用語集（SAS用語-1）

用語	説明
アセスメント	SAS Enterprise Minerの評価機能。モデルを比較し効果的なモデルを特定することができます。
インポート	他のアプリケーションで作成した形式の異なるデータやファイル等を変換して読み込むことです。
オブザベーション	行、レコードのことです。
解析	テキストを成分語、フレーズ、マルチワード語、句読点、およびその他のタイプの情報に分割する目的でテキストを分析することです。
カテゴリ変数	1つ1つのデータを区別・分類するために用いられている変数のこと。それぞれのデータに大小関係や優劣はなく、単純にデータを分類するために用いられます。
クエリビルダー	クエリとは、1つ以上のデータソースからデータを取得するリクエストで、クエリを使用してデータを操作する方法のことです。
クレンジング	データ分析しやすいデータに整形することです。
欠損値	無回答や非該当など集計から除去する値のことです。
正規分布	確率分布の一種で、データの分布が平均値を頂点とした左右対称の山形で表示されるもの。平均±標準偏差の範囲に全体の約68%、平均±標準偏差×2の範囲に約95%、平均±標準偏差×3の範囲に約99%が含まれる等の特長があります。
精度	精密さの度合です。
選択ペイン	SAS Enterprise Guideでは各タスクウィンドウの左側には、選択ペインがあります。選択ペインを使用して一連のオプションを選択することが出来ます。すべてのタスクには[データ]領域があり、ここでタスクの役割に変数を割り当てます。
タスク	SAS Enterprise Guideの機能でタスクの役割ボックスがあります。そのタスクで変数を割り当てられる役割がリストされます。タスクではデータの操作、分析プロシジャの実行、レポートの作成といったすべての作業が行われます。また、タスクの多くは、ウィザードとして使用できます。

用語集（SAS用語-2）

用語	説明
ターゲット変数	モデルを作成するには、予測に使用可能な履歴イベントと特性を表す入力データが必要です。その予測対象のイベントまたは値を表す変数をターゲットデータ（変数）と呼び必要になります。
トランザクションデータ	業務に伴って発生した出来事の詳細を記録したデータのこと。登録や変更、削除等の手続き処理で蓄積されていくもの。（例：受注データ・履歴データなど）
ツリーノード	SAS Enterprise Minerにおいて、一連のノードをツリーとして表示します。
外れ値	統計において他の値から大きく外れた値である。 測定ミス・記録ミス等に起因する異常値とは概念的には異なるが、実用上は区別できないこともある。
ハンドリング	取り扱い、処理、操作、対処、対応などの意味を持つ英単語。ソフトウェアやプログラミングなどの分野で、特定の状況や対象について、対応する処理を行うことをハンドリングという。
変数	SASデータセットまたはSASデータビュー内の列。各変数のデータ値は、すべてのオブザベーションの単一の特性を表します。各SAS変数は、名前、データタイプ(文字または数値)、長さ、出力形式、入力形式、ラベルという属性を持ちます。
モデル	入力から出力を計算する公式またはアルゴリズムです。データマイニングモデルには、入力変数が与えられた場合、ターゲット変数の条件付き分布に関する情報が含まれています。
ライブラリ参照名	SASライブラリに一時的に関連付けられる名前。SASファイルの完全名は、ピリオドで区切られた2つの語から構成されます。最初の語はライブラリ参照名であり、これはライブラリを表します。2番目の語は、特定のSASファイルの名前になります。たとえば、VLIB.NEWBDAYの場合、ライブラリ参照名VLIBは、ファイルNEWBDAYが格納されているライブラリを表しています。ライブラリ参照名を割り当てるには、LIBNAMEステートメントを使用するか、またはオペレーティングシステムのコマンドを使用します。
ラベル	変数より細かい説明を記述する目的として使われるもの。SAS変数名とは別にラベルを指定することができる。SAS変数名はアルファベット、数字、アンダーバーで構成しなければいけない一方で、ラベルは日本語も使用可能です。

用語集（SAS用語-3）

用語	説明
連続変数	値として表すことができ、四則演算が可能で連続的な値をとる変数のこと（例：長さ、時間、温度など）。
SASデータセット	SAS固有のいずれかのファイル形式で内容が格納されたファイル。SASデータセットには次の2種類があります。SASデータファイルとSASデータビューです。SASデータファイルは、データ値に加えて、そのデータに関連付けられているディスクリプタ情報を含みます。SASデータビューには、ディスクリプタ情報と、他のSASデータセットまたはソフトウェアベンダのファイル形式で格納されたファイルからデータ値を取り出すために必要となるその他の情報のみが含まれます。
SASライブラリ	SASデータセットが集まったデータの貯蔵庫のようなものです。SASライブラリに対応するのはWINDOWSのフォルダになります。
WORKライブラリ	一時データライブラリのこと。WORKライブラリに保存したSASデータセットは、SASを一旦終了すると消去される。SAS終了後も保存しておきたいデータセットは、WORK以外のライブラリ（永久データライブラリ）に保存する必要があります。

用語集（統計用語-1）

用語	説明
因子分析	因子は、ある結果をひき起こすもとになる要素であり、因子分析はある観測された変数が、どのような潜在的な変数から影響を受けているかを探る手法で、多変量解析の手法の一つです。
オッズ比	2つの異なる群においてある事象が起こる確率をそれぞれ、としたとき、2つの群のオッズの比をオッズ比、あるいは見込み比と言います。
オーバーフィッティング	過剰適合と同意となります。
過剰適合	統計学や機械学習において、訓練データに対して学習されているが、未知データ（テストデータ）に対しては適合できていない、汎化できていない状態を指します。
カイ2乗検定	帰無仮説が正しければ検定統計量が漸近的にカイ二乗分布に従うような統計学的検定法の総称。カイ二乗分布は正規分布に従いくつかの変数があるとき、それらの二乗和が従う分布のことです。
回帰分析 回帰モデル	結果の数値と、その要因の数値から、それぞれの関係を予測する分析手法のこと。比較的容易な分析手法で、ひとつの要因から結果を予測する「単回帰分析」と、複数の要因からひとつの結果を求める「重回帰分析」がある。 たとえば、既存顧客に新たなダイレクトメールを送る際、過去のダイレクトメールへの反応履歴（結果）と、送付対象者の年齢、収入、住居地域、購入金額、購入履歴など複数の顧客属性（要因）との因果関係を重回帰分析することで今回発送するダイレクトメールの反応数を推定することができます。なお、結果の数値は「目的変数」や「従属変数」と呼ばれ、要因となる数値は「説明変数」と呼ばれる。回帰モデルには線形（一般化線形モデル、一般線形モデルなど）と非線形もでるがあります。
期待信頼度	全てのデータの中で結果（Bを買う）の割合。B単独の人気を判断します。
帰無仮説	統計的仮説検定の際にとりあえず立てる仮説のことで、対立仮説の方が重要であることが多いです。 例えば、帰無仮説として「差がない」という仮説が立てられた場合、これが棄却されることにより、対立仮説の「差がある」を結論とします。
記述統計	データを整理し、そのデータの持つ特徴をできるだけ簡潔で明確に記述する方法を研究するもので推測統計と対比して用いられることが多い。数値や表、グラフ、図などを用いてデータの特徴を表現します。

用語集（統計用語-2）

用語	説明
擬似相関	2つの事象に因果関係がないのに、見えない要因（潜伏変数）によって因果関係があるかのように推測されることです。
クラスター	英語で「集団」「群れ」のことで、似た性質のものが集まっている様子を表す。クラスター分析とは、異なる性質のものが混ざり合った集団から、互いに似た性質を持つものを集め、クラスターを作る方法です。
クロス集計	質問項目を2つ以上かけ合わせて集計する手法です。
交互作用	2つの因子が組み合わさることで初めて現れる相乗効果のことです。
誤分類率	測定誤差により、優良部品が不適合と分類され、不良部品が適合と分類される場合があります。これを誤分類といいます。
信頼区間	区間推定において、ある確率（信頼係数）のもとで母数がその内に含まれると推定された区間のことで信頼限界とも言います。 信頼区間は主に95%信頼区間として使用されます。
重回帰分析	重回帰分析とは、ある結果（目的変数）を説明する際に、関連する複数の要因（説明変数）のうち、どの変数がどの程度、結果を左右しているのかを関数の形で数値化し両者の関係を表し、それを元にして将来の予測を行う統計手法のことです。
説明変数	因果関係における原因、関数における入力、y=f(x)のxを説明変数と言い、独立変数とも言います。
相関係数	2つの確率変数間の相関（類似性の度合い）を示す統計学的な指標のこと。Aの値とそれに対応するBの値をグラフ化した場合に、右上がりの直線となるものを「正の相関」、右下がりの直線となるものを「負の相関」と呼ぶ。数値は−1から+1の間となり、数値が0に近づくほど相関関係が希薄になる。数値が0の場合は「相関がない」、つまりAの数値が変化してもBの数値に影響がないということになる ただし、相関係数は順序尺度であり、間隔尺度ではない。このため相関係数を単純に比較することは意味がない。 たとえば自動車では、搭載するガソリンの量が多いほど走行可能距離が伸びる「正の相関」、走行距離が伸びるほどガソリンは減るので「負の相関」となる。なお、基本的に単位はつけない。相関係数を把握することで、Aの数値によってBの数値を予測することができます。

用語集（統計用語-3）

用語	説明
対数変換	対数正規分布に従う変数の対数をとり，正規分布に従う変数を作ることです。
多重共線性	重回帰分析において、いくつかの説明変数間で相関関係が認められる場合、多重共線性があるという。多重共線性が認められると、回帰係数の結果が不安定となり、一般的には解析結果として妥当なものは得られません。
多変量解析	多数の変数間の相互の関係性をとらえるために使われる統計的手法の総称。重回帰分析、判別分析、因子分析、クラスター分析など多岐に渡る分析手法があります。
単回帰分析	回帰分析のうち、単回帰分析というのは1つの目的変数を1つの説明変数で予測するもので、その2変量の間の関係性をY＝aX＋bという一次方程式の形で表します。
ツリー分析（決定木）	観察対象データの集団を、従属変数（結果：購買の有無、解約の有無など）に対し最も効率よく分類できる独立変数（原因）によって次々と分割し、木の枝のように分岐・整理していく分析手法。データの集団を効率よく分類・整理し、ルール抽出・生成や予測モデル構築などに利用される。たとえば、商品を購入する／しないに最も強く影響する要素を探る際などに用いられる。 マーケティング分野では、最も高い反応が期待できる顧客グループに対して販促計画を練るなどの目的で使われる。 ツリー分析では、視覚的に分析結果を把握できるとともに計算方法が比較的簡単で、モデル作成しやすいことが特徴とされます。
ヒストグラム	統計で、度数分布を示すグラフの一つです。横軸に階級、縦軸に度数をとり、各階級の度数を長方形の柱で示します。 （柱状グラフ）
標準偏差	分布の拡がりを表す統計量の一つで、分散の正の平方根に等しいことになります。
平均平方誤差（ASE）	平均平方誤差(ASE)は、平方誤差(SSE)の合計をオブザベーションの数で除算したものです。小さな値ほど、望ましい値です。
変数変換	統計手法の多くは（1）母分布が正規分布であること，（2）各群の分散が等しい ことを仮定している。対数正規分布，二項分布，ポアソン分布などに従う変数はいずれの条件も満たさないので生のデータを使用して分析ができない。このような変数に数学的な変換をほどこし，新しく得られる変数が条件を満たすようにするための方法です。その具体的な方法のひとつが対数変換です。

用語集（統計用語-4）

用語	説明
変数増減法 （ステップワイズ法）	独立変数の候補から，予測や判別に有用な順に独立変数を採用するための方法。
目的変数	因果関係における結果、関数における出力、y=f(x)のyを目的変数と言い、従属変数や被説明変数とも言います。
有意水準	統計的仮説検定において第一種の過誤を犯す確率のことで、P値の小ささの基準である。P値が有意水準よりも小さい場合は帰無仮説は棄却される。α「（アルファ）」として表され、一般的にα = 0.05か α = 0.01と設定されることが多いです。
要約統計量	基本的なデータ特性を表す統計値。平均値や最大、最小値、標準偏差などがある。基本統計量とも呼ばれます。
リフト値	前提（Aを買う）が起きた場合に結果（Bを買う）が起きる割合は、全てのデータの中で結果（Bを買う）の割合よりどれだけ多いかを倍率で示したもの。リフト値が低ければ、商品Bは単独（の理由）で売れており、商品Aの商品との関連性よりも商品B特有の理由で売れていると考えられる。
連続変数	身長や体重のように，精度の高い測定法によればいくらでも正確な値が得られるデータ。実際は離散量であるが連続量として取り扱ってもかまわないようなものもあります。
ロジスティック回帰分析	見込み顧客が製品を買ってくれるかどうか、キャンペーンに反応するかどうか、など 将来のYES/NO を予測するときに使える手法です。
P値	統計的仮説検定において、帰無仮説の元で検定統計量がその値となる確率のこと。P値が小さいほど、検定統計量がその値となることはあまり起こりえないことを意味する。 一般的にP値が5%または1%以下の場合に帰無仮説を偽として棄却し、対立仮説を採択する。
ROC (Receiver Operating Characteristic)曲線	二値変数（YES/NO　例. 実際に購買した／しなかった）と連続変数（例. 購買可能性%予測値）との関係の強さを評価する方法。例えば連続変数のあるカットオフの値を設定し、それ以上をYES=購買する、それ未満をNO=購買しない、と予測した場合の陽性率（予測=YES、実際=YES）、偽陽性率（予測=YES、実際=NO）を取得する。カットオフの値を動かすことで陽性率、偽陽性率がどのように変化するかをグラフ上に曲線として表現し、その曲線で連続変数と二値変数の関係の強さを評価する。縦軸に感度を、横軸に1−特異性をとった場合に、曲線の左上方向へのふくらみ（ROC曲線下面積）が大きいほど、変数間の関係が強いと判断できる。 ROC曲線下面積は、SAS Enterprise Guideでロジスティック回帰を実行した際のc統計量と一致する。
Waldによる信頼区間	ロジスティック回帰分析などで、推定された偏回帰係数の有意性を確認するために用いられる検定の一つです。偏回帰係数の推定値を標準誤差で割ったものを2乗した値が自由度1のカイ二乗分布に従うことを用いて、帰無仮説「偏回帰係数は0である」のもとで検定を行います。