用語集（SAS用語-1）

SAS用語	説明
オブザベーション	行、レコードのこと。
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。
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。
クエリビルダ	Enterprise Guideにおいて、データセット同士の結合や加工、抽出、並べ替えなどを行うタスクのこと。
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。
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。
精度	精密さの度合。
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。
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。
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。
プロジェクト	Enterprise Guideにおいて、データやタスク、プログラム、結果、操作などを保存するファイルのこと。

用語集（SAS用語-2）

SAS用語	説明
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。
ノード	Enterprise Minerにおいて、ダイアグラム上で使用するデータやデータ加工・分析を行うツールのこと。ノード間を矢印でつなぎ、プロセスを作成する。
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。変数名とは別にラベルを指定することができ、日本語も使用可能。
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。

用語集（統計用語-1）

統計用語	説明
因子	ある結果をひき起こすもとになる要素。現象の要因を構成している作用素または力。
オッズ比	ある事象の起こりやすさを2つの群で比較して示す統計学的な尺度のこと。2つの群にどれくらい特定の要因があったかを比較する。オッズ比が1の場合、2つの群の間に差がないと判断する。
回帰分析	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。
帰無仮説（きむかせつ）	「ある仮説」に対して、それが正しいのか否かを統計学的に検証する仮説のこと。たいていは否定されることを期待して立てられるもの。
交互作用	2つ以上の要因が考えらえる時、要因が組み合わさった時にだけ現れる相乗効果のこと。要因の効果が別の要因によって変化することを指す。
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る。

用語集（統計用語-2）

統計用語	説明
最尤法（さいゆうほう）	統計的推定の際、実際に得られた標本があるとき、それが得られる確率が最大になるような母数の値をその推定値とする手法。
ステップワイズ法	一定の基準で、説明変数を増やしたり減らしたりしながら当てはまりの良いモデルに近づける説明変数の選択方法のこと。
対数変換	対数正規分布（右に裾の長い非対称分布、標準偏差は平均値に比例する）に従う変数の対数を取り、正規分布に従う変数を作ること。
多重共線性 （たじゅうきょうせんせい）	説明変数同士が強く相関してしまっているケースで発生する問題のこと。多重共線性の影響下では、回帰式の信頼性低下や得られた結果が真の値に反するなどの問題が起きる。英語で「multicollinerarity」と言われるため、略して「マルチコ」とも呼ばれる。
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る。
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。

用語集（統計用語-3）

統計用語	説明
Waldによる信頼区間	Wald検定（推定手法）における信頼区間のこと。
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。縦軸に陽性率を、横軸に偽陽性率をとった場合に、曲線の左上方向へのふくらみが大きいほど、変数間の関係が強いと判断できる。