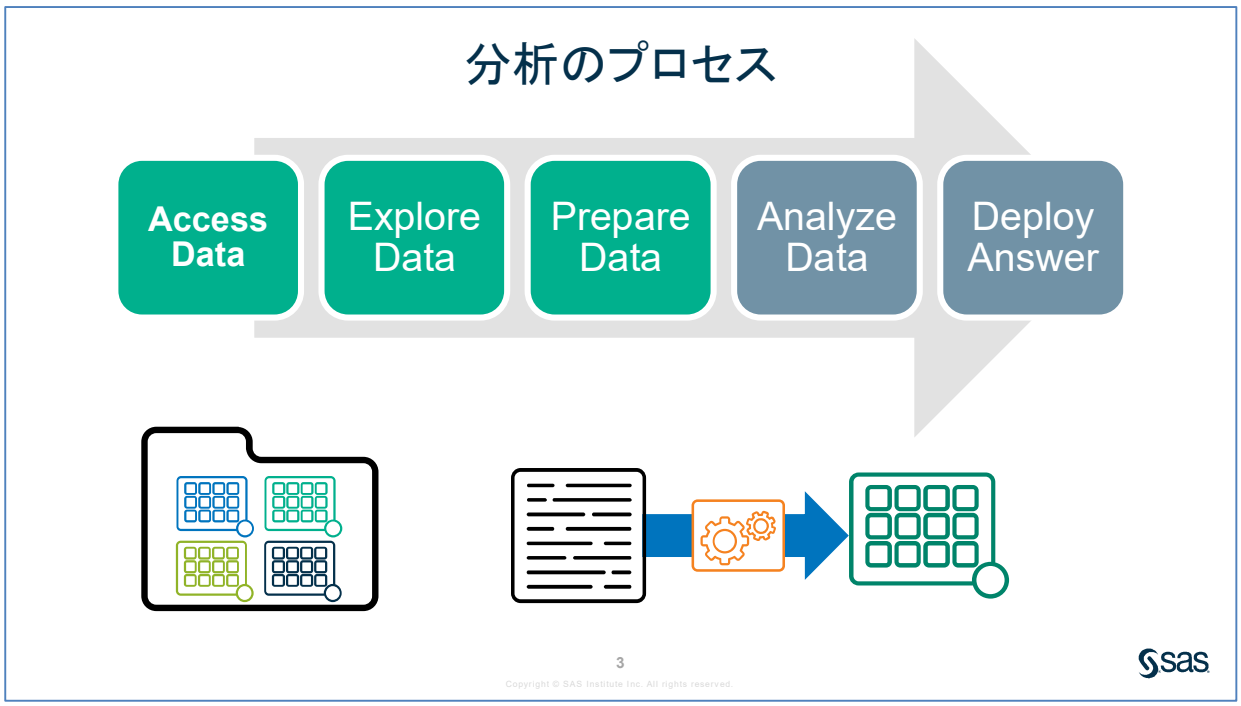
あなたはモデルを「収束」「推定」しているのであって、モデルのパラメータを正確に計算しているわけではありません。アルゴリズムによっては、プロセスにランダム性が含まれているため、結果の再現性は期待できません。

しかし、結果は収束していきます。ビッグデータのために設計された優れたコンピューティング環境だからこそ、この再現性のなさが代償となっているのです。

参考：「Data Science’s Reproducibility Crisis」https://towardsdatascience.com/data-sciences-reproducibility-crisis-b87792d88513

Lesson 2 分析データのハンドリング

2.1 本章の学習目標



Access Data：様々なファイル形式のデータを、分析ツールで利用できるようにします。

Explore Data：データの構造や関係性、保存されている値について把握します。

Prepare Data：分析に適したデータの形に加工します。また必要な項目の作成や選択を行います。

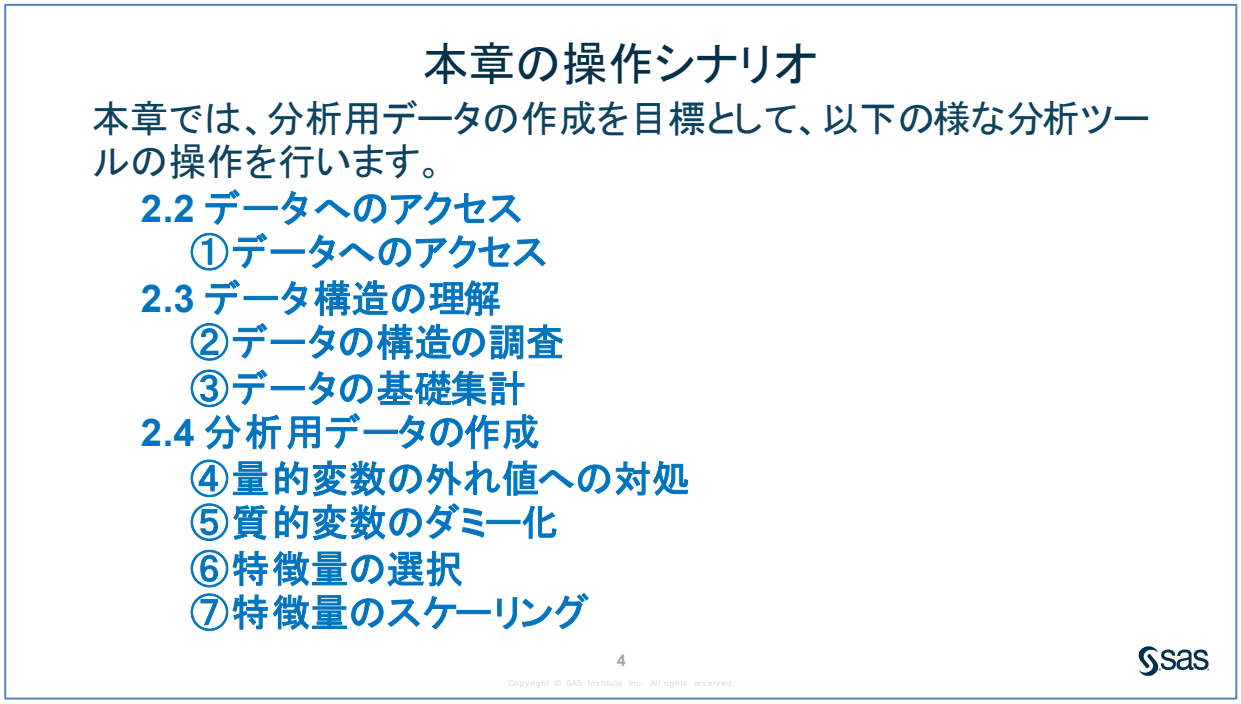
この章では、データへのアクセス、データの探索、データの準備について学習します。

良く知られているように、分析を行うプロセスの中で、分析用のデータを作成する過程は、全体のおよそ 80%を占めるとも言われます。手元にあるデータは、そのまま分析に利用できるとは限りません。むしろ、そのまま利用できるケースは非常に稀です。 “Garbage in, garbage out.” という言葉にもあるように、質の悪いデータから、良い分析結果は生まれません。異なるファイル形式のデータを統合して内容を精査し、必要なデータを抽出・加工して、また必要な項目を作成・選択して、分析に耐えうるデータを作成します。

分析用データの作成プロセスでは、以下の様な作業が想定されます：

- ファイルの種別の確認
- 列の種別の確認
- ファイルのインポート
- データのプロファイリング
- データの一意性の確認
- キーの設定
- 列の選択
- 行の抽出
- 並べ替え
- 計算列の作成
- 列の再コード化
- 列の型変換
- データ構造の変換
- データの結合
- 変数型の分類
- 変数尺度の分類
- 分布の確認
- データの分割
- 欠損値の補完
- 次元の削減
- 変数の選択
- 変数の変換
- その他・・・

上記の例は、基礎となる分析用のデータを作成するための、データ加工トピックの一部です。
これらの作業を行って、基礎となるデータ構造が作成されますが、実際に分析を行う前には、
選択した分析手法に適したデータに変更するための二次的な加工が発生することも一般的です。



2.2 データへのアクセス

①データへのアクセス
分析ツールでデータを使用できるように、データの読み込みを行います。

2.3 データ構造の理解

②データの構造の調査
データの行数や列数、各列の属性や格納されている値について確認を行います。

③データの基礎集計
量的変数、質的変数の基礎集計を行い、データ値の特徴について把握します。

2.4 分析用データの作成

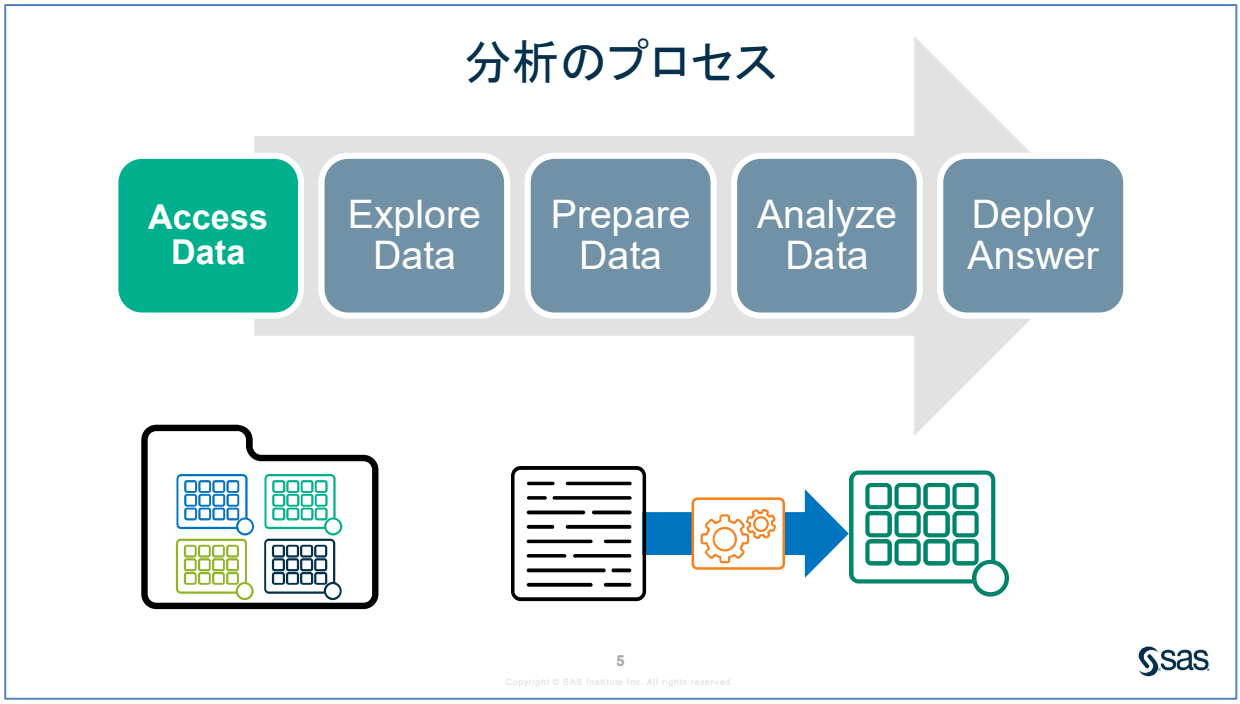
④量的変数の外れ値への対処
外れ値に対する、値の補完を行います。

⑤質的変数のダミー化
カテゴリ変数を、数値化します。

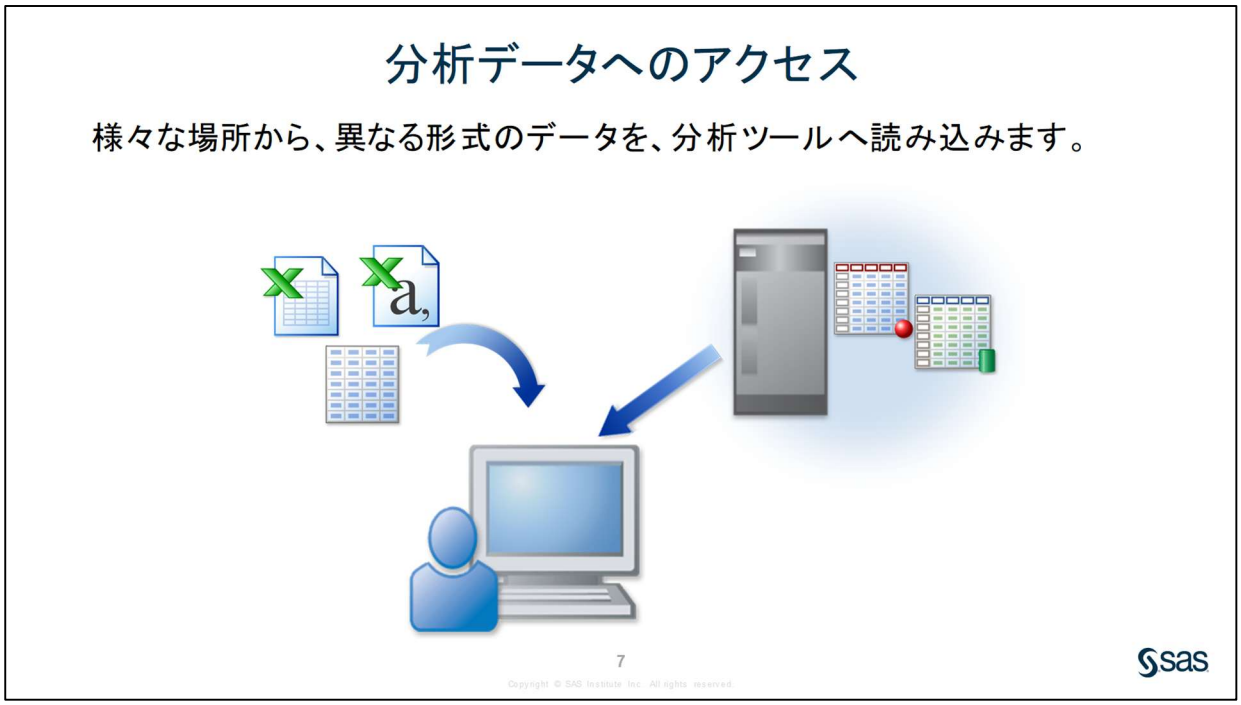
⑥特徴量の選択
多くの特徴量から、重要な特徴量に絞り込みます。

⑦特徴量のスケーリング
特徴量の範囲を、一定の範囲に変換します。

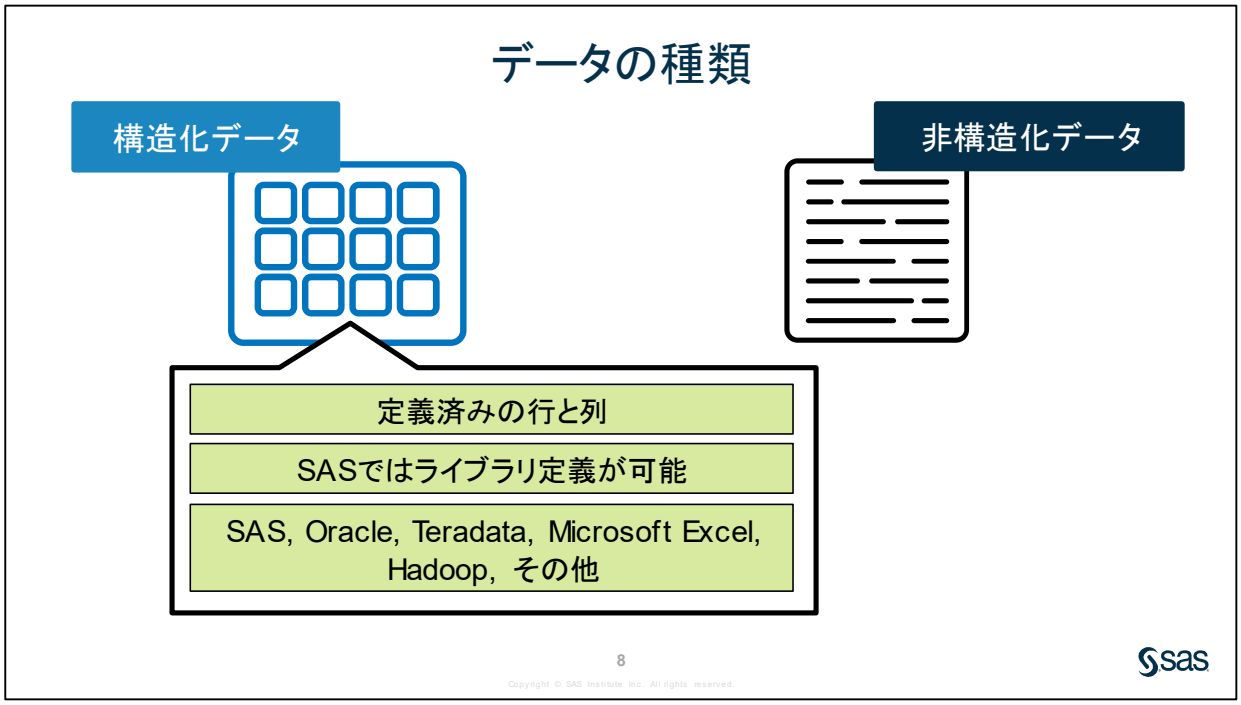
2.2 データへのアクセス

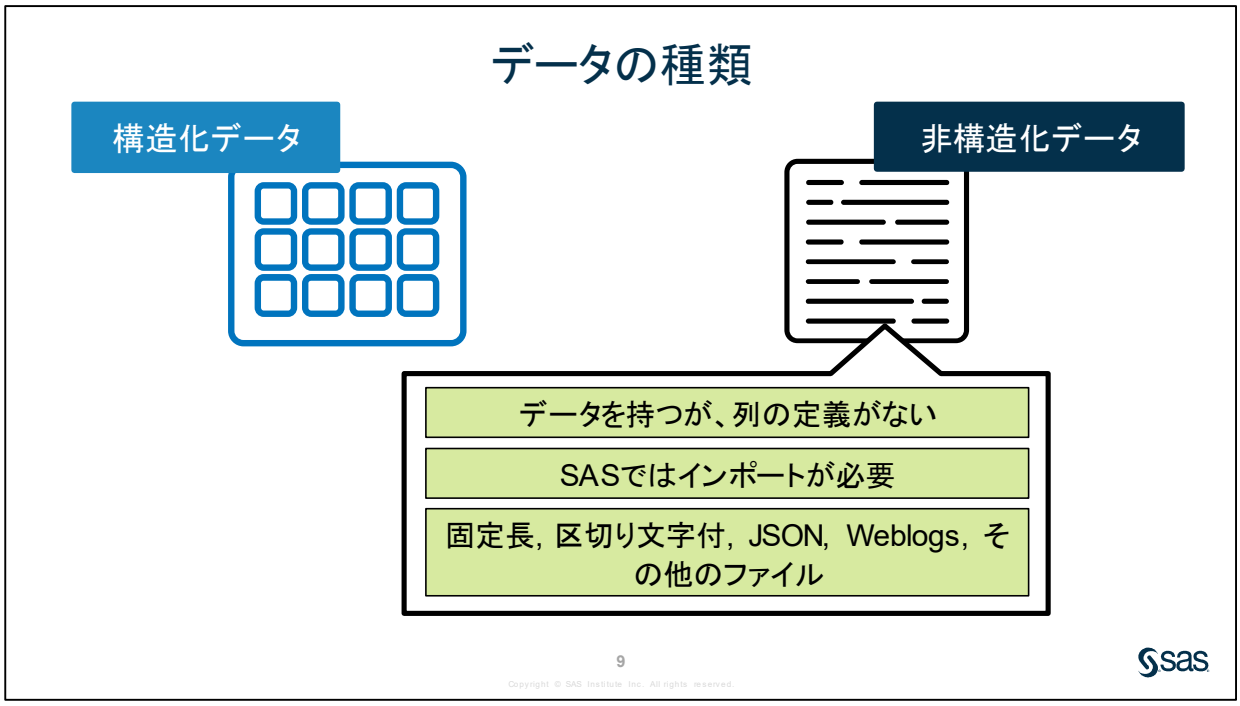


Access Data：様々なファイル形式のデータを、分析ツールで利用できるようにします。



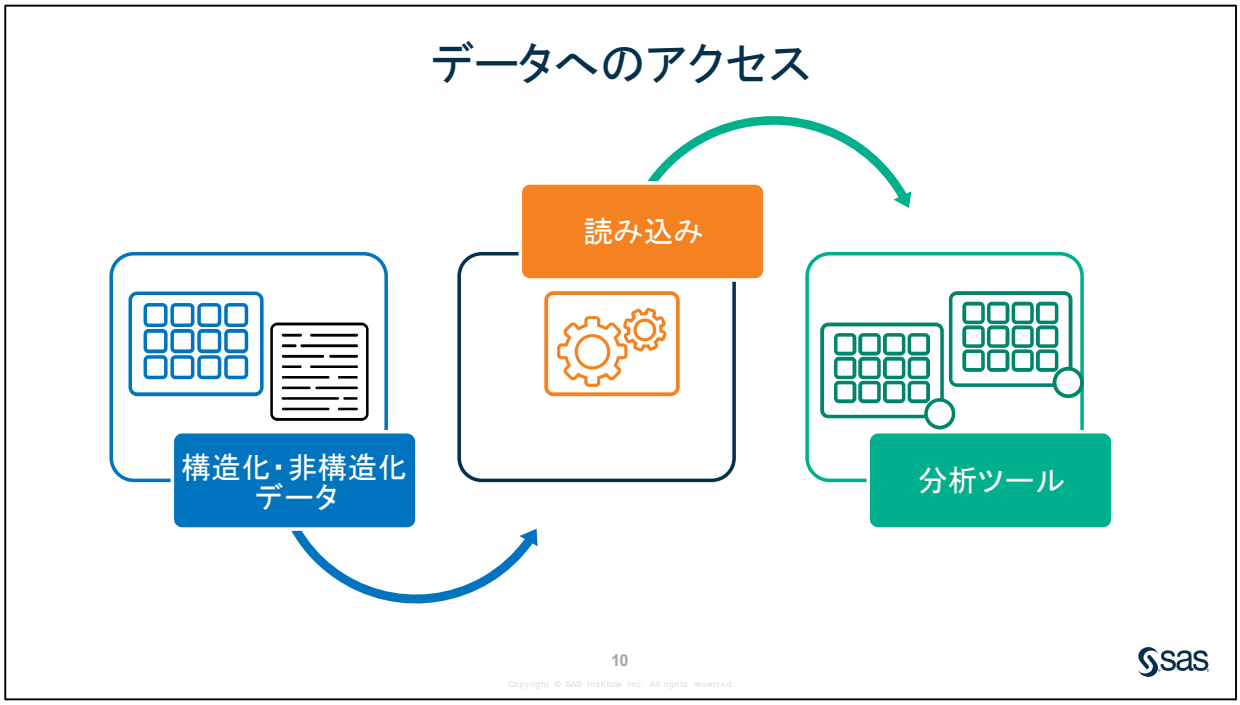
企業内のデータは、組織をまたいで、様々なファイル形式で保存されているかもしれません。それらのデータの中から、必要なファイルを、ご自身が使用する分析環境（ツール）で利用できるように、その場所へのアクセスの確立を行い、ファイルのインポートを行います。



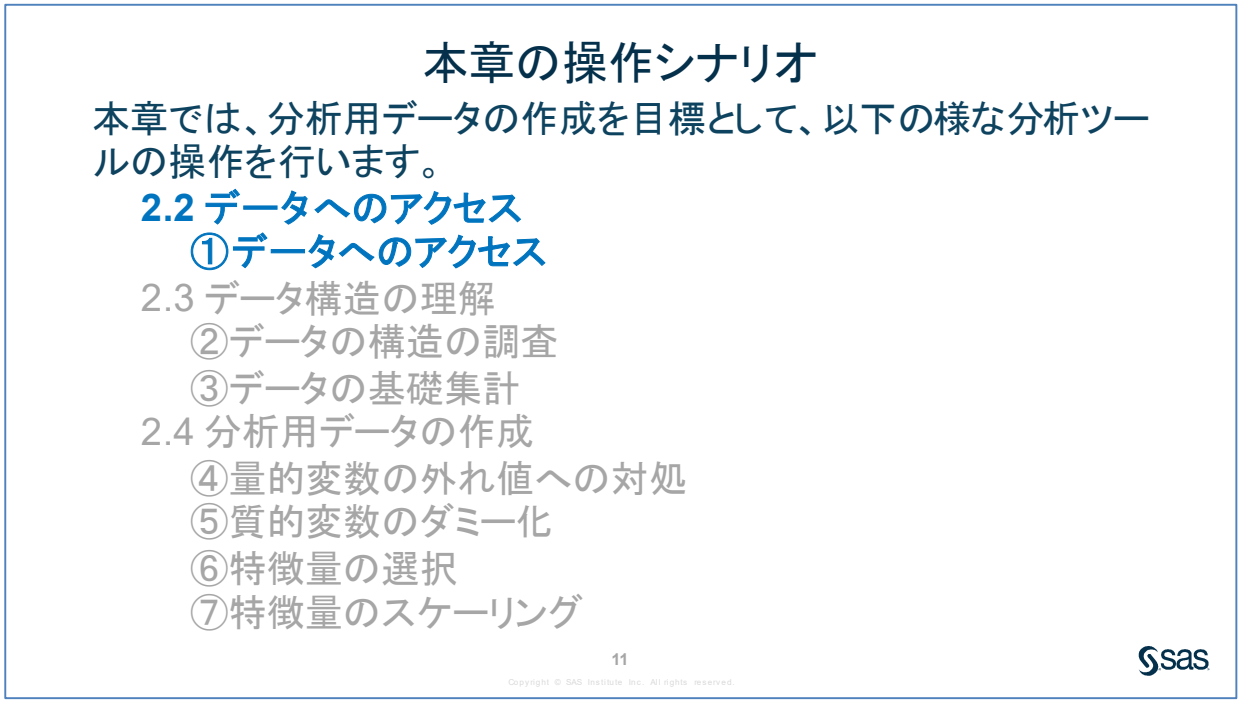


構造化データ、非構造化データに対する、ファイル形式の区別には様々な解釈があります。一般的には、リレーショナルデータベースに格納できるテーブル、行や列の概念を持つ CSV や Excel ファイルを構造化データ、文書や音声、画像、テキストファイルを非構造化データ、と分けることが多いようです。

SAS では、データの保存場所を直接的に参照する、ライブラリの割り当てが可能なファイルを構造化データ、インポートして SAS データの形式に変換が必要なファイルを非構造化データと呼んでいます。



分析のプロセスを進めるスターティングポイントとして、データを分析ツールで利用できるようにするために、ファイルへのアクセスを行います。



実際に分析ツールを使用して、ファイルを読み込んでいきましょう。

データへのアクセス

このデモでは、分析に使用する区切り文字付きファイルを分析ツールにインポートします。

ここでは、SAS Viya で分析が出来るように、SAS Studio を使用して操作をしていきます。

・区切り文字付きファイル：bank.csv

SAS Viya で区切り文字付き（CSV）ファイルや Excel のファイルを使用するためには、データを CAS サーバーにインポートして、インメモリテーブルとしてアップロードしなければなりません。また、CAS サーバーにインポートするには、Viya 環境内のフォルダへ該当のファイルをアップロードする必要があります。

1. Windows エクスプローラーを使用してファイルの場所へ移動して、ファイルを開いて内容を確認します。ファイルの場所は、講師の指示に従ってください。

   bank.csv は、セミコロンで区切られたデータです。ファイルの 1 行目には列ラベルが入っており、実際のデータは 2 行目から始まっています。

```
age;job;marital;education;default;balance;housing;loan;contact;day;month;campaign;
42;技術職;離婚;高校卒業;いいえ;305;はい;いいえ;携帯電話;28;aug;4;-1;0;不明;いいえ
52;ブルーカラー;離婚;中学卒業;いいえ;2800;いいえ;いいえ;不明;19;jun;1;-1;0;不明;いいえ
55;失業者;未婚;中学卒業;いいえ;274;いいえ;いいえ;携帯電話;9;feb;5;-1;0;不明;いいえ
41;ブルーカラー;既婚;中学卒業;いいえ;1461;はい;いいえ;不明;6;jun;2;-1;0;不明;はい
28;管理職;未婚;大学卒業;いいえ;187;いいえ;いいえ;携帯電話;9;mar;1;-1;0;不明;はい
60;事務職;既婚;高校卒業;いいえ;106;いいえ;いいえ;携帯電話;24;feb;1;187;4;成功;はい
44;事務職;離婚;高校卒業;いいえ;2999;はい;いいえ;携帯電話;14;may;1;-1;0;不明;はい
30;技術職;離婚;高校卒業;いいえ;3100;はい;いいえ;携帯電話;20;いいえv;1;183;4;失敗;いいえ
40;管理職;既婚;高校卒業;いいえ;643;はい;いいえ;携帯電話;17;apr;2;256;1;失敗;いいえ
41;技術職;既婚;高校卒業;いいえ;2152;いいえ;いいえ;携帯電話;30;sep;1;121;1;その他;いいえ
45;サービス業;既婚;高校卒業;いいえ;0;いいえ;いいえ;不明;4;jun;6;-1;0;不明;いいえ
```

   また、dic_col_bank.csv は、bank.csv の列構造を含むファイルです。

```
Name,SASColumnType,BegionPosition,EndPosition,ReadFlag,Label,Length,SASFormat,SASInformat
age,n,,,y,顧客の年齢,8,BEST12.,BEST32.
job,c,,,y,顧客の職業,18,$18.,$18.
marital,c,,,y,顧客の婚姻状況,6,$6.,$6.
education,c,,,y,顧客の最終学歴,12,$12.,$12.
default,c,,,y,債務不履行の有無,9,$9.,$9.
balance,n,,,y,残高,8,BEST12.,BEST32.
housing,c,,,y,住宅ローンの有無,9,$9.,$9.
loan,c,,,y,個人ローンの有無,9,$9.,$9.
contact,c,,,y,最後の連絡手段,12,$12.,$12.
day,n,,,y,顧客へ最後に連絡した日付,8,BEST12.,BEST32.
month,c,,,y,顧客へ最後に連絡した月,10,$10.,$10.
campaign,n,,,y,連絡回数(現在),8,BEST12.,BEST32.
pdays,n,,,y,経過日数,8,BEST12.,BEST32.
previous,n,,,y,連絡回数(以前),8,BEST12.,BEST32.
poutcome,c,,,y,結果(以前),9,$9.,$9.
deposit,c,,,y,定期預金購入の有無,9,$9.,$9.
```

2. このファイルを SAS Studio のフローを使用してインポートします。デスクトップから Chrome を起動して、ブックマークバーから、**SAS Studio** を選択します。

3. 入力されているユーザー名が **Student** であることを確認し、**サインイン**をクリックします。

4. 「すべての想定されるグループに参加しますか？ ― SAS Administrators」のメッセージには、**はい**をクリックします。

5. 以下のように SAS Studo が開きます。

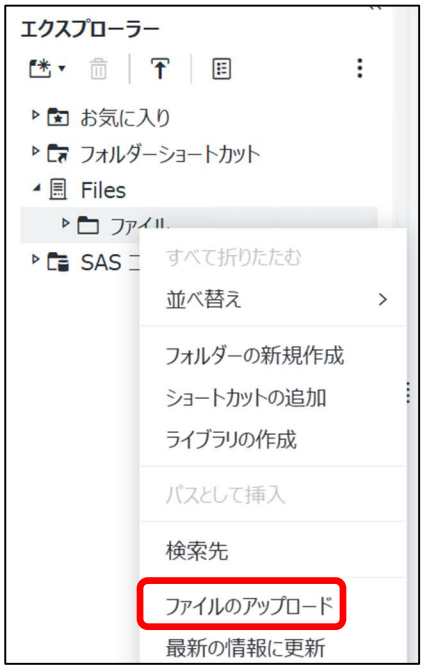
ナビゲーションペインには、以下のセクションが含まれます。左端のアイコンをクリックして各セクションを開くことができます。

セクション	説明
開いているファイル	作業領域に開いているタブのリストにアクセスできる

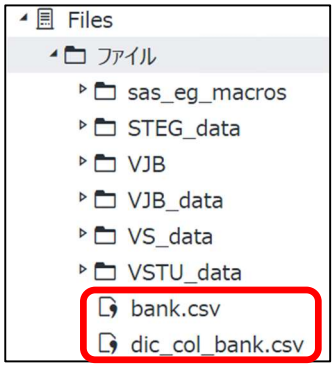
エクスプローラー	SAS Viya 環境内のフォルダやファイルにアクセスできる
ステップ	フローで使用できるステップの一覧にアクセスできる
スニペット	事前定義済みおよび保存したカスタムスニペット(コードテンプレート)の一覧にアクセスできる
ライブラリ	SAS ライブラリおよび caslib の一覧にアクセスできる
Git リポジトリ	SAS Studio 内から Git 機能にアクセスできる

作業領域は、データ、コード、ログ、結果、フローを表示するために使用されます。これらのアイテムを開くと、タブ形式のインターフェイスのウィンドウとして作業領域に追加されます。

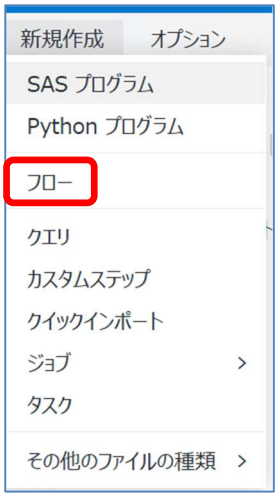
6. エクスプローラーセクションで、**Files → ファイル**を右クリックし、先ほどの 2 つのファイルを Viya 環境内のフォルダにアップロードします。ファイルのアップロードウィンドウでは、+アイコンをクリックし、2 つのファイルを選択し、**アップロード**をクリックします。

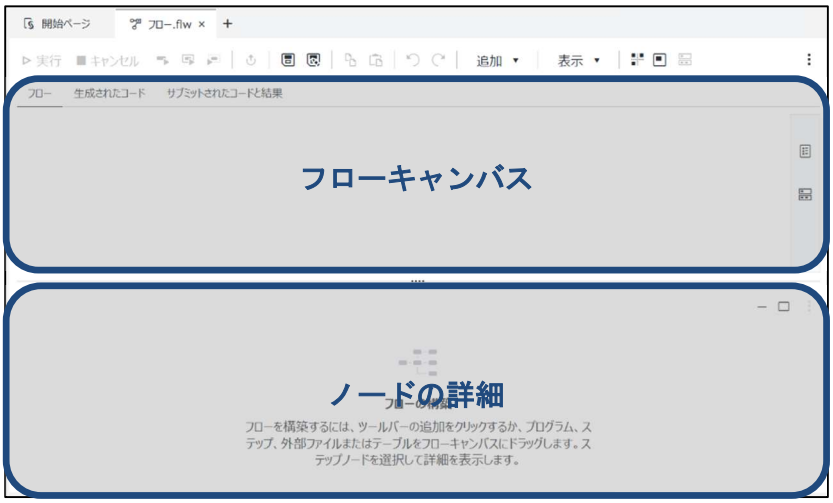


Viya 環境内にアップロードすることができました。



7. 左上にある、**新規作成→フロー**を選択し、SAS Studio フローを作成します。フロー内で csv ファイルの CAS へのインポート含め、様々なステップを使用して作業します。





フローキャンバスにはノードを使用したデータ処理の流れを作成し、ノードの詳細ではフローキャンバスで選択したノードの属性を管理します。

8. エクスプローラーセクションの先ほど Viya 環境内にアップロードした bank.csv ファイルをフローキャンバスにドラッグ&ドロップします。

bank ファイルノードを選択し、ノードの詳細のオプションタブを確認し、ファイルの内容に合わせて以下の設定に変更します：

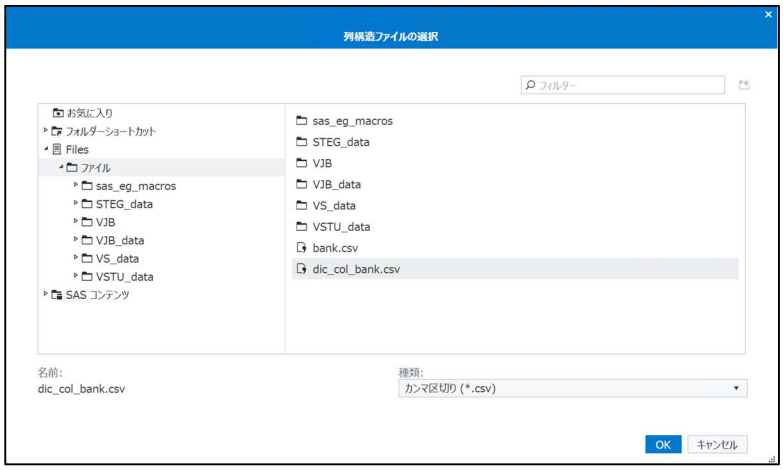
ファイルの種類：区切りファイル(.dlm)

区切り記号：セミコロン

9. フローキャンバスの bank ファイルノードを右クリックし、**インポートの追加**をクリックします。

10. フローキャンバスのインポートノードをクリックし、ノードの詳細にて、**構造→構造のロード**を選択します。

先ほどアップロードした dic_col_bank.csv を選択、**OK** をクリックします。

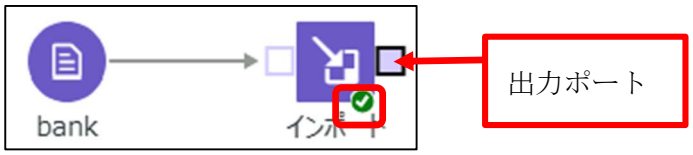


インポートノードの列構造に列定義が読み込まれ、出力データプレビューにインポートするデータを確認できました。


インポート
オプション
ノード
メモ
bank.csv
/workshop/
分析
オプション
列構造
新規列
すべて削除
列の移動
構造
(フィルターなし)
フィルター
名前
ラベル
種類
長さ
出力形式
1 age 顧客の年齢 数値 8 BEST12.
2 job 顧客の職業 文字 18 $18.
3 marital 顧客の婚姻状況 文字 6 $6.
出力データプレビュー
bank
更新
式の入力
age job marital education default balan… housing
1 42 技術職 離婚 高校卒業 いいえ 305 はい
2 52 ブルーカラー 離婚 中学卒業 いいえ 2800 いいえ
3 55 失業者 未婚 中学卒業 いいえ 274 いいえ
4 41 ブルーカラー 既婚 中学卒業 いいえ 1461 はい

11. フローキャンバスでインポートノードを右クリックし、**ノードの実行**を選択し、インポートを実行します。

インポートが正常に実行されると、インポートノードに緑のチェックが付きます。また、インポートノードの右端の四角(出力ポート) を選択します。

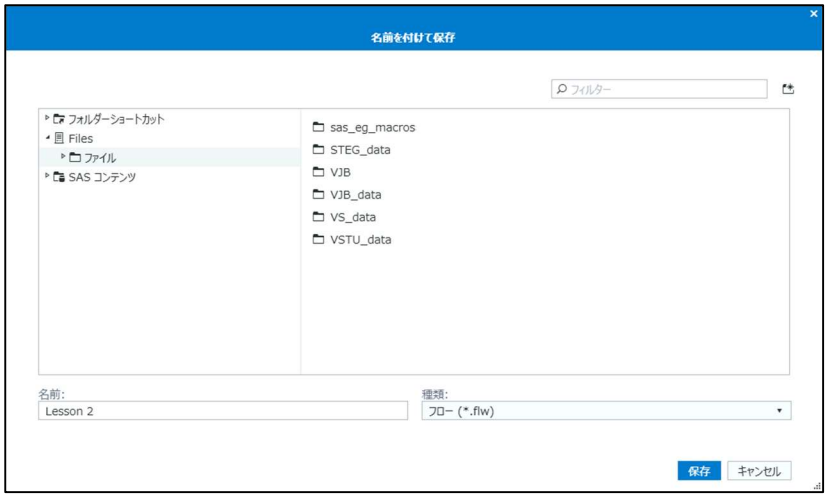


ノードの詳細の、データのプレビュータブで、インポートされたデータを確認します。

	age	job	marital	education	default	balan…	housing	loan	contact
1	42	技術職	離婚	高校卒業	いいえ	305	はい	いいえ	携帯電話
2	52	ブルーカラー	離婚	中学卒業	いいえ	2800	いいえ	いいえ	不明
3	55	失業者	未婚	中学卒業	いいえ	274	いいえ	いいえ	携帯電話
4	41	ブルーカラー	既婚	中学卒業	いいえ	1461	はい	いいえ	不明
5	28	管理職	未婚	大学卒業	いいえ	187	いいえ	いいえ	携帯電話
6	60	事務職	既婚	高校卒業	いいえ	106	いいえ	いいえ	携帯電話

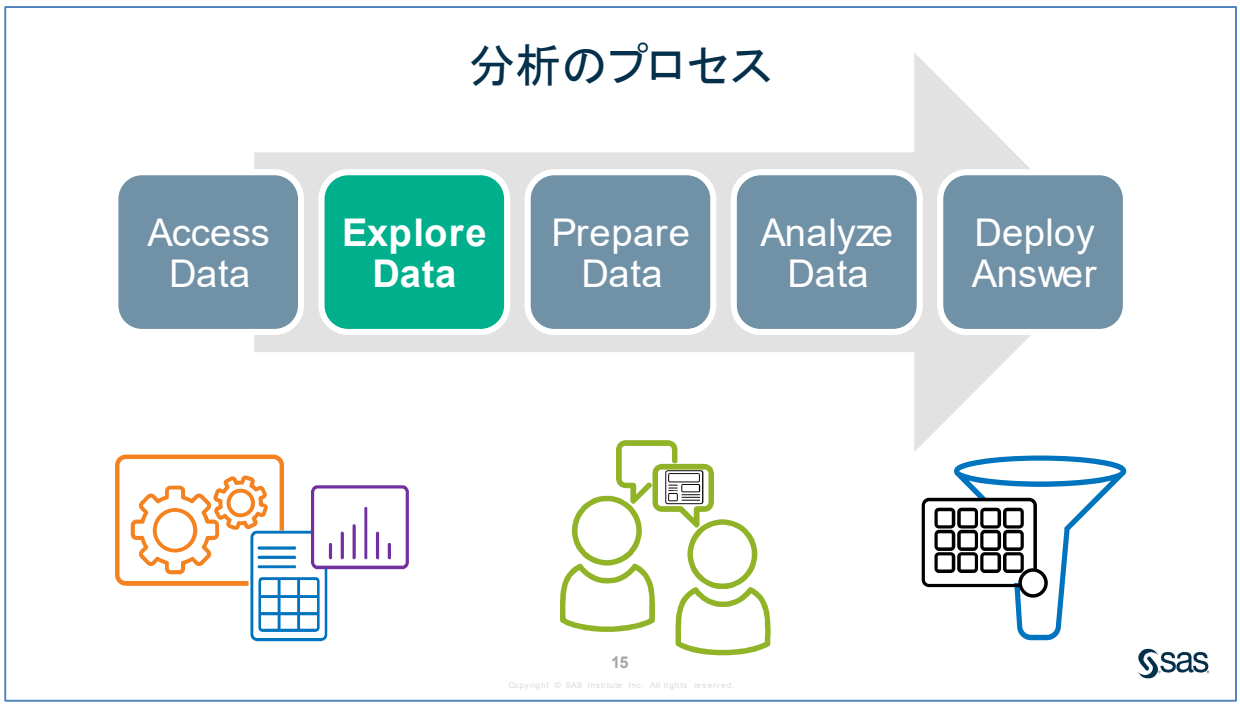
12. **保存**ボタンをクリックして、フローを保存します。

13. **Files → ファイル**に、**Lesson 2** という名前で保存します。

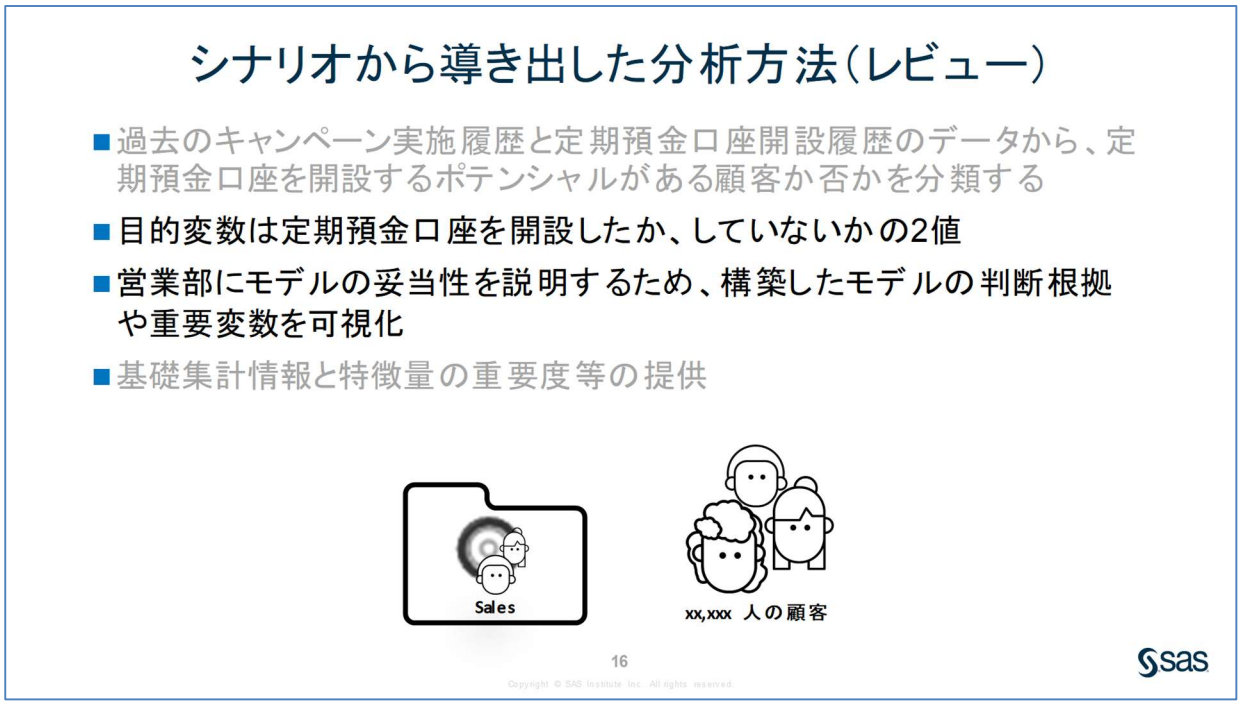


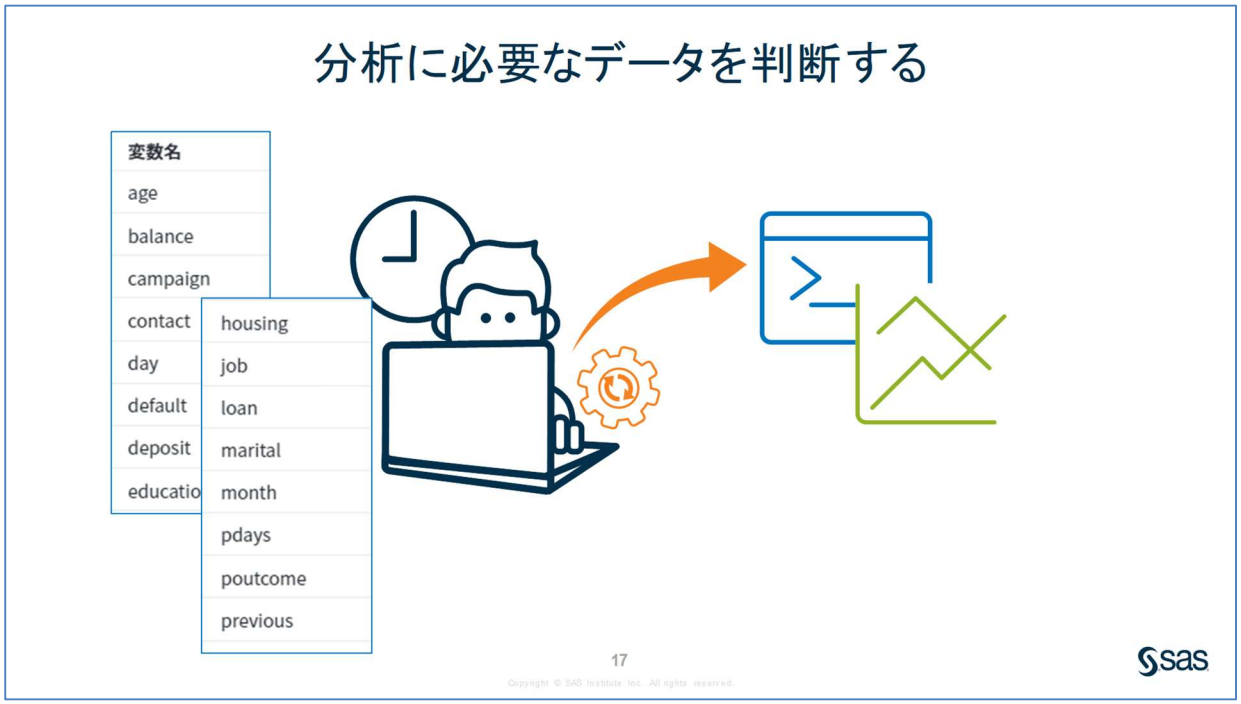
End of Demonstration

2.3 データ構造の理解



Explore Data：データの構造や関係性、保存されている値について把握します。





データから、予測の対象（目的）となる変数、予測の根拠（説明）となる変数を選択します。
今回利用するデータでは、分析に利用する変数があらかじめ絞り込まれていますが、通常は多くの変数から、より予測結果に影響を与えそうな変数を選択することが必要になります。言い換えると、予測結果に強く影響を及ぼす、データの特徴を量的にあらわす変数、すなわち特徴量の選択が重要です。
例えば、トマトの収穫高を予測（目的変数と）したい場合に、予測の根拠となる特徴量（説明変数）として、気温、降水量、市場価格があったとします。気温と降水量は収穫高に影響を与えそうですが、市場価格はあまり影響を与えないかもしれません。そのような場合には、市場価格は分析から除外する必要があるでしょう。
実際の場面では、予測に影響を与える特徴量を判断するためには、ドメイン知識が必要・重要になります。自身が分析者であり業務知識に乏しい場合には、実務者と十分ディスカッションをして特徴量を決定していく必要があります。

データ構造や値の把握

- データ構造
  レコード数、列の数、
  列の属性:
  - 名前、タイプ、長さ、
  - キー、インデックス、Null、並べ替え、etc.
- データ値
  列の値の種類、欠損値の存在
  列の値の水準数
- テーブル間の関係
  - 結合キー、カーディナリティ(一意性)

18

分析用のデータを作成する上でも、分析に使用する変数を選択する上でも、データの構造や特性を理解しておくことは重要です。データに対する仕様書がある場合には、値の確認等にそれを参考にすることができます。

bank.csv の変数：
顧客データ：:
1 – **age**：年齢［数値］

2 – **job**：職業［カテゴリ］
　値の種類："事務職.","管理職","家政婦","起業家","学生","ブルーカラー",
　"自営業","定年退職者","技術職","サービス業","失業者",,"不明"

3 – **marital**：婚姻状況［カテゴリ］
　値の種類："既婚","離婚","未婚"; 注意: "離婚" は離婚または未亡人を含む

4 – **education**：最終学歴［カテゴリ］
　値の種類："不明","中学卒業","高校卒業","大学卒業"

5 – **default**：債務不履行(デフォルト)の有無［二値］
　値の種類："はい","いいえ"

6 – **balance**：年間残高平均（ユーロ）［数値］

7 – **housing**：住宅ローンの有無［二値］
　値の種類："はい","いいえ"

8 – **loan**：個人ローンの有無［二値］
　値の種類："はい","いいえ"

＃ 現在のキャンペーンの直近のコンタクト：

9 – **contact**：コンタクトの方法［カテゴリ］
　値の種類： "固定電話","携帯電話","不明"

10 – **day**：直近のコンタクト日［数値］

11 – **month**：直近のコンタクト月［カテゴリ］
　値の種類：（"jan", "feb", "mar", ..., "nov", "dec"）

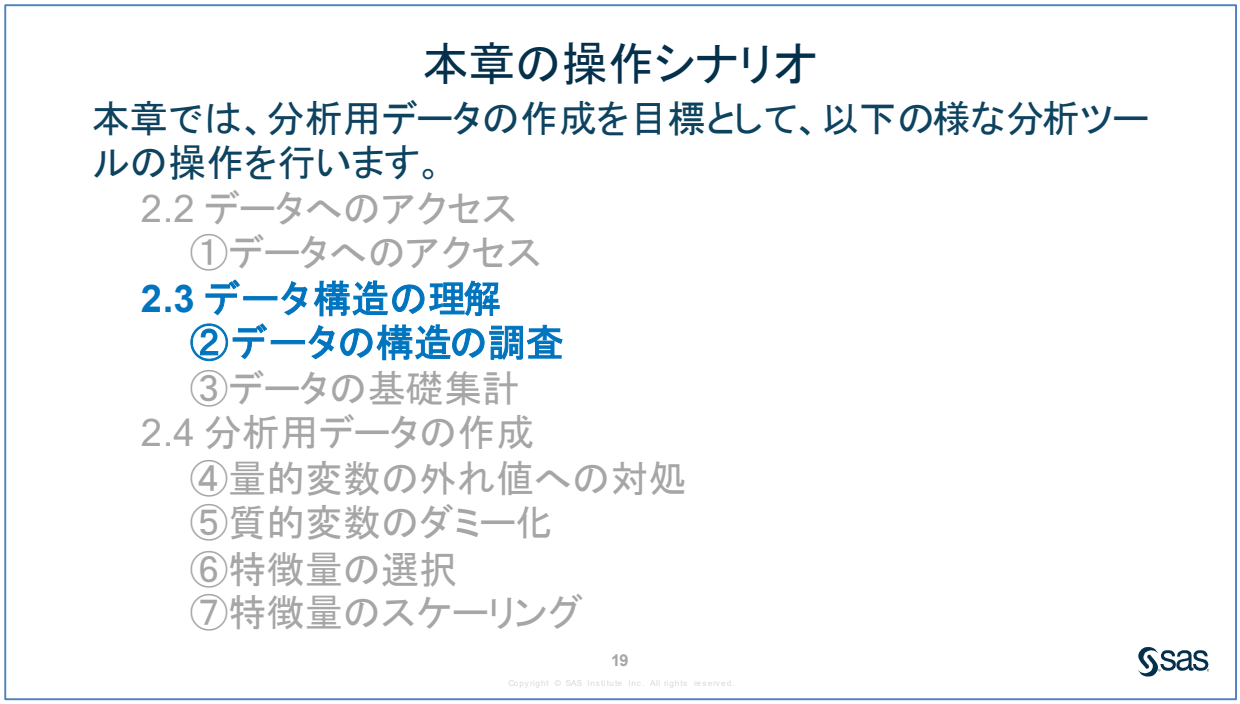
＃ その他の属性：
12 – **campaign**：このキャンペーン中に顧客に対して行われたコンタクトの数［数値］

13 – **pdays**：前回のキャンペーンで最後のコンタクトから経過した日数［数値］
　値の種類：-1 はクライアントへの以前のコンタクトが無かったことを意味します。

14 – **previous**：このキャンペーンの前に顧客に対して行われたコンタクトの数［数値］

15 – **poutcome**：：前回のマーケティングキャンペーンの結果［カテゴリ］
　値の種類："成功","失敗","その他","不明"

＃ 出力変数（目的のターゲット）
16 – **deposit**：クライアントは定期預金を申し込みしたか？［二値］
　値の種類： "はい","いいえ"

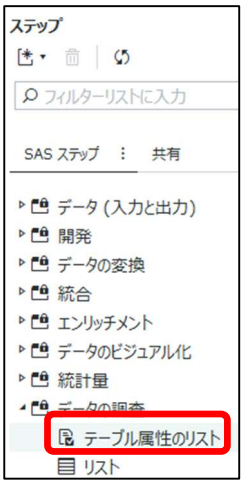


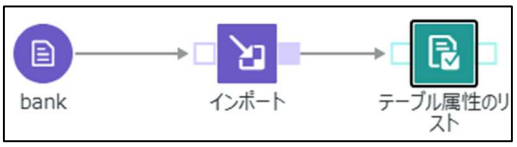
実際に分析ツールを使用して、データ構造を確認しましょう。

データ構造の調査

このデモでは、分析に必要となるデータの構造を理解します。各データのレコード数、列の数や属性、各列に保存されている値について調査を行います。

1. SAS Studio で作成した **Lesson 2** フローを引き続き使用して、インポートした Bank データの情報を確認します。**ステップ**セクションから**データの調査 → テーブル属性のリスト**を選択し、フローキャンバス内のインポートノードの出力ポートにドラッグします。





2. テーブル属性のリストノードの、ノードの詳細の、オプションタブで**変数順序**を**データセットの位置**に変更します。

3. フローキャンバスの**テーブル属性のリスト**ノードを右クリックし、**ノードの実行**を選択します。

4. サブミットされたコードと結果タブの、結果タブから生成されたレポートを確認します。

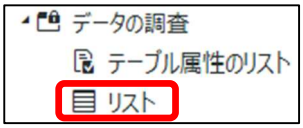
データセット名	WORK_FLW0011750122267518574000O	オブザベーション数	6888
メンバータイプ	DATA	変数の数	16
エンジン	V9	インデックス数	0
作成日時	2025/06/17 10:05:17	オブザベーションのバッファー長	152
更新日時	2025/06/17 10:05:17	削除済みオブザベーション数	0
保護		圧縮済み	NO
データセットタイプ		ソート済み	NO
ラベル			
データ表現	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64, LINUX_POWER_64		
エンコード	utf-8 Unicode (UTF-8)		

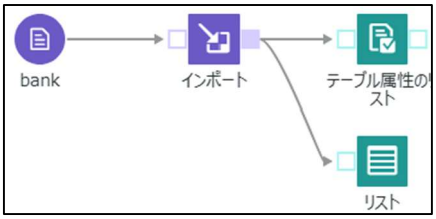
作成順の変数

#	変数	タイプ	長さ	出力形式	入力形式	ラベル
1	age	数値	8	BEST12.	BEST32.	顧客の年齢
2	job	文字	18	$18.	$18.	顧客の職業
3	marital	文字	6	$6.	$6.	顧客の婚姻状況
4	education	文字	12	$12.	$12.	顧客の最終学歴
5	default	文字	9	$9.	$9.	債務不履行の有無
6	balance	数値	8	BEST12.	BEST32.	年間残高平均
7	housing	文字	9	$9.	$9.	住宅ローンの有無
8	loan	文字	9	$9.	$9.	個人ローンの有無
9	contact	文字	12	$12.	$12.	直近のコンタクト方法
10	day	数値	8	BEST12.	BEST32.	直近のコンタクト日付
11	month	文字	10	$10.	$10.	直近のコンタクト月
12	campaign	数値	8	BEST12.	BEST32.	コンタクト回数(今回)
13	pdays	数値	8	BEST12.	BEST32.	経過日数
14	previous	数値	8	BEST12.	BEST32.	コンタクト回数(前回)
15	poutcome	文字	9	$9.	$9.	結果(前回)
16	deposit	文字	9	$9.	$9.	定期預金申込の有無

オブザベーション数の欄から、このデータが 6888 レコード持っていることがわかります。また、変数の数から、16 の列を持っていることがわかります。また、列の名前やデータタイプ等の属性については、作成順の変数の表から確認できます。

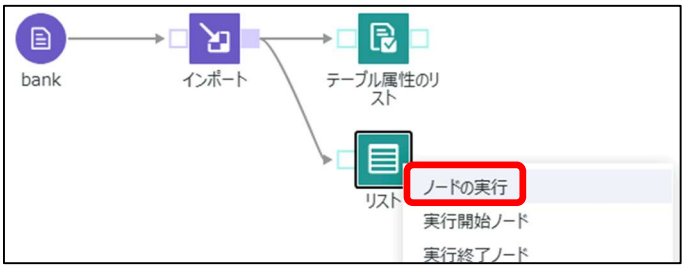
5. フロータブに戻り、**ステップ**セクションから**データの調査 → リストステップ**を選択し、フローキャンバス内のインポートノードの出力ポートにドラッグします。





6. フローキャンバスでリストノードを選択し、ノードの詳細の、データタブで**リスト変数**の、**列の追加**ボタンからすべての変数を追加します。

7. フローキャンバスの**リスト**ノードを右クリックし、**ノードの実行**を選択します。



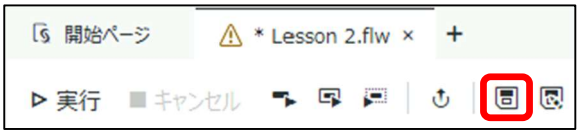
8. サブミットされたコードと結果タブの、結果タブから生成されたレポートを確認します。

WORK._flw001175066002399657800000 のリスト

OBS	顧客の年齢	顧客の職業	顧客の婚姻状況	顧客の最終学歴	債務不履行の有無	年間残高平均	住宅ローンの有無	個人ローンの有無	直近のコンタクト方法	直近のコンタクト日付	直近のコンタクト月	コンタクト回数(今回)	経過日数	コンタクト回数(前回)	結果(前回)	定期預金申込の有無
1	42	技術職	離婚	高校卒業	いいえ	305	はい	いいえ	携帯電話	28	aug	4	-1	0	不明	いいえ
2	52	ブルーカラー	離婚	中学卒業	いいえ	2800	いいえ	いいえ	不明	19	jun	1	-1	0	不明	いいえ
3	55	失業者	未婚	中学卒業	いいえ	274	いいえ	いいえ	携帯電話	9	feb	5	-1	0	不明	いいえ
4	41	ブルーカラー	既婚	中学卒業	いいえ	1461	はい	いいえ	不明	6	jun	2	-1	0	不明	はい
5	28	管理職	未婚	大学卒業	いいえ	187	いいえ	いいえ	携帯電話	9	mar	1	-1	0	不明	はい

各列の実際のデータ値を確認できます。

9. **保存**ボタンをクリックして、フローを上書き保存します。

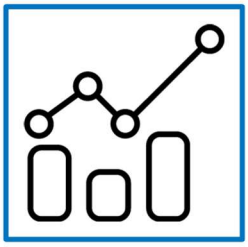


End of Demonstration

データの探索

データについて
知りましょう。

グラフィカルな結果
数値的な結果



- 外れ値
- 最小値
- 最大値
- 欠損の割合
- 平均値
- 範囲
- 標準偏差
- 分布
- …

21

sas

データの傾向を事前に把握するため、度数、平均、分散、標準偏差などの統計量、欠損値の数などを求め、またチャートやプロットを描いて傾向を調べます。

本章の操作シナリオ

本章では、分析用データの作成を目標として、以下の様な分析ツールの操作を行います。

2.2 データへのアクセス
　①データへのアクセス
2.3 データ構造の理解
　②データの構造の調査
　③データの基礎集計
2.4 分析用データの作成
　④量的変数の外れ値への対処
　⑤質的変数のダミー化
　⑥特徴量の選択
　⑦特徴量のスケーリング

22

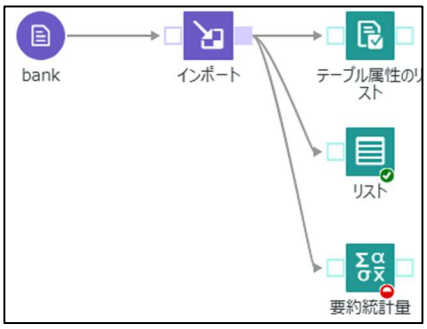
sas

実際に分析ツールを使用して、データの探索を行いましょう。

データの基礎集計

このデモでは、分析データの各変数の値を探索します。基礎統計量の算出や、グラフを使用してデータの傾向について把握します。

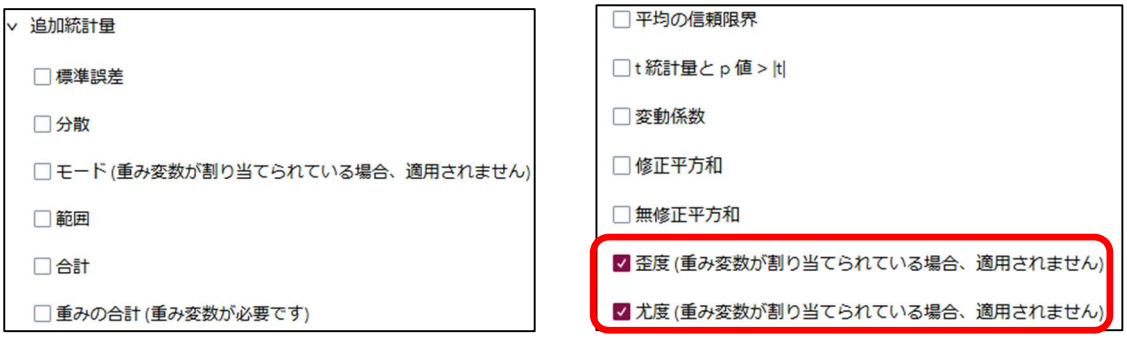
1. SAS Studio で作成した **Lesson 2** フローを引き続き使用して、Bank データの統計量を算出します。フロータブに戻り、**ステップ**セクションから**統計量** → **要約統計量**ステップを選択し、フローキャンバス内のインポートノードの出力ポートにドラッグします。



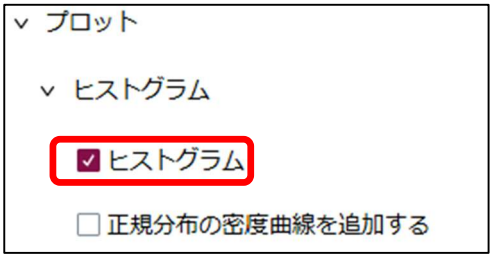
2. フローキャンバスで要約統計量ノードを選択し、ノードの詳細の、データタブで**分析変数**の、**列の追加**ボタンからすべての数値変数を追加します。

3. オプションタブで、**基本統計量**の、**欠損値数**にチェックを入れます。

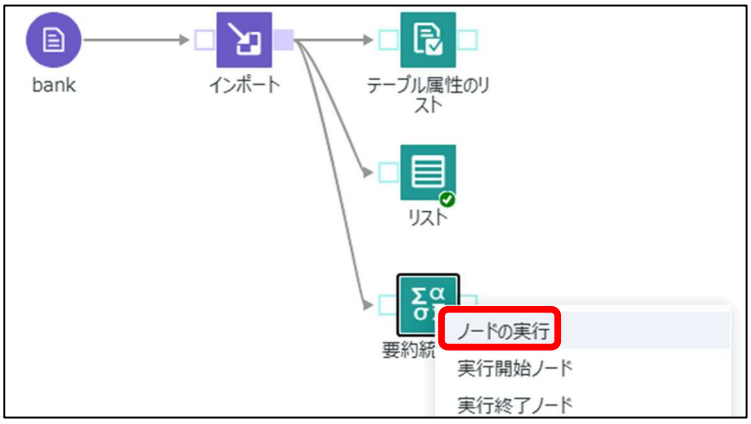
ほかに算出したい統計量があれば、**追加統計量**から選択します。今回は**歪度**と**尖度**にチェックを入れます。



4. プロットで、ヒストグラムにチェックを入れます。



5. フローキャンバスの**要約統計量**ノードを右クリックし、**ノードの実行**を選択します。

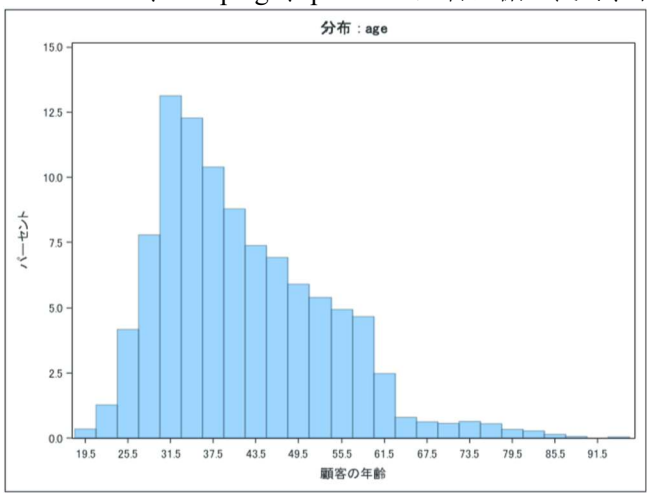


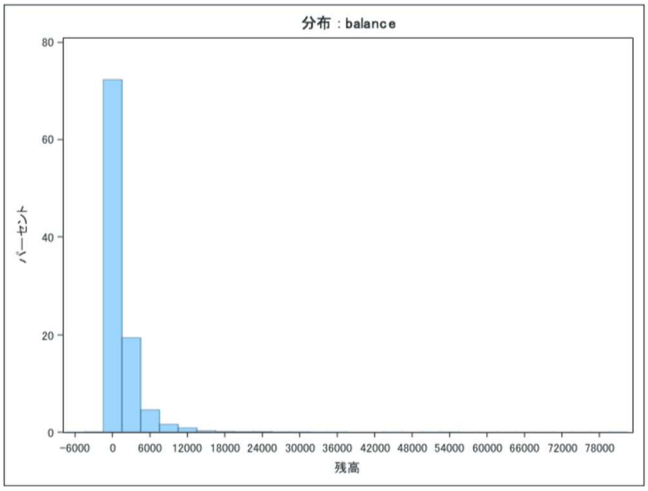
6. サブミットされたコードと結果タブの、結果タブから生成されたレポートを確認します。

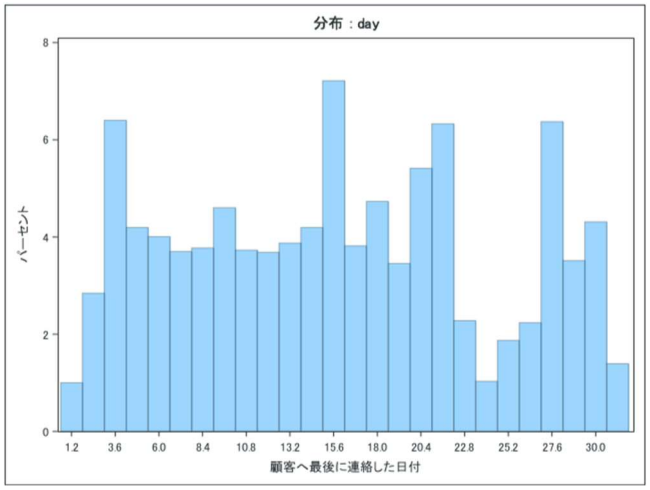
変数	ラベル	平均	標準偏差	最小値	最大値	N	欠損値の数	歪度	尖度
age	顧客の年齢	41.1325494	11.7053511	18.0000000	95.0000000	6888	0	0.8682182	0.6961385
balance	年間残高平均	1513.56	3181.15	-6847.00	81204.00	6888	0	7.6670427	109.9965353
day	直近のコンタクト日付	15.7128339	8.4386251	1.0000000	31.0000000	6888	0	0.1137017	-1.0708734
campaign	コンタクト回数(今回)	2.5617015	2.8323203	1.0000000	63.0000000	6888	0	5.8434754	65.1132748
pdays	経過日数	50.8620790	110.2942057	-1.0000000	854.0000000	6888	0	2.5058553	7.1636075
previous	コンタクト回数(前回)	0.8244774	2.4172314	0	58.0000000	6888	0	8.4124021	129.2042463

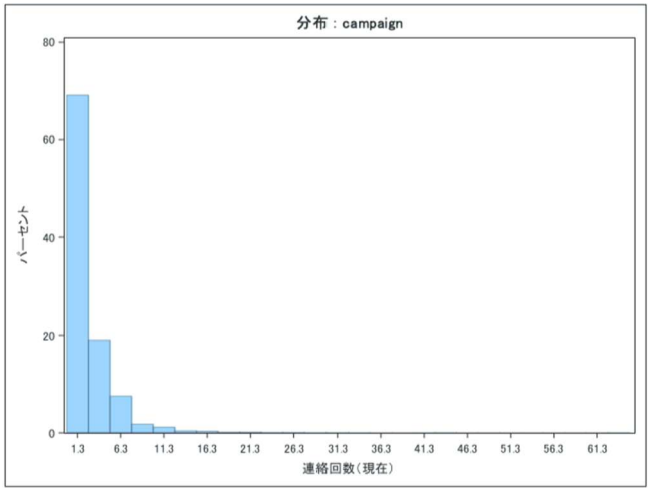
各数値変数の、記述統計量を確認できます。今回の数値データには欠損がないことがわかります。変数 previous の最大値は 58 であること、変数 balance、campaign、previous では歪度および尖度が高いことがわかります。

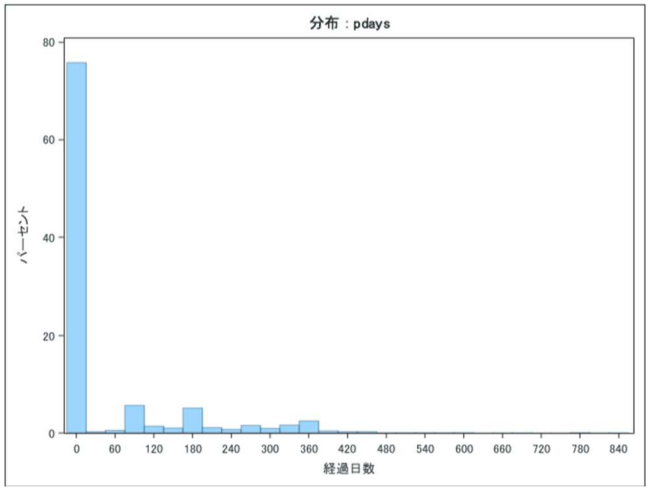
また、各ヒストグラムでぞれぞれの変数の値の傾向や分布状況を確認できます。
例えば、変数 age のボリュームゾーンは 30 台前後であることがわかります。また、変数 balance、campaign、previous は右に裾が長く、極端な値があるようです。

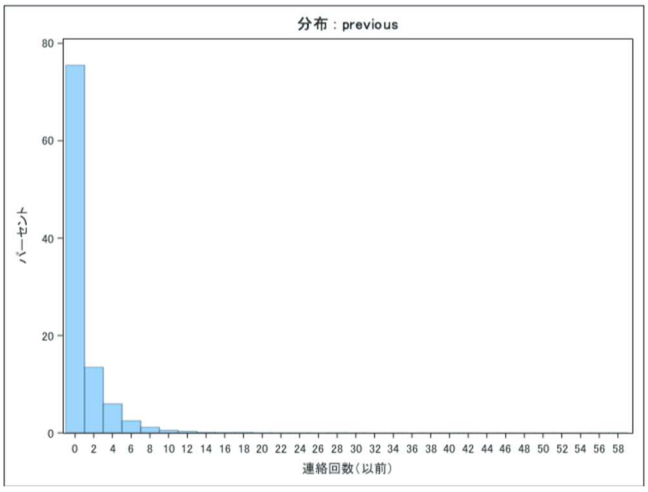




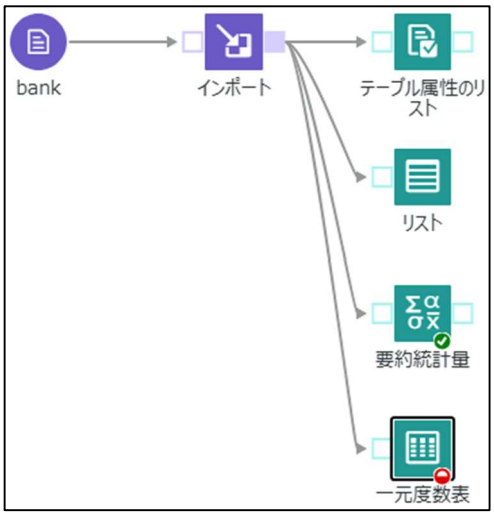




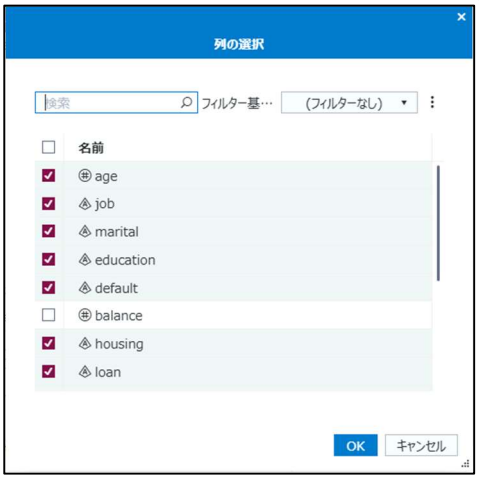




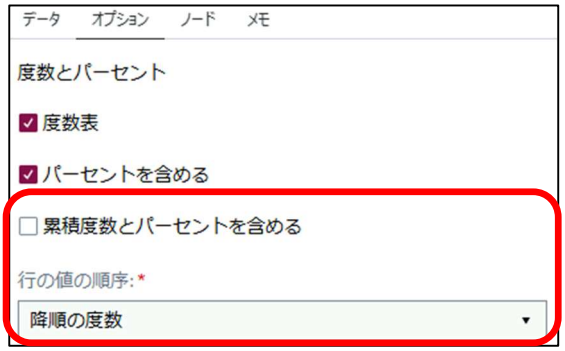
7. フロータブに戻り、**ステップ**セクションから**統計量** → **一元度数表**ステップを選択し、フローキャンバス内のインポートノードの出力ポートにドラッグします。



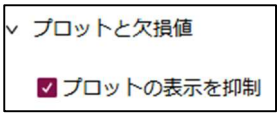
8. フローキャンバスで**一元度数表**ノードを選択し、ノードの詳細の、データタブで**分析変数**の、**列の追加**ボタンから、カテゴリ変数追加します。(例えば、変数 balance は、数値変数で、一意の値が多すぎるため除外します。)



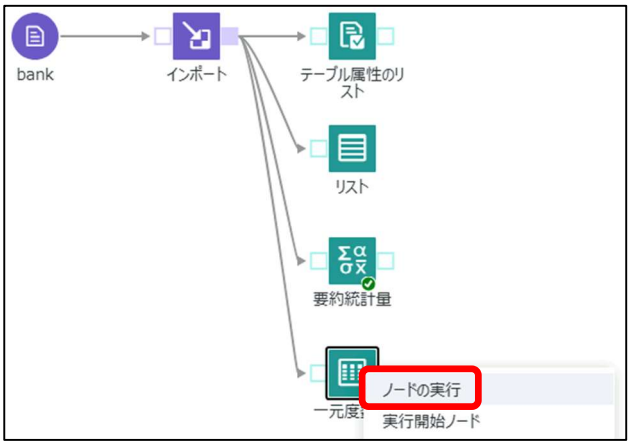
9. オプションタブで、**累積度数とパーセントを含める**チェックを外し、**行の値の順序**で、**降順の度数**を選択します。



10. **プロットと欠損値**の、**プロットの表示を抑制**を選択します。



11. フローキャンバスの**一元度数表**ノードを右クリックし、**ノードの実行**を選択します。



12. サブミットされたコードと結果タブの、結果タブから生成されたレポートを確認します。

顧客の年齢

age	度数	パーセント
31	315	4.57
32	302	4.38
34	291	4.22
30	288	4.18
35	283	4.11
33	272	3.95
36	267	3.88
38	230	3.34
29	219	3.18
37	219	3.18
39	215	3.12

顧客の職業

job	度数	パーセント
管理職	1551	22.52
ブルーカラー	1247	18.10
技術職	1139	16.54
事務職	814	11.82
サービス業	549	7.97
定年退職者	464	6.74
自営業	256	3.72
失業者	219	3.18
学生	219	3.18
起業家	212	3.08
家政婦	176	2.56
不明	42	0.61

顧客の婚姻状況

marital	度数	パーセント
既婚	3990	57.93
未婚	2121	30.79
離婚	777	11.28

顧客の最終学歴

education	度数	パーセント
高校卒業	3405	49.43
大学卒業	2237	32.48
中学卒業	957	13.89
不明	289	4.20

債務不履行の有無

default	度数	パーセント
いいえ	6776	98.37
はい	112	1.63

住宅ローンの有無

housing	度数	パーセント
いいえ	3531	51.26
はい	3357	48.74

個人ローンの有無

loan	度数	パーセント
いいえ	5983	86.86
はい	905	13.14

直近のコンタクト方法

contact	度数	パーセント
携帯電話	4892	71.02
不明	1522	22.10
固定電話	474	6.88

直近のコンタクト日付

day	度数	パーセント
20	373	5.42
18	326	4.73
30	297	4.31
5	289	4.20
14	289	4.20
15	286	4.15
6	276	4.01
21	269	3.91

直近のコンタクト月

month	度数	パーセント
may	1817	26.38
jul	953	13.84
aug	934	13.56
jun	745	10.82
nov	586	8.51
apr	552	8.01
feb	471	6.84
jan	219	3.18

コンタクト回数(今回)

campaign	度数	パーセント
1	2929	42.52
2	1834	26.63
3	829	12.04
4	481	6.98
5	244	3.54
6	180	2.61
7	94	1.36
8	81	1.18

経過日数

pdays	度数	パーセント
-1	5201	75.51
92	58	0.84
181	50	0.73
182	48	0.70
91	45	0.65
183	32	0.46
184	28	0.41
94	26	0.38
95	24	0.35
93	23	0.33

コンタクト回数(前回)

previous	度数	パーセント
0	5201	75.51
1	504	7.32
2	424	6.16
3	256	3.72
4	156	2.26
5	98	1.42
6	74	1.07
7	43	0.62
8	36	0.52
9	20	0.29
10	16	0.23

結果(前回)

poutcome	度数	パーセント
不明	5202	75.52
失敗	751	10.90
成功	599	8.70
その他	336	4.88

定期預金申込の有無

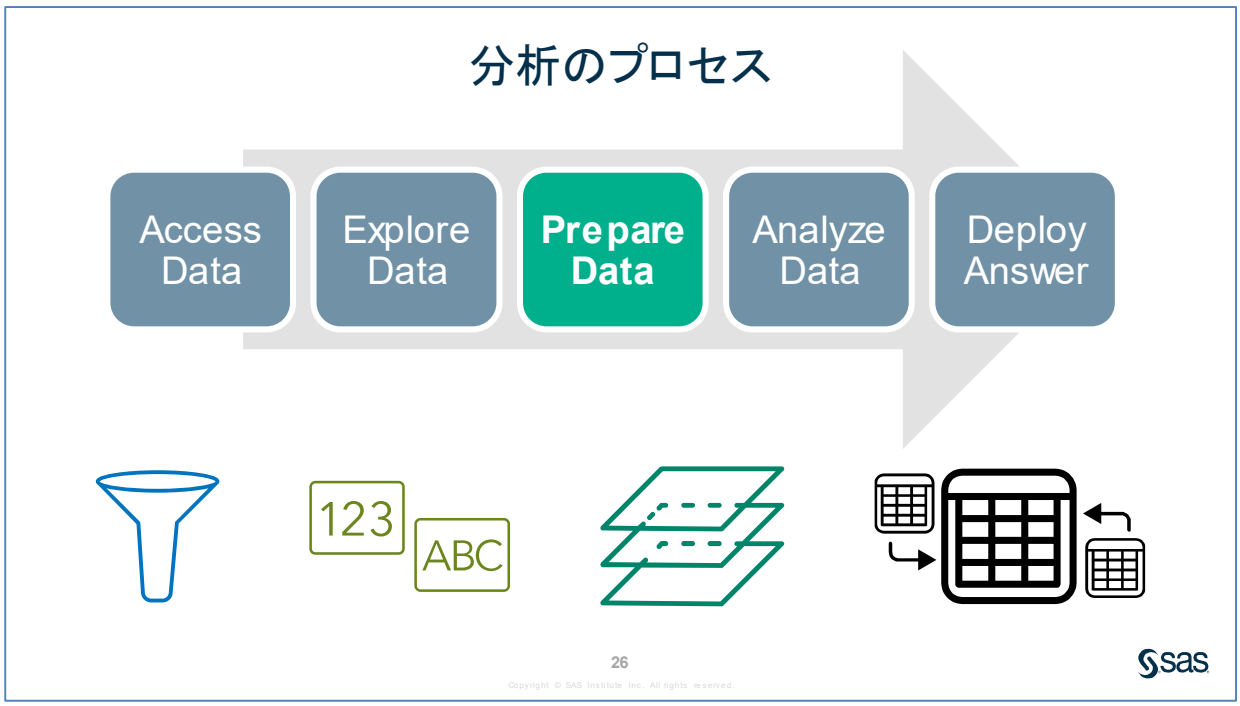
deposit	度数	パーセント
いいえ	4126	59.90
はい	2762	40.10

すべての変数の一意な値の一覧およびその度数(頻度値)が表示されます。おかしな値が含まれていないか、値の偏りの状況などを具体的な値で確認できます。
変数 pdays は、全体の 75%が-1、つまり、コンタクト履歴が存在しないようです。また、変数 previous は、全体の 75%が 0、かつ、10 未満の値が全体の 99%を占めていることがわかります。

13. 保存ボタンをクリックして、フローを上書き保存します。

End of Demonstration

2.4 分析用データの作成

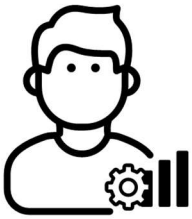


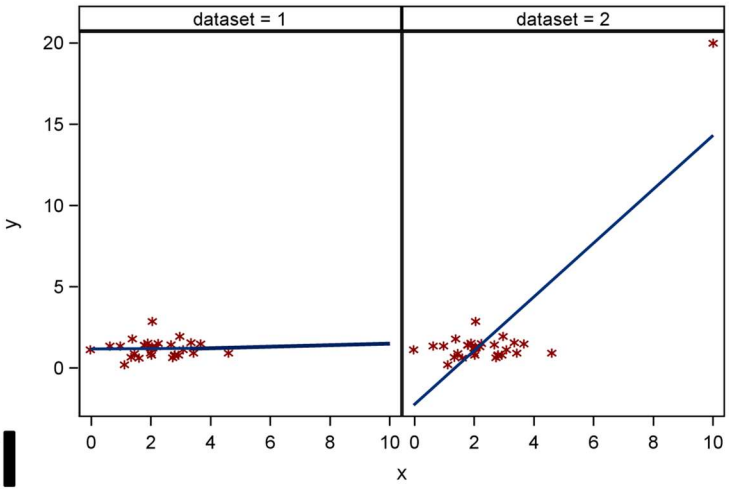
Prepare Data：分析に適したデータの形に加工します。また必要な項目を作成します。

先述の通り様々なデータ加工があります。本節では、量的変数の外れ値への対処、質的変数のダミー化、特徴量の選択、特徴量のスケーリングについて紹介し、シナリオデータを用いて演習を行います。

※コースの進行状況によっては、本セクションは概念の紹介のみにとどめ、演習は各自で学習していただく場合があります。あらかじめご了承ください。

極端なデータ値





27

極端な値が少し存在するだけで、選択したモデルの種類によってはその影響を強く受けてしまいます。上の散布図では、一つの点によって直線関係の大部分が決められてしまっています。
そこで、極端な値に対して、除去や代入を行って対応することがあります。
データの探索を通じて、campaign、balance、pdays、previous に極端な値（外れ値）のあることが分かりました。これらの変数にも対応を行ってみましょう。

本章の操作シナリオ

本章では、分析用データの作成を目標として、以下の様な分析ツールの操作を行います。

- 2.2 データへのアクセス
  - ①データへのアクセス
- 2.3 データ構造の理解
  - ②データの構造の調査
  - ③データの基礎集計
- **2.4 分析用データの作成**
  - **④量的変数の外れ値への対処**
  - ⑤質的変数のダミー化
  - ⑥特徴量の選択
  - ⑦特徴量のスケーリング

28

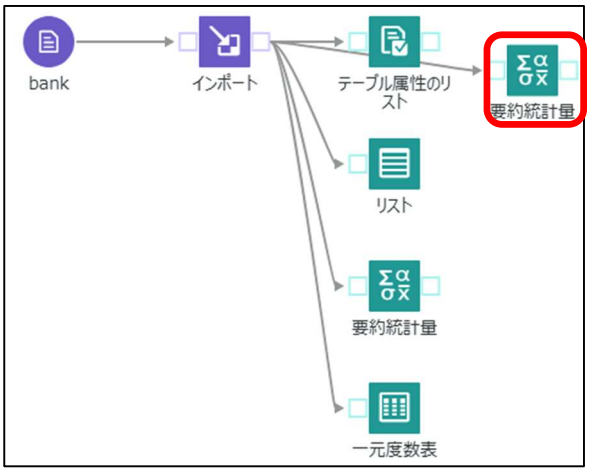
実際に分析ツールを使用して、外れ値への対処について確認します。

量的変数の外れ値への対処①

このデモでは、極端な値のある変数に対してその範囲を特定し、平均値を求めて代入します。
ここからしばらくの間、SAS Studio を使用して、特徴量エンジニアリングについて見ていきます。

1. 引き続き Lesson 2 フローを引き続き使用します。フローキャンバスで**要約統計量**ノードを右クリックし、**コピー**を選択し、白紙の部分で右クリックし、**貼り付け**を選択します。

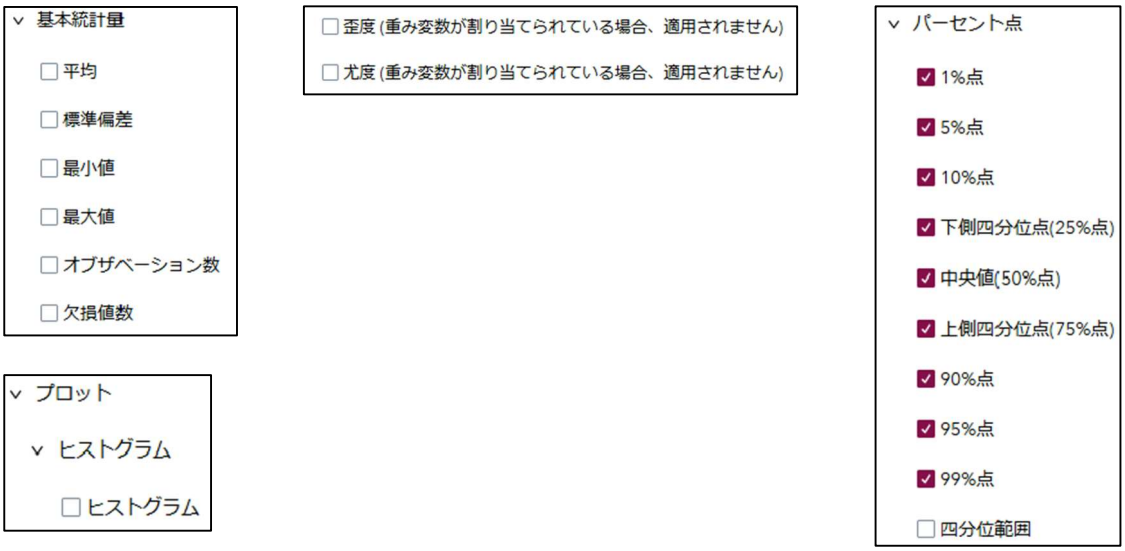
2. 新しく追加された要約統計量ノードをフローキャンバスで選択します。



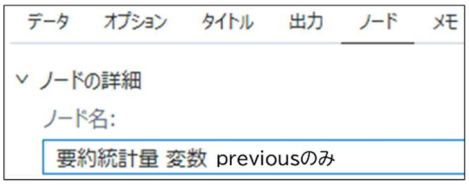
3. ノードの詳細の、データタブで**分析変数**の、変数 previous 以外のすべての変数を選択し、**列の削除**ボタンをクリックします。

これで分析変数に previous のみ残りました。

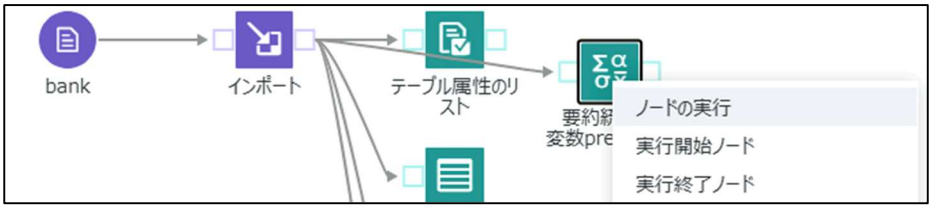
4. オプションタブで、**基本統計量**および**追加統計量**からすべてのチェックを外し、**パーセント点**の、**1%点**、**5%点**、**10%点**、**下側四分位点(25%点)**、**中央値(50%点)**、**上側四分位点(75%点)**、**90%点**、**95%点**、**99%点**にチェックを入れます。また、**プロット**の**ヒストグラム**のチェックも外します。



5. ノードタブで、ノード名に、「**要約統計量 変数 previous のみ**」と入力します。



6. フローキャンバスの、**要約統計量 変数 previous のみ**ノードを右クリックし、**ノードの実行**を選択します。



7. サブミットされたコードと結果タブの、結果タブから生成されたレポートを確認します。

分析変数：previous コンタクト回数(前回)

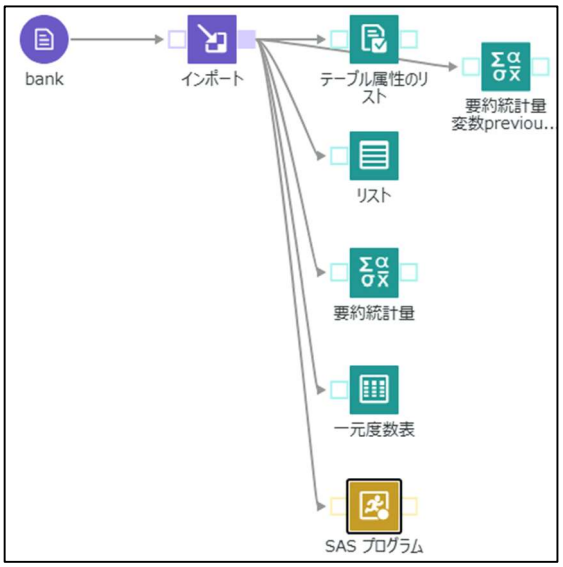
中央値	1 パーセント点	5 パーセント点	10 パーセント点	下側四分位点	上側四分位点	90 パーセント点	95 パーセント点	99 パーセント点
0	0	0	0	0	0	3.0000000	5.0000000	10.0000000

先ほどデータ探索を行った際、変数 previous の最大値は 58 であることがわかっています。また、先ほどの要約統計量の結果と同様に、この結果からほとんどのデータ値は 10 未満であることがわかります。そこで、99%点以上のデータの平均値を求め、その範囲のデータを平均値で置き換えることにします。

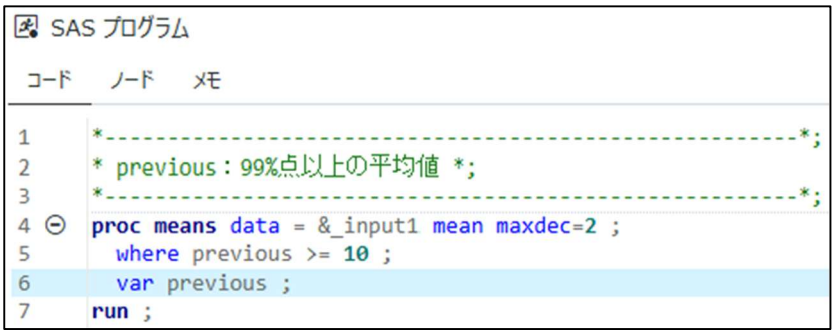
8. 今回は SAS プログラムを使用して算出するため、講師の指示に従ってファイルの場所に移動します。「**2.3.1 量的変数の外れ値への対処①.sas**」を右クリックし、**Edit with Notepad++** を選択します。ファイルが開いたら、以下部分を選択し、Ctrl+C でコピーします。

```
*------------------------------------------------------------*;
* previous：99%点以上の平均値 *;
*------------------------------------------------------------*;
proc means data = &_input1 mean maxdec=2 ;
  where previous >= 10 ;
  var previous ;
run ;
```

9. フローキャンバスに戻り、**ステップ**セクションから**開発→SAS プログラム**ステップを選択し、フローキャンバス内のインポートノードの出力ポートにドラッグします。



10. フローキャンバスで **SAS プログラム**ノードを選択し、ノードの詳細の、コードタブで Ctrl+V で先ほどのプログラムを貼り付けます。



このプログラムでは、99%点(値 10)以上のデータを抽出して平均値を求めます。

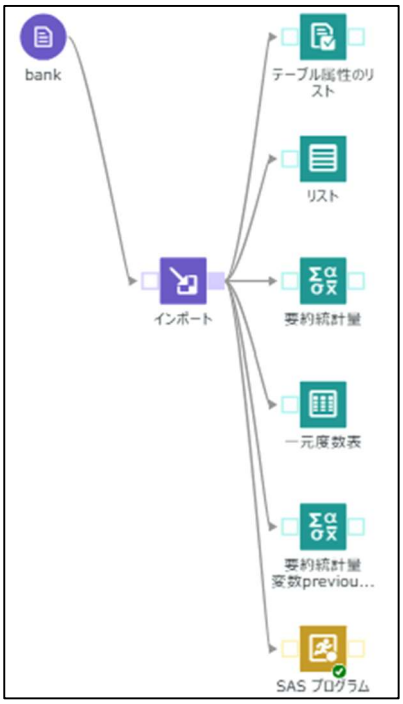
11. フローキャンバスの、**SAS プログラム**ノードを右クリックし、**ノードの実行**を選択します。

12. サブミットされたコードと結果タブの、結果タブから生成されたレポートを確認します。

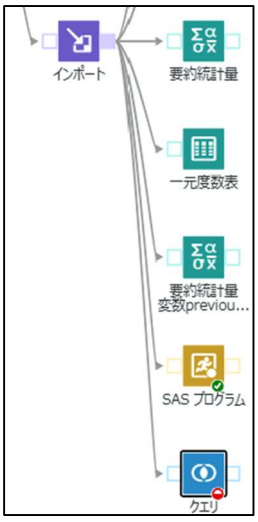
分析変数：previous コンタクト回数(前回)
平均
16.21

平均値は 16.21 です。

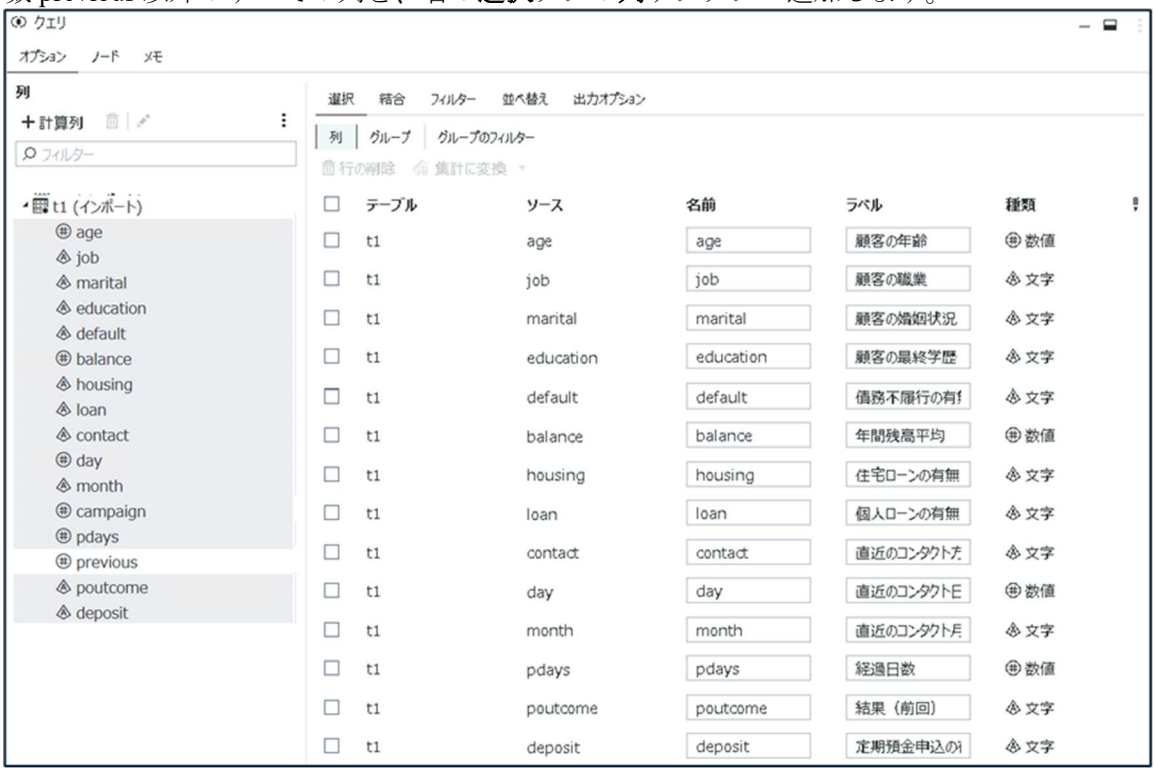
13. フローを整形するために、フローのツールバーのノードの配置ボタンをクリックします。



14. 求めた平均値に置き換えた値を持つ新しい変数 rep_previous を作成し、Bank データから変数 campaign を削除するデータ加工作業を実施します。フロータブに戻り、**ステップ**セクションから**データの変換** → **クエリ**ステップを選択し、フローキャンバス内のインポートノードの出力ポートにドラッグします



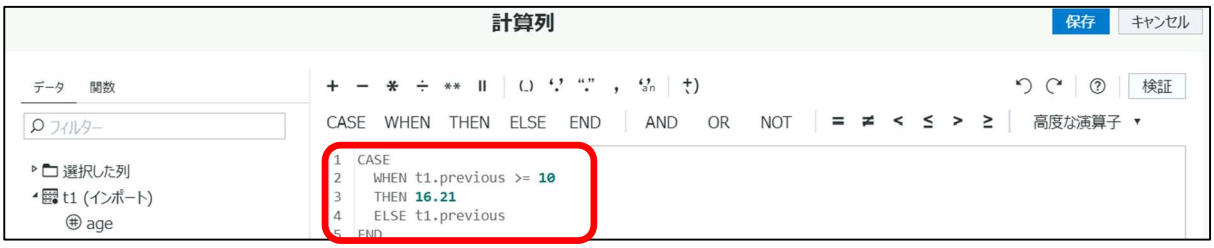
15. フローキャンバスで**クエリ**ノードを選択し、ノードの詳細の、左の**列**ウィンドウから、変数 previous 以外のすべての列を、右の**選択**タブの**列**サブタブへ追加します。



16. **列**ウィンドウにある**計算列**アイコンをクリックし、先ほど求めた平均値に置き換えた値を持つ新しい変数 rep_previous を作成します。

式には以下を入力します：

```
CASE
  WHEN t1.previous >= 10
  THEN 16.21
  ELSE t1.previous
END
```

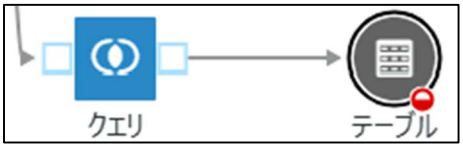


下部の**プロパティタブ**には以下を入力します：

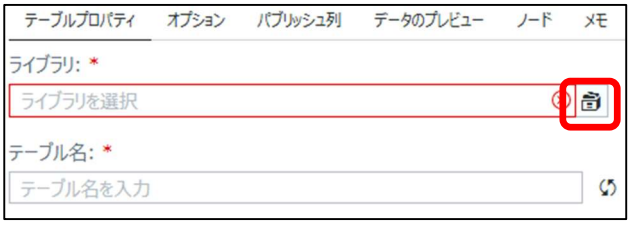
列名：rep_previous
列ラベル：コンタクト回数（前回）
種類：数値

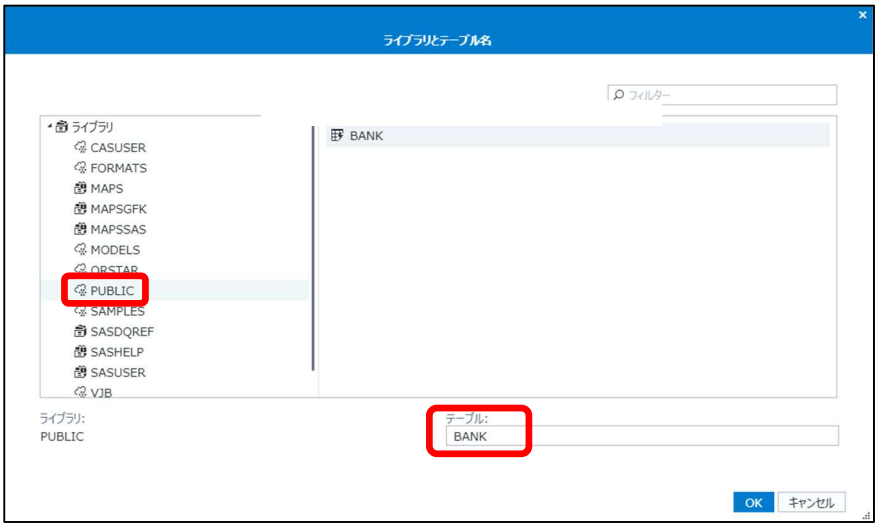
右上の**保存**ボタンをクリックします。
新しい変数が**列**サブタブの一番下に追加されました。

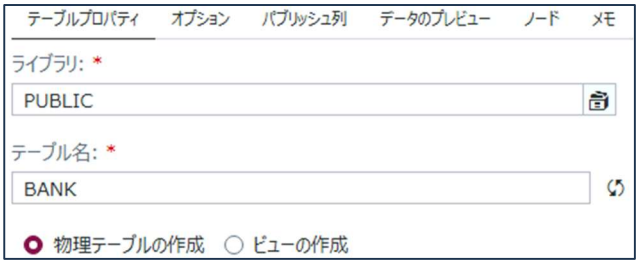
17. フローキャンバスのクエリ**ノードの右端の四角(出力ポート)**を右クリックし、テーブルの追加を選択します。テーブルノードが追加されます。



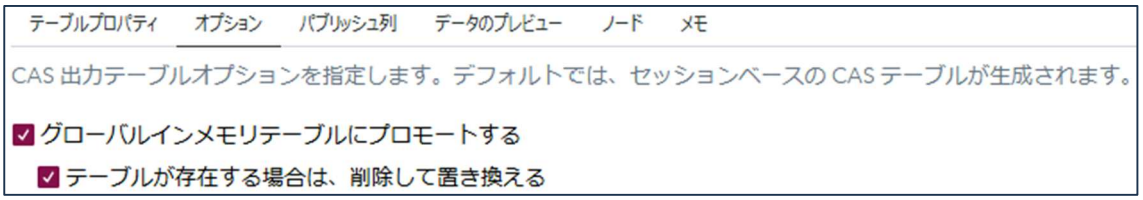
18. 追加されたテーブルノードを選択し、ノードの詳細のテーブルプロパティでライブラリアイコンをクリックして以下を設定します。
ライブラリ名：public
テーブル名：bank







19. オプションタブで、グローバルインメモリテーブルにプロモートするにチェックを入れます。あわせて、テーブルが存在する場合は、削除して置き換えるにもチェックを入れます。



20. フローキャンバスの、クエリノードを右クリックし、ノードの実行を選択します。

21. BANK テーブルノードを選択します。

22. ノードの詳細の、データのプレビュータブで、加工されたデータを確認します。

	day	month	pdays	poutcome	deposit	rep_previous
43	21	may	-1	不明	いいえ	0
44	21	may	-1	不明	いいえ	0
45	18	may	355	失敗	いいえ	5
46	4	jun	-1	不明	いいえ	0
47	30	jul	-1	不明	いいえ	0
48	11	feb	297	失敗	はい	2
49	2	mar	89	失敗	はい	16.21
50	15	jul	-1	不明	いいえ	0
51	12	nov	-1	不明	はい	0
52	28	jan	-1	不明	いいえ	0
53	17	jul	-1	不明	いいえ	0

新しい変数 rep_previous が作成されていること、変数 campaign が削除されていることを確認できます。

23. 保存ボタンをクリックして、フローを上書き保存します。

End of Demonstration

量的変数の外れ値への対処②

このデモでは、コンタクト経過日数の値（-1 とそれ以外）から、コンタクトのあり・なしの二値の変数を作成します。

1. Lesson 2 フローを引き続き使用します。フローキャンバスでクエリノードを選択し、ノードの詳細の、列ウィンドウにある計算列アイコンをクリックし、新しい変数 bin_pdays を作成します。

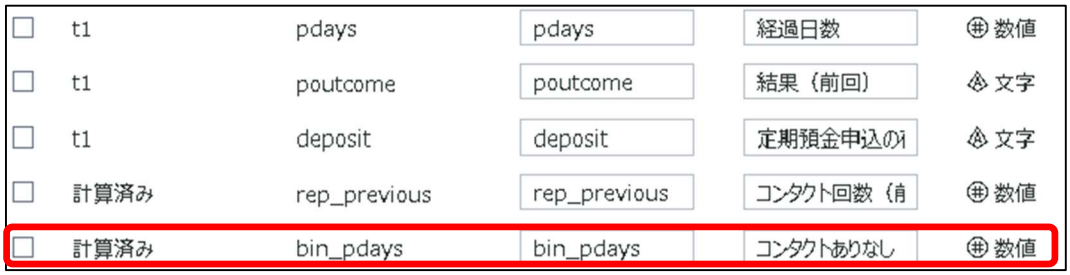
式には以下を入力します：

```
CASE
  WHEN pdays = -1
  THEN 0
  ELSE 1
END
```

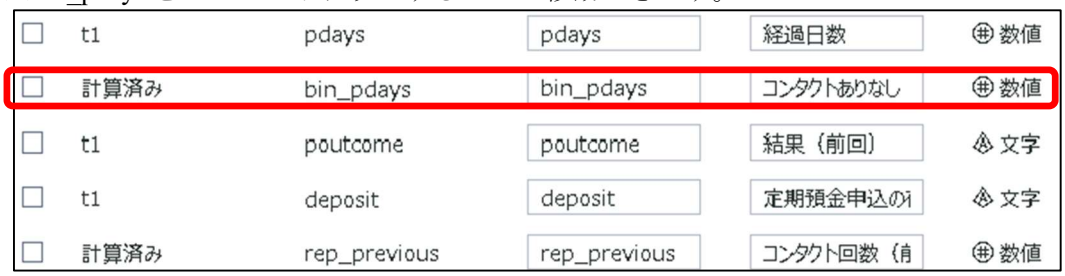
下部のプロパティタブには以下を入力します：

列名：bin_pdays
列ラベル：コンタクトありなし（0/1）
種類：数値

右上の**保存**ボタンをクリックします。
新しい変数が**列**サブタブの一番下に追加されました。



2. 列サブタブで、作成された新規変数 bin_pdays を、変数 pdays の下に移動します。
※bin_pdays をつかんでドラッグすることで移動できます。

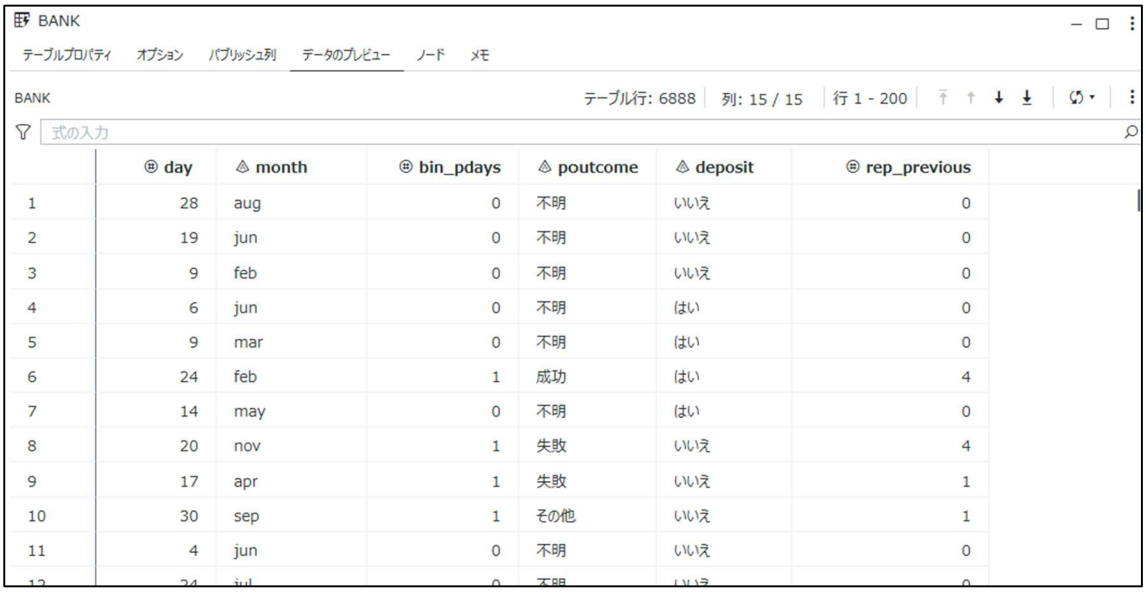


3. 変数 pdays のチェックをいれ、行の削除をクリックし、変数 pdays を削除します。

4. フローキャンバスの、クエリノードを右クリックし、ノードの実行を選択します。

5. BANK テーブルノードを選択します。

6. ノードの詳細の、データのプレビュータブで、加工されたデータを確認します。



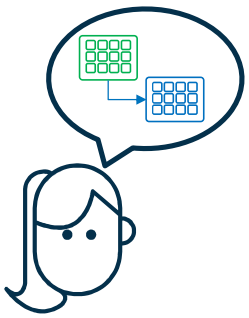
	day	month	bin_pdays	poutcome	deposit	rep_previous
1	28	aug	0	不明	いいえ	0
2	19	jun	0	不明	いいえ	0
3	9	feb	0	不明	いいえ	0
4	6	jun	0	不明	はい	0
5	9	mar	0	不明	はい	0
6	24	feb	1	成功	はい	4
7	14	may	0	不明	はい	0
8	20	nov	1	失敗	いいえ	4
9	17	apr	1	失敗	いいえ	1
10	30	sep	1	その他	いいえ	1
11	4	jun	0	不明	いいえ	0

新しい変数 bin_pdays が作成されていること、変数 pdays が削除されていることを確認できます。

7. 保存ボタンをクリックして、フローを上書き保存します。

End of Demonstration

ダミー変数化



Produce
Apple
Orange
Grape

Produce	Apple	Orange	Grape
Apple	1	0	0
Orange	0	1	0
Grape	0	0	1

31

sas

ダミー変数化とは、質的変数を量的変数に変換することです。言い換えると、文字変数を数値変数に変換して、モデル構築の際に計算に利用できる状態にすることです。
代表的な方法に、One-Hot エンコーディングがあります。One-Hot エンコーディングでは、元の文字変数をカテゴリごとに複数の変数に分割して、該当する行には 1 を、それ以外には 0 を割り当てます。

本章の操作シナリオ

本章では、分析用データの作成を目標として、以下の様な分析ツールの操作を行います。

2.2 データへのアクセス
　①データへのアクセス
2.3 データ構造の理解
　②データの構造の調査
　③データの基礎集計
2.4 分析用データの作成
　④量的変数の外れ値への対処
　⑤質的変数のダミー化
　⑥特徴量の選択
　⑦特徴量のスケーリング

32

sas

実際に分析ツールを使用して、カテゴリ変数のダミー化を行います。

質的変数のダミー化

このデモでは、カテゴリ変数を量的変数に変換します。

1. SAS Studio の画面左上部の、**新規作成→SAS プログラム**を選択して、新しいプログラムタブ開きます。

2. 「2.3.3 質的変数のダミー化.sas」を右クリックして、**Edit with Notepad++**を選択します。ファイルが開いたら、Ctrl＋A で全体を選択し、Ctrl＋C でコピーして、SAS Studio の先ほど開いた「SAS プログラム.sas」タブに Ctrl＋V で貼り付けます。
※講師の指示に従ってファイルの場所に移動します。

3. コードタブをクリックして、以下のプログラムを範囲選択して実行します。

```
*------------------------------------------------------------*;
* カテゴリ変数のダミー化 *;
*------------------------------------------------------------*;
data work.bank ;
  retain FakeY 0 ;
  set public.bank ;
run;

proc glmselect data = work.bank outdesign( addinputvars prefix =
dmy_ ) = work.bank2( drop = FakeY ) noprint ;
   class job marital education default housing loan contact month
poutcome ;
   model FakeY = job marital education default housing loan contact
month poutcome / noint selection=none;
run;

data work.bank3 ;
  set work.bank2 ;
  drop job marital education default housing loan contact month
poutcome ;
run ;
```

このプログラムでは、ダミー変数化を行う二つ目のステップで一時的に使用する偽変数 FakeY を一つ目のステップで作成しています。二つ目のステップで、ダミー変数を作成し、三つ目のステップで、不要となったカテゴリ変数をすべて削除しています。

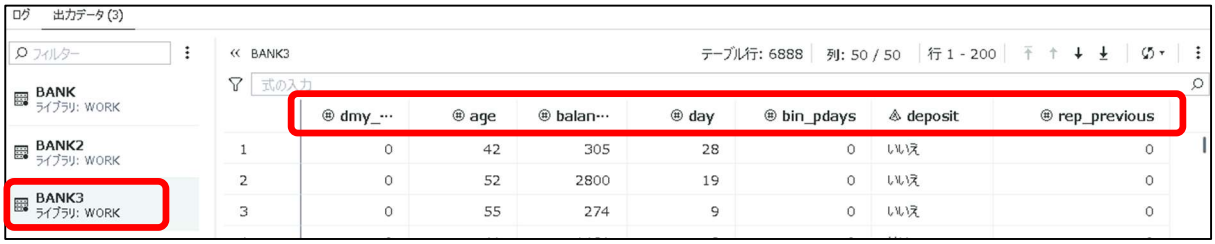
4. 出力データタブで、WORK ライブラリの、BANK データをクリックして表示し確認します。

偽変数 FakeY が作成されていることを確認できます。

5. WORK ライブラリの、BANK2 データをクリックして表示し確認します。

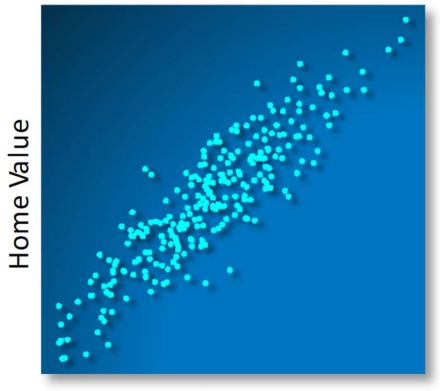
ダミー変数が作成されたことを確認できます。

6. WORK ライブラリから、BANK3 データをクリックして表示し確認します。
変数 job、marital、education、default、housing、loan、contact、month、poutcome が削除されたことを確認できます。

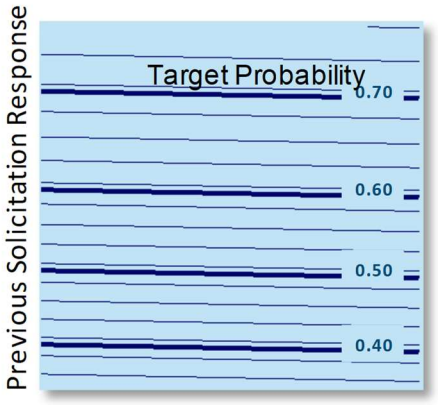


End of Demonstration

特徴量選択の戦略



冗長性



非関連性

34

特徴量（変数）の数が多い高次元のデータでモデルを作成すると、良い結果が得られなくなることがあります。例えば、モデルが過度に学習され未知のデータに適合できない、また計算コストが増大するという可能性があります。そこで、分析に利用する特徴量を選択する必要があります。その際のポイントは主に 2 つ、冗長的な値を持つ特徴量を特定し選択する、また予測結果に関連のない特徴量を特定し削除することです。高次元のデータセットでは、関連のない変数を特定することは、冗長な変数を特定することよりも困難です。最初に冗長性を減らし、次に低次元の空間で非関連性の問題に取り組むことです。

本章の操作シナリオ

本章では、分析用データの作成を目標として、以下の様な分析ツールの操作を行います。

- 2.2 データへのアクセス
  - ①データへのアクセス
- 2.3 データ構造の理解
  - ②データの構造の調査
  - ③データの基礎集計
- **2.4 分析用データの作成**
  - ④量的変数の外れ値への対処
  - ⑤質的変数のダミー化
  - **⑥特徴量の選択**
  - ⑦特徴量のスケーリング

35

実際に分析ツールを使用して、特徴量の選択を行います。

特徴量の選択①

このデモでは、高相関の特徴量を特定して、分析に使用する特徴量の選択を行います。

1. SAS Studio の画面左上部の、**新規作成→SAS プログラム**を選択して、新しいプログラムタブ開きます。

2. 「2.3.4 特徴量の選択①.sas」を右クリックして、**Edit with Notepad++**を選択します。ファイルが開いたら、Ctrl＋A で全体を選択し、Ctrl＋C でコピーして、SAS Studio の「プログラム 1.sas」タブに Ctrl＋V で貼り付けます。
※講師の指示に従ってファイルの場所に移動します。

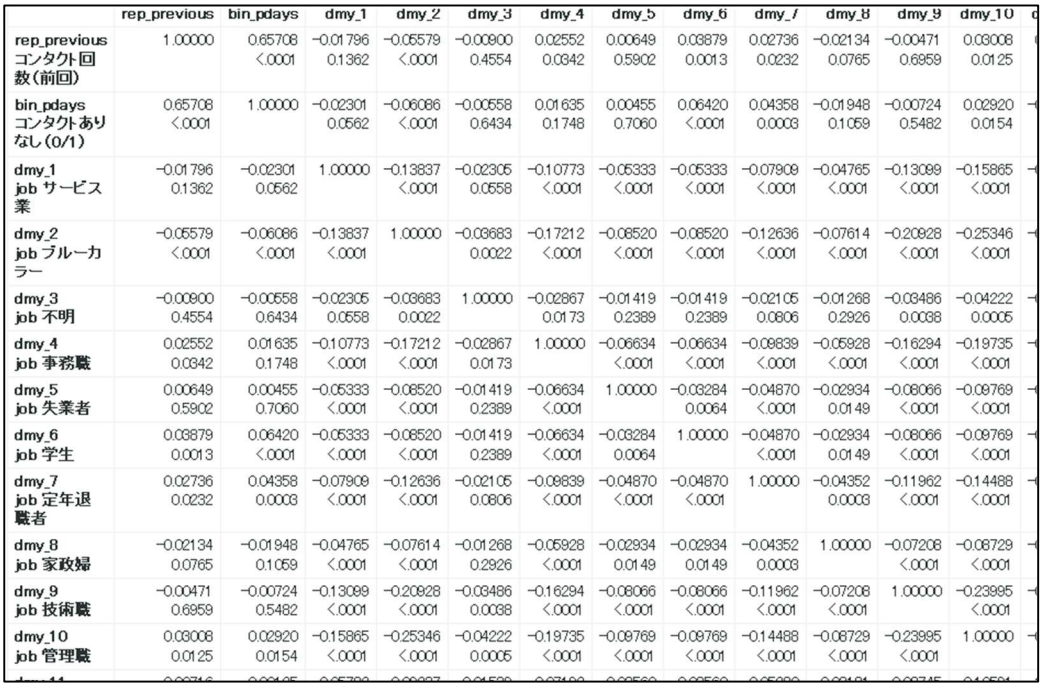
3. コードタブをクリックして、以下のプログラムを範囲選択して実行します。

```
*------------------------------------------------------------*;
* 高相関の確認 *;
*------------------------------------------------------------*;
proc corr data = work.bank3 outp = work.Pearson ;
  var rep_: bin_: dmy_: ;
run ;

data work.Pearson2 ;
  set work.Pearson ;
  where _type_ = 'CORR' ;
  array P{ * } rep_: bin_: dmy_: ;
  do i = 1 to dim( P ) ;
    if _name_ = vname( P{ i } ) then _label_ = vlabel( P{ i } ) ;
  end ;
  do i = 1 to dim( P ) ;
    if _name_ ^= vname( P{ i } ) and ( P{ i } <= -0.8 or P{ i } >=
0.8 ) then do ;
      _name2_  = vname( P{ i } ) ;
      _label2_ = vlabel( P{ i } ) ;
      PCC = P{ i } ;
      output ;
    end ;
  end ;
  keep _type_ _name_ _label_ _name2_ _label2_ PCC ;
run ;

data bank4 ;
  set bank3 ;
  drop dmy_20 /* default no        */
       dmy_22 /* housing no        */
       dmy_24 /* loan no           */
       dmy_28 /* contact unknown   */
       dmy_44 /* poutcome unknown */
       ;
```

```
run ;
```

このプログラムでは、一つ目のステップで、変数の相関係数を出力しています。二つ目のステップで、高い相関係数を持つ項目を選択しています。三つ目のステップで、冗長している変数を削除しています。

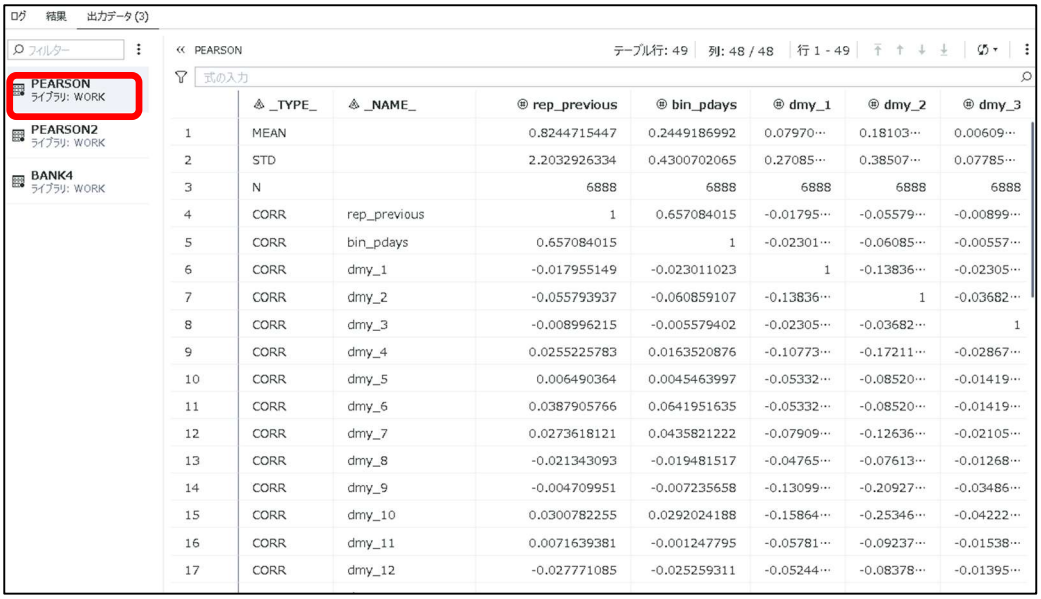
4. まず、結果タブから以下の表を確認します。



	rep_previous	bin_pdays	dmy_1	dmy_2	dmy_3	dmy_4	dmy_5	dmy_6	dmy_7	dmy_8	dmy_9	dmy_10
rep_previous コンタクト回数(前回)	1.00000	0.65708 <.0001	−0.01796 0.1362	−0.05579 <.0001	−0.00900 0.4554	0.02552 0.0342	0.00649 0.5902	0.03879 0.0013	0.02736 0.0232	−0.02134 0.0765	−0.00471 0.6959	0.03008 0.0125
bin_pdays コンタクトありなし(0/1)	0.65708 <.0001	1.00000	−0.02301 0.0562	−0.06086 <.0001	−0.00558 0.6434	0.01635 0.1748	0.00455 0.7060	0.06420 <.0001	0.04358 0.0003	−0.01948 0.1059	−0.00724 0.5482	0.02920 0.0154
dmy_1 job サービス業	−0.01796 0.1362	−0.02301 0.0562	1.00000	−0.13837 <.0001	−0.02305 0.0558	−0.10773 <.0001	−0.05333 <.0001	−0.05333 <.0001	−0.07909 <.0001	−0.04765 <.0001	−0.13089 <.0001	−0.15865 <.0001
dmy_2 job ブルーカラー	−0.05579 <.0001	−0.06086 <.0001	−0.13837 <.0001	1.00000	−0.03683 0.0022	−0.17212 <.0001	−0.08520 <.0001	−0.08520 <.0001	−0.12636 <.0001	−0.07614 <.0001	−0.20928 <.0001	−0.25346 <.0001
dmy_3 job 不明	−0.00900 0.4554	−0.00558 0.6434	−0.02305 0.0558	−0.03683 0.0022	1.00000	−0.02867 0.0173	−0.01419 0.2389	−0.01419 0.2389	−0.02105 0.0806	−0.01268 0.2926	−0.03486 0.0038	−0.04222 0.0005
dmy_4 job 事務職	0.02552 0.0342	0.01635 0.1748	−0.10773 <.0001	−0.17212 <.0001	−0.02867 0.0173	1.00000	−0.06634 <.0001	−0.06634 <.0001	−0.09839 <.0001	−0.05928 <.0001	−0.16294 <.0001	−0.19735 <.0001
dmy_5 job 失業者	0.00649 0.5902	0.00455 0.7060	−0.05333 <.0001	−0.08520 <.0001	−0.01419 0.2389	−0.06634 <.0001	1.00000	−0.03284 0.0064	−0.04870 <.0001	−0.02934 0.0149	−0.08066 <.0001	−0.09769 <.0001
dmy_6 job 学生	0.03879 0.0013	0.06420 <.0001	−0.05333 <.0001	−0.08520 <.0001	−0.01419 0.2389	−0.06634 <.0001	−0.03284 0.0064	1.00000	−0.04870 <.0001	−0.02934 0.0149	−0.08066 <.0001	−0.09769 <.0001
dmy_7 job 定年退職者	0.02736 0.0232	0.04358 0.0003	−0.07909 <.0001	−0.12636 <.0001	−0.02105 0.0806	−0.09839 <.0001	−0.04870 <.0001	−0.04870 <.0001	1.00000	−0.04352 0.0003	−0.11962 <.0001	−0.14488 <.0001
dmy_8 job 家政婦	−0.02134 0.0765	−0.01948 0.1059	−0.04765 <.0001	−0.07614 <.0001	−0.01268 0.2926	−0.05928 <.0001	−0.02934 0.0149	−0.02934 0.0149	−0.04352 0.0003	1.00000	−0.07208 <.0001	−0.08729 <.0001
dmy_9 job 技術職	−0.00471 0.6959	−0.00724 0.5482	−0.13089 <.0001	−0.20928 <.0001	−0.03486 0.0038	−0.16294 <.0001	−0.08066 <.0001	−0.08066 <.0001	−0.11962 <.0001	−0.07208 <.0001	1.00000	−0.23995 <.0001
dmy_10 job 管理職	0.03008 0.0125	0.02920 0.0154	−0.15865 <.0001	−0.25346 <.0001	−0.04222 0.0005	−0.19735 <.0001	−0.09769 <.0001	−0.09769 <.0001	−0.14488 <.0001	−0.08729 <.0001	−0.23995 <.0001	1.00000

相関係数は-1 から 1 までの値を取り、-1 に近いほど負の相関があり、1 に近いほど正の相関があります。負の相関とは、一方の変数の値が高いほど、もう一方の変数の値が低い傾向にあること、正の相関とは、一方の変数の値が高いほど、もう一方の変数の値が高い傾向にあることです。-0.8 以下、または 0.8 以上の場合に、変数間に強い相関があると考えることができます。しかしながら、この表から強い相関を持つ変数を探すのは骨が折れそうです。

5. 出力データタブで、WORK ライブラリの、Pearson データをクリックして表示し確認します。



	TYPE	_NAME_	rep_previous	bin_pdays	dmy_1	dmy_2	dmy_3
1	MEAN		0.8244715447	0.2449186992	0.07970…	0.18103…	0.00609…
2	STD		2.2032926334	0.4300702065	0.27085…	0.38507…	0.07785…
3	N		6888	6888	6888	6888	6888
4	CORR	rep_previous	1	0.657084015	-0.01795…	-0.05579…	-0.00899…
5	CORR	bin_pdays	0.657084015	1	-0.02301…	-0.06085…	-0.00557…
6	CORR	dmy_1	-0.017955149	-0.023011023	1	-0.13836…	-0.02305…
7	CORR	dmy_2	-0.055793937	-0.060859107	-0.13836…	1	-0.03682…
8	CORR	dmy_3	-0.008996215	-0.005579402	-0.02305…	-0.03682…	1
9	CORR	dmy_4	0.0255225783	0.0163520876	-0.10773…	-0.17211…	-0.02867…
10	CORR	dmy_5	0.006490364	0.0045463997	-0.05332…	-0.08520…	-0.01419…
11	CORR	dmy_6	0.0387905766	0.0641951635	-0.05332…	-0.08520…	-0.01419…
12	CORR	dmy_7	0.0273618121	0.0435821222	-0.07909…	-0.12636…	-0.02105…
13	CORR	dmy_8	-0.021343093	-0.019481517	-0.04765…	-0.07613…	-0.01268…
14	CORR	dmy_9	-0.004709951	-0.007235658	-0.13099…	-0.20927…	-0.03486…
15	CORR	dmy_10	0.0300782255	0.0292024188	-0.15864…	-0.25346…	-0.04222…
16	CORR	dmy_11	0.0071639381	-0.001247795	-0.05781…	-0.09237…	-0.01538…
17	CORR	dmy_12	-0.027771085	-0.025259311	-0.05244…	-0.08378…	-0.01395…

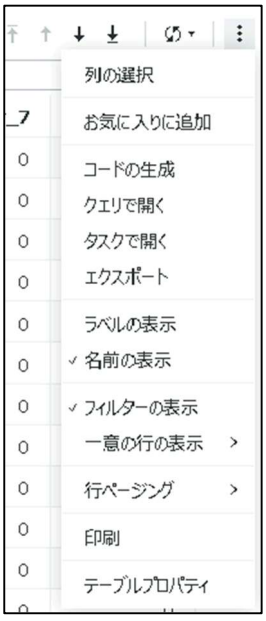
これは上の結果の表をデータに保存したものです。プログラムを使用して、このデータから強い相関がある項目を抽出します。

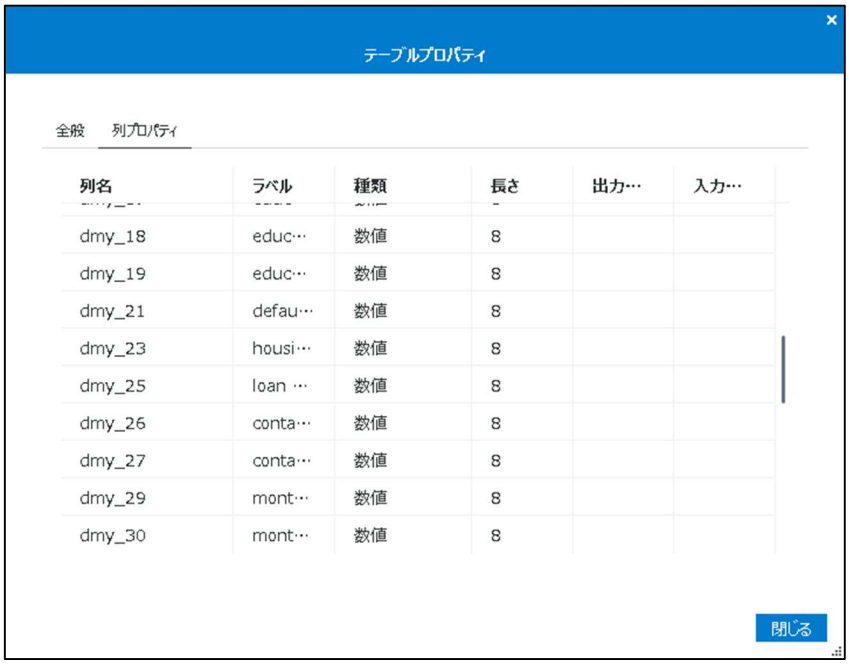
6. WORK ライブラリの、Pearson2 データをクリックして表示し確認します。

	TYPE	_NAME_	_label_	_name2_	_label2_	PCC
1	CORR	bin_pdays	コンタクトありなし (0/1)	dmy_42	poutcome 不明	-0.999607…
2	CORR	dmy_20	default いいえ	dmy_21	default はい	-1
3	CORR	dmy_21	default はい	dmy_20	default いいえ	-1
4	CORR	dmy_22	housing いいえ	dmy_23	housing はい	-1
5	CORR	dmy_23	housing はい	dmy_22	housing いいえ	-1
6	CORR	dmy_24	loan いいえ	dmy_25	loan はい	-1
7	CORR	dmy_25	loan はい	dmy_24	loan いいえ	-1
8	CORR	dmy_26	contact 不明	dmy_28	contact 携帯電話	-0.833767…
9	CORR	dmy_28	contact 携帯電話	dmy_26	contact 不明	-0.833767…
10	CORR	dmy_42	poutcome 不明	bin_pdays	コンタクトありなし (0/1)	-0.999607…

相関係数が-0.8 以下または 0.8 以上の変数の組み合わせが取得されました。
このデータにある、以下の特徴量の組み合わせは、バイナリ（0、1）のダミー変数なのでいずれかを削除します。
bin_pday と poutcome 不明
default いいえと default はい
housing いいえと housing はい
loan いいえと loan はい
contact 不明と contact 携帯電話

7. WORK ライブラリの BANK4 データをクリックし、右上端の詳細オプション→テーブルプロパティを選択し、WORK ライブラリの BANK4 データの列のプロパティを表示します。





列名	ラベル	種類	長さ	出力…	入力…
dmy_18	educ…	数値	8		
dmy_19	educ…	数値	8		
dmy_21	defau…	数値	8		
dmy_23	housi…	数値	8		
dmy_25	loan …	数値	8		
dmy_26	conta…	数値	8		
dmy_27	conta…	数値	8		
dmy_29	mont…	数値	8		
dmy_30	mont…	数値	8		

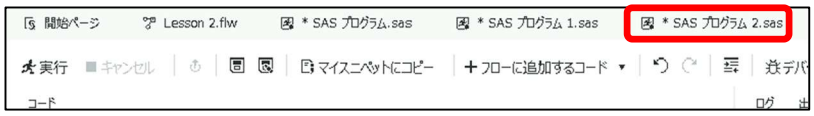
default いいえ（dmy_20）、housing いいえ（dmy_22）、loan いいえ（dmy_24）、contact 不明（dmy_28）、poutcome 不明（dmy_44）を削除しました。

End of Demonstration

特徴量の選択②

このデモでは、低分散の特徴量を特定して、分析に使用する特徴量の選択を行います。

1. SAS Studio の画面左上部の、**新規作成→SAS プログラム**を選択して、新しいプログラムタブ開きます。



2. 「2.3.5 特徴量の選択②.sas」を右クリックして、**Edit with Notepad++**を選択します。ファイルが開いたら、Ctrl＋A で全体を選択し、Ctrl＋C でコピーして、SAS Studio の「プログラム 2.sas」タブに Ctrl＋V で貼り付けます。
※講師の指示に従ってファイルの場所に移動します。

3. コードタブをクリックして、以下のプログラムを範囲選択して実行します。

```
*------------------------------------------------------------*;
* 低分散の確認 *;
*------------------------------------------------------------*;
proc means data = work.bank4 var ;
  var rep_: bin_: dmy_: ;
  output out = work.var
         var =
         / autoname ;
run ;

data work.var2 ;
  set var ;
  array V{ * } rep_: bin_: dmy_: ;
  do i = 1 to dim( V ) ;
    if V{ i } < 0.01 then do ;
      name  = vname( V{ i } ) ;
      label = vlabel( V{ i } ) ;
      var = V{ i } ;
    end ;
  end ;
  keep name label var ;
run ;
```

このプログラムでは、一つ目のステップで、変数の分散を出力しています。二つ目のステップで、低い分散を持つ項目を選択しています。

4. まず、結果タブから以下の表を確認します。

変数	ラベル	分散
rep_previous	コンタクト回数(前回)	4.8544984
bin_pdays	コンタクトありなし(0/1)	0.1849604
dmy_1	job サービス業	0.0733618
dmy_2	job ブルーカラー	0.1482857
dmy_3	job 不明	0.0060613
dmy_4	job 事務職	0.1042260
dmy_5	job 失業者	0.0307880
dmy_6	job 学生	0.0307880
dmy_7	job 定年退職者	0.0628348
dmy_8	job 家政婦	0.0249024
dmy_9	job 技術職	0.1380361
dmy_10	job 管理職	0.1744961
dmy_11	job 自営業	0.0357900
dmy_12	job 起業家	0.0298352
dmy_13	marital 既婚	0.2437519
dmy_14	marital 未婚	0.2131388
dmy_15	marital 離婚	0.1000945
dmy_16	education 不明	0.0402025
dmy_17	education 中学卒業	0.1196511
dmy_18	education 大学卒業	0.2193255
dmy_19	education 高校卒業	0.2500042
dmy_21	default はい	0.0159981
dmy_23	housing はい	0.2498767
dmy_25	loan はい	0.1141417
dmy_26	contact 不明	0.1721639
dmy_27	contact 固定電話	0.0640891
dmy_29	month apr	0.0737278

分散とは数値データのばらつきを表すための指標です。ある数値データにおける、平均値と個々のデータの差の二乗の平均を計算することによって求めることができます。
この分散の値がごく小さい変数は、除外しても基本的に問題がありません。例えば、分散が 0 の変数は、すべての変数で同じ値が入っていることになり特徴量として機能しないからです。
上記の表から、低い分散の値を探すには、やはり骨が折れそうです。

5. 出力データタブで、WORK ライブラリの、VAR データをクリックして表示し確認します。

これは上の結果の表をデータに保存したものです。プログラムを使用して、このデータから低い分散がある項目を抽出します。

6. WORK ライブラリの、VAR2 データをクリックして表示し確認します。

分散が 0 の特徴量は削除しても基本問題ありませんが、今回は存在しませんでした。

分散がほぼ 0 の場合、分散が極端に低い場合は、説明変数としての意味を持たないと考え、特徴から削除するという方法も取れます。ただし、これらの方法は必ずしも正解ではなく、元のデータを観察して削除するかどうかの判断が必要になります。

End of Demonstration

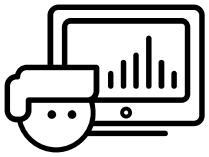
標準化・正規化

標準化

- $\mu = 0,\ \sigma = 1$
- $x_{std} = \frac{x-\mu}{\sigma}$

正規化

- $min >= 0,\ max <= 1$
- $x_{\mathrm{norm}} = \frac{x-x\mathrm{mi}}{x max - x\mathrm{min}}$



38

異なる単位を持つ特徴量、例えば、身長、体重、年齢などを、そのまま使用してしまうと、尺度（スケール）が異なるため、モデルの学習がうまく行かない場合があります。そこで、ある基準に従ってデータを変換し、尺度を統一することを行うことがあります。尺度を統一することをスケーリングと呼びます。スケーリングには、標準化、正規化といった方法があります。標準化は Standardization もしくは Z-score Normalization、正規化は Min-Max Normalization の日本語です。

本章の操作シナリオ

本章では、分析用データの作成を目標として、以下の様な分析ツールの操作を行います。

2.2 データへのアクセス
　①データへのアクセス
2.3 データ構造の理解
　②データの構造の調査
　③データの基礎集計
2.4 分析用データの作成
　④量的変数の外れ値への対処
　⑤質的変数のダミー化
　⑥特徴量の選択
　⑦特徴量のスケーリング

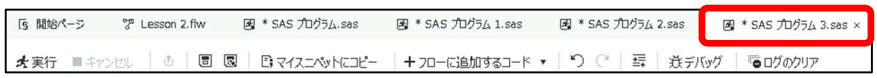
39

実際に分析ツールを使用して、特徴量のスケーリングを行います。

特徴量のスケーリング

このデモでは、標準化による特徴量のスケーリングを行います。

1. SAS Studio の画面左上部の、**新規作成→SAS プログラム**を選択して、新しいプログラムタブ開きます。



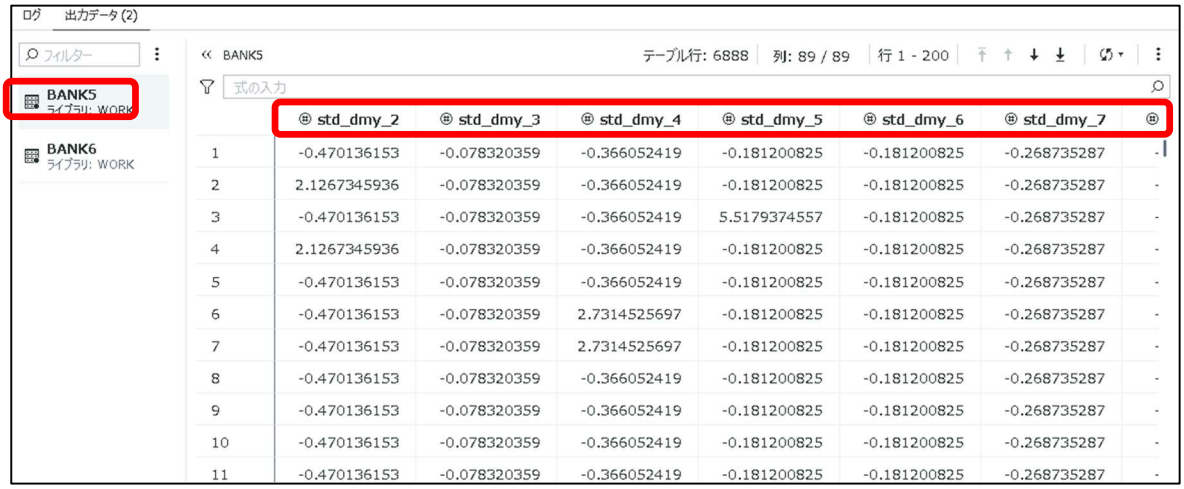
2. 「2.3.6 特徴量のスケーリング.sas」を右クリックして、**Edit with Notepad++**を選択します。ファイルが開いたら、Ctrl＋A で全体を選択し、Ctrl＋C でコピーして、SAS Studio の「プログラム 3.sas」タブに Ctrl＋V で貼り付けます。
※講師の指示に従ってファイルの場所に移動します。

コードタブをクリックして、以下のプログラムを範囲選択して実行します。

```
*------------------------------------------------------------*;
* 標準化 *;
*------------------------------------------------------------*;
proc stdize data=work.bank4 out=work.bank5 oprefix sprefix=std_
method=std ;
  var _numeric_ ;
run ;

*------------------------------------------------------------*;
* 分析用データ *;
*------------------------------------------------------------*;
data work.bank6 ;
  set work.bank5 ;
  keep deposit std_: ;
run ;
```

このプログラムでは、一つ目のステップで、すべての数値変数を標準化しています。
これでデータの前処理が完了となりますので、二つ目のステップで、出来上がったデータを分析用のデータとして使用するため、必要な変数に絞り込んでいます。

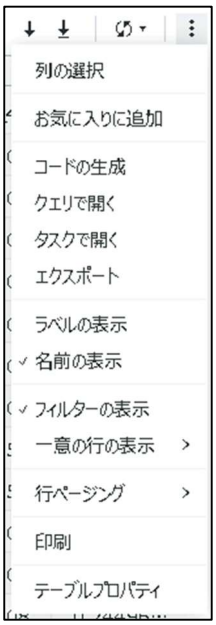
3. 出力データタブで、WORK ライブラリの、BANK5 データをクリックして表示し確認します。



	std_dmy_2	std_dmy_3	std_dmy_4	std_dmy_5	std_dmy_6	std_dmy_7
1	-0.470136153	-0.078320359	-0.366052419	-0.181200825	-0.181200825	-0.268735287
2	2.1267345936	-0.078320359	-0.366052419	-0.181200825	-0.181200825	-0.268735287
3	-0.470136153	-0.078320359	-0.366052419	5.5179374557	-0.181200825	-0.268735287
4	2.1267345936	-0.078320359	-0.366052419	-0.181200825	-0.181200825	-0.268735287
5	-0.470136153	-0.078320359	-0.366052419	-0.181200825	-0.181200825	-0.268735287
6	-0.470136153	-0.078320359	2.7314525697	-0.181200825	-0.181200825	-0.268735287
7	-0.470136153	-0.078320359	2.7314525697	-0.181200825	-0.181200825	-0.268735287
8	-0.470136153	-0.078320359	-0.366052419	-0.181200825	-0.181200825	-0.268735287
9	-0.470136153	-0.078320359	-0.366052419	-0.181200825	-0.181200825	-0.268735287
10	-0.470136153	-0.078320359	-0.366052419	-0.181200825	-0.181200825	-0.268735287
11	-0.470136153	-0.078320359	-0.366052419	-0.181200825	-0.181200825	-0.268735287

名前の接頭辞に std_を持つ、標準化された値が割り当てられた変数が出来上がっていることが確認できます。

4. 出力データタブで、WORK ライブラリの、BANK5 データをクリックします。右上端の詳細オプション→テーブルプロパティを選択し、WORK ライブラリの BANK6 データの列プロパティを表示します。



列名	ラベル	種類	長さ	出力…	入力…
deposit	定期…	文字	9	$9.	
std_dmy_1	job サ…	数値	8		
std_dmy_2	job ブ…	数値	8		
std_dmy_3	job …	数値	8		
std_dmy_4	job …	数値	8		
std_dmy_5	job …	数値	8		
std_dmy_6	job …	数値	8		
std_dmy_7	job …	数値	8		
std_dmy_8	job …	数値	8		

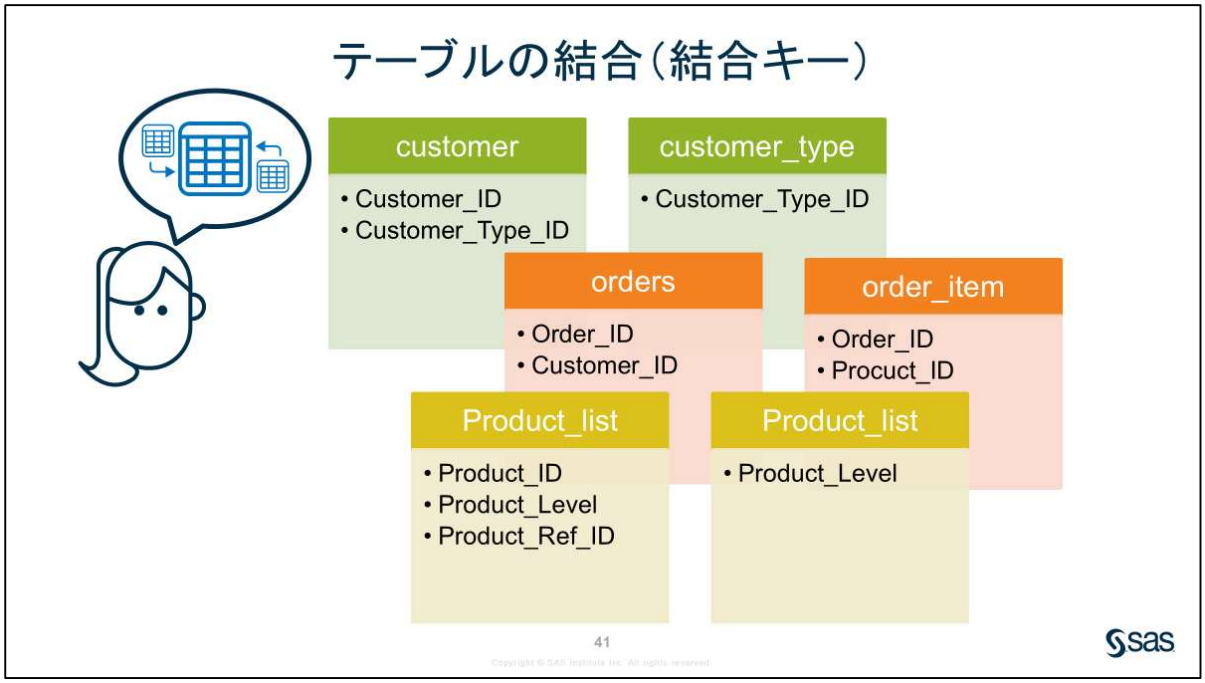
ターゲットの変数 deposit と、標準化された変数のみが保存されていることを確認できます。

End of Demonstration

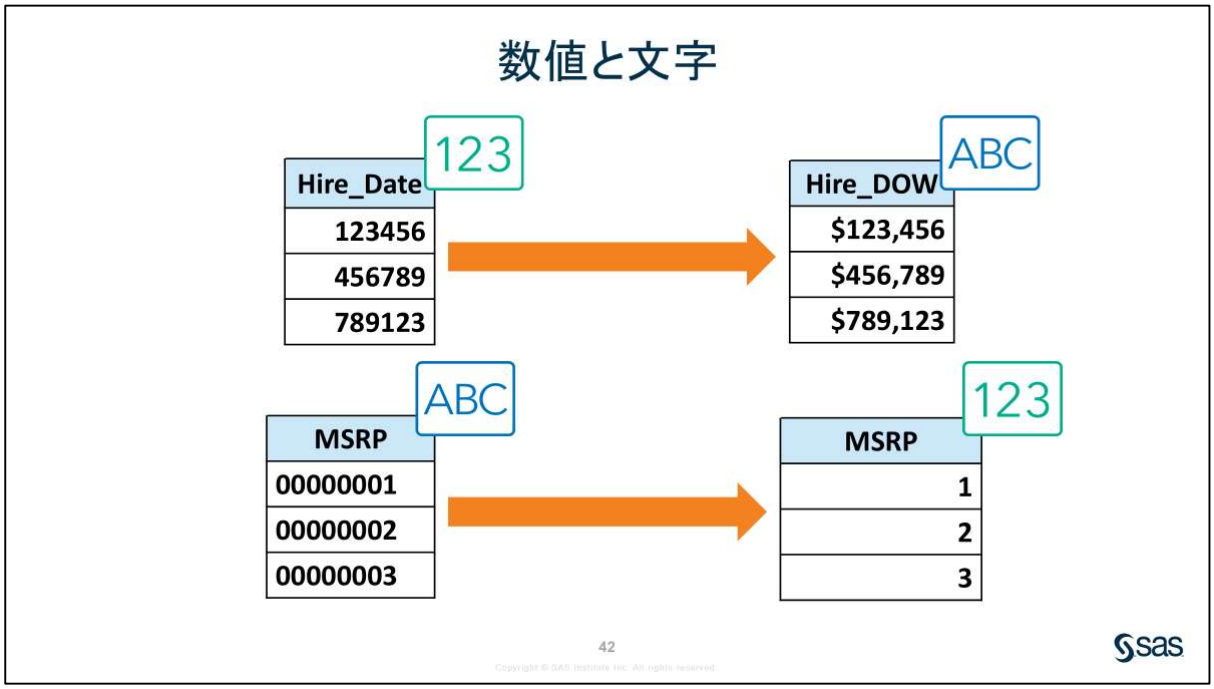
分析データの作成には、他に、データの結合、欠損値の処理があります。

データの結合

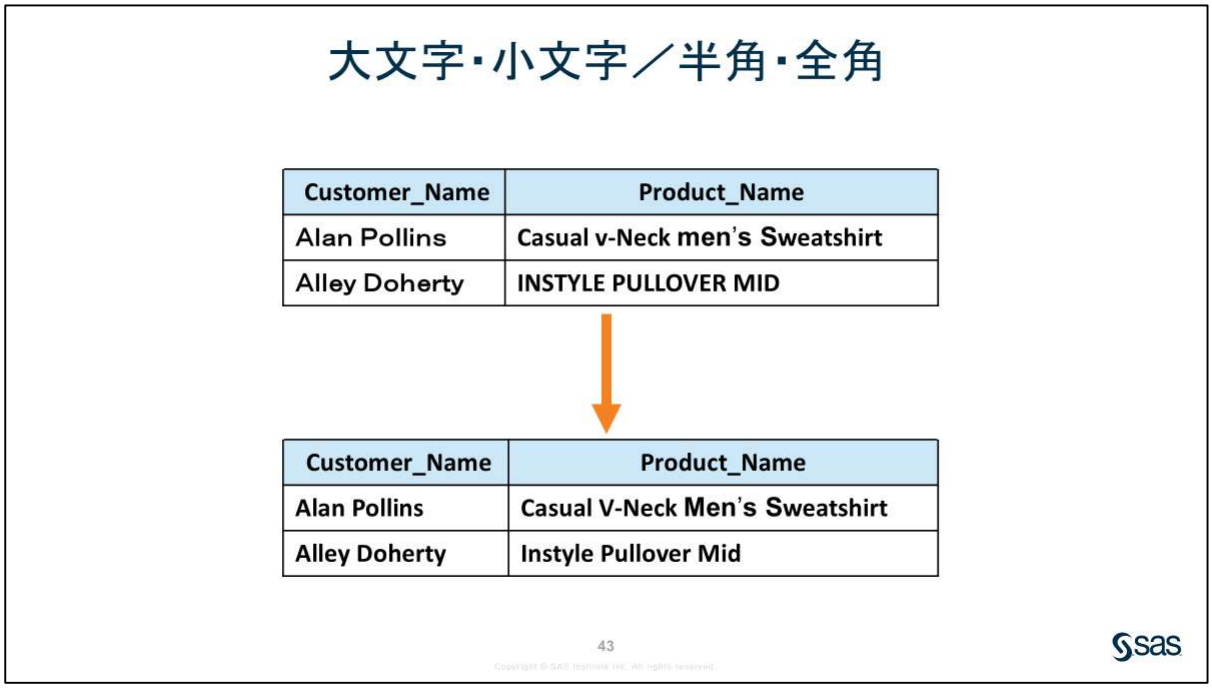
データ分析において、必要なデータが複数の場所に散在している場合に、これらを統合する作業が必要です。例えば、顧客に関する情報（顧客 ID、名前、性別、年齢、学歴、収入、住所など）は「顧客マスタ」、商品に関する情報（商品 ID、商品コード、商品名、商品カテゴリ、サイズ、形状、カラーなど）は「商品マスタ」、販売に関する情報（取引日、商品コード、顧客コード、販売員コード、販売価格、販売個数など）は「トランザクショナルデータ」として、それぞれ別々のデータセットとして蓄積・管理されていることが多いです。これらを分析の際に、ユニークなキー（例：顧客 ID、商品 ID など）を用いて結合し、必要な情報を一つにまとめます。



テーブル同士を結合するために使用する、キーとなる列を特定しましょう。



結合キーとして使用する列の型が、片方のテーブルで文字、もう一方で数値の場合、結合キーとして利用することが出来ません。また、文字のままでは計算に利用できない場合や、数値のままでは値の記号などの修飾文字が挿入できない場合には、数値から文字もしくは文字から数値への型変換を行います。



結合キーとして使用する列の文字列が、片方のテーブルで大文字（全角）、もう一方で小文字（半角）の場合、結合キーとして利用することが出来ません。また、名寄せやカテゴライズを行う場合にも、同じ水準の値として扱うために文字列変換を行います。

欠損や重複

Employee ID	Start Date	Gender	Department
120101	01JUL2007	M	Sales
120102	01JUN1993		Sales
	01JAN1985	F	Marketing
120104	01JAN1999	F	Legal

Employee ID	Employee Name	State	Country
120144	Abbott, Ray	FL	US
120144	Abbott, Ray	FL	US
120761	Akinfolarin, Tameaka	PA	US

44

結合キーとして使用する列に、欠損や重複がある場合には、期待した結果が得られない可能性がありますので注意が必要です。また、通常のデータ値としての欠損が、何等かの理由により発生している場合は、分析の前にその取扱いを考慮する必要があるかもしれません。重複行の値が、すべて完全に一致しているような場合には、その行は削除すべきかもしれません。

テーブルの結合（カーディナリティ）

1

X	Y
1	a
2	b

2

X	Z
1	f
2	g

1対1マッチ

X	Y	Z
1	a	f
2	b	g

1

X	Y
1	a
2	b

2

X	Z
1	f
1	r
2	g

1対多（多対1）マッチ

X	Y	Z
1	a	f
1	a	r
2	b	g

1

X	Y
1	a
1	c
2	b

2

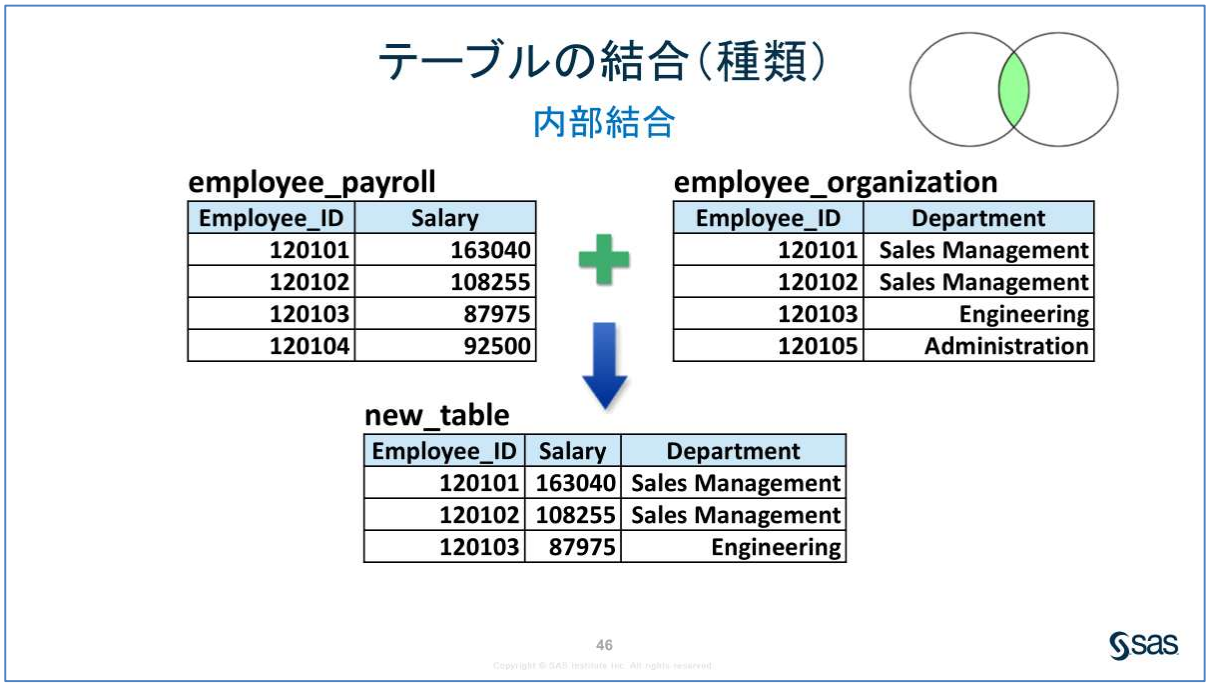
X	Z
1	f
1	r
2	g

多対多マッチ

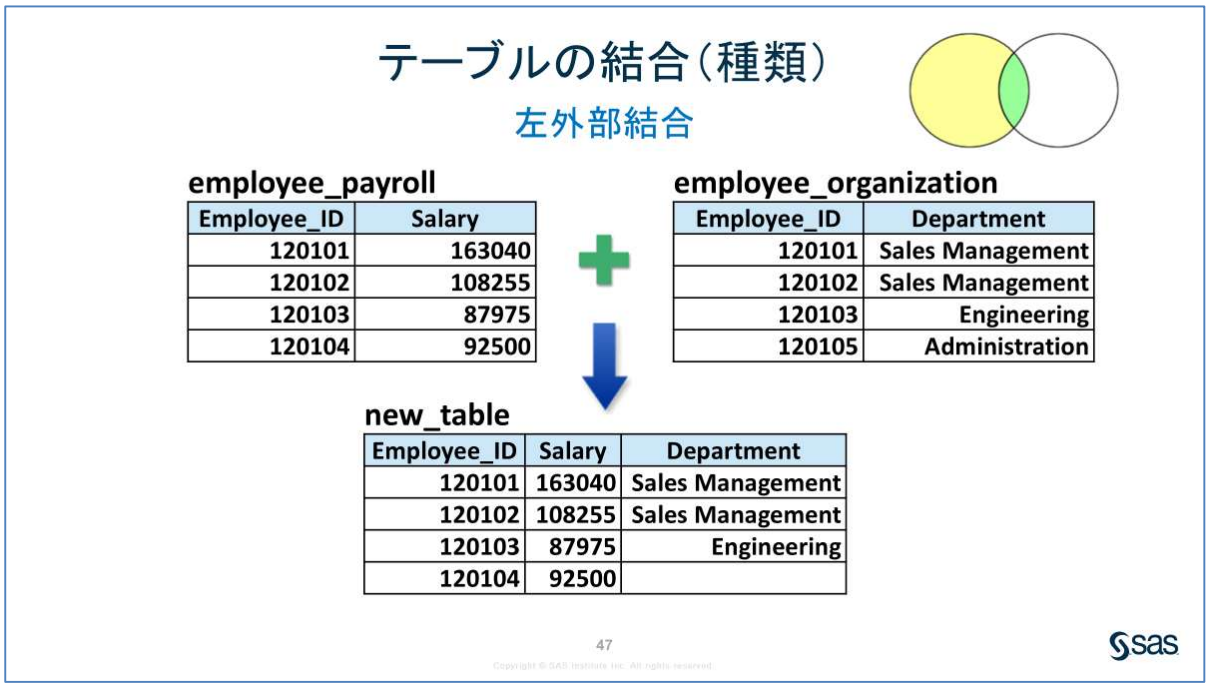
X	Y	Z
1	a	f
1	a	r
1	c	f
1	c	r
2	b	g

6

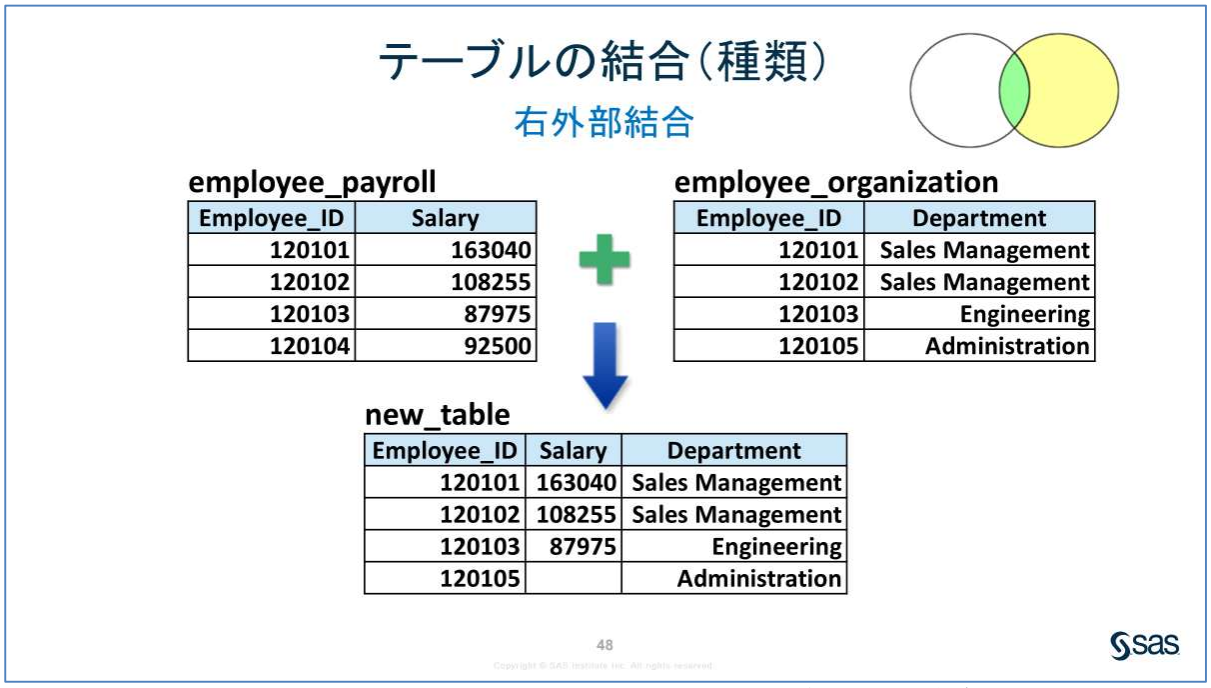
結合に使用するキー列の、テーブル間における関係を理解しておきましょう。基本的にテーブルを水平方向へ結合する場合、多対多の関係を持つテーブル同士は結合しないようにします。多対多の関係を持つテーブルを結合する場合には、一対一、一対多の関係になるように、片方もしくは両方のデータの構造を変更したり、データを抽出したり、データを要約したりします。



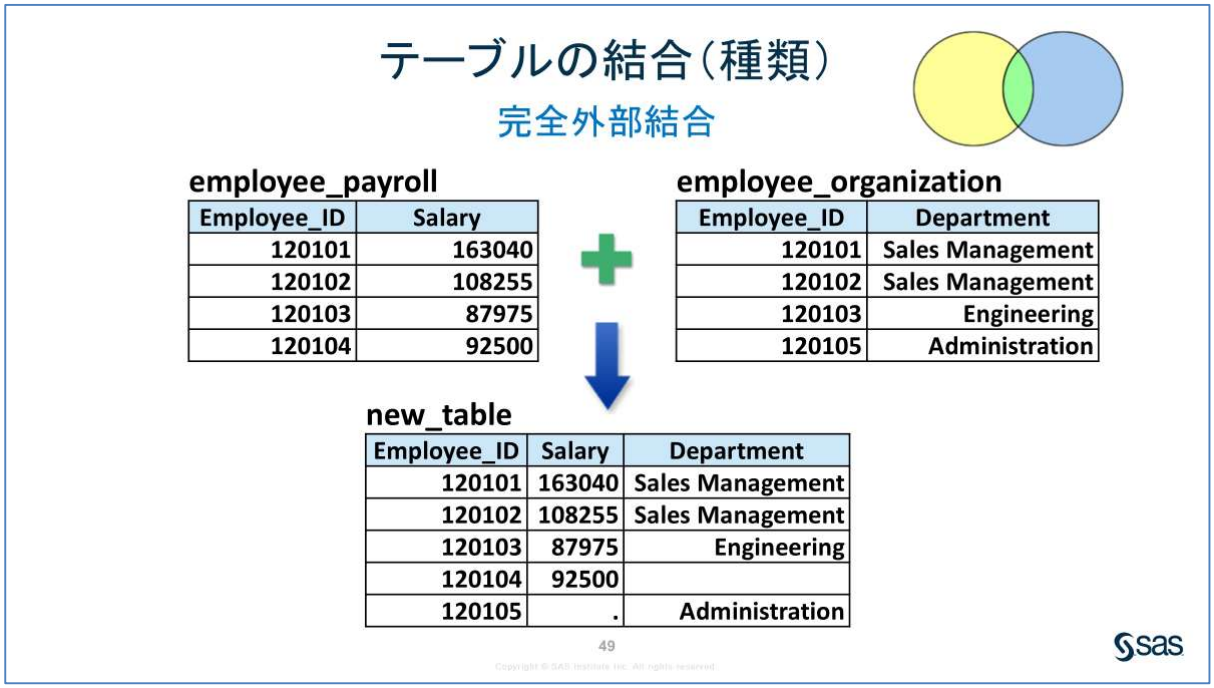
内部結合は、キー列の値で一致する行のみを出力します。



左外部結合は、キー列の値で一致する行に加えて、左側のテーブルにだけ存在するキー列の値を持つ行を出力します。



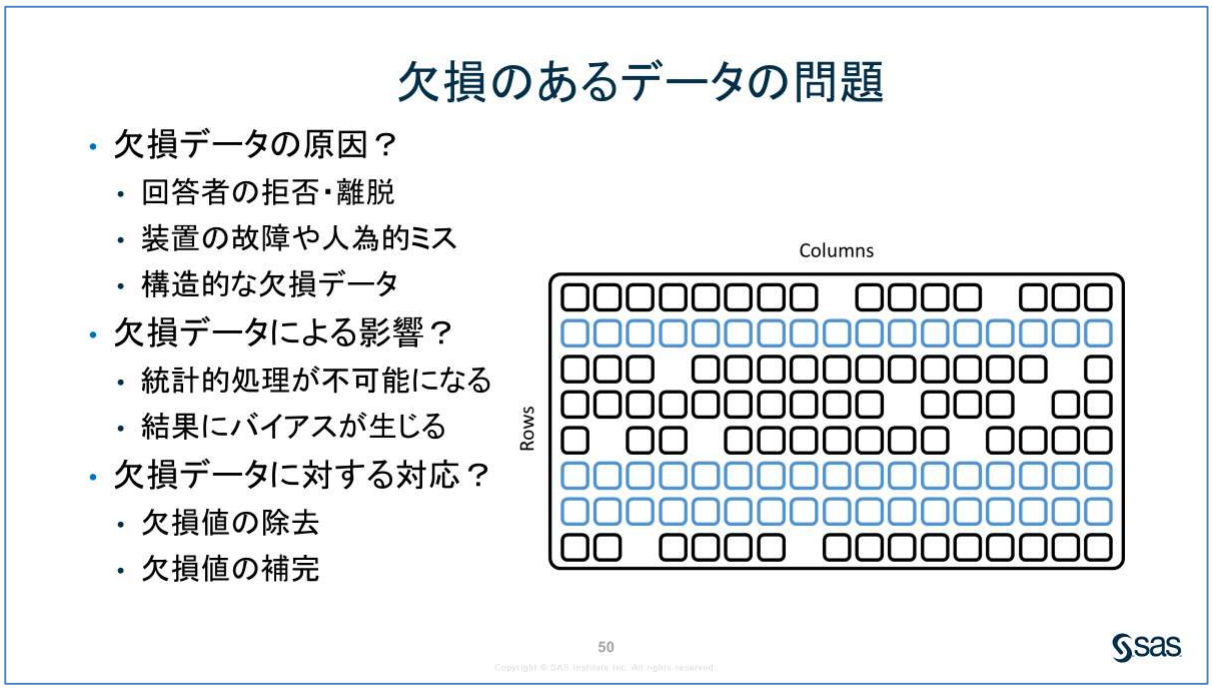
右外部結合は、キー列の値で一致する行に加えて、右側のテーブルにだけ存在するキー列の値を持つ行を出力します。



完全外部結合は、キー列の値で一致する行に加えて、左側・右側それぞれのテーブルにだけ存在するキー列の値を持つ行を含め、すべての行を完全に出力します。

欠損値の処理

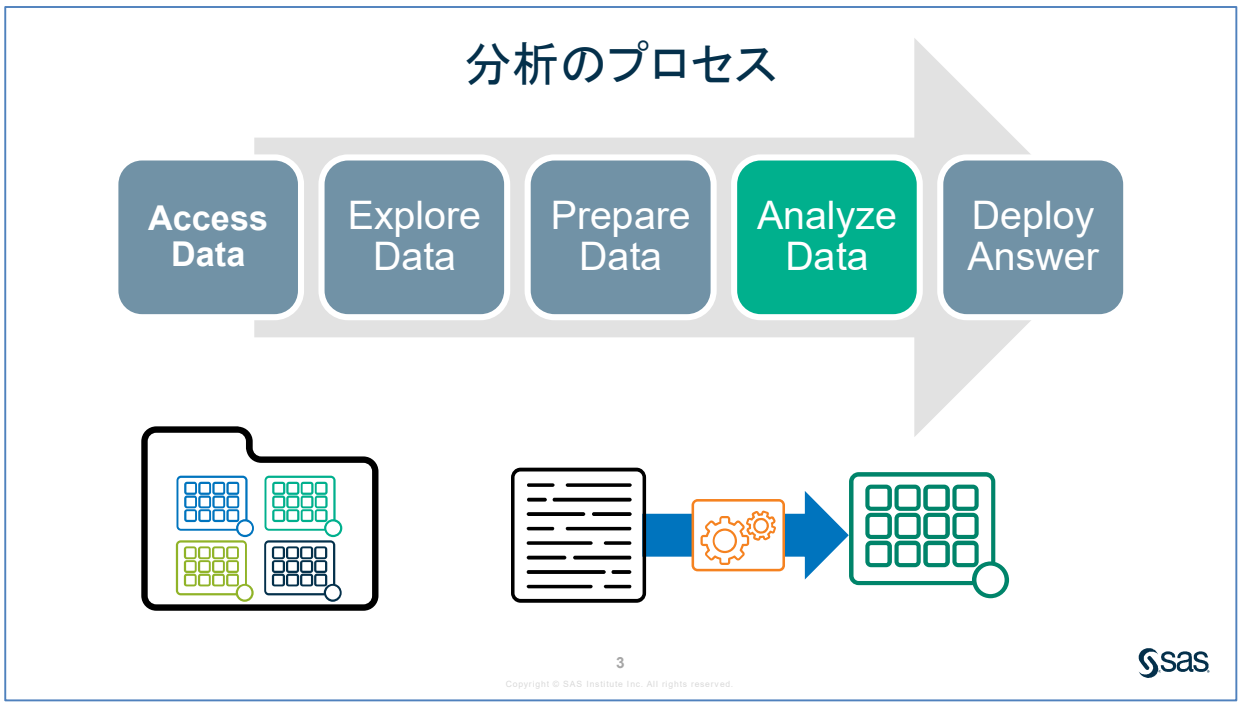
欠損値とは、何らかの理由で観測されなかったデータのことです。分析を行う際には、欠損が発生した原因やその影響を考慮し、適切な対応を取る必要があります。欠損値を他のデータから推定して補うことを「欠損値の補完」と言います。この補完方法としては、他の関連データから導き出された値を使う方法や、分析者が事前に設定した値で置き換える方法があります。



様々な理由で、データには欠損値が含まれます。欠損値の問題は（ほぼ）常に存在し、常に懸念事項となります。
欠損値に直面した場合、欠損値がランダムに分布しているかどうか、または欠損値が何らかの形で予測できるかどうかを考慮する必要があります。 欠損値が入力データにランダムに現れる場合、欠落値を含む行はモデルにバイアスを生じることなく分析から削除できます。 ただ、いくつかの分析手法では完全なケースが利用されるため、わずかな欠損値の存在が、膨大な量のデータの損失につながり、モデルの予測精度が低下する可能性があります。
そのため、欠損している値を、データの欠損していない値から導出された情報に、補完（置き換え）することが考えられます。一つの方法として、入力変数の欠損値を、その変数の非欠損値の平均や最頻値に置き換えることができます。また、統計的に適切なモデルを使用して、値を導出するような方法もあります。

Lesson 3　分析モデルの構築

3.1 本章の学習目標



Analyze Data：分析モデルを生成して評価を行います。

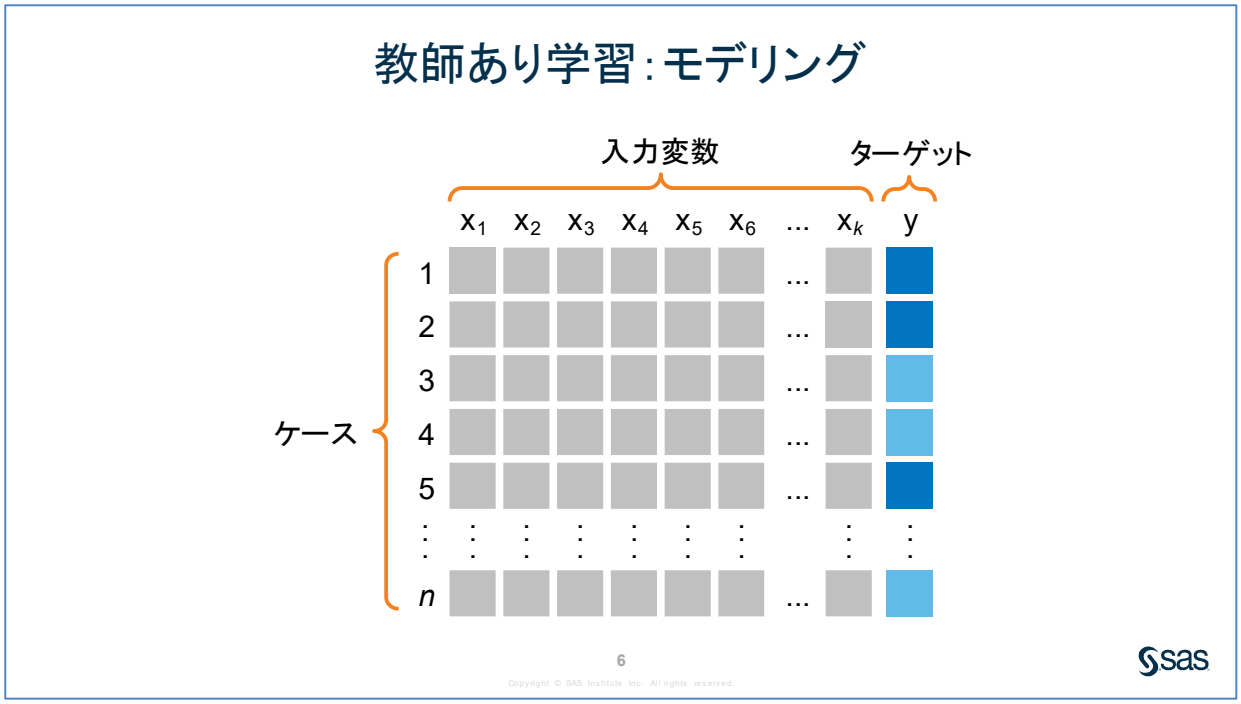
この章では、データの分析を行い、ビジネス課題に答えるためのヒントを探し出します。

様々な分析のアルゴリズムやテクニックを使用することができますが、その課題と目的にあった分析手法を選択する必要もあるでしょう。機械学習は大きく教師ありと教師なしに分けられ、実に様々な分析手法があります。今回は教師あり学習の手法の中でも基本的なロジスティック回帰分析と決定木分析を使用します。機械学習の分析手法についての簡単な紹介は Appendix をご参考ください。

分析のプロセスでは、以下の様な作業が想定されます：

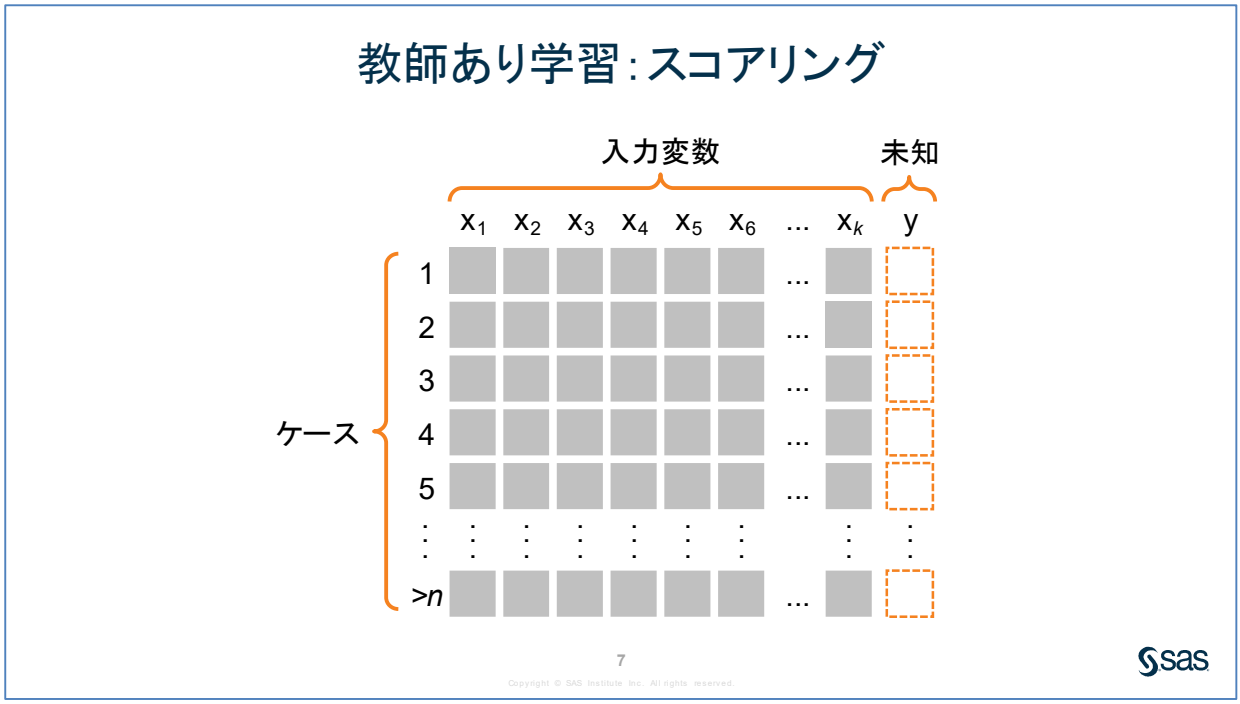
・分析モデル作成　　・分析モデル評価
・分析モデル精度向上　　・モデルの比較
・モデルの選択　　・その他・・・

上記の例は、分析モデル作成トピックの一部です。モデルの作成は複数の手法を用いて行い、それぞれのモデルのチューニングを行って、結果の比較を行いチャンピオンモデルの選択を行います。



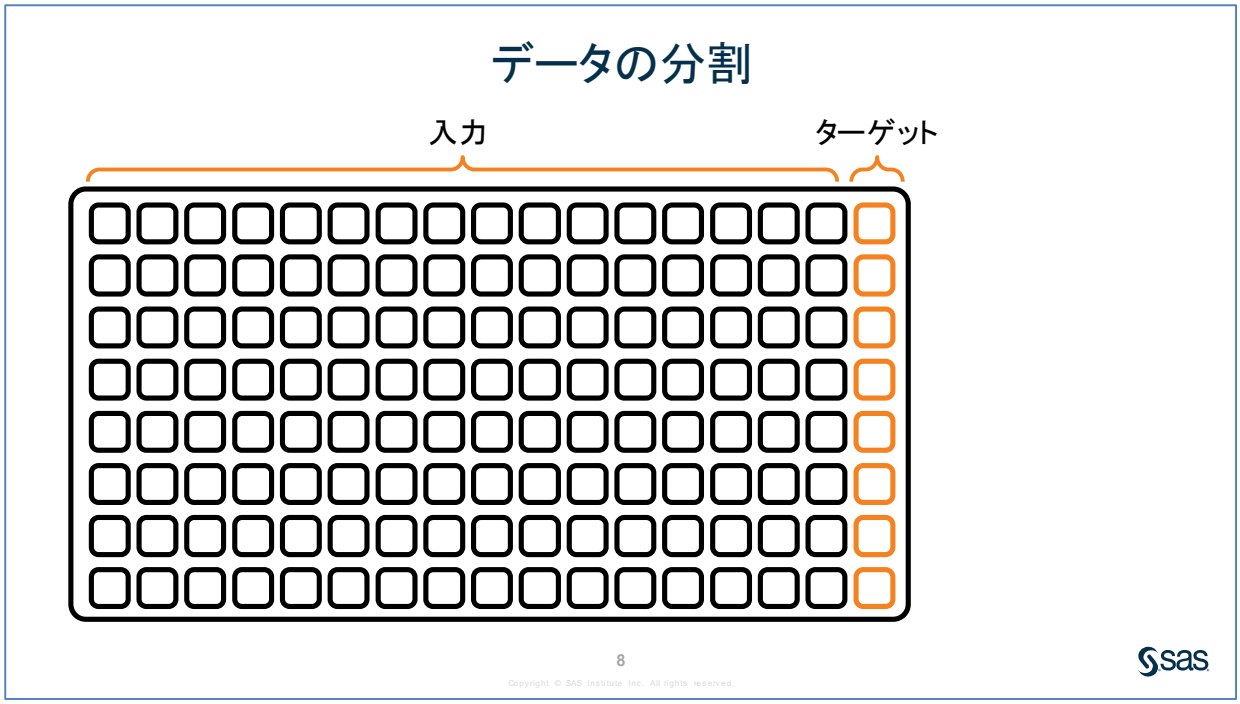
教師あり機械学習モデルの開発に使用されるデータは、観測値としても知られる一連のケース、行で構成されます。 各ケースには、予測変数、説明変数、独立変数、および特徴とも呼ばれる入力変数が列で構成されます。また、 各ケースには、応答変数、目的変数、従属変数とも呼ばれるターゲット変数が列で構成されます。 機械学習モデル作成（モデリング）は、入力変数のベクトル（列の並び）をターゲット変数に（式で）マッピングします。ケースは予測が行われる単位で、目標は予測される結果を求めることです。

ケースごとにターゲット変数の値（正解ラベル）がわかっている場合、機械学習モデルは教師ありと呼ばれます。しかし、ターゲット変数の値がわかっているならば、なぜ予測モデルを構築するのでしょうか。

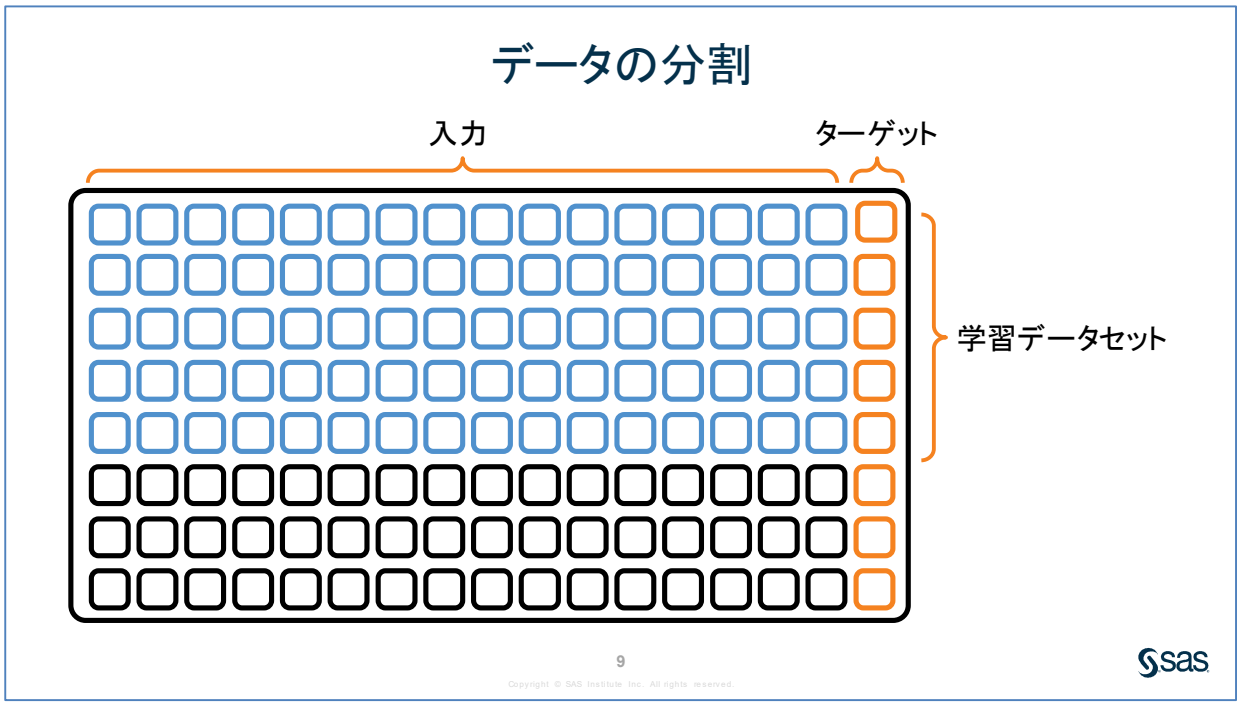


作成（構築）された教師あり機械学習モデルは、入力変数の値はわかっているが、ターゲット変数の値が不明（未知）な新しいケース（データ）で使用されます。 教師あり機械学習モデルの主な目的は、一般化、つまり新しいケースの結果を予測（スコアリング）することです。

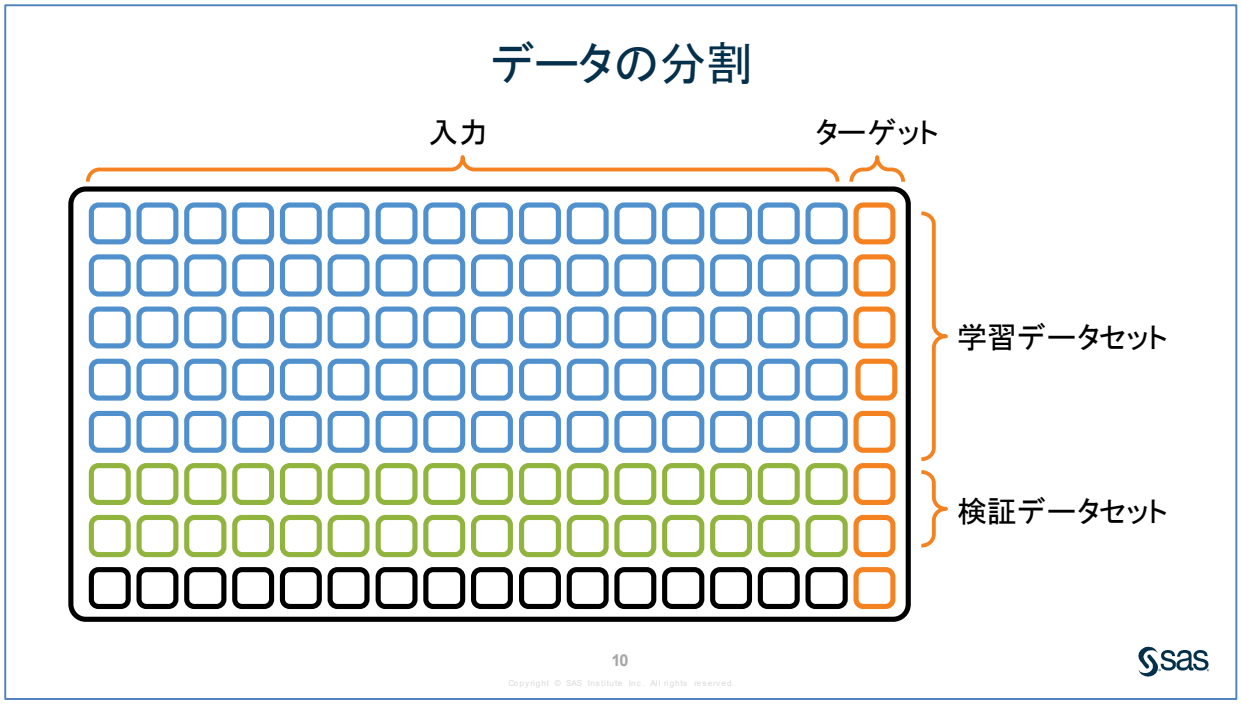
一般化はモデル評価にも関係しています。 モデルはトレーニング（学習）データに適合し、バリデーション（検証）データで予測値をターゲットの観測値と比較することにより、パフォーマンスが評価されます。



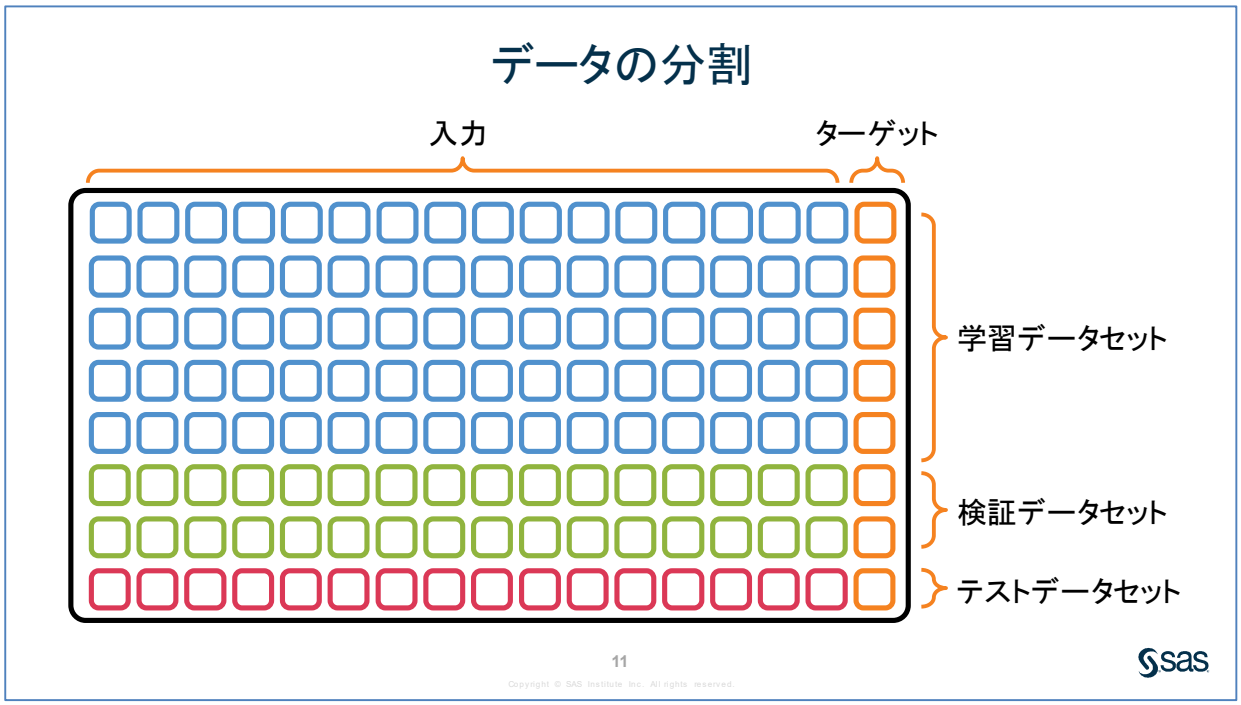
モデリングを行うためのデータがあるとします。モデルを構築する際には、その予測結果の精度を客観的に評価するために、データの分割を行います。



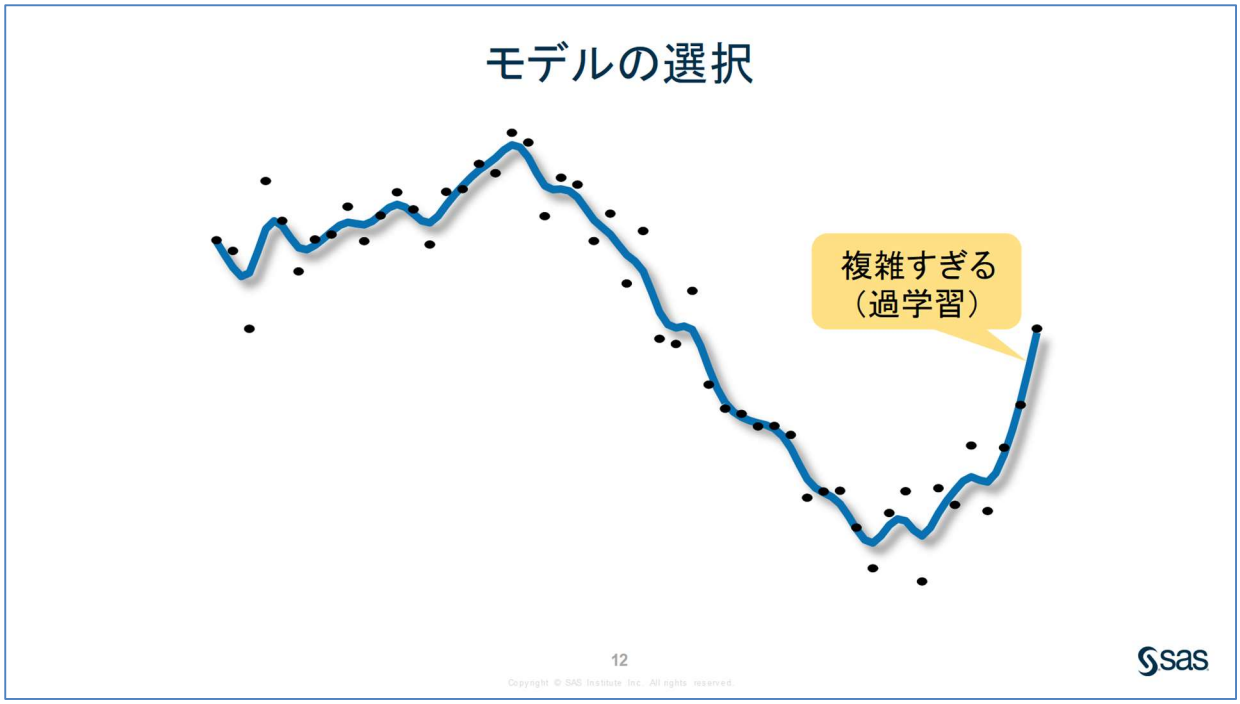
まず、学習データは、モデルを作成するために部分的に分割（選択）されたケースです。



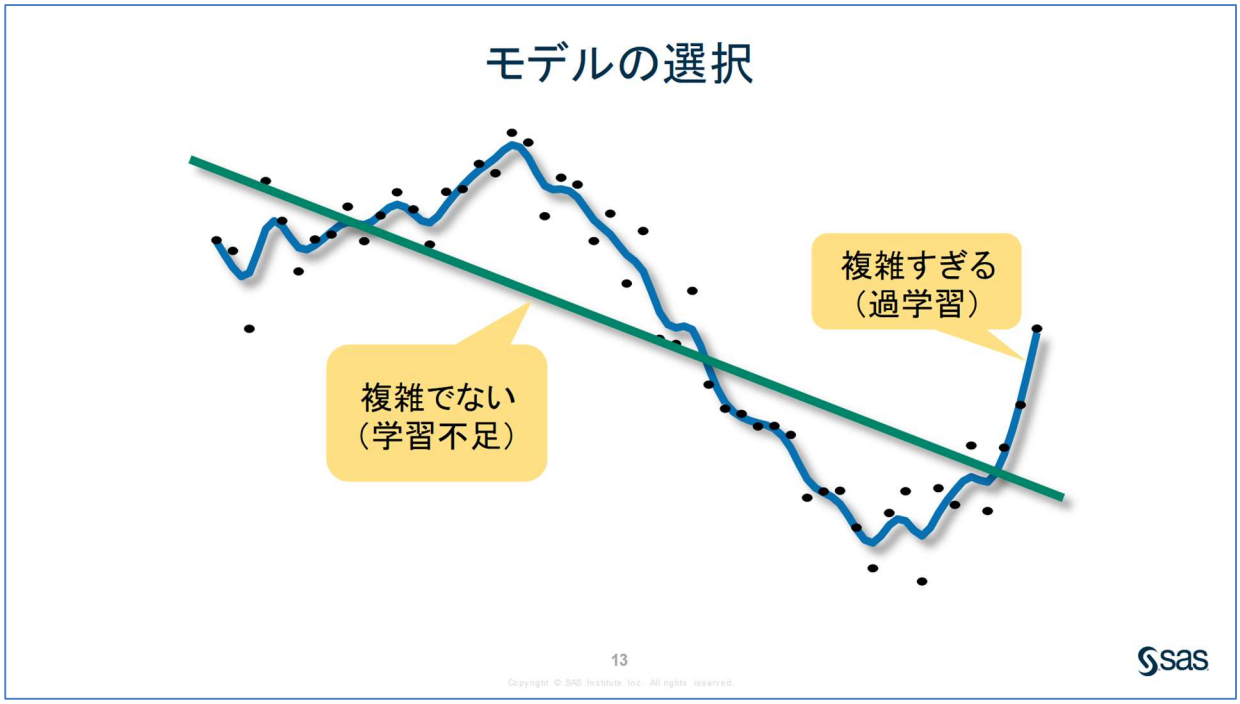
次に、検証データと呼ばれる別の部分（ケース）で、作成されたモデルのパフォーマンスが評価されます。また、 検証データはモデルの比較、選択、および変更に使用されます。



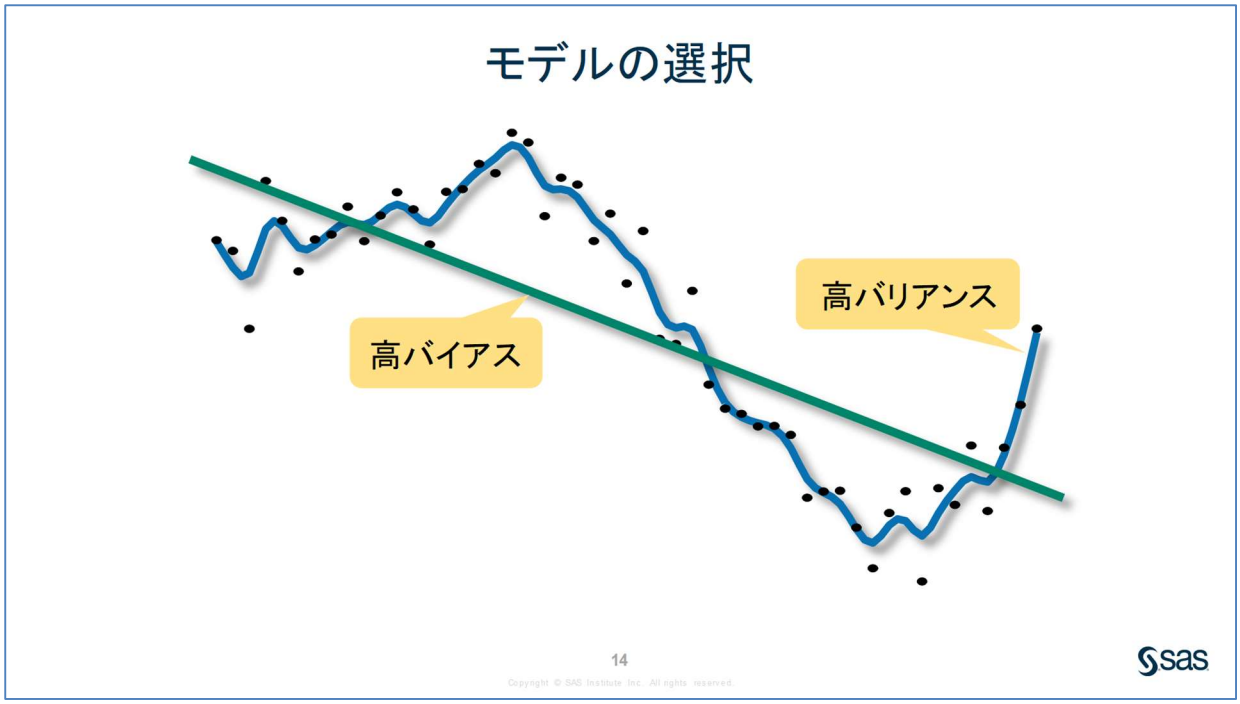
そして、テストデータを別途用意し、最終評価に使用する場合があります。



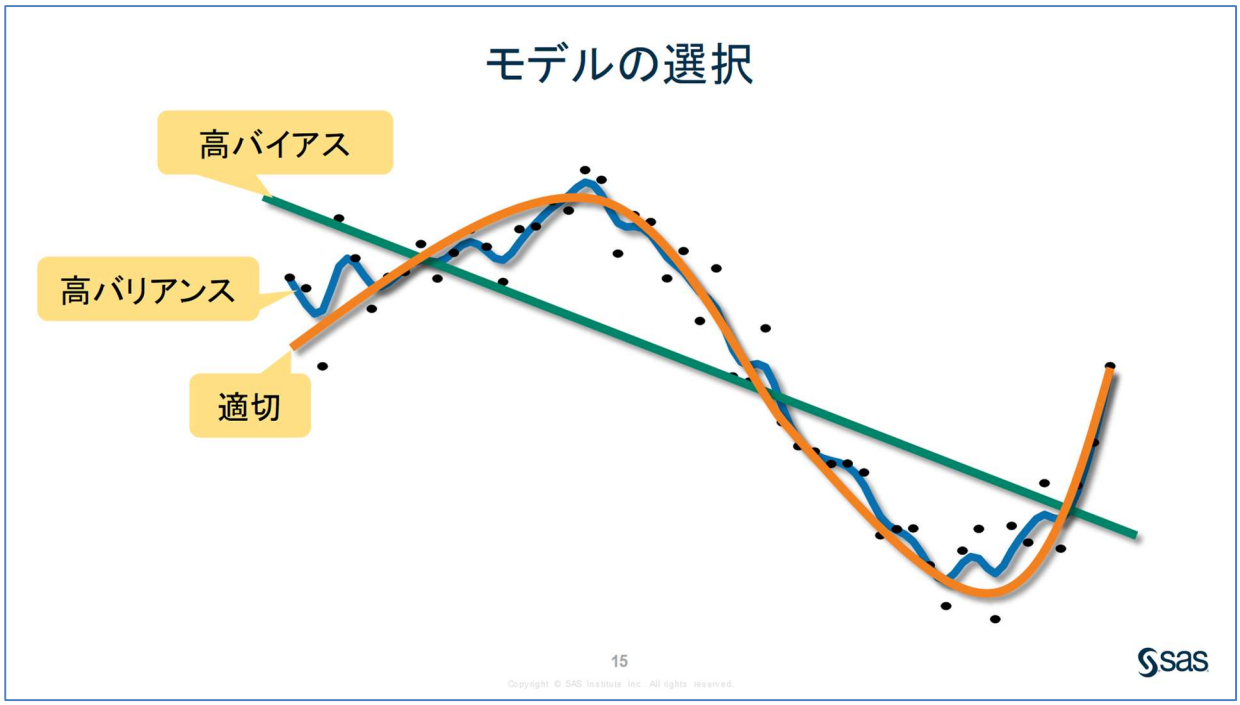
モデルはその複雑さに基づいて、いくつかのモデルから選択することができます。 よくある落とし穴は、選択したモデルが複雑すぎる場合に、（学習）データに過剰適合してしまうことです。 過度に複雑なモデルは、データのノイズに敏感すぎて、新しいデータにうまく一般化されない可能性があります。



一方で単純すぎるモデルを使用すると、適合性が低下する可能性があります。つまり、特徴がつかめていない状態です。

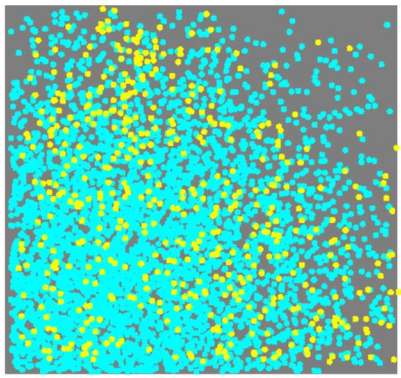


目標は、バイアス（偏り）が低くバリアンス（分散）が低いモデルを適合させることです。 バイアスの高いモデルでは、入力とターゲットの間の重要な関係が失われます。これは学習不足（アンダーフィッティング）の例です。バリアンスの高いモデルは、ノイズをモデルに組み込んでしまいます。これは過学習（オーバーフィッティング）の例です。



（学習）データ内の関係を正確に表すだけでなく、新しいデータにうまく一般化するモデルを選択するために、バイアスとバリアンスのトレードオフを考慮します。

イベントベースサンプリング



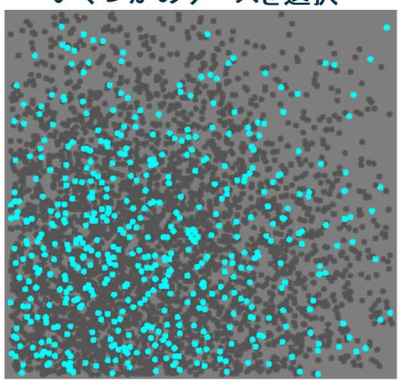
プライマリとセカンダリの結果
（二値の結果）

16

実際のビジネスの場面では、関心のあるイベントが稀な（非常に少ない）ケースであることが多くあります。 例えば、ある事業における解約率は 1～3%の範囲である可能性があります。

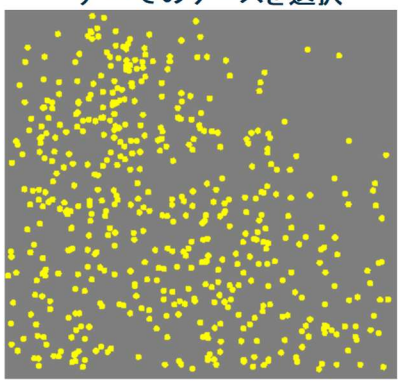
イベントベースサンプリング

いくつかのケースを選択



セカンダリ結果

すべてのケースを選択

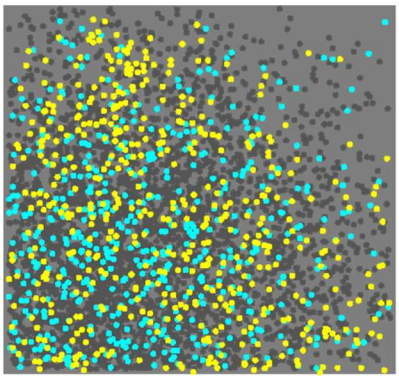


プライマリ結果

17

このように結果の割合に大きな偏りのある不均衡データでは、関心のある稀なイベントを予測するために、関心のあるイベントのすべてと、非イベントをサンプリングしたデータに基づいてモデルを構築することです。 イベントベースのサンプリングの利点は、関心のあるイベントのケース数が少ないデータと、同様の予測力のモデルを平均的に得られることです。

イベントベースサンプリング



セカンダリとプライマリの結果

18

このスライドは、一つ前のスライドのプライマリ結果と、サンプリング後のセカンダリ結果のデータを重ね合わせたイメージです。

本章の操作シナリオ

本章では、ロジスティック回帰分析と決定木分析、それぞれの実行を行うため、以下の様な分析ツールの操作を行います。

3.2 ロジスティック回帰分析
①ロジスティック回帰ノードの設定と実行
3.3 決定木分析
②ディシジョンツリーノードの設定と実行

4

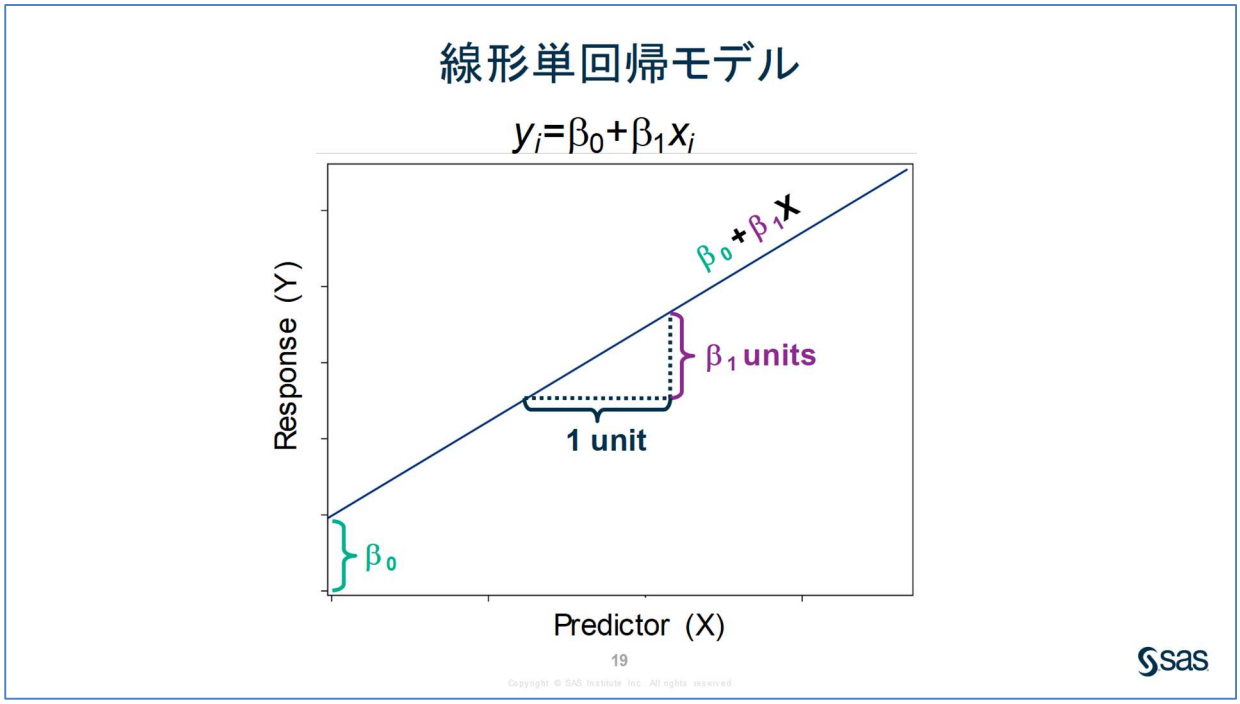
3.2 ロジスティック回帰分析

①ロジスティック回帰ノードの設定と実行
SAS Model Studio のパイプラインで、ロジスティック回帰ノードを設定して実行し、結果を確認します。

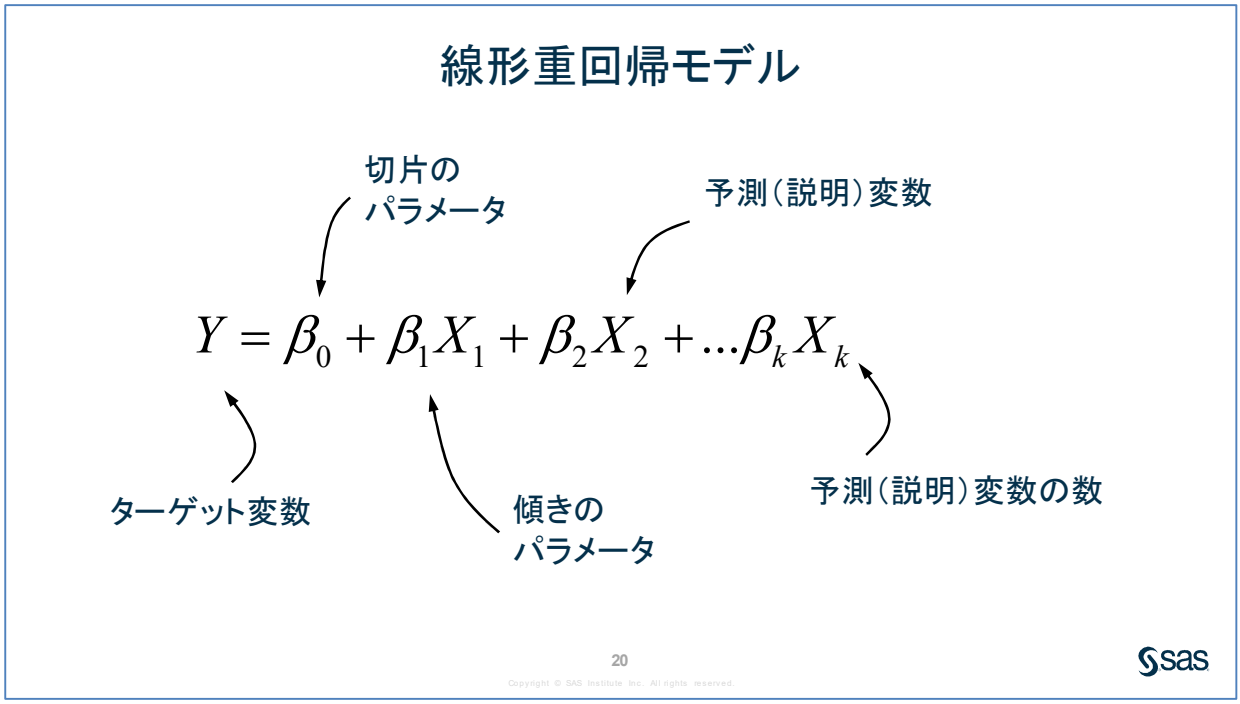
3.3 決定木分析

②ディシジョンツリーノードの設定と実行
SAS Model Studio のパイプラインで、ディシジョンツリーノードを設定して実行し、結果を確認します。

3.2 ロジスティック回帰分析



ターゲット（応答）変数と予測（説明）変数の関係は、方程式 $y_i=\beta_0+\beta_1 x_i$（i = 1 … n）で特徴付けることができます。



線形単回帰とは異なり、多重線形重回帰を使用すると、ターゲット変数（Y）と複数の予測変数（X_1、X_2、…X_k）の間の関係を同時に調査できます。重回帰の傾きのパラメータ（β_1、β_2、…β_k）は、モデル内の他のすべての変数を制御しながら、応答に対する 1 つの変数の影響を表します。

ロジスティック回帰モデル

事後確率

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}$$

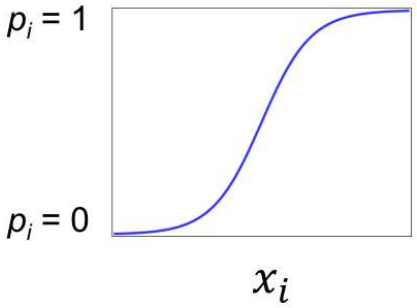
傾きのパラメータ　予測変数

21

ロジスティック回帰モデルは、事後確率（一連の予測変数値が与えられた場合の予測確率）のロジットが予測変数の線形結合であると想定しています。 パラメータβ_0、…、β_kは未知の定数であり、データから推定する必要があります。 傾きのパラメータは、予測変数が 1 単位増加した場合のロジットの変化を示します。

ロジットリンク関数

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} \Leftrightarrow \quad p_i = \frac{1}{1+e^{-(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki})}}$$



22

ロジスティック回帰モデルは、確率にロジット変換を適用します。 その理由は、線形結合（$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$）は任意の値を取ることができるためです。 ただし、確率は 0 から 1 の間でなければなりません。 ロジット変換（オッズの対数である $\ln(p_i / (1-p_i))$ ）は、確率スケールを実数直線（-∞、+∞）に変換します。 したがって、ロジットは線形結合でモデル化できます。

プロットは、予測子と結果の確率との関係のモデルを示しています。 この関係を直接モデル化するには、非線形関数 $1 / (1 + \exp\{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)\}$を使用する必要があります。 この関数は、0 から 1 の間に制限されている推定確率を提供します。

本章の操作シナリオ

本章では、ロジスティック回帰分析と決定木分析、それぞれの実行を行うため、以下の様な分析ツールの操作を行います。

3.2 ロジスティック回帰分析

①ロジスティック回帰ノードの設定と実行

3.3 決定木分析

②ディシジョンツリーノードの設定と実行

23

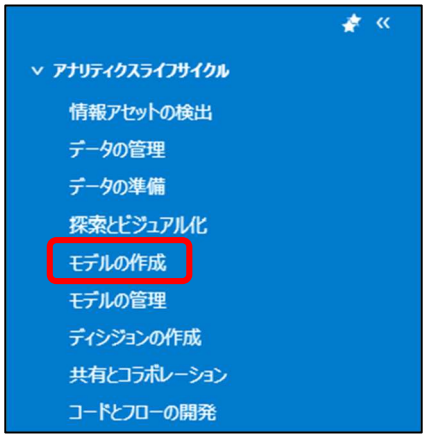
実際に分析ツールを使用して、ロジスティック回帰分析を利用してみましょう。

ロジスティック回帰分析

このデモでは、Model Studio を使用して、ロジスティック回帰ノードを設定し実行して、結果を確認します。

ここまでは、分析の基礎となるプロセスを学習する目的で、SAS Studio 上でプログラムを利用して、データの前処理を行ってきました。しかしながら、実は Model Studio では、カテゴリ変数のダミー化などの処理を自動的に行ってくれます。ここからは、Model Studio を使用して、実際に分析を行ってみましょう。

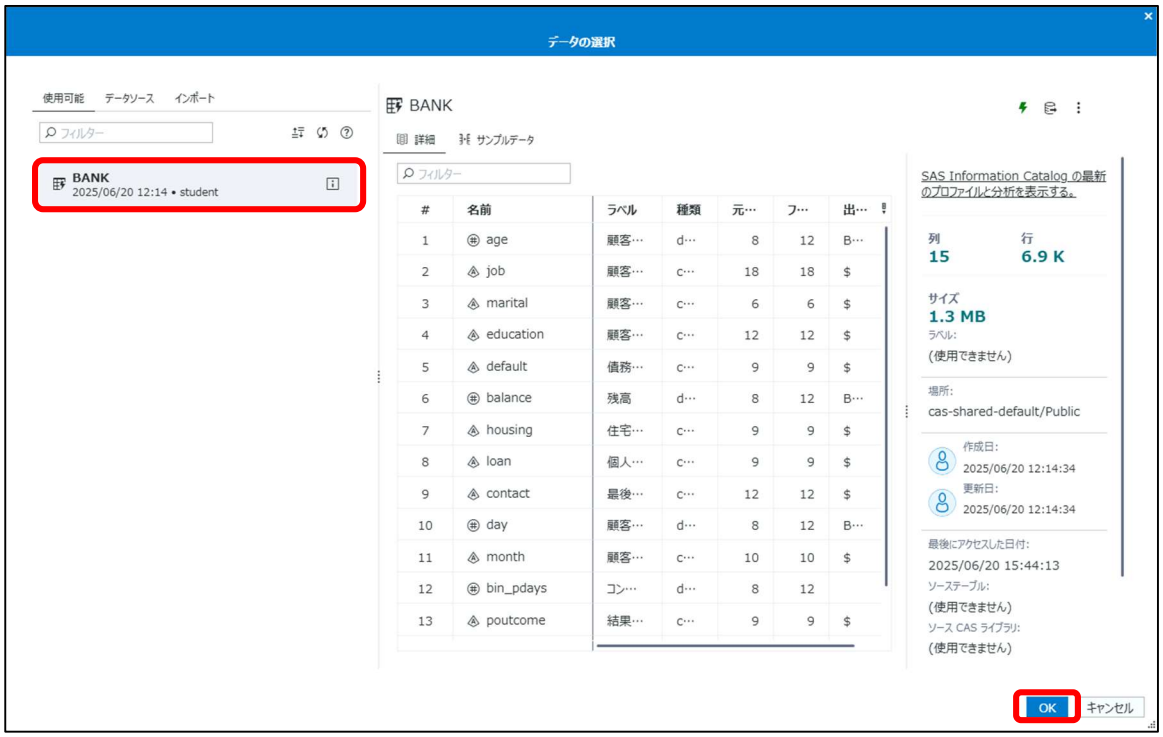
1. Model Studio を起動するために、画面左上の 三本線のボタンをクリックして、表示されたメニューから、**モデルの作成**を選択します。



2. プロジェクトの新規作成ボタンをクリックします。

3. 名前に Bank、データのセクションの参照ボタンをクリックします。

使用可能タブに、Lesson 2 で準備した Bank テーブルがあるので、そちらを選択し、OK をクリックします。



4. 保存ボタンをクリックします。

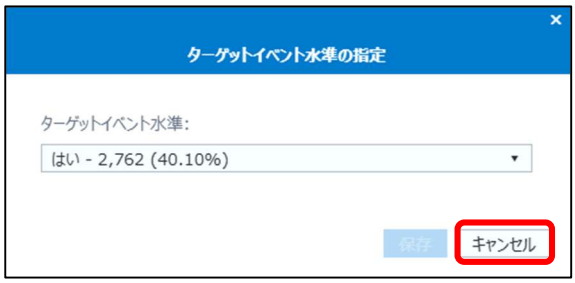
Bank プロジェクトが開き、データタブが表示されます。

※こちらの画面からでも変数の属性、データの行数、列数を確認できます。

5. 分析処理フロー(パイプライン)の実行に、ターゲットとなる目的変数の設定が必要なため、変数 deposit にチェックを入れます。

6. 画面右手で、deposit の役割にターゲットを選択します。

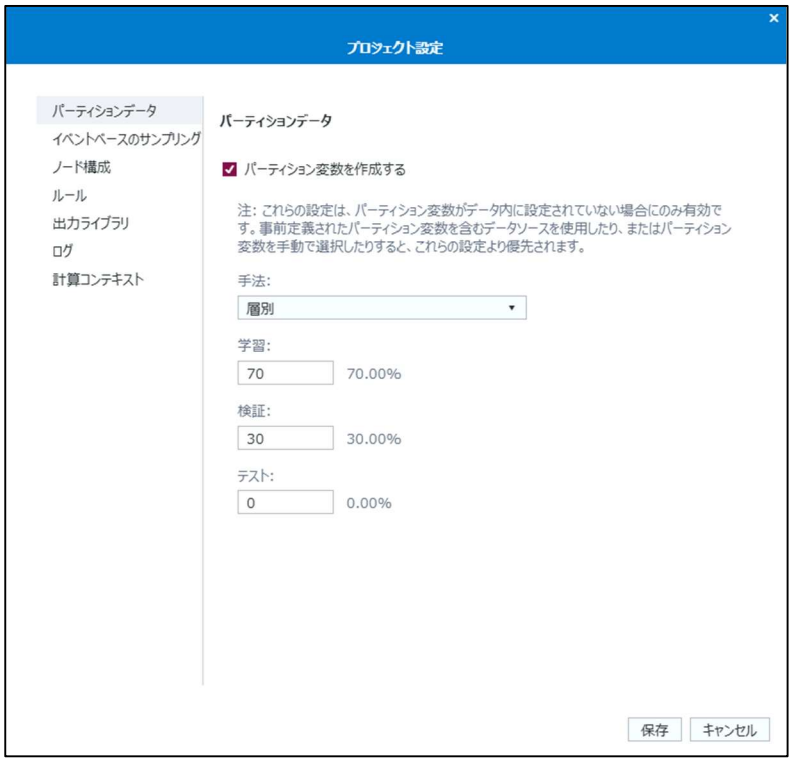
7. ターゲットイベント水準の指定を選択します。ターゲットイベント水準の指定ウィンドウで、「はい」が選択されていることを確認し、キャンセルをクリックします。



8. データタブで、campaign にチェックを入れ、画面右手から、役割を**リジェクト**に変更します

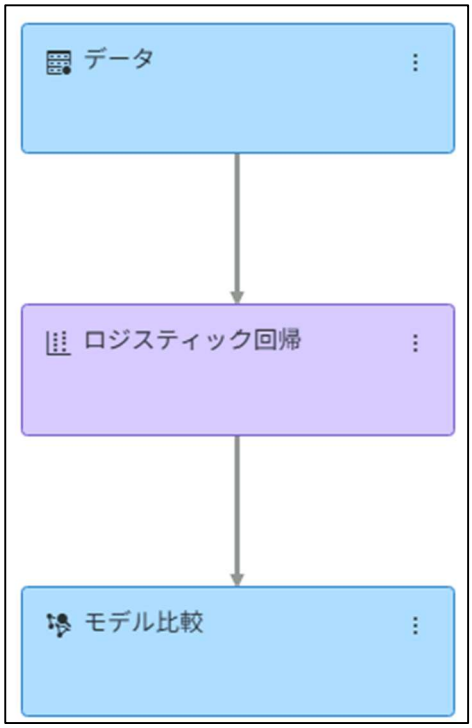
9. 画面右上の歯車のボタンをクリックし、プロジェクト設定を選択します。

10. パーティションデータで、学習 70、検証 30、テスト 0 に設定し、保存をクリックします。



11. パイプランタブを選択します。画面左手のノード( )をクリックして、「教師あり学習」を展開します。

12. ロジスティック回帰を、パイプラインタブの「データ」ノードにドラッグアンドドロップします。

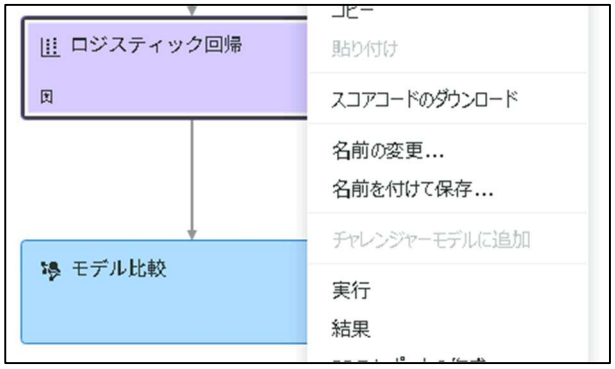


設定した分析用のノードはまだ一つですが、分析のノードを設定すると、パイプラインには必ず「モデルの比較」ノードが付いてきます。

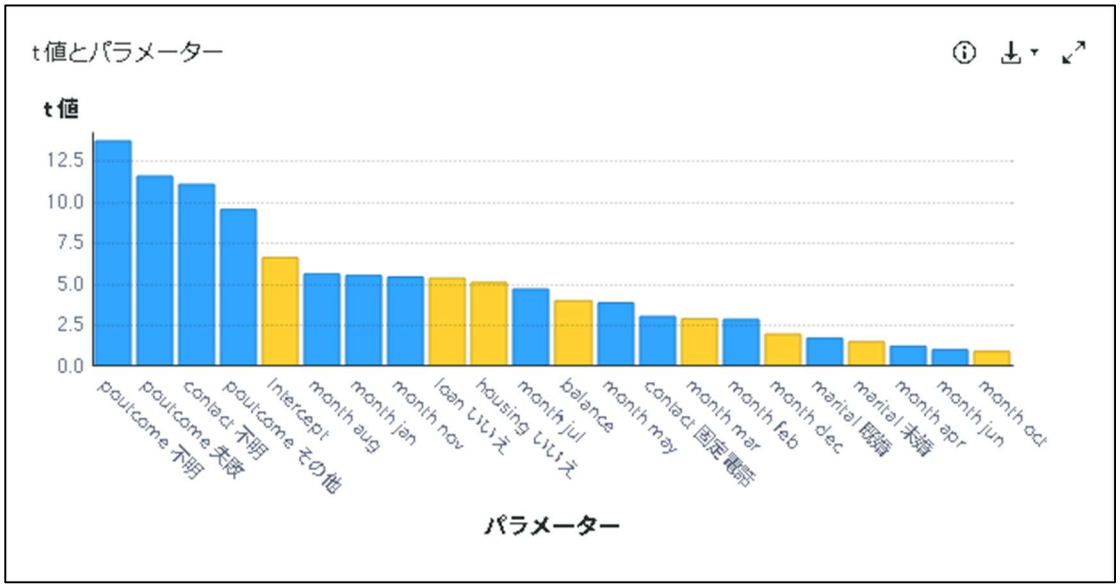
13. 画面右上から、**パイプラインの実行**をクリックします。

実行が正常に完了すると、ノードの右下に緑のチェックマークが現れます。

14. 「ロジスティック回帰」ノードを右クリックして、**結果**を選択します。



15. 「t 値とパラメータ」ウィンドウを確認します。



このプロットは、ロジスティック回帰モデルの各パラメータ推定値の t 値の絶対値を表示します。 値が大きいほど、より重要なパラメータであることを示します。 パラメータを表すバーは、推定値の符号によって色分けされています。 正（+）で色付けされたバーは、正のパラメータ推定値に対応します。これは、パラメータ値が増加するにつれて、イベントの予測確率が増加することを示します。 負（-）で色付けされたバーは、負のパラメータ推定値に対応します。これは、パラメータ値が増加するにつれて、イベントの予測確率が減少することを示します。

16. 「パラメータ推定値」ウィンドウを確認します。


パラメーター推定値

効果	パラメ...	t値	符号	推定	絶対推定	標
poutcome	poutcome 不明	13.7761	-	-2.2751	2.2751	(
poutcome	poutcome 失敗	11.6234	-	-2.1777	2.1777	(
contact	contact 不明	11.1191	-	-1.3176	1.3176	(
poutcome	poutcome その他	9.5836	-	-2.0717	2.0717	(
Intercept	Intercept	6.6571	+	2.0600	2.0600	(
month	month	5.6464		1.3276	1.3276	(

モデルパラメータの推定値を表示します。

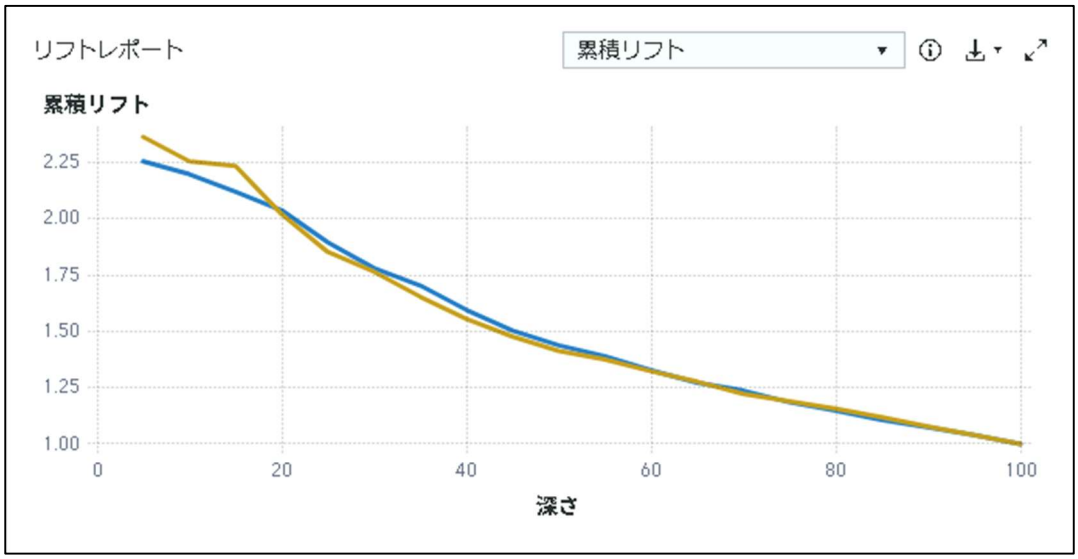
17. 「選択要約」ウィンドウを確認します。

選択要約

ステップ	入力された...	効果数	SBC	最適 SBC
0	Intercept	1	6,501.3712	0
1	poutcome	2	6,039.3041	0
2	month	3	5,763.7132	0
3	contact	4	5,626.0607	0
4	loan	5	5,596.7608	0
5	housing	6	5,573.5183	0
6	balance	7	5,567.1559	0
7	marital	8	5,559.5532	1

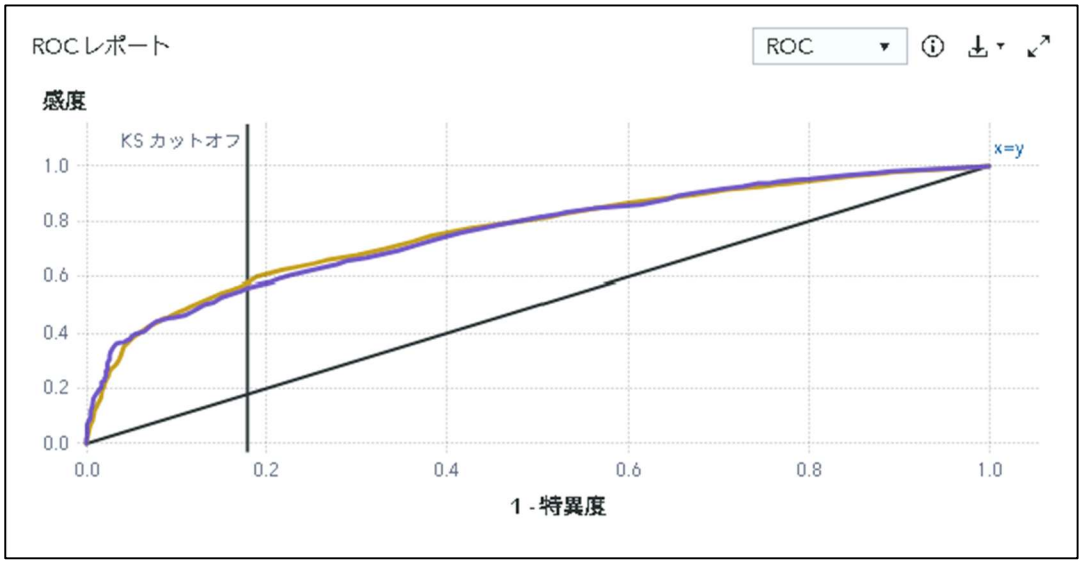
モデルに選択された変数が表示されます。

18. 画面の中央右から**アセスメント**をクリックして、「リフトレポート」ウィンドウを確認します。



累積リフトは、ターゲットイベントの予測確率の降順で各パーティションを並べ替えることによって計算されます。データは 20 分位数（各データの 5%）に分割され、各分位数のイベント数が計算されます。累積リフトは、観測値をランダムに選択するよりも、分位数でイベントを観測する可能性がどれだけ高いかを測定します。

19. 「ROC レポート」ウィンドウを確認します。



ROC 曲線は、1－特異度（偽陽性率）に対する感度（真陽性率）のプロットであり、どちらも混同行列に基づく分類の尺度です。混同行列については Appendix をご参考ください。

これらの測定値は、さまざまなカットオフ値で計算されます。データをスコアリングするときに使用する最適なカットオフを特定しやすくするために、KS カットオフ参照線が描画されます。グラフのカーブが左上隅に急速に近づく ROC 曲線はより正確なモデルを示しています。対角線はランダムモデルを示します。

End of Demonstration

3.3 決定木分析

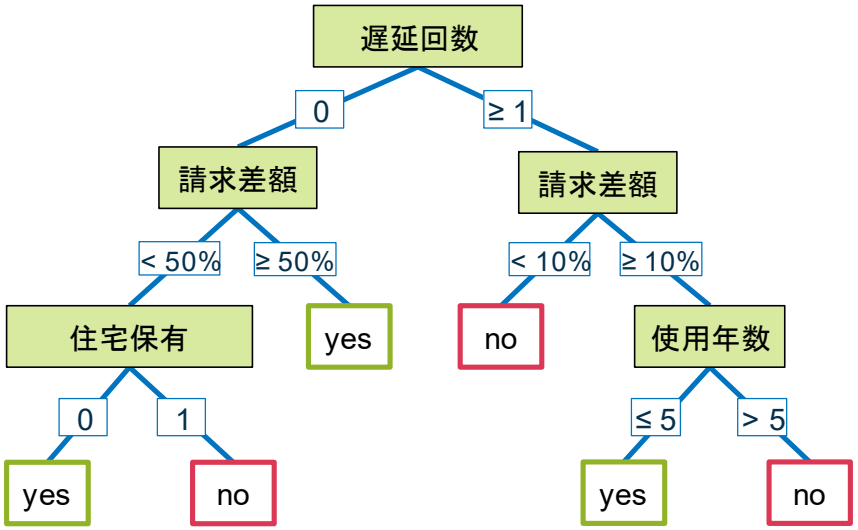
不良債権の例

遅延回数	請求差額	製品数	使用年数	不良債権
0	+5%	3	1	No
0	-7%	2	3	No
1	-2%	2	8	No
1	+32%	5	0	Yes
0	+78%	1	1	Yes
0	+23%	1	9	No
0	-8%	2	4	No
2	-17%	3	2	No
6	+9%	2	7	Yes

27

ターゲット変数は不良債権であり、顧客がデフォルトしたかどうかを示します。予測変数は、過去の支払い遅延回数、平均請求額との差、使用している製品の数、および製品の使用年数です。請求差額は請求額÷平均請求額です。使用年数は顧客が最初に製品を購入してからの年数です。

不良債権の決定木



28

ツリーのような構造で表現できるため、決定木、デシジョンツリーと呼ばれます。決定木は、実際の木を逆さにしたような状態で、一番上のルートノードから下に読み取られていきます。

ツリーの末端のノードはリーフノードと呼ばれます。リーフは予測されたターゲットの値を表します。

Gini指数と不純度

$$1-\sum_{j=1}^{r} p_j^2 = 2\sum_{j<k} p_j p_k$$

高多様性、低純度

Pr(interspecific encounter) = 1-2(3/8)^2-2(1/8)^2 = .69

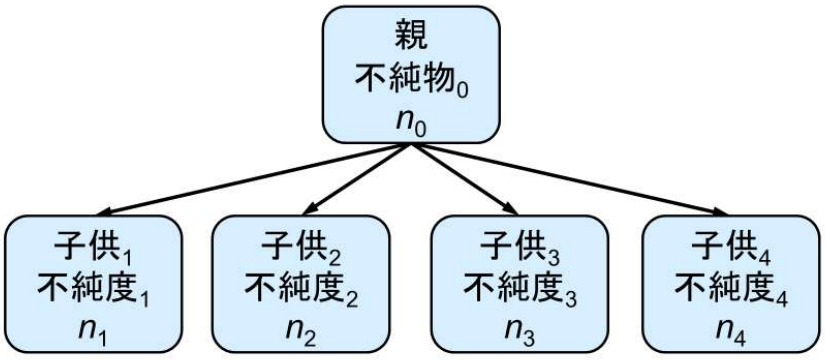
低多様性、高純度

Pr(interspecific encounter) = 1-(6/7)^2-(1/7)^2 = .24

29

分割基準として典型的なものが Gini 指数です。Gini 指数は（1912 年にイタリアの著名な統計学者 Corrado Gini によって開発された）、カテゴリデータの変動性の尺度です。Gini 指数は、ノードの不純度の尺度として使用でき、$p_1, p_2, \ldots, p_r$ はノード内の各ターゲットクラスの割合です。Gini 分割基準は、Breiman らによって提案されました（BFOS 1984）。純粋なノードの Gini 指数は 0 です。均等に分配されたクラスの数が増えると、Gini 指数は 1 に近づきます。

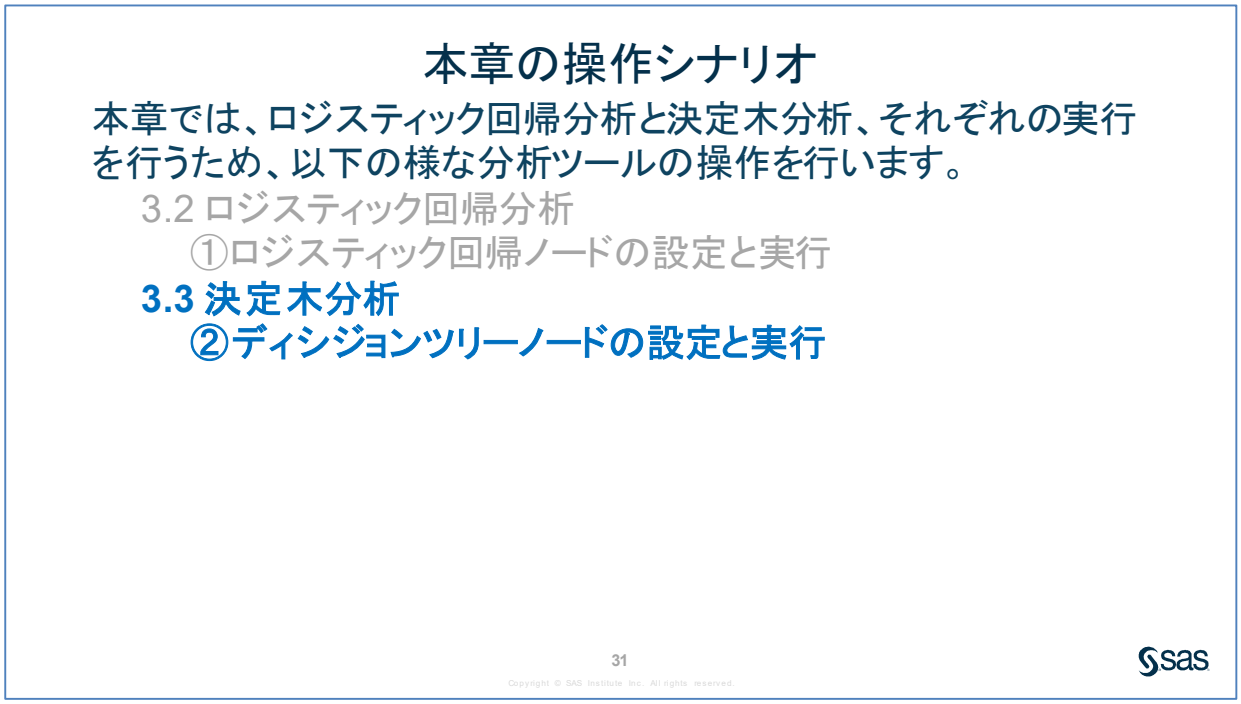
分割基準:不純度測定



$$\Delta i = i(0) - \left(\frac{n_1}{n_0} i(1) + \frac{n_2}{n_0} i(2) + \frac{n_3}{n_0} i(3) + \frac{n_4}{n_0} i(4) \right)$$

30

i（.）をノード内の不純度の測定値とし、Δi をツリーの不純度の全体的な減少を表すとします。多くの分割基準（Gini およびエントロピーを含む）は、分割によって引き起こされるノードの不純度の減少（つまり、ノード内の変動の減少）に基づいています。

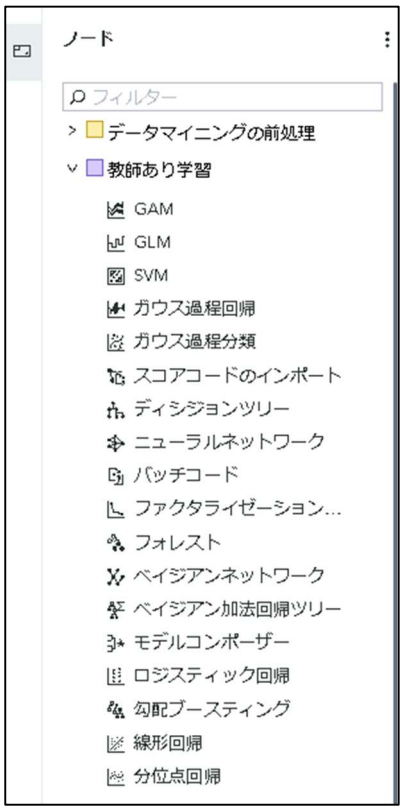


実際に分析ツールを使用して、決定木分析を利用してみましょう。

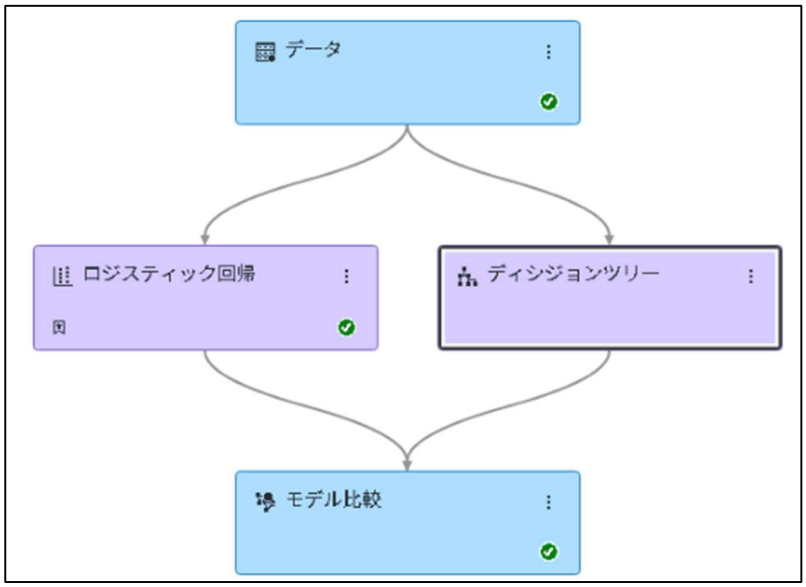
決定木分析

このデモでは、Model Studio を使用して、ディシジョンツリーノードを設定し実行して、結果を確認します。

1. 画面左手から、 ノードボタンをクリックして、「教師あり学習」を展開します。



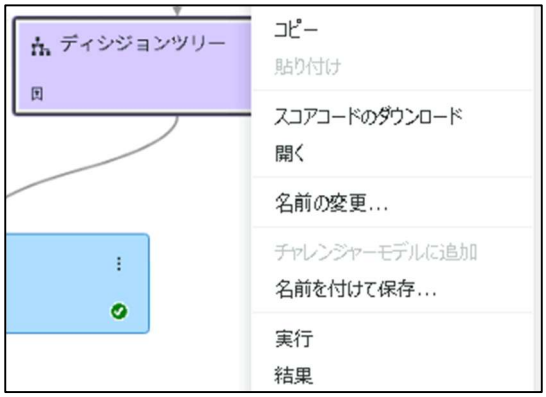
2. ディシジョンツリーを、パイプラインの「データ」ノードにドラッグアンドドロップします。



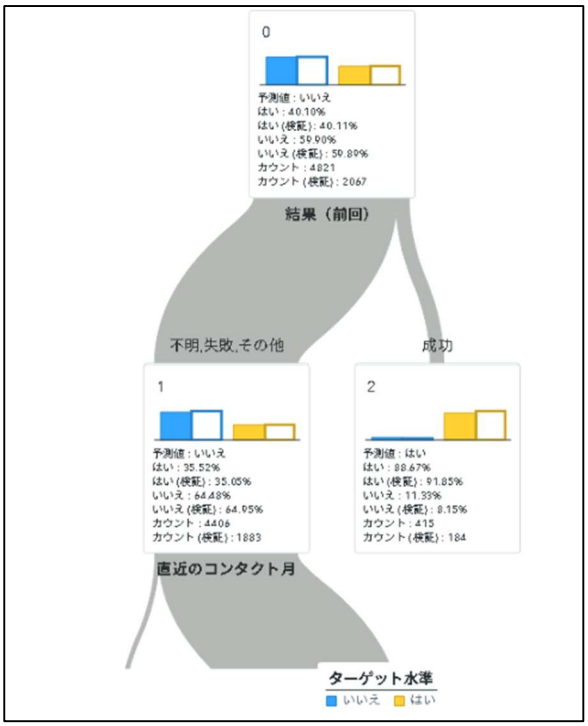
3. 画面右上から、**パイプラインの実行**をクリックします。

実行が正常に完了すると、ノードの右下に緑のチェックマークが現れます。

4. 「ディシジョンツリー」ノードを右クリックして、**結果**を選択します。

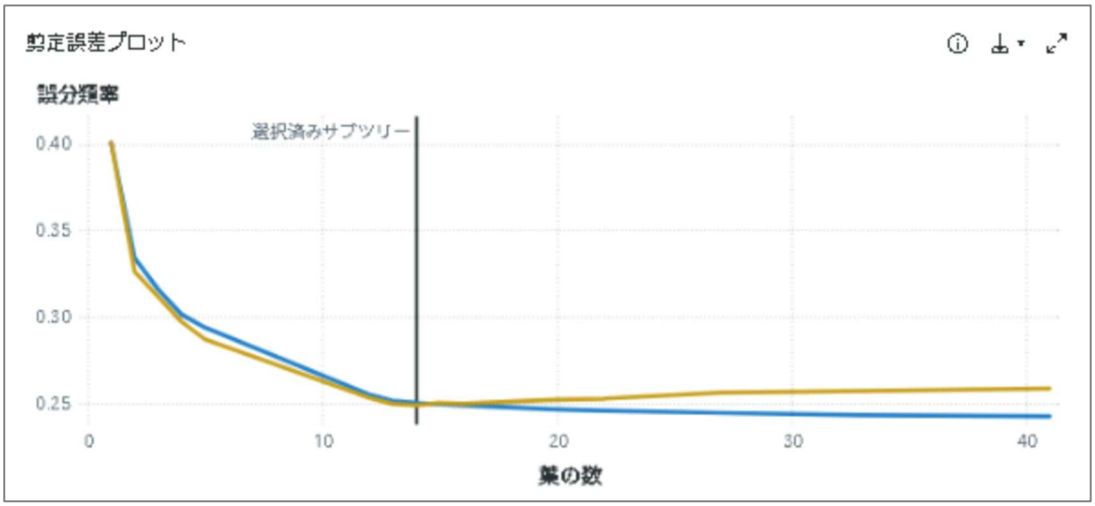


5. 「ツリーダイアグラム」ウィンドウを確認します。



ツリーが大きくなると、ダイアグラムは必ずしも見やすいものになりません。

6. 「剪定誤差プロット」ウィンドウを確認します。



このプロットは、最大の決定木をさまざまな葉の数に剪定することによって作成された、サブツリーの誤分類率がどのように変化するかを示しています。 葉の数が増えると学習データでの誤差が減少するため、検証データを使用してツリーを剪定し過学習を防ぎます。

7. 「変数の重要度」ウィンドウを確認します。

変数の重要度

変数名	学習の重要度	学習の相対...	検証の相対...	カウント	検証の重要度
poutcome	214.3250	1	1	1	107.6935
month	201.9012	0.9420	0.6093	6	65.6137
age	56.5991	0.2641	0.2356	2	25.3764
contact	45.7212	0.2133	0.1841	1	19.8297
day	19.4547	0.0908	0.0957	2	10.3100
housing	32.0168	0.1494	0.0890	1	9.5805

モデルに選択された変数が表示されます。

8. 画面右上の**閉じる**をクリックして、結果を閉じます。

9. パイプラインから「モデルの比較」ノードを右クリックして、**結果**を選択します。


モデルの比較

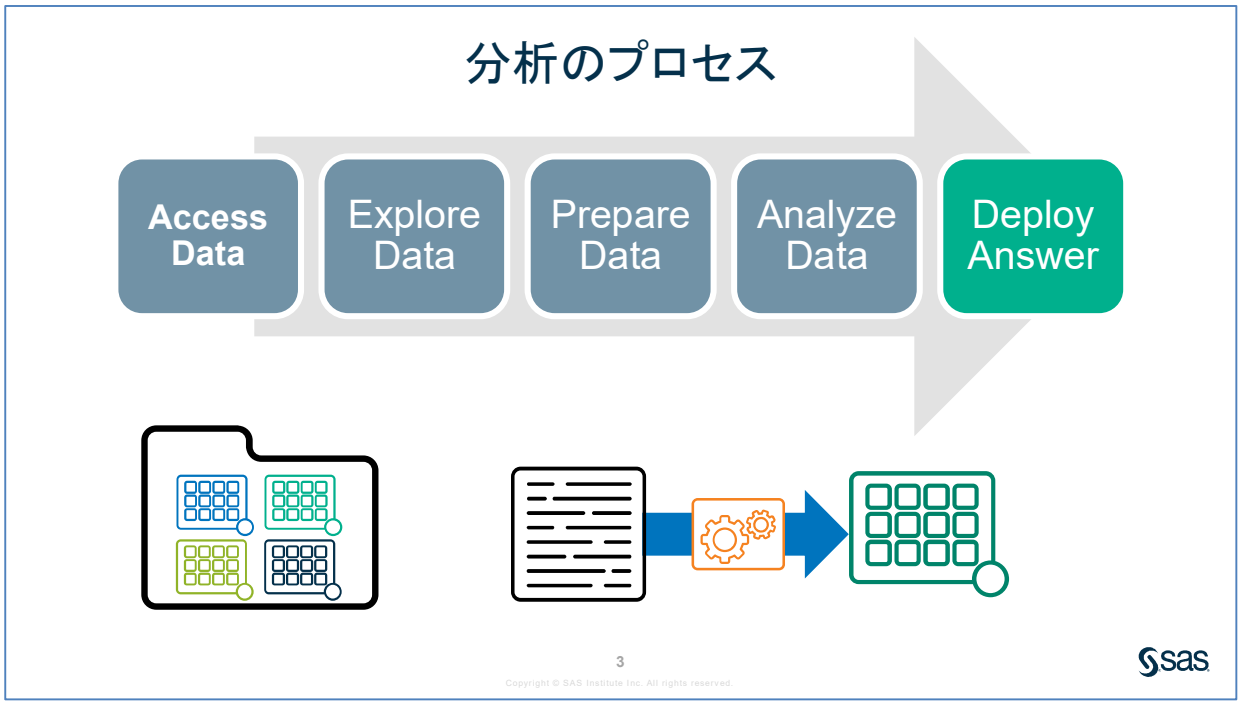
チャンピオン	名前	アルゴリズム名	KS (Youden)
✓	ロジスティック回帰	ロジスティック回帰	0.3673
	ディシジョンツリー	ディシジョンツリー	0.3615

ロジスティック回帰モデルがチャンピオンとして選ばれました。

End of Demonstration

Lesson 4　分析結果の考察と展開

4.1 本章の学習目標

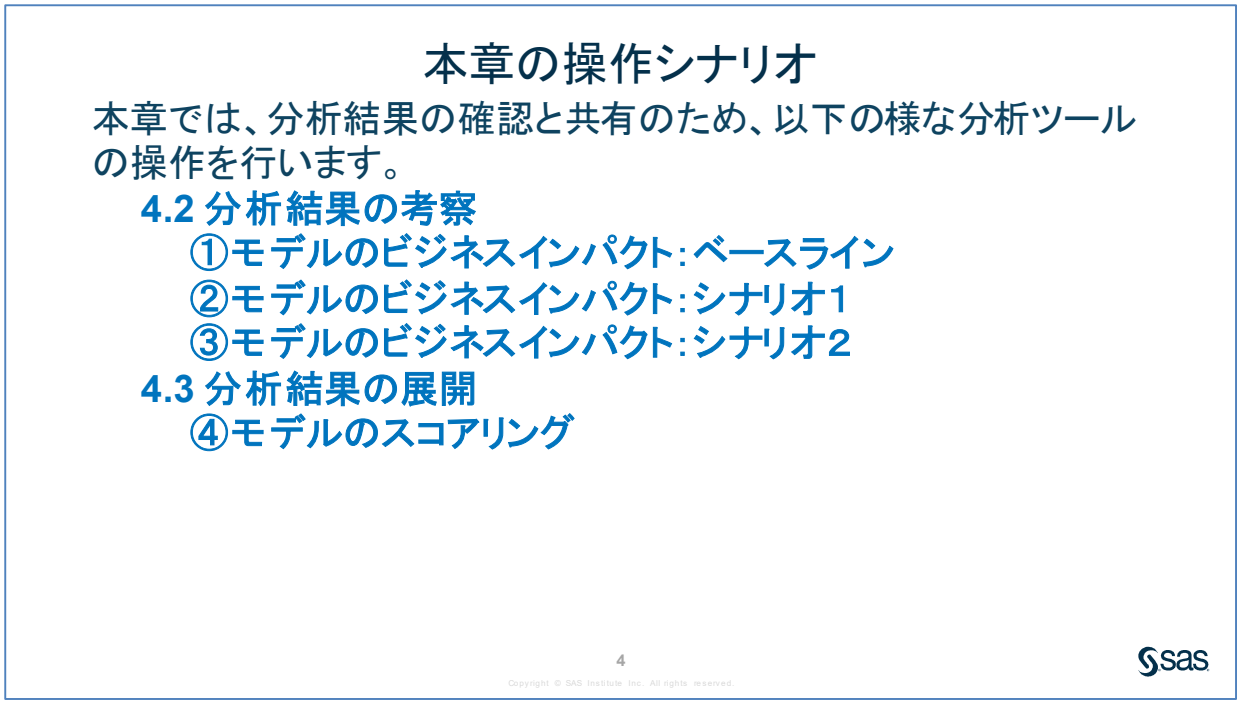


Deploy Answer：分析結果を考察し、ビジネス課題への答えを見つけ、意思決定をサポートするストーリーを作成します。

この章では、分析した結果から答えを導き出して展開します。

ビジネス課題に対する答えを見つけ出し、そこからビジネスの意思決定をサポートすることのできるストーリーを検討します。

結果の展開のプロセスでは、報告書などのドキュメント作成が必要になります。その際には、分析の課題と目的、結論の記述、分析結果を導き出すために利用したデータ、変数、手法が選択された理由、またこれらの結果をできるだけ受け手の理解しやすい内容で、共通の用語や視覚的な素材を用いることを意識してみましょう。



4.2 分析結果の考察

①モデルのビジネスインパクト：ベースライン
検証データ中の実績値を基に ROI を算出します。

②モデルのビジネスインパクト：シナリオ１
顧客のコンタクト数を削減することによる ROI の変化を確認します。

③モデルのビジネスインパクト：シナリオ２
顧客のコンタクトリストが増えた場合を想定した ROI の変化を確認します。

4.3 分析結果の展開

④モデルのスコアリング

報告用ドキュメント作成のヒントとして、レポートを作成する際に、参考となるファイル作成の方法をご紹介します。

4.2 分析結果の考察

本コースの分析シナリオ（レビュー）

ポルトガル銀行では、顧客に対して定期預金の開設キャンペーンを実施しています。過去、現場の判断で多数の顧客に対して営業を実施していたため、販促費が増加しています。そこで、定期預金の開設見込みのある顧客に対してのみ重点的に営業を実施したいと考えています。

そこで、過去実績データを用いて、分析を活用することで効率的にアプローチが可能かを検証してほしいと、データサイエンス部に依頼がありました。また、結果を用いることで期待できる増収益の推定も期待されています。

※1回のコンタクトにかかる費用はおおよそ 3 EURと見積もっています。

※定期預金を1件成約した際の粗利はおよそ 45 EURと見積もっています。

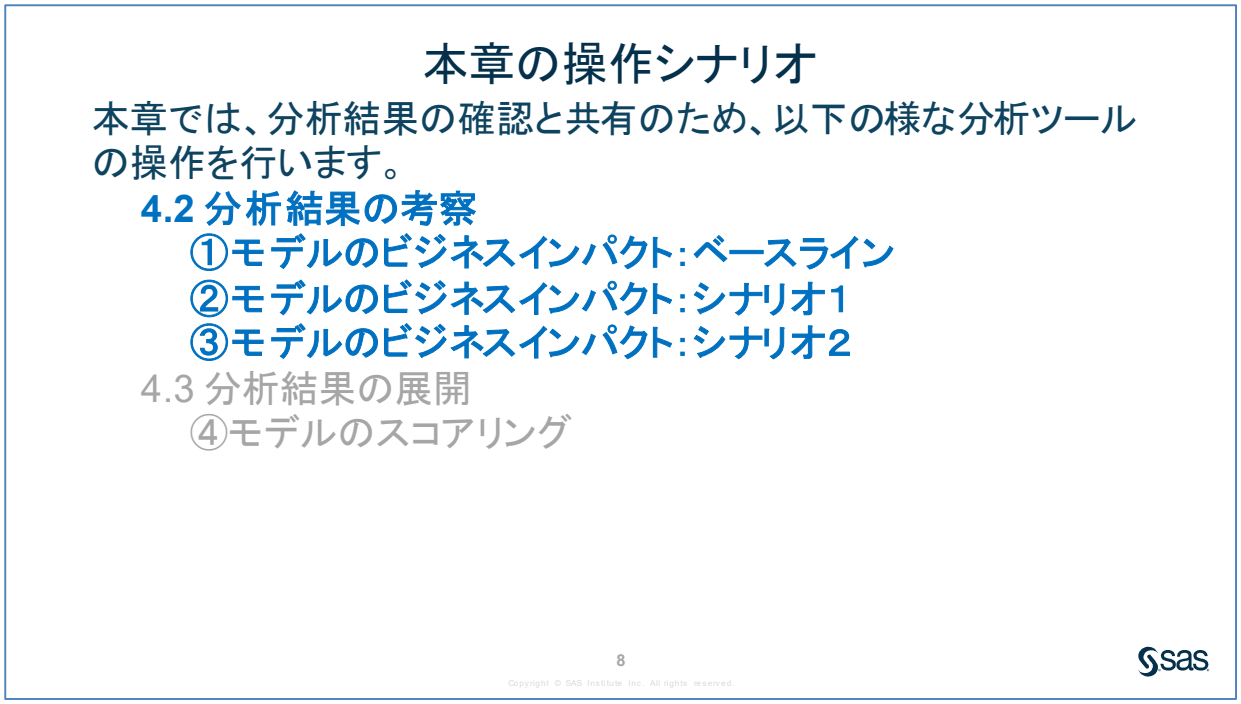
6

口座開設見込みのある顧客に対して重点的なアプローチを行い、かつ費用対効果の向上を目指すために必要な指標を算出してみましょう。

モデルのビジネスインパクト推定

7

モデルが実際に運用された場合のビジネスインパクトについて推定します。ビジネスインパクトの推定時に設定する仮説やシナリオの構築にはドメイン知識が重要になりますので、実務者と十分ディスカッションをして決定していく必要があります。



実際に分析ツールを使用して、ビジネスインパクトを検証するために必要なメトリックを算出します。

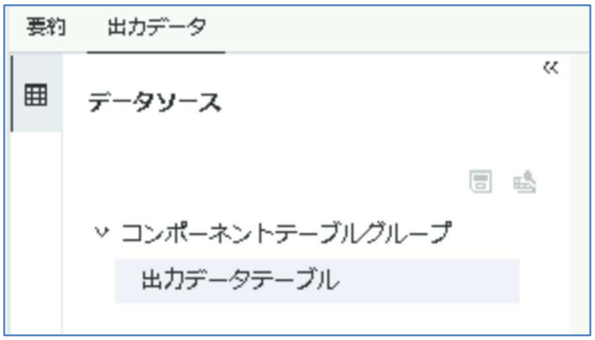
モデルのビジネスインパクト：ベースライン

このデモでは、検証データ中の口座開設数や成約率、コンタクト数、費用、利益などを算出し、費用対効果の実績値を求めます。

まず、Model Studio でチャンピオンモデルとなった、予測結果のテーブルを保存します。そして、そこから先は、また SAS Studio でプログラムを用いて計算を行います。

1. チャンピオンモデルのノードを右クリックして、**結果**を選択します。

2. 画面左上から、**出力データ**タブをクリックします。



3. 画面中央の**出力データの表示**をクリックします。

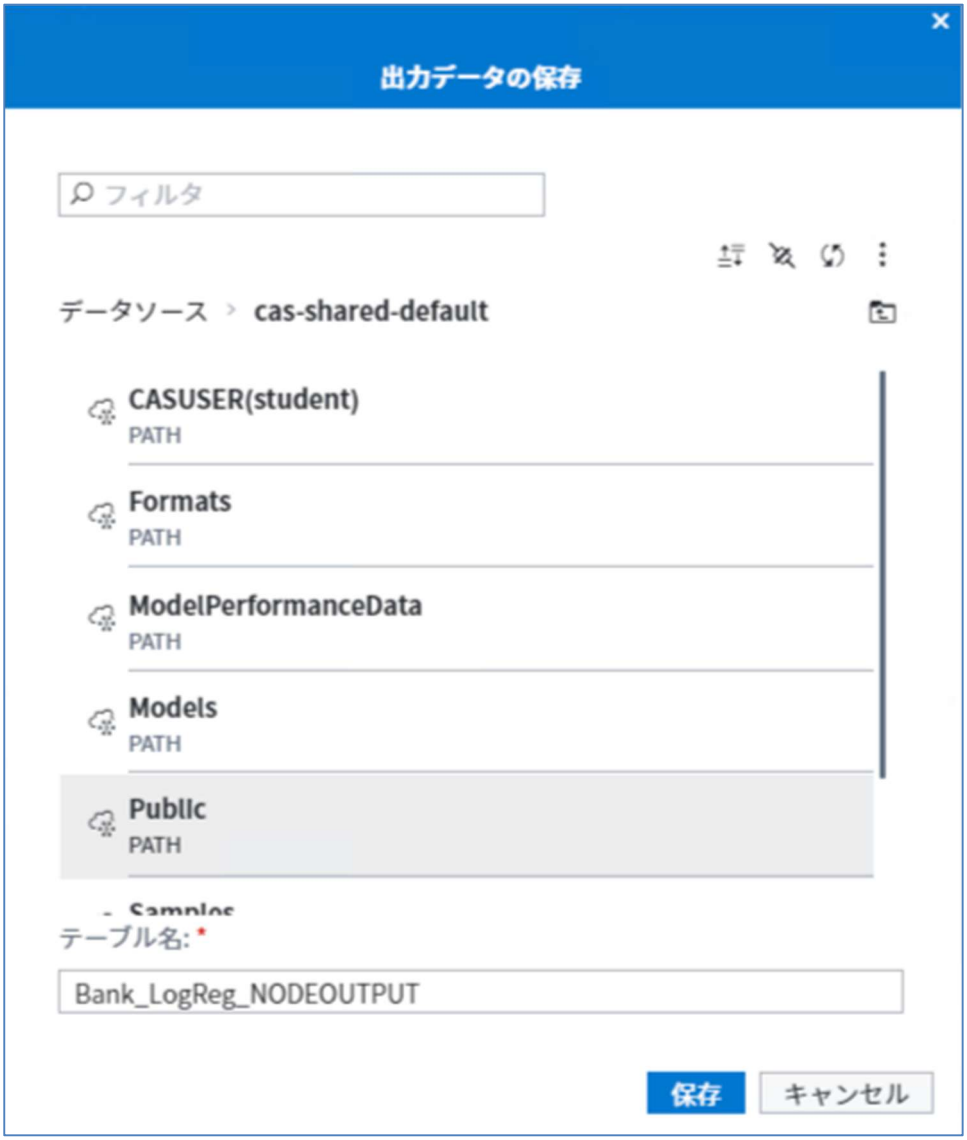
4. 表示された画面で、**出力データの表示**をクリックします。

5. 表示されたデータには、分析の実行によって出力された予測結果の列が追加されています。

パーティション…	ターゲット: de…	予測: deposit=…	P_depositno	Probability for…	Predic
1	no	0.2892854484	0.7107145516	0.2892854484	no
0	no	0.3216036595	0.6783963405	0.3216036595	no

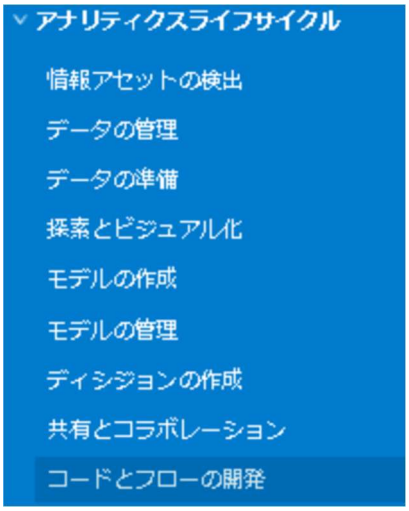
6. 画面左手のデータ名「NODEOUTPUT」を右クリックして、**保存**を選択します。

7. データソースの **cas-shared-default** をダブルクリックして、**PUBLIC** を選択します。
テーブル名に「Bank_LogReg_NODEOUTPUT」と入力して、**保存**をクリックします。



「データは正常に保存されました。」のメッセージを確認します。

8. 画面左上の三本線 のボタンをクリックして、**コードとフローの開発**を選択します。



9. 開始ページタブから、**SAS プログラムの新規作成**をクリックします。
※これまでに使用していたプログラムのタブがある場合には、事前にすべて保存せずに閉じてください。

10. SAS プログラムを開くため、講師の指示に従ってファイルの場所に移動します。「3.1.1 ベースライン.sas」を右クリックして、プログラムを開く⇒Edit with Notepad++を選択します。ファイルが開いたら、Ctrl＋A で全体を選択し、Ctrl＋C でコピーして、SAS Studio の「プログラム.sas」タブに Ctrl＋V で貼り付けます。

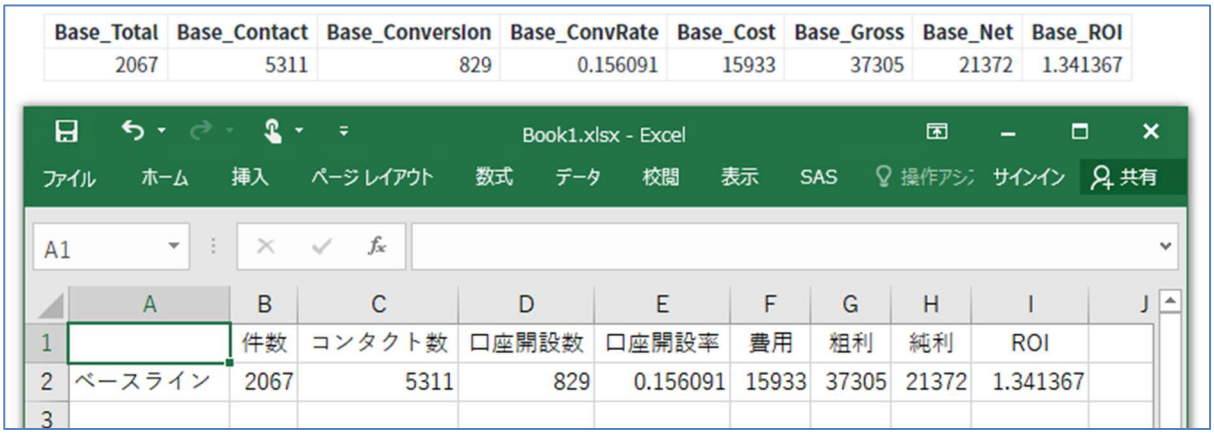
11. コードタブをクリックして、以下のプログラムを範囲選択して実行します。

```
*-------------------------------------------------------*;
* 分析結果データ *;
*-------------------------------------------------------*;
proc casutil ;
  load incaslib  = 'public'
       casdata = 'Bank_LogReg_NODEOUTPUT.sashdat'
       outcaslib = 'public'
       casout  = 'Bank_LogReg_NODEOUTPUT'
       promote
       ;
run ;

*-------------------------------------------------------*;
* 検証データ *;
*-------------------------------------------------------*;
data work.validation ;
  set public.Bank_LogReg_NODEOUTPUT ;
  where _PartInd_ = 0 ;
run ;

*-------------------------------------------------------*;
* 検証データ：ベースライン *;
*-------------------------------------------------------*;
proc sql ;
  select count( * ) as Base_Total
       , sum( campaign ) as Base_Contact
       , sum( ( deposit = 'はい' ) ) as Base_Conversion
       , calculated Base_Conversion / calculated Base_Contact
         as Base_ConvRate
       , calculated Base_Contact * 3 as Base_Cost
       , calculated Base_Conversion * 45 as Base_Gross
       , calculated Base_Gross - calculated Base_Cost as Base_Net
       , calculated Base_Net / calculated Base_Cost as Base_ROI
    from work.validation
    ;
quit ;
```

このプログラムでは、初めのステップで分析結果のデータを CAS へロードしています。次のステップで、学習と検証に分割されたパーティションから検証データを抽出しています。そして、最後のステップで、データ件数（Base_Total）、コンタクト数（Base_Contact）、口座開設数（Base_Conversion）、口座開設率（Base_ConvRate）、コンタクト費用（Base_Cost）、口座開設の粗利（Base_Gross）、純利（Base_Net）、費用対効果（Base_ROI）を求めています。

12. 表示された結果を、Excel にまとめます。

Base_Total	Base_Contact	Base_Conversion	Base_ConvRate	Base_Cost	Base_Gross	Base_Net	Base_ROI
2067	5311	829	0.156091	15933	37305	21372	1.341367



	A	B	C	D	E	F	G	H	I
1		件数	コンタクト数	口座開設数	口座開設率	費用	粗利	純利	ROI
2	ベースライン	2067	5311	829	0.156091	15933	37305	21372	1.341367
3									

End of Demonstration

モデルのビジネスインパクト：シナリオ１

このデモでは、構築したモデルのデータを用いて、現状の顧客リストから最適な顧客のみを選択し、コンタクト数を削減して利益が最大となるポイントをシミュレーションします。

1. SAS Studio の画面上部、「*プログラム.sas」の右にあるプラス「＋」ボタンをクリックして、新しいプログラムエディタを立ち上げます。

2. SAS プログラムを開くため、講師の指示に従ってファイルの場所に移動します。「3.1.2 シナリオ１.sas」を右クリックして、プログラムを開く⇒Edit with Notepad++を選択します。ファイルが開いたら、Ctrl＋A で全体を選択し、Ctrl＋C でコピーして、SAS Studio の「プログラム 1.sas」タブに Ctrl＋V で貼り付けます。

3. コードタブをクリックして、以下のプログラムを範囲選択して実行します。

```
*-------------------------------------------------------*;
* 検証データ：シナリオ１ *;
*-------------------------------------------------------*;
%macro Loop ;
  %do i = 1 %to 1000 ;
    proc sql ;
      create table work.tmp_Biz1 as
        select sum( campaign ) as Biz1_Contact
             , sum( ( deposit = 'はい' ) ) as Biz1_Conversion
             , calculated BIz1_Contact * 3 as Biz1_Cost
             , calculated BIz1_Conversion * 45 as Biz1_Gross
             , calculated BIz1_Gross - calculated BIz1_Cost
               as Biz1_Net
             , &i / 1000 as Biz1_ThresHld
          from work.validation
          where em_eventprobability >= &i / 1000
          ;
    quit ;

    proc append base = work.Biz1 data = work.tmp_Biz1 ;
    run ;
  %end ;
%mend Loop ;
%Loop

ods select ExtremeObs ;
proc univariate data = work.Biz1 ;
  var Biz1_Net ;
  id Biz1_ThresHld ;
run ;
ods select all ;

proc sgplot data = work.Biz1 ;
  series x = Biz1_ThresHld y = Biz1_Net ;
run ;

proc sql ;
  select count( * ) as Biz1_Total
       , sum( campaign ) as Biz1_Contact
       , sum( ( deposit = 'はい' ) ) as Biz1_Conversion
       , calculated Biz1_Conversion / calculated Biz1_Contact
         as Biz1_ConvRate
       , calculated Biz1_Contact * 3 as Biz1_Cost
       , calculated Biz1_Conversion * 45 as Biz1_Gross
       , calculated Biz1_Gross - calculated Biz1_Cost
         as Biz1_Net
       , calculated Biz1_Net / calculated Biz1_Cost as Biz1_ROI
    from work.validation
    where em_eventprobability >= 0.115
    ;
quit ;
```

このプログラムでは、初めのステップで分析結果のデータから、予測確率を元に 0.1～100%の範囲で 0.1%ごとにデータを抽出してそれぞれで利益を求め work.biz1 データに蓄積して保存しています。次のステップで、利益が最大となるデータをレポート出力しています。さらに次のステップで、利益と予測確率でグラフを描いています。最後のステップで

は、先ほどのベースラインを算出した場合と同様に、ROI を求めるための各種指標を算出しています。

4. 表を確認します。

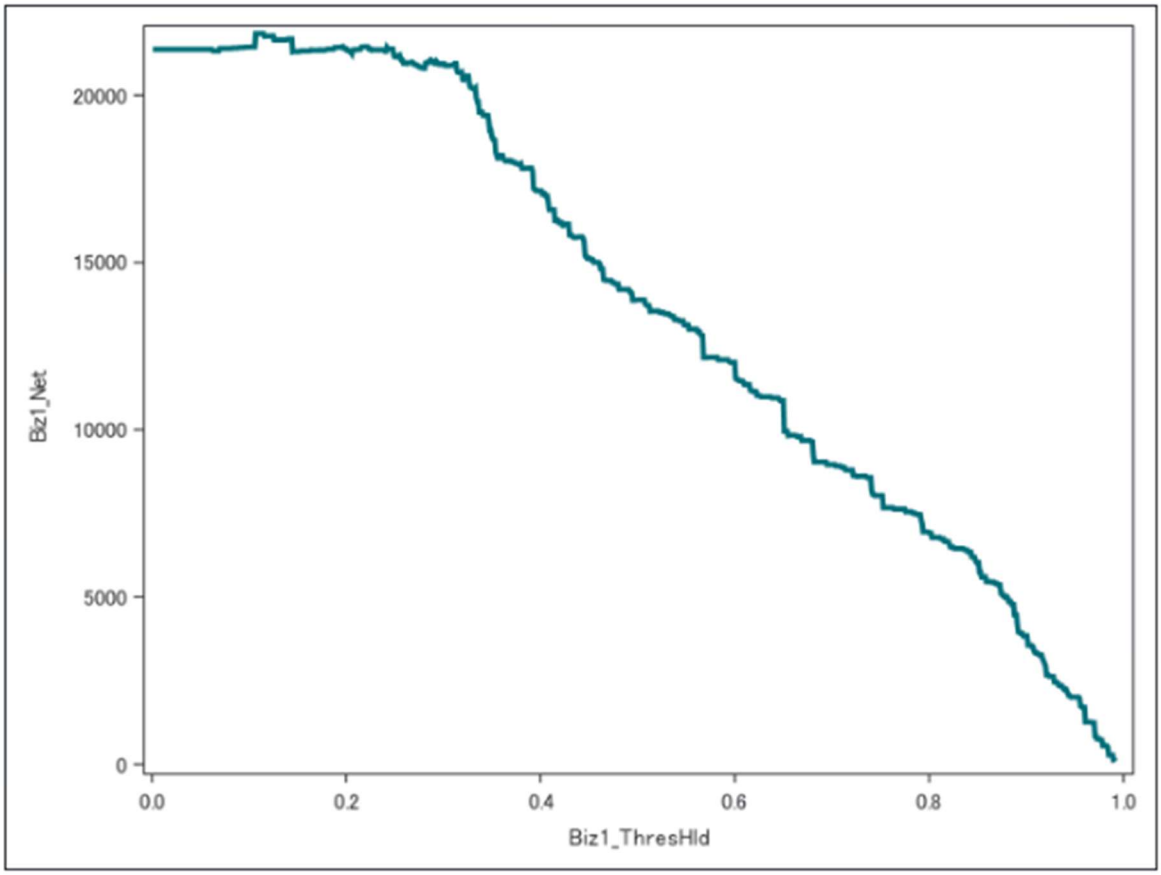
UNIVARIATE プロシジャ
変数 : Biz1_Net

極値

最小値			最大値		
値	Biz1_ThresHld	Obs	値	Biz1_ThresHld	Obs
84	0.992	992	21843	0.111	111
126	0.991	991	21843	0.112	112
126	0.990	990	21843	0.113	113
279	0.989	989	21843	0.114	114
279	0.988	988	21843	0.115	115

純利が最大になった時の、閾値を確認することができます。

5. グラフを確認します。



一定の閾値を超えると、利益が下がっていくことを確認できます。

6. 表示された結果を、Excel にまとめます。

Biz1_Total	Biz1_Contact	Biz1_Conversion	Biz1_ConvRate	Biz1_Cost	Biz1_Gross	Biz1_Net	Biz1_ROI
1657	4205	755	0.179548	12615	33975	21360	1.693222

	A	B	C	D	E	F	G	H	I	J
1		件数	コンタクト数	口座開設数	口座開設率	費用	粗利	純利	ROI	
2	ベースライン	2067	5311	829	0.156091	15933	37305	21372	1.341367	
3	シナリオ 1	1657	4205	755	0.179548	12615	33975	21360	1.693222	

モデルから出力された予測確率を利用することにより、費用対効果（ROI）が上がったことが確認できました。

End of Demonstration

モデルのビジネスインパクト：シナリオ２

このデモでは、構築したモデルのデータを用いて、顧客をランダムに選択して顧客リストを増加して、利益の変化をシミュレーションします。

1. SAS Studio の画面上部、「*プログラム 1.sas」の右にあるプラス「＋」ボタンをクリックして、新しいプログラムエディタを立ち上げます。

2. SAS プログラムを開くため、講師の指示に従ってファイルの場所に移動します。「3.1.3 シナリオ２.sas」を右クリックして、プログラムを開く⇒Edit with Notepad++を選択します。ファイルが開いたら、Ctrl＋A で全体を選択し、Ctrl＋C でコピーして、SAS Studio の「プログラム 2.sas」タブに Ctrl＋V で貼り付けます。

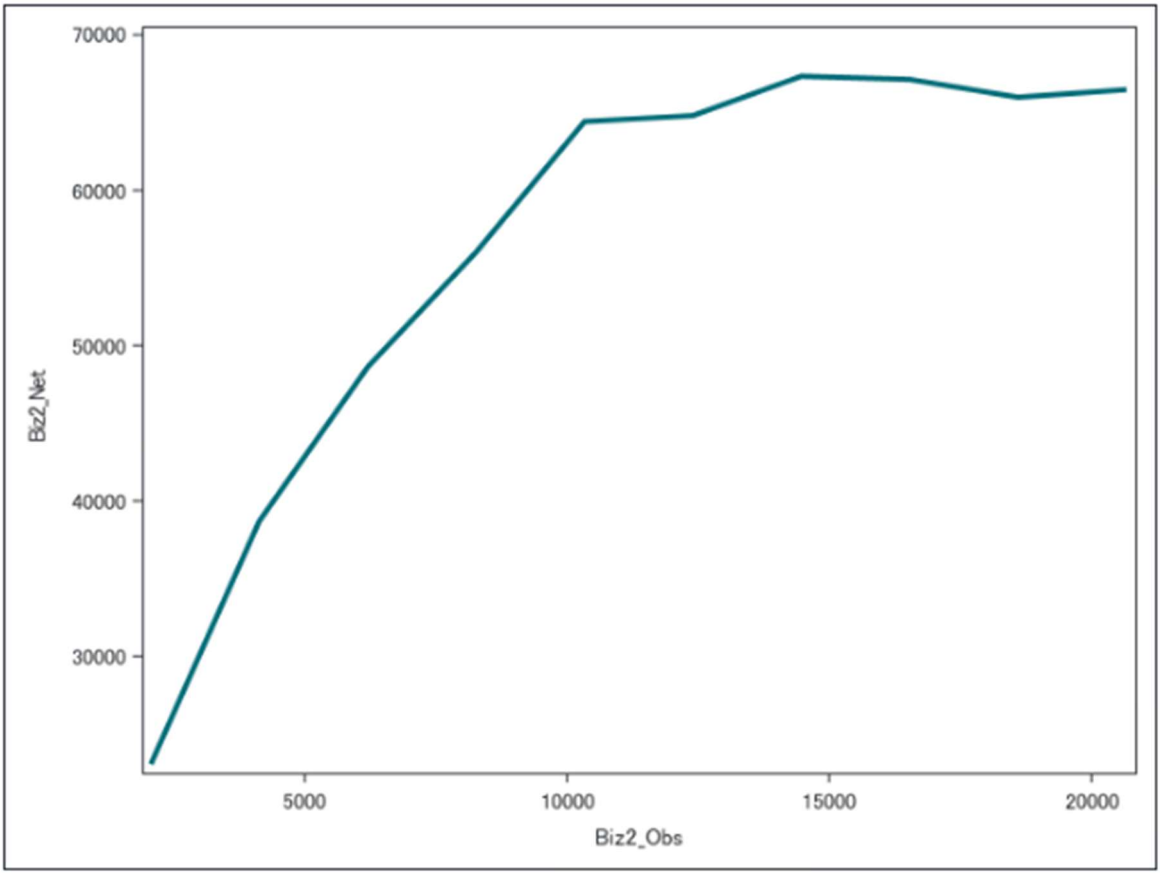
3. コードタブをクリックして、以下のプログラムを範囲選択して実行します。

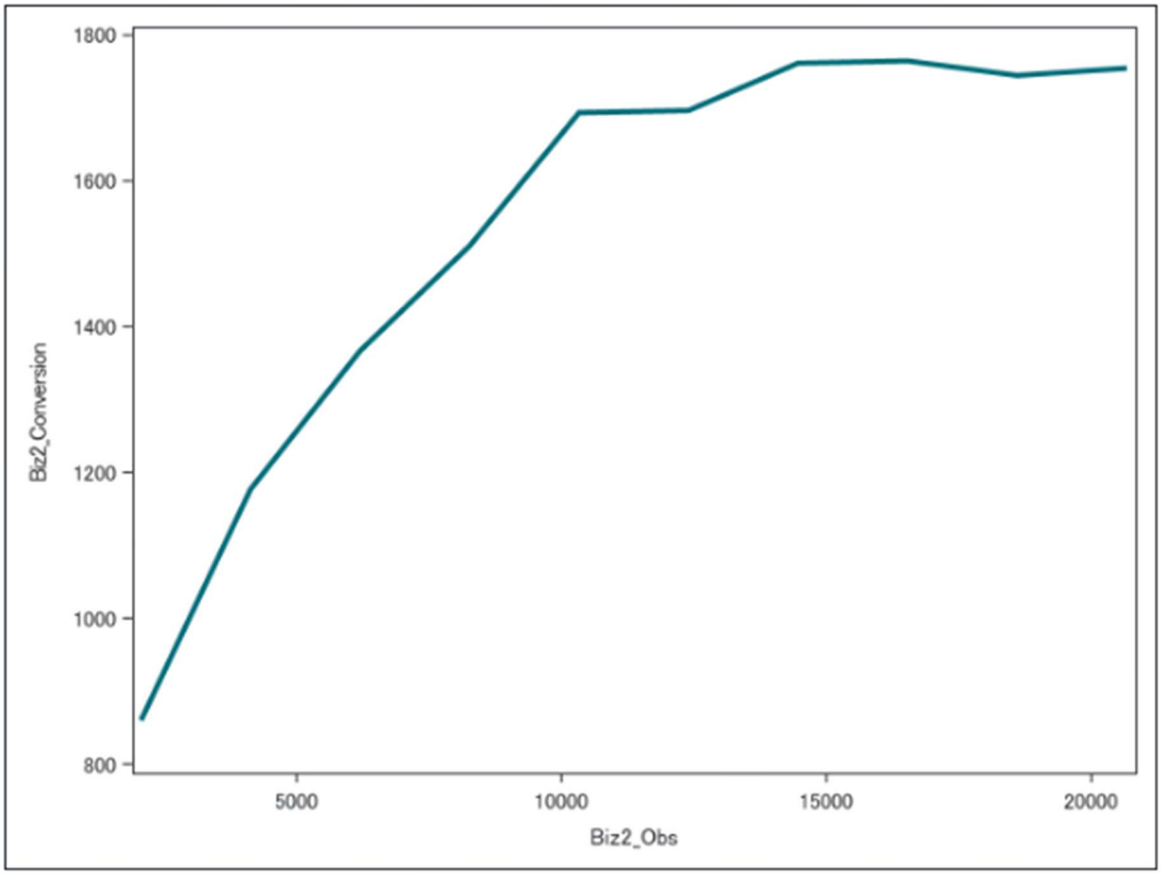
```
*-----------------------------------------------------------*;
* 検証データ：シナリオ２ *;
*-----------------------------------------------------------*;
%macro Loop ;
  %do i = 1 %to 10 ;
    %let obs = %eval( 2067 * &i ) ;

    data work.smp_Biz2 ;
      sampsize = &obs ;
      do i = 1 to sampsize ;
        PickIt = ceil( ranuni( 1234 ) * TotObs ) ;
        set work.validation point = PickIt nobs = TotObs ;
        output ;
      end ;
      stop ;
      drop i ;
    run ;

    proc sort data = work.smp_Biz2 out = work.top_Biz2 ;
      by descending em_eventprobability ;
    run ;

    proc sql ;
      create table work.tmp_Biz2 as
        select &obs as Biz2_Obs
             , sum( campaign ) as Biz2_Contact
             , sum( ( deposit = 'はい' ) ) as Biz2_Conversion
             , calculated Biz2_Conversion / calculated Biz2_Contact
               as Biz2_ConvRate
             , calculated BIz2_Contact * 3 as Biz2_Cost
             , calculated BIz2_Conversion * 45 as Biz2_Gross
             , calculated BIz2_Gross - calculated BIz2_Cost
               as Biz2_Net
             , calculated Biz2_Net / calculated Biz2_Cost
               as Biz2_ROI
          from work.top_Biz2( obs = 2067 )
          ;
    quit ;

    proc append base = work.Biz2 data = work.tmp_Biz2 ;
    run ;
  %end ;
%mend Loop ;
%Loop

proc sgplot data = work.Biz2 ;
  series x = Biz2_Obs y = Biz2_Net ;
run ;

proc sgplot data = work.Biz2 ;
  series x = Biz2_Obs y = Biz2_Conversion ;
run ;

proc sgplot data = work.Biz2 ;
  series x = Biz2_Obs y = Biz2_ROI ;
run ;
```

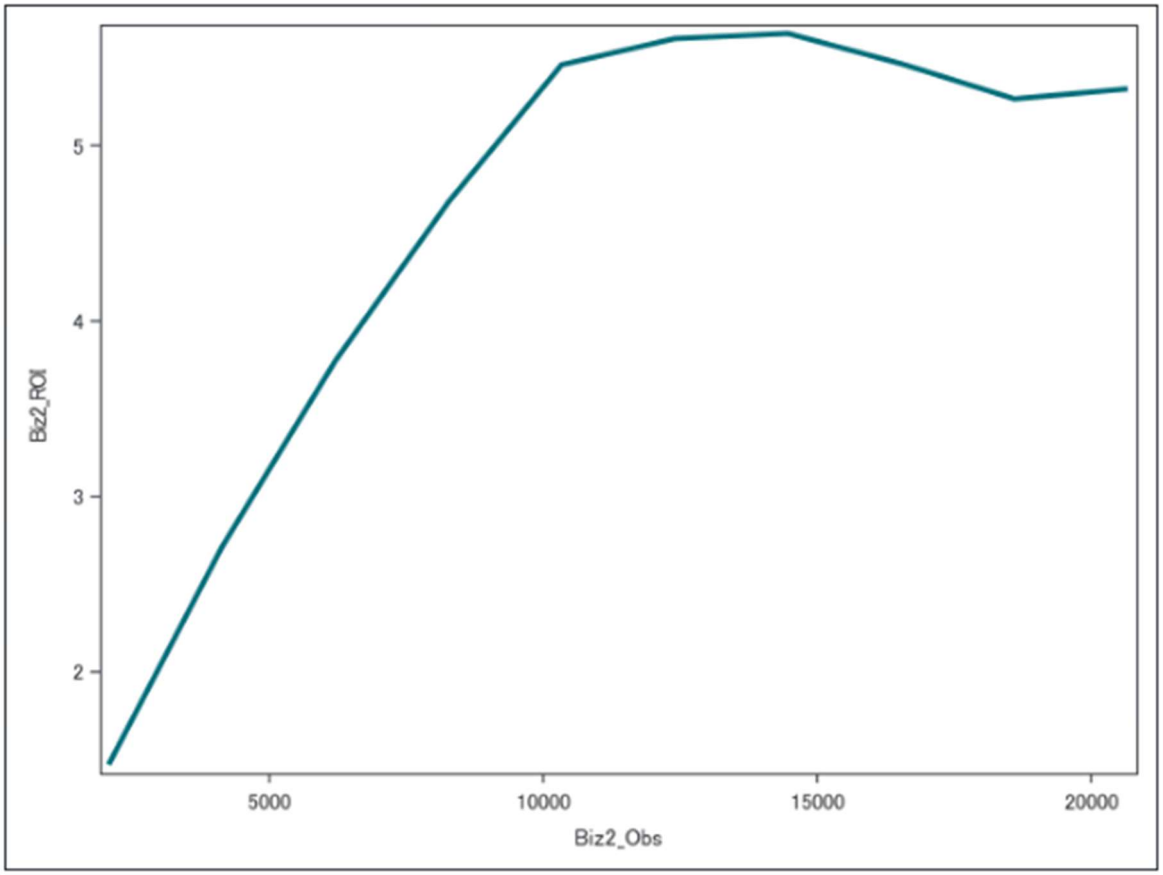
```
proc print data = work.Biz2 ;
run ;
```

このプログラムでは、初めのステップで分析結果のデータから、重複ありでランダムサンプリングを行い、1～10 倍までデータを増やし、それぞれで ROI 算出のための指標を求めて work.biz2 データに蓄積して保存しています。次のステップで、純利、口座開設数、ROI と人数でグラフを描き、各指標を一覧表示しています。

4. グラフを確認します。



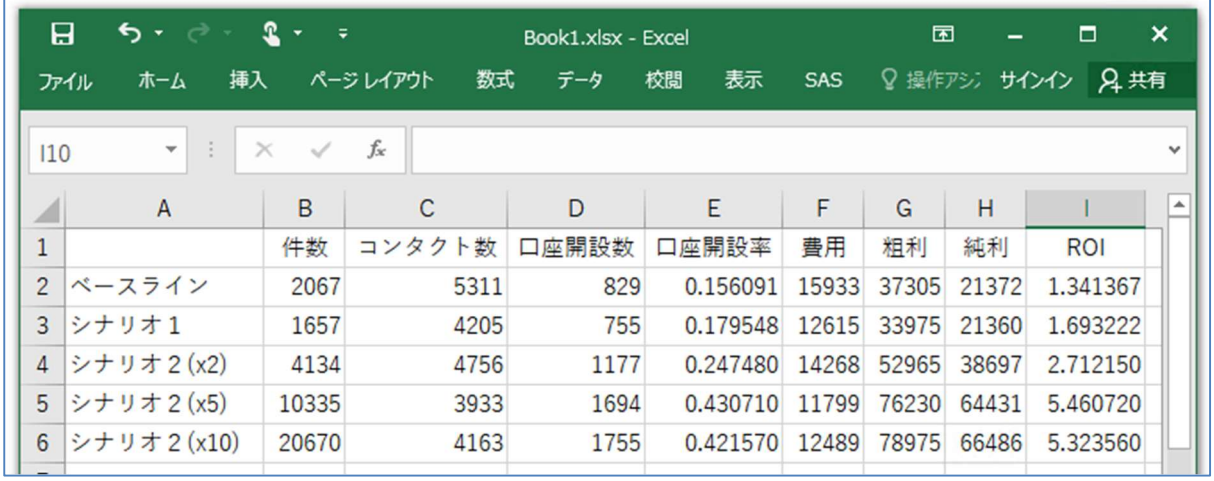




一定の閾値を超えると、サチュレーションの起きることがわかります。

5. 表示された結果を、Excel にまとめます。

Obs	Biz2_Obs	Biz2_Contact	Biz2_Conversion	Biz2_ConvRate	Biz2_Cost	Biz2_Gross	Biz2_Net	Biz2_ROI
1	2067	5216	860	0.16488	15648	38700	23052	1.47316
2	4134	4756	1177	0.24748	14268	52965	38697	2.71215
3	6201	4297	1367	0.31813	12891	61515	48624	3.77193
4	8268	3992	1511	0.37851	11976	67995	56019	4.67761
5	10335	3933	1694	0.43071	11799	76230	64431	5.46072
6	12402	3851	1697	0.44066	11553	76365	64812	5.60997
7	14469	3981	1762	0.44260	11943	79290	67347	5.63904
8	16536	4094	1765	0.43112	12282	79425	67143	5.46678
9	18603	4177	1745	0.41776	12531	78525	65994	5.26646
10	20670	4163	1755	0.42157	12489	78975	66486	5.32356



	A	B	C	D	E	F	G	H	I
1		件数	コンタクト数	口座開設数	口座開設率	費用	粗利	純利	ROI
2	ベースライン	2067	5311	829	0.156091	15933	37305	21372	1.341367
3	シナリオ 1	1657	4205	755	0.179548	12615	33975	21360	1.693222
4	シナリオ 2 (x2)	4134	4756	1177	0.247480	14268	52965	38697	2.712150
5	シナリオ 2 (x5)	10335	3933	1694	0.430710	11799	76230	64431	5.460720
6	シナリオ 2 (x10)	20670	4163	1755	0.421570	12489	78975	66486	5.323560

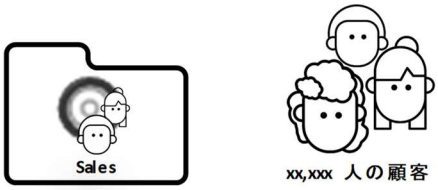
モデルから出力された予測確率を利用することにより、費用対効果（ROI）が上がったことが確認できました。

End of Demonstration

4.3 分析結果の展開

シナリオから導き出した分析方法（レビュー）

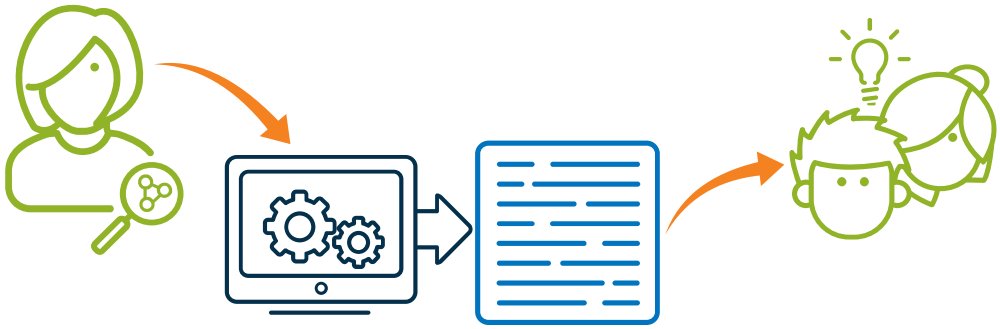
- 過去のキャンペーン実施履歴と定期預金口座開設履歴のデータから、定期預金口座を開設するポテンシャルがある顧客か否かを分類する
- 目的変数は定期預金口座を開設したか、していないかの2値
- 営業部にモデルの妥当性を説明するため、構築したモデルの判断根拠や重要変数を可視化
- 基礎集計情報と特徴量の重要度等の提供



13

想定していた分析方法の内容に基づき、分析結果の報告をビジネスサイドへ行います。

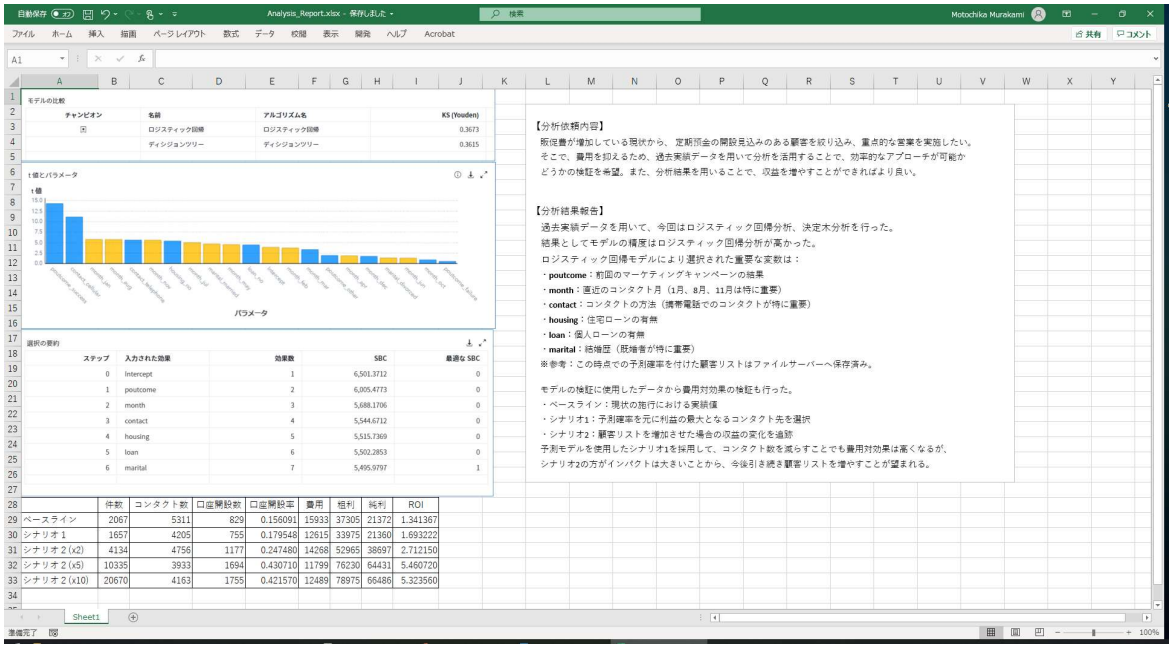
報告用ドキュメントの作成



14

分析した結果から導き出された内容を報告するために、レポート用のドキュメントを作成する必要があるでしょう。

※報告用レポートのイメージ例



モデルの比較

チャンピオン	名前	アルゴリズム名	KS (Youden)
	ロジスティック回帰	ロジスティック回帰	0.3673
	ディシジョンツリー	ディシジョンツリー	0.3615

t値とパラメータ

選択の要約

ステップ	入力された効果	効果数	SBC	最適な SBC
0	Intercept	1	6,501.3712	0
1	poutcome	2	6,005.4773	0
2	month	3	5,688.1706	0
3	contact	4	5,544.6712	0
4	housing	5	5,515.7369	0
5	loan	6	5,502.2853	0
6	marital	7	5,495.9797	1

【分析依頼内容】
販促費が増加している現状から、定期預金の開設見込みのある顧客を絞り込み、重点的な営業を実施したい。
そこで、費用を抑えるため、過去実績データを用いて分析を活用することで、効率的なアプローチが可能か
どうかの検証を希望。また、分析結果を用いることで、収益を増やすことができればより良い。

【分析結果報告】
過去実績データを用いて、今回はロジスティック回帰分析、決定木分析を行った。
結果としてモデルの精度はロジスティック回帰分析が高かった。
ロジスティック回帰モデルにより選択された重要な変数は：
・**poutcome**：前回のマーケティングキャンペーンの結果
・**month**：直近のコンタクト月（1月、8月、11月は特に重要）
・**contact**：コンタクトの方法（携帯電話でのコンタクトが特に重要）
・**housing**：住宅ローンの有無
・**loan**：個人ローンの有無
・**marital**：結婚歴（既婚者が特に重要）
※参考：この時点での予測確率を付けた顧客リストはファイルサーバーへ保存済み。

モデルの検証に使用したデータから費用対効果の検証も行った。
・ベースライン：現状の施行における実績値
・シナリオ1：予測確率を元に利益の最大となるコンタクト先を選択
・シナリオ2：顧客リストを増加させた場合の収益の変化を追跡
予測モデルを使用したシナリオ1を採用して、コンタクト数を減らすことでも費用対効果は高くなるが、
シナリオ2の方がインパクトは大きいことから、今後引き続き顧客リストを増やすことが望まれる。

	件数	コンタクト数	口座開設数	口座開設率	費用	粗利	純利	ROI
ベースライン	2067	5311	829	0.156091	15933	37305	21372	1.341367
シナリオ 1	1657	4205	755	0.179548	12615	33975	21360	1.693222
シナリオ 2 (x2)	4134	4756	1177	0.247480	14268	52965	38697	2.712150
シナリオ 2 (x5)	10335	3933	1694	0.430710	11799	76230	64431	5.460720
シナリオ 2 (x10)	20670	4163	1755	0.421570	12489	78975	66486	5.323560

今回のケースでは、販促費が増加している現状から、 定期預金の開設見込みのある顧客を絞り込み重点的な営業を実施して費用を抑えるため、過去実績データを用いて分析を活用することで、効率的なアプローチが可能かどうかの検討を依頼されました。また、分析結果を用いることで、収益を増やすことができればより良いという依頼もありました。

そこで、今回はロジスティック回帰分析、決定木分析を行いました。結果としてモデルの精度はロジスティック回帰分析が高かったため、このモデルから選択された重要な変数、予測確率付の顧客のリスト、また仮想のシナリオに基づいたシミュレーション結果をドキュメントにまとめて報告を行うことにしました。

■ロジスティック回帰モデルにより選択された重要な変数：

- poutcome：前回のマーケティングキャンペーンの結果
- month：直近のコンタクト月（1 月、8 月、11 月は特に重要）
- contact：コンタクトの方法（携帯電話でのコンタクトが特に重要）
- housing：住宅ローンの有無
- loan：個人ローンの有無
- marital：結婚歴（既婚者が特に重要）

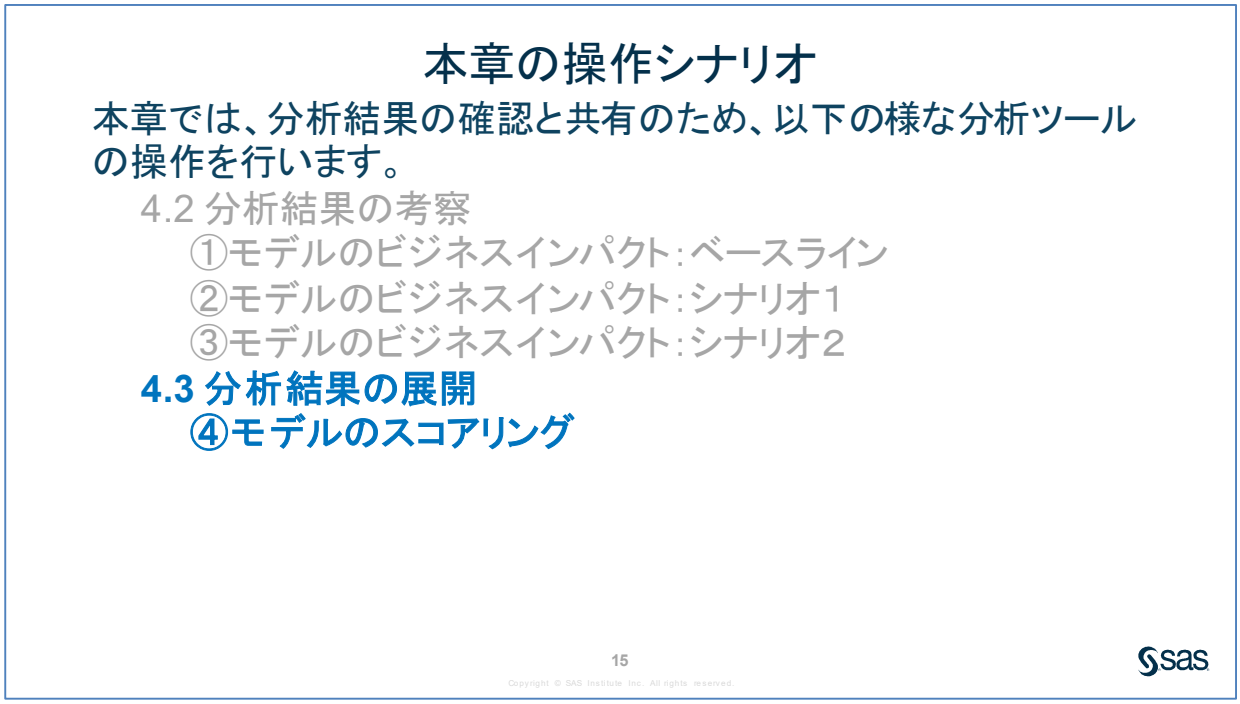
※参考：この時点での予測確率を付けた顧客リストはファイルサーバーへ保存済みとします。

■モデルの結果から実績値と仮想シナリオで費用対効果をシミュレート：

- ベースライン：現状の施行における実績値
- シナリオ 1：予測確率を元に利益の最大となるコンタクト数を選択
- シナリオ 2：顧客リストを増加させた場合の収益の変化を追跡

予測モデルを使用したシナリオ 1 を採用してコンタクト数を減らすことでも費用対効果は高くなるが、シナリオ 2 の方がインパクトは大きいことから、今後引き続き顧客リストを増やすことが望まれることがわかりました。

報告内容に了承が得られれば、モデルを展開して施策を実施し、その効果を検証します。

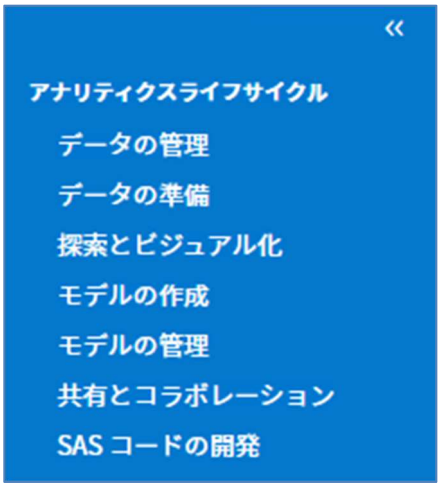


実際に分析ツールを使用して、モデルのスコアリングを行います。

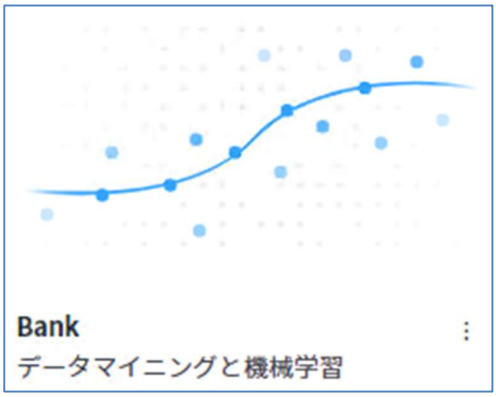
モデルのスコアリング

このデモでは、構築したモデルを新しいデータに適用して、口座開設の見込みのある将来的な顧客をスコアリングします。

1. 画面左上の三本線のボタンをクリックして、**モデルの作成**を選択します。

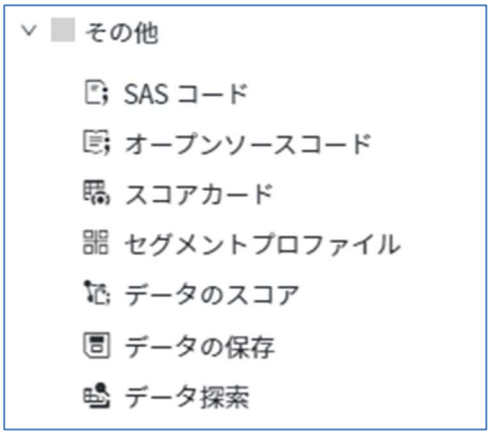


2. **Bank** プロジェクトを選択して開きます。

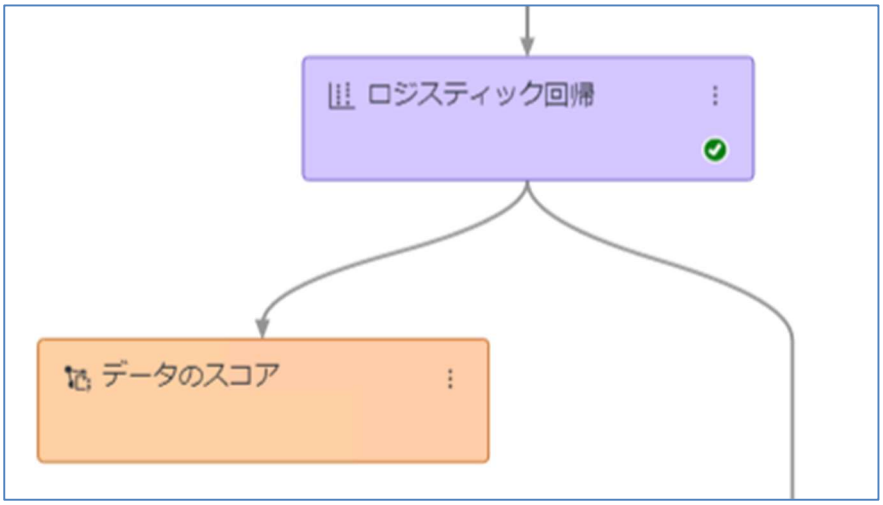


3. **パイプライン**タブをクリックして、「パイプライン」タブを選択します。

4. 画面左手からノードボタンをクリックして、「その他」を展開します。



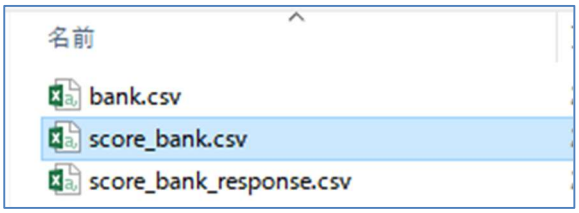
5. **データのスコア**ノードをロジスティック回帰ノードにドラッグアンドドロップします。



6. データのスコアノードを選択して、画面右手のテーブル名から**参照**をクリックします。

7. インポートタブをクリックして、ローカルファイル⇒ローカルファイルを選択します。

8. 講師の指示に従ってファイルの場所へ移動して、score_bank.csv を選択して**開く**をクリックします。

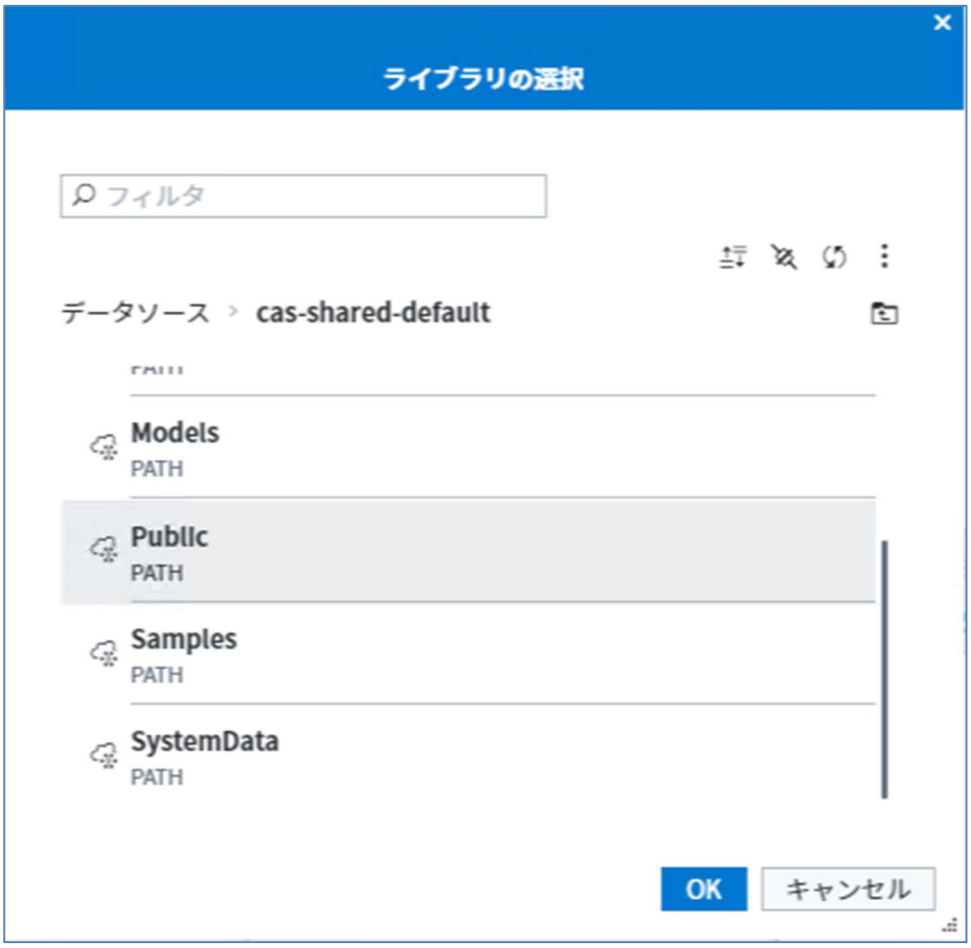


9. 「ファイルの区切り文字を入力：」セクションから、**カスタム**を選択して、区切り文字に**;**を入力します。

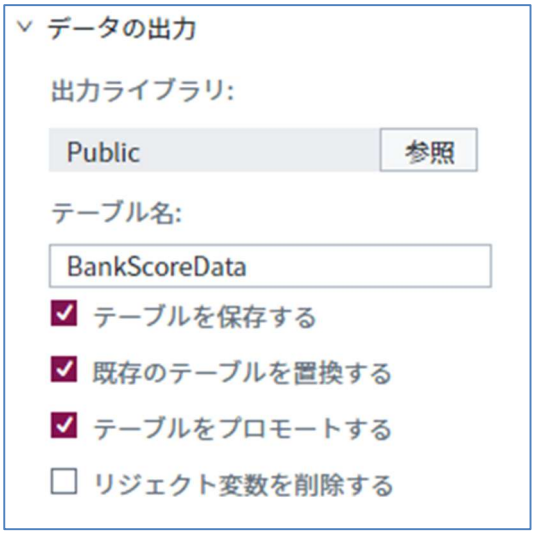
アイテムのインポートをクリックして、「テーブルが正常にインポートされ使用できます」のメッセージが現れたら、画面右下の **OK** ボタンをクリックします。

10. 画面右手の出力ライブラリから、**参照**をクリックします。

11. **cas-shared-default** をダブルクリックして、**Public** を選択して **OK** をクリックします。



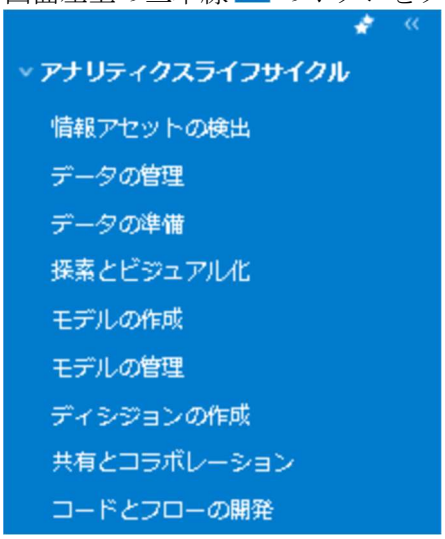
12. テーブル名を「BankScoreData」に変更し、「既存のテーブルを置換する」と「テーブルをプロモートする」にチェックを入れます。



※既存のテーブルを置換する：既に存在する同名のテーブルを置き換えます。
※テーブルをプロモートする：他のセッションからのデータアクセスを可能にします。

13. パイプラインの右上から、**パイプラインの実行**をクリックします。

14. 画面左上の三本線 のボタンをクリックして、**データの管理**を選択します。



15. **BANKSCOREDATA** を右クリックして、**テーブルのダウンロード**を選択します。

16. **ダウンロード**をクリックします。

17. 画面右上の table_BANKSCOREDATA.csv ファイルをクリックして開いてください。

18. 開いたファイルから、スコアリングによって追加された項目を確認します。

I_deposit	P_deposityes	P_depositno	EM_EVENTPROBABILITY	EM_CLASSIFICATION	EM_PROBABILITY
no	0.106116083	0.893883917	0.106116083	no	0.893883917
yes	0.848411341	0.151588659	0.848411341	yes	0.848411341
no	0.238864648	0.761135352	0.238864648	no	0.761135352
no	0.268825311	0.731174689	0.268825311	no	0.731174689
no	0.205795431	0.794204569	0.205795431	no	0.794204569
no	0.069262006	0.930737994	0.069262006	no	0.930737994
no	0.137559319	0.862440681	0.137559319	no	0.862440681
no	0.46178135	0.53821865	0.46178135	no	0.53821865
no	0.205795431	0.794204569	0.205795431	no	0.794204569
yes	0.890384317	0.109615683	0.890384317	yes	0.890384317
no	0.32651293	0.67348707	0.32651293	no	0.67348707
no	0.184861123	0.815138877	0.184861123	no	0.815138877
no	0.313306882	0.686693118	0.313306882	no	0.686693118
no	0.422749409	0.577250591	0.422749409	no	0.577250591
yes	0.786309563	0.213690437	0.786309563	yes	0.786309563

スコアリングした結果を確認できます。このファイルから対象顧客を選択し施策を実施することができます。

End of Demonstration

Appendix A　用語集

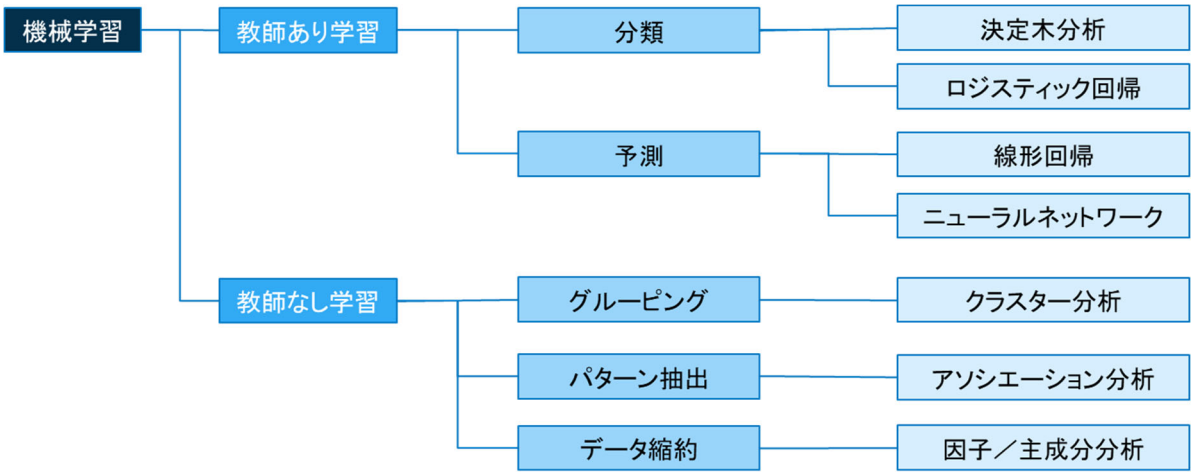
A.1 用語集

※Lesson 番号は、用語が初めて登場する場所を指しています。

【Lesson 1】

教師あり学習

教師あり学習（Supervised Learning）は機械学習の手法の一つで、あらかじめ用意された正解（ラベル）に基づき、データから機械が学習して値の予測を行う手法。これに対して教師なし学習（Unsupervied Learning）は、既知の正解（ラベル）を与えずに、データから機械が学習してパターンを推察する手法。機械学習は、データからルールやパラーンを機械が見つける技術で、大きく教師ありと教師なしに分けられ、実に様々な分析手法がある。



【Lesson 2】

特徴量

二次元のデータテーブルにおける列（変数）を指す。機械学習の教師あり学習において、変数は目的変数、説明変数に分けられる。目的変数は予測の対象となる列で、説明変数（特徴量）は予測の根拠となる列にあたる。また、特徴量の数を次元と呼ぶ。

量的変数・質的変数

「量的変数」とは、値の計測が可能な変数のことで、年齢や収入、人数などがこれにあたり、「間隔尺度の変数」や「連続量の変数」、「数値変数」と呼ばれることもある。対して「質的変数」は、例えば、性別、職業、配偶者の有無などがあたり、「名義尺度の変数」や「カテゴリ変数」、「文字変数」と呼ばれることもある。

オブザベーション

データセットの行（レコード）。

変数

データセットの列（カラム）。

カーディナリティ

列の中に入っている値の種類の数。値の種類の多いものをカーディナリティが高い、少ないものをカーディナリティが低いと言う。また、テーブル結合の場面では、テーブル間の関係性を示す言葉として用いられる。一対一、一対多（多対一）、多対多の関係がある。一対一とは、両方のデータの結合キーの値がすべて重複することなく、1 つの種類の値が 1 行にだけ保存されている状態を指す。一対多とは、片方のテーブルの結合キーの値の種類は重複せずに保存されていて、もう一方のテーブルの結合キーの値は重複して存在し、1 つの種類の値が複数の行に保存されている状態を指す。多対多とは、両方のテーブルで結合キーの値が重複して保存されている状態を指す。

質的変数のダミー化

質的変数を量的変数に変換すること。代表的な方法に One-Hot エンコーディングがあげられる。One-Hot エンコーディングでは、対象の列を複数の列に分割して 0 と 1 の値を割り当てる。例えば、性別の列を男性・女性の二つの列に分けて、男性の列では男性の値に 1、その他の値に 0 を割り当てる。女性の列で同様である。

特徴量のスケーリング

特徴量間の単位の違いや値が極端に違う場合にそれらの尺度を整えること。正規化、標準化といった方法があり、標準化は平均を 0，分散を 1 とする方法で、正規化は最小値を 0，最大値を 1 とする方法。

構造化データ

列の定義を持つデータ。データベースが代表的だが、CSV、Excel ファイルも含まれる。

非構造化データ

列の定義を持たないデータ。XML、JSON、PDF、音声、画像、映像などのファイル。

特徴量エンジニアリング

機械学習モデルを構築するために、収集されたデータを適切な特徴量に変換すること。例えば、新しい特徴量の作成、集計方法の変更、特徴量の変換、値の補完、特徴量の選択などがあげられる。

SAS データセット

SAS 形式のファイル。二次元の表（テーブル）構造を持ち、ファイルシステム上（フォルダ、ディレクトリ）に直接的に保存できる。拡張子は、sas7bdat。

SAS ライブラリ

SAS で扱うことのできるデータファイルなどが保存されている場所を指す。SAS データに対してはフォルダやディレクトリが、データベースではスキーマがそれにあたり、その場所をライブラリと呼ぶ。また、SAS Viya では、CAS サーバー上のインメモリ空間におけるデータの論理的な保存場所を CAS ライブラリ（caslib）と呼ぶ。

PUBLIC ライブラリ

SAS 上のデフォルトの、共有可能、書き込み可能なデータの保存場所。デフォルトでは、ファイルがパブリック caslib にインポートされている場合、インポートされたすべてのファイルに誰でもアクセスできる。

WORK ライブラリ

SAS でデータを保存する先として用意されている、一時的な作業領域。一時ライブラリと呼ばれ、WORK ライブラリに保存されているデータは、利用しているアプリケーション（セッション）を終了すると、自動的に削除される。WORK 以外のライブラリは、永久ライブラリと呼び、アプリケーションを終了しても、データは自動的に削除されることはない。

ライブラリの名前

ライブラリの名前は、ライブラリ参照名と呼ばれ、データの保存先である SAS ライブラリへのショートカット名である。命名規則として、半角の英字・数字。アンダースコアで記述し、先頭文字の数字の利用は禁止され、8 文字以内で記述する。

結合キー

テーブルとテーブルを水平方向に結合する際に使用する。どの行同士を結合するかの基準となる値が入っている列。

内部結合

テーブルの結合の際、結合キーの一致した行のみを結果として返す結合。

左外部結合

テーブルの結合の際、結合キーの一致した行と、左側に配置したテーブルの結合キーにのみ存在する値の行を結果として返す結合。

右外部結合

テーブルの結合の際、結合キーの一致した行と、右側に配置したテーブルの結合キーにのみ存在する値の行を結果として返す結合。

完全外部結合

テーブルの結合の際、結合キーの一致した行と、左側、右側それぞれのテーブルの結合キーにのみ存在する値の行を結果として返す結合。

【Lesson 3】

モデリング、スコアリング

過去のデータから機械学習モデルを構築することをモデリングと呼び、作成されたモデルを新しいデータに割り当てて未知の値を予測することをスコアリングと呼ぶ。

データの分割（学習、検証、テストデータ）

モデリングの際に、すべてのケースを使ってモデルを作成するのではなく、データを学習と検証の 2 分割、もしくは学習、検証、テストの 3 分割に分け、モデルの作成と評価を別のデータで行うことをホールドアウト法と呼ぶ。2 分割の場合は、学習データでモデルを作成し、検証データでモデルのチューニングと評価を行い、3 分割の場合は、学習データでモデルを作成し、検証データでモデルのチューニング、テストデータで評価を行う。

過学習、学習不足

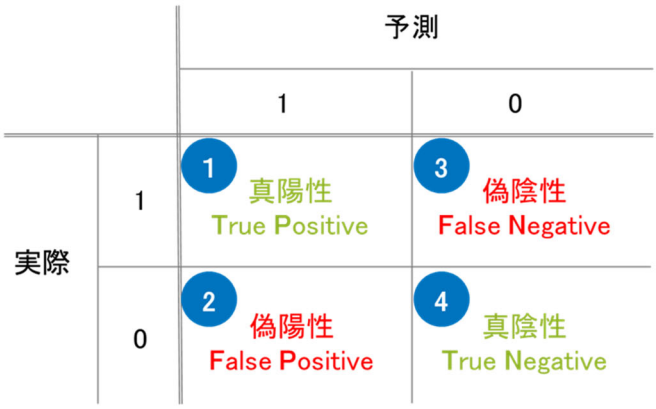
学習データに過度に当てはまりの良い（複雑な）モデルを作成すると、新しいデータへの適合（一般化、汎化）がうまく行かない場合があり、これを過学習（オーバーフィッティング）と呼ぶ。逆に学習データの特徴をとらえることのできてない（単純な）モデルは、学習不足（アンダーフィッティング）と呼ばれる。

イベントベースサンプリング

予測対象のイベントの割合が極端に異なる場合、例えばデータ全体で正常が 1%、異常が 99%を占めるような、不均衡なデータの場合は予測モデルの性能が上がらないことがある。その場合に、イベントの割合を 50:50 に調整してサンプリングする方法をイベントベースサンプリングと呼ぶ。

混同行列

下図の通り、2 値分類の問題においてモデルによる予測と実際との違いを表すもの。例えば、予測結果が 1 の場合に実際が 1 なら真陽性（①）、0 の場合は偽陽性（②）、と呼ぶ。モデルの精度を評価する指標は、正解率、混合率、再現率、特異度、F 値がある。



正解率
＝{❶+❹}÷{❶+❷+❸+❹}

適合率
＝{❶}÷{❶+❷+③+④}

再現率（感度）
＝{❶}÷{❶+②+❸+④}

特異度
＝{❹}÷{①+❷+③+❹}

F値 ＝{2×適合率×再現率}
÷{適合率＋再現率}

実践! ビジネス課題へのアナリティクス活用基礎講座

2025年7月23日　初版 第1刷

発行元　SAS Institute Japan株式会社

〒106-6661　東京都港区六本木6-10-1 六本木ヒルズ森タワー11F
TEL 03(6434)3690