

実務直結! 分析力向上ウェビナーシリーズ
機械学習によるビッグデータ分析の手法

#1 ロジスティック回帰分析

2022年12月01日



Agenda

- グラフによるデータ観察
 - 散布図、散布図行列の活用方法
 - 箱ひげ図の活用方法
- ロジスティック回帰分析
 - 教師あり学習と分類手法
 - ロジスティック回帰の基本とその他手法との違い
 - マーケティングデータを用いてロジスティック回帰により契約確度の高い顧客を予測する
 - 回帰係数により契約の影響因子を探索する

ビッグデータとは

- ビッグデータとは、「**4V**」* の特徴を持つデータ *Veracity を除いて 3V と呼ばれることもある
- つまり、「大量かつ、様々な種類のデータが、次々と蓄積されていること」

Volume

データ量

Velocity

生成速度

Variety

データ種

Veracity

正確性

ビッグデータからの新知識獲得＝「ビッグデータ分析」

- ビッグデータ分析は、膨大なデータ（＝事実、結果）に潜むデータ間の関係性や特徴（＝新知識）をデータドリブンで抽出する試みである
- データから「**一見関連のなさそうな新たな洞察を得る**」手法と言える

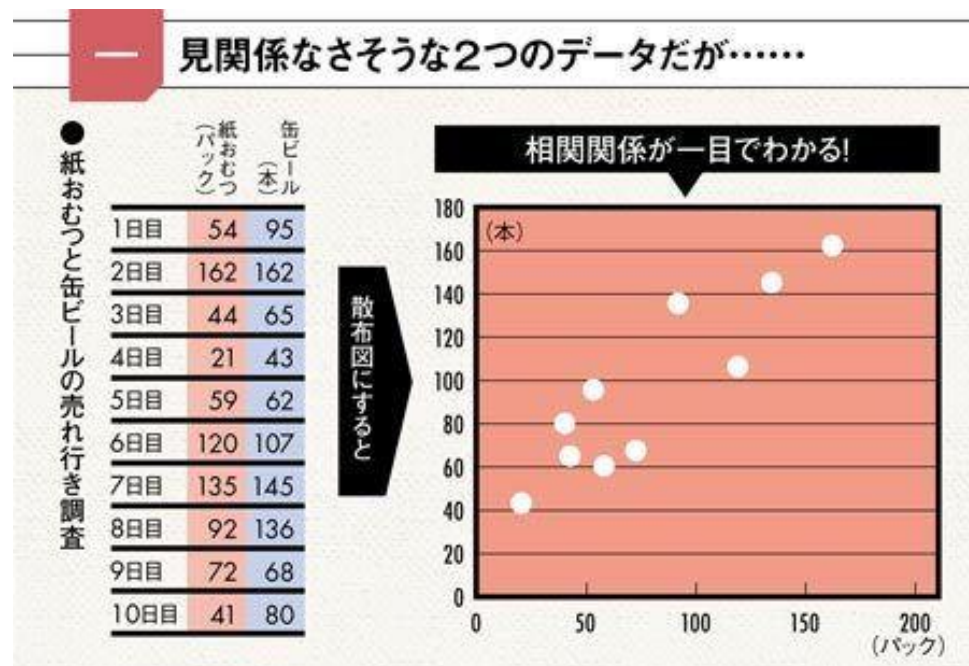
ビール



オムツ



X



Source: <https://president.jp/articles/-/15536>

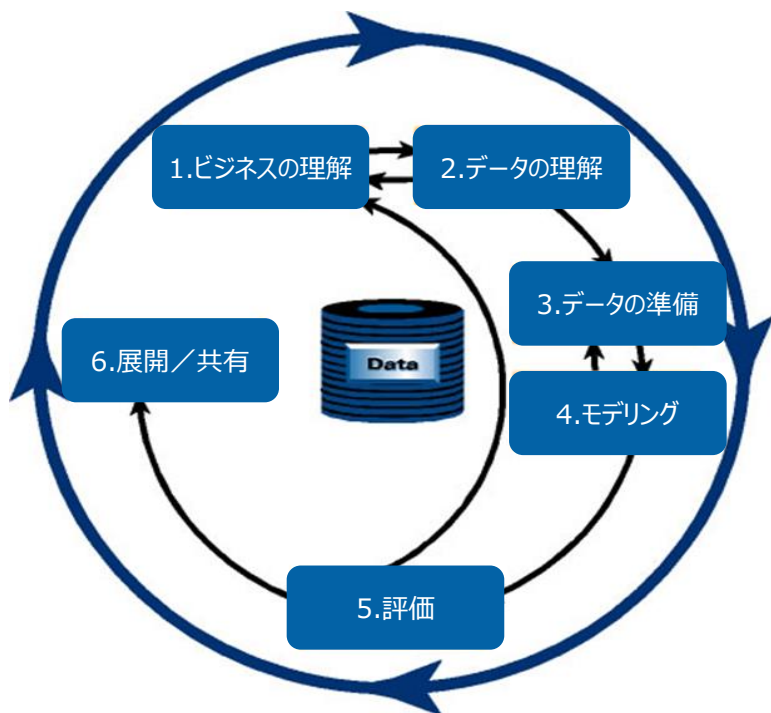


マイニング = 採掘 という意味合いから、「**データマイニング**」とも呼ばれる
(mining)

ビッグデータ分析の進め方

- データマイニングの進め方に関する方法論「**CRISP-DM**」に基づいて、分析と評価を繰り返して試行錯誤しながら進めるのが一般的である

CRISP-DM: データマイニング方法論



(CRoss Industry Standard Process for Data Mining)

1. ビジネスの理解

- ビジネス、データマイニング目標の決定
- プロジェクトの立ち上げ

2. データの理解

- データの収集
- データの調査
- データ品質の検証

**本セミナーでの
主なターゲット**

3. データの準備

- データの選択や除外
- データのクリーニング
- データの構築や統合

4. モデル作成

- モデリング手法の選択
- モデルの作成
- モデルの評価

5. 評価

- データマイニングの結果の評価
- プロセスの見直し
- 実行可能なアクションリストの作成

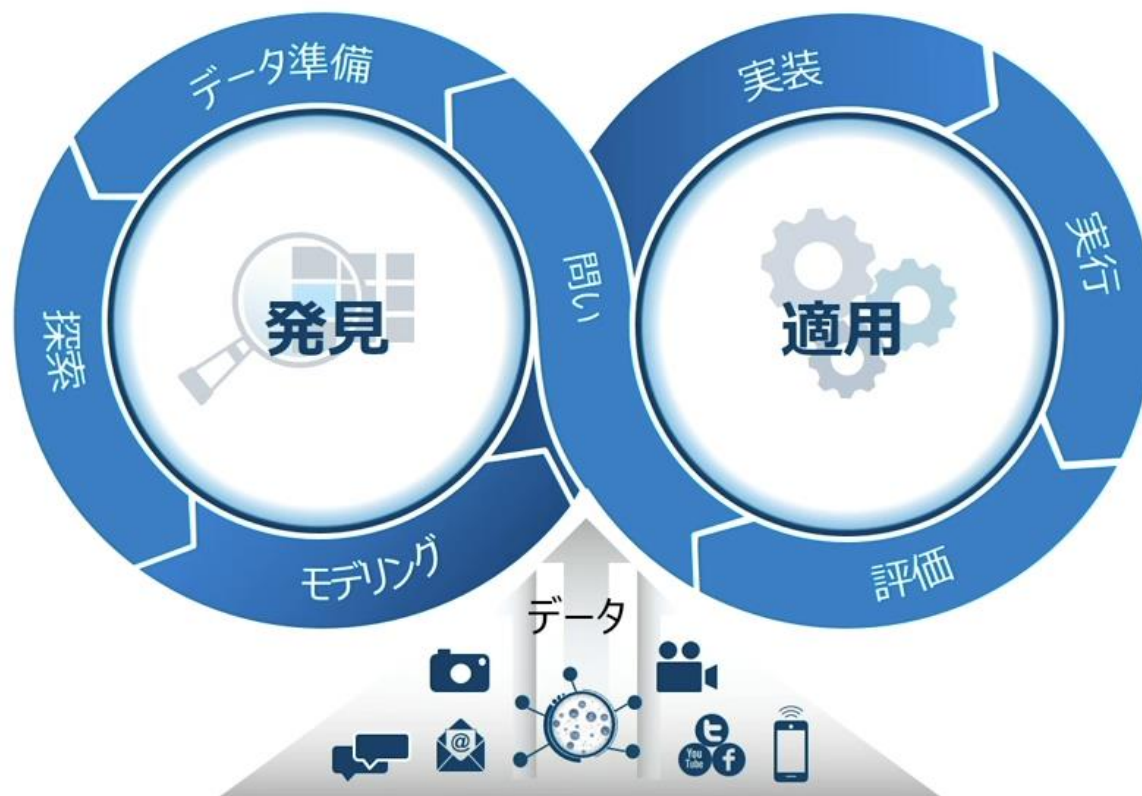
6. 展開／共有

- 業務への導入計画
- モニタリング、メンテナンスの計画

参考：SASの方法論 “SAS Analytics Lifecycle”

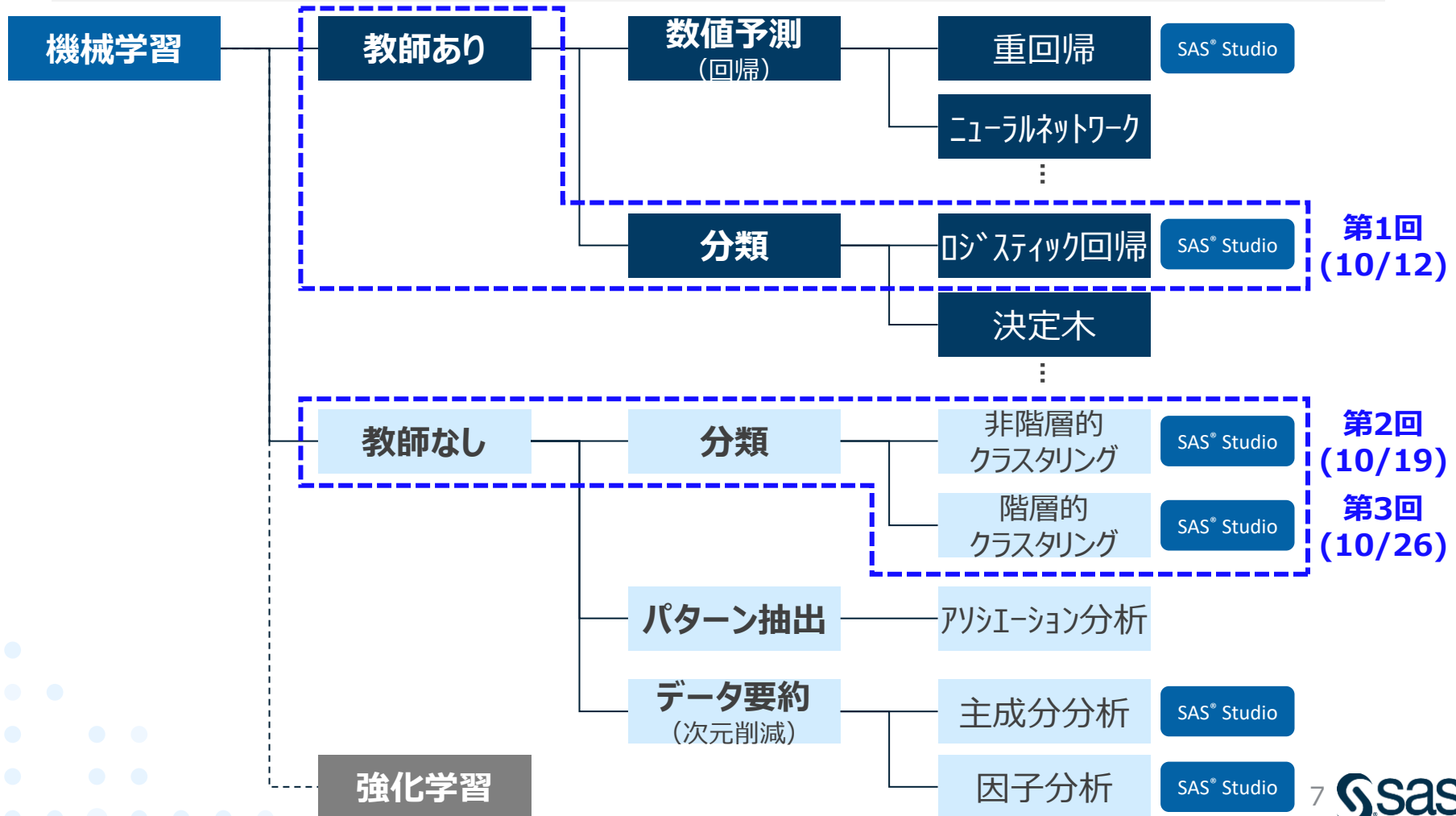
- SASの分析フレームワーク “SAS Analytics Lifecycle” においても、大きな流れは同じであり、データ準備の重要性や、反復的に進めることの必要性を説いている

SAS Analytics Lifecycle



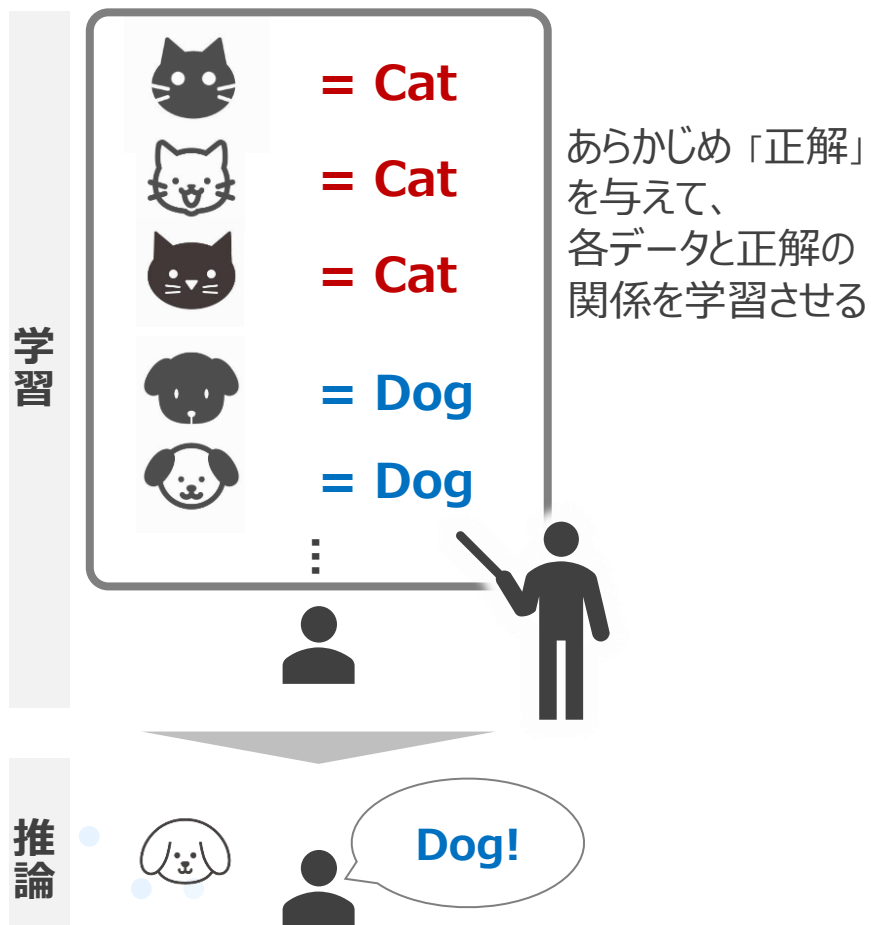
代表的な機械学習手法

- 機械学習手法は、教師あり、教師なし、強化学習に大別される
- なかでも、**教師あり分類**、**教師なし分類**は極めて基本的かつ頻用される手法である



教師あり学習 と 教師なし学習

教師あり学習



教師なし学習



教師あり学習のイメージ（数値予測と分類）

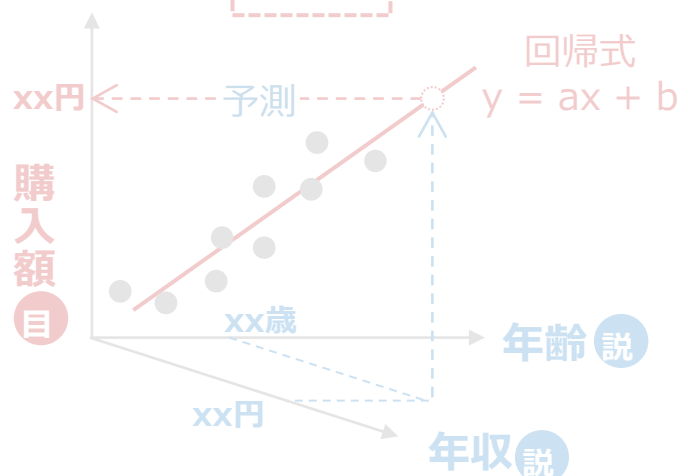
- 各顧客レコードに対して数値 or カテゴリー値（クラス）の解答を与え関係性を学習

数値予測

（回帰）

説明変数 目的変数

顧客ID	名前	年齢	年収	購入額	購入有無	...
0001	xx	25	300万	35,000	購入	...
0002	xx	35	600万	68,000	購入	...
0003	xx	18	120万	0	非購入	...
0004	xx	42	820万	85,000	購入	...
⋮	⋮	⋮	⋮	⋮	⋮	...



- 小売：売上の予測
- 小売：需要予測
- 製造：不良発生率の予測

本日の内容

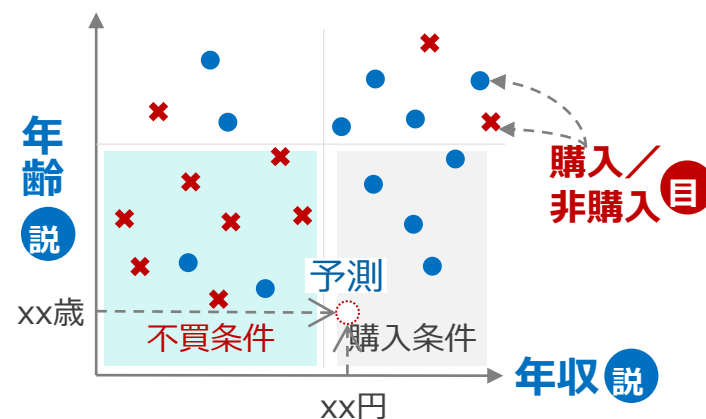
クラス分類

（2クラス or 多クラス分類）

説明変数

目的変数

顧客ID	名前	年齢	年収	購入額	購入有無	...
0001	xx	25	300万	35,000	購入	...
0002	xx	35	600万	68,000	購入	...
0003	xx	18	120万	0	非購入	...
0004	xx	42	820万	85,000	購入	...
⋮	⋮	⋮	⋮	⋮	⋮	...



- 小売：購入／非購入顧客の分類
- 医療：生存条件の分析
- 製造：故障種類の分類

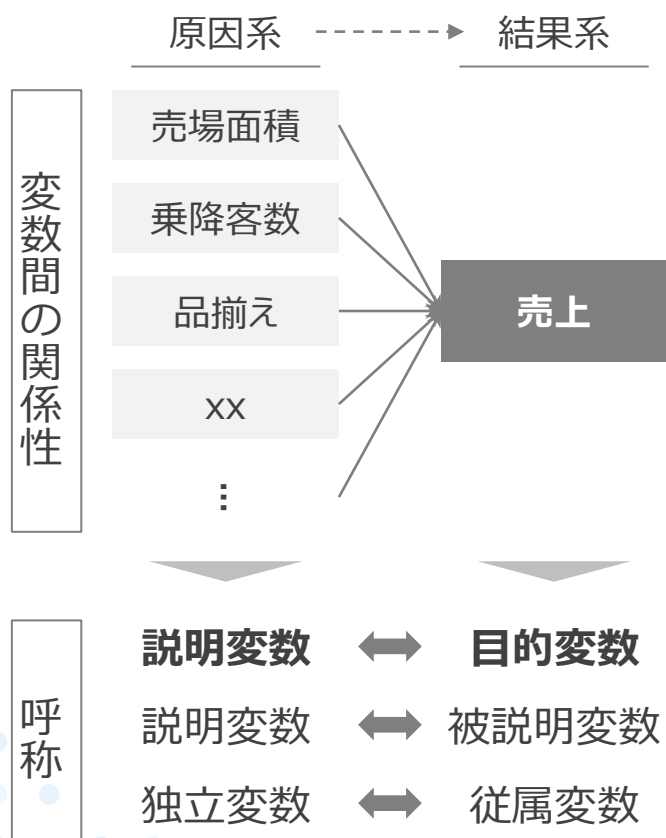
データ例

イメージ

分析例

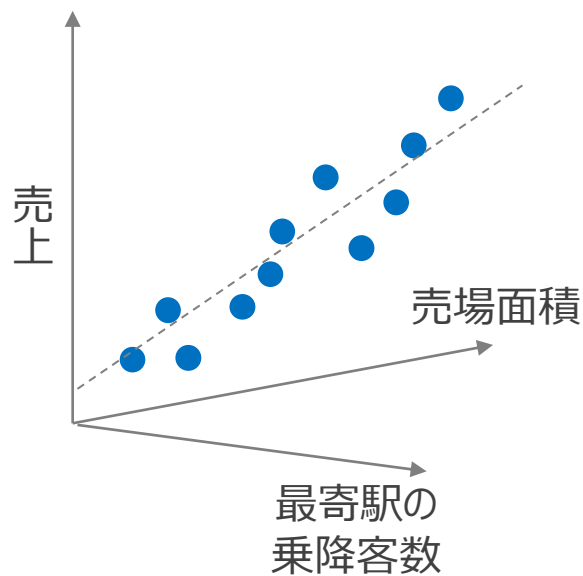
参考：説明変数と目的変数

- 原因となる変数のことを「**説明変数**」、その原因を受けて変動する結果系の変数のことを「**目的変数**」と呼ぶ
- 書籍や講師、ツールなどによって呼称が異なる場合もあるため、慣れておく必要がある



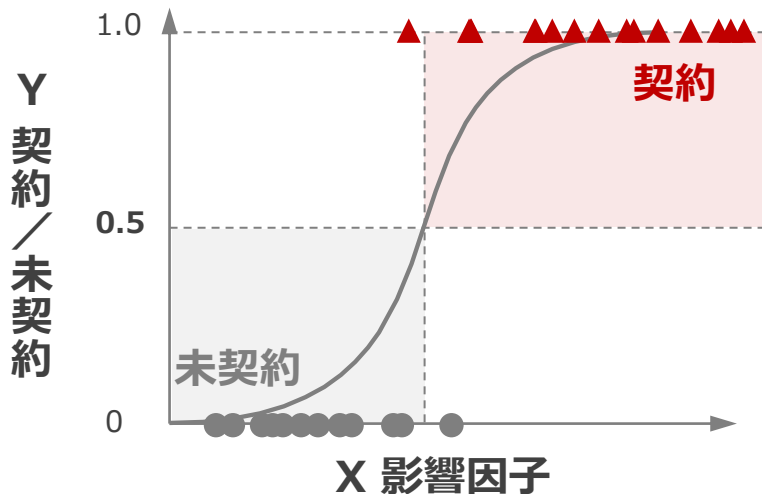
重回帰の場合：

$$\text{売上} = \alpha + \beta_1 \cdot \text{売場面積} + \beta_2 \cdot \text{乗降客数} + \dots$$



ロジスティック回帰分析の基本

- ロジスティック回帰は、回帰させる関数にS字カーブ状のロジット関数を用いることで、**二クラス分類**（購入する／しない、良品／不良品、生／死 など）を可能とする回帰



オッズ = 発生しない確率 (1-p) に対する発生する確率 (p) の比

$$\ln \left[\frac{P}{1-P} \right] = a_1 X_1 + a_2 X_2 + \dots + a_n X_n + b$$

ロジット

P: ある事象の発生確率

※ 目的変数にロジット関数を用いる以外は重回帰と同様の式

ロジスティック回帰のバリエーション

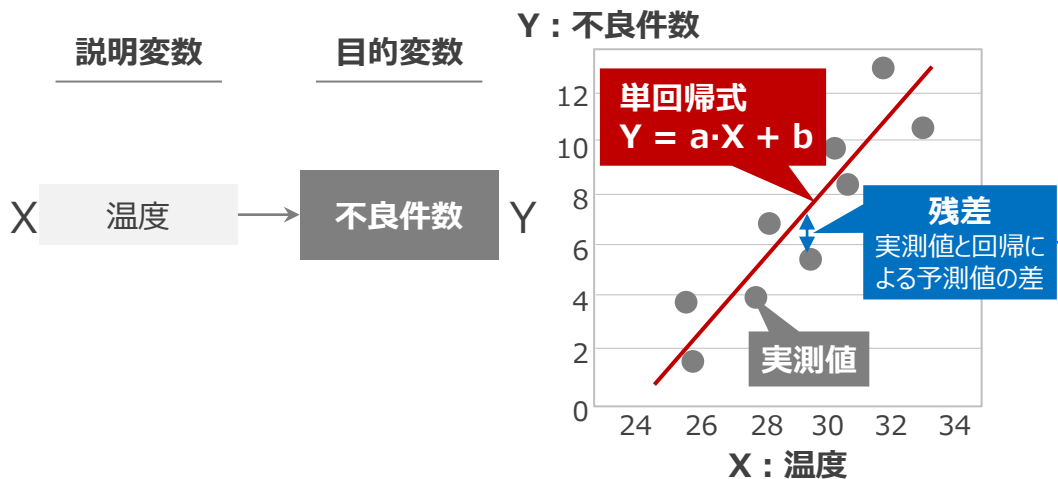
順序ロジスティック回帰：順序尺度（1,2,3 / 大,中,小 など）の予測に用いる

多項ロジスティック回帰：3クラス以上の名義尺度（犬,猫,人 / 病気の種類 など）の予測に用いる

参考：重回帰モデル

- 目的変数Yとその影響因子である説明変数Xとで関数 ($Y=f(X)$) に当てはめることを**回帰**という
- 特に $f(X)$ が線形関数の場合、**線形回帰**と呼び、1変数では**単回帰**、多変数では**重回帰**という

単回帰

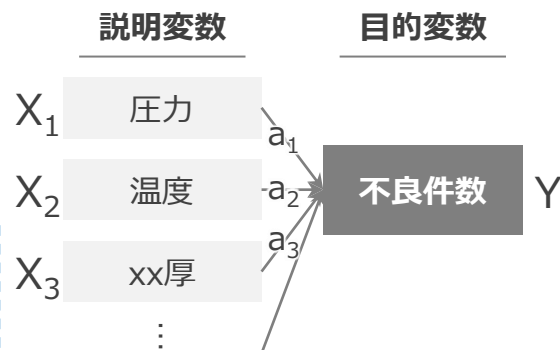


残差の2乗和が最小になるように
係数a, 定数項bを決める（最小二乗法）

$$Y = a \cdot X + b$$

目的変数	回帰係数 (傾き)	説明変数	定数項 (Y切片)
結果となる変数	Xの影響の大きさ	原因となる変数	X=0のときのY

重回帰



$$Y = a_1 X_1 + a_2 X_2 + \dots + a_n X_n + b$$

各回帰係数は、
各説明変数の影響の大きさを表す

参考：教師あり分類手法における決定境界の比較

- 分類手法は、データの空間内で、いかに異なるクラスを綺麗に分割できるかで性能が決まる。
この分割の境目を「**決定境界**」と呼ぶ
- ロジスティック回帰は、直線的な決定境界**を描くため、非線形に入り組んだ分類には向かない

入力データ形式

ロジスティック回帰

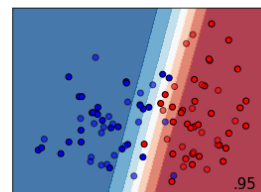
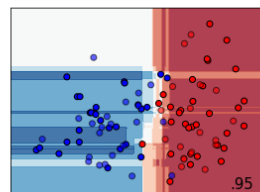
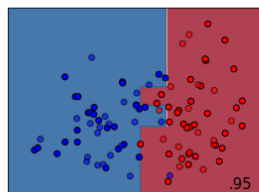
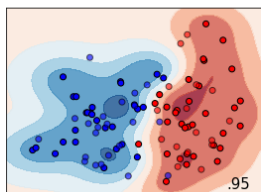
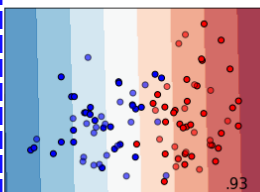
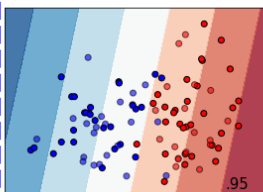
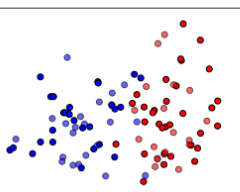
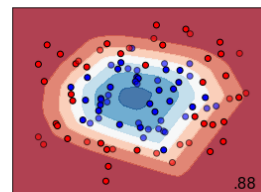
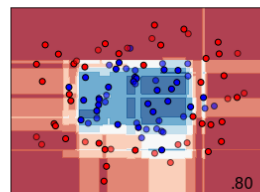
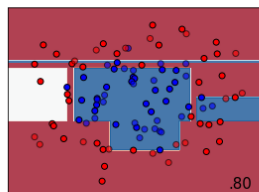
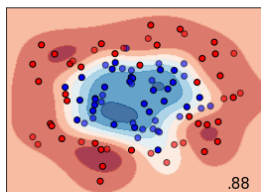
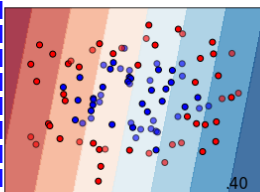
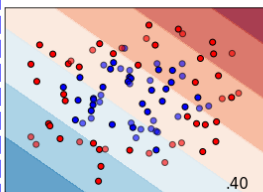
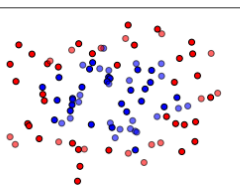
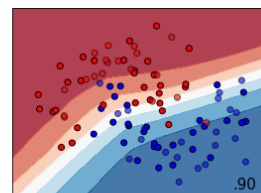
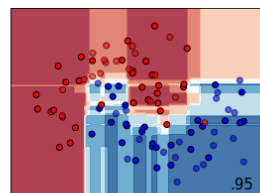
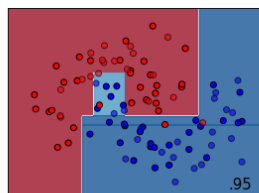
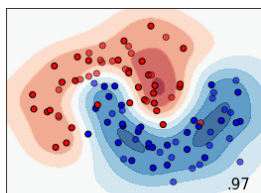
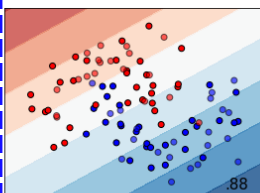
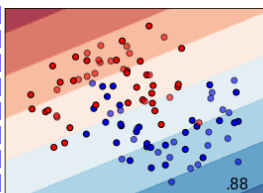
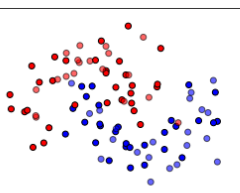
線形SVM

カーネルSVM

決定木

ランダムフォレスト

ニューラルネットワーク



直線的
(斜線)

直線的
(斜線)

複雑

直線的
(水平/垂直)

複雑

複雑

ビッグデータ分析の進め方

- データマイニングの進め方に関する方法論「**CRISP-DM**」に基づいて、分析と評価を繰り返して試行錯誤しながら進めるのが一般的である

CRISP-DM: データマイニング方法論



(CRoss Industry Standard Process for Data Mining)

1. ビジネスの理解

- ビジネス、データマイニング目標の決定
- プロジェクトの立ち上げ

2. データの理解

- データの収集
- データの調査
- データ品質の検証

3. データの準備

- データの選択や除外
- データのクリーニング
- データの構築や統合

4. モデル作成

- モデリング手法の選択
- モデルの作成
- モデルの評価

5. 評価

- データマイニングの結果の評価
- プロセスの見直し
- 実行可能なアクションリストの作成

6. 展開／共有

- 業務への導入計画
- モニタリング、メンテナンスの計画

使用データ

- UCI Machine Learning Repositoryでは様々な分野のデータが公開
- 今回は、**銀行のマーケティングデータ**を活用し、分析を行う



Machine Learning Repository
Center for Machine Learning and Intelligent Systems

Bank Marketing Data Set

Download: [Data Folder](#), [Data Set Description](#)

Abstract: The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y).

Data Set Characteristics:	Multivariate	Number of Instances:	45211	Area:	Business
Attribute Characteristics:	Real	Number of Attributes:	17	Date Donated	2012-02-14
Associated Tasks:	Classification	Missing Values?	N/A	Number of Web Hits:	1577437

Source:

[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

Data Set Information:

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was require ('yes') or not ('no') subscribed.

There are four datasets:

- 1) bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014]
- 2) bank-additional.csv with 10% of the examples (4119), randomly selected from 1), and 20 inputs.
- 3) bank-full.csv with all examples and 17 inputs, ordered by date (older version of this dataset with less inputs).
- 4) bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).

The smallest datasets are provided to test more computationally demanding machine learning algorithms (e.g., SVM).

The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

<https://archive.ics.uci.edu/ml/datasets/bank+marketing>

データの概要

- 4,521人分の顧客について、顧客情報や営業アプローチ状況、最終的な狙いである「定期預金の契約有無」に関する情報（計17列）が格納されている

※クラウド型のSAS Studio (SAS OnDemand for Academics) において
列名を日本語にする場合、全角6文字以内を推奨

		クレジットカード 債務不履行の有無			年間平均残高 (ユーロ)		最終連絡時の 会話時間 (秒)		キャンペーン中の 連絡回数		最終連絡からの 経過日数		キャンペーン前の 連絡回数		前回キャンペーン の結果	
年齢	職業	結婚歴	学歴	クレカ債務	年間平均 残高	住宅 ローン	個人 ローン	連絡手段	最終連 絡日	最終連 絡月	最終会話 時間	CP中連絡 回数	最終連絡 日数	CP前連絡 回数	前回CP結果	定期預金 契約
30	unemployed	married	primary	no	1787	no	no	cellular	19	oct	79	1	-1	0	unknown	no
33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220	1	339	4	failure	no
35	management	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	330	1	failure	no
30	management	married	tertiary	no	1476	yes	yes	unknown	3	jun	199	4	-1	0	unknown	no
59	blue-collar	married	secondary	no	0	yes	no	unknown	5	may	226	1	-1	0	unknown	no
35	management	single	tertiary	no	747	no	no	cellular	23	feb	141	2	176	3	failure	no
36	self-employed	married	tertiary	no	307	yes	no	cellular	14	may	341	1	330	2	other	no
39	technician	married	secondary	no	147	yes	no	cellular	6	may	151	2	-1	0	unknown	no
41	entrepreneur	married	tertiary	no	221	yes	no	unknown	14	may	57	2	-1	0	unknown	no
43	services	married	primary	no	-88	yes	yes	cellular	17	apr	313	1	147	2	failure	no
39	services	married	secondary	no	9374	yes	no	unknown	20	may	273	1	-1	0	unknown	no
43	admin.	married	secondary	no	264	yes	no	cellular	17	apr	113	2	-1	0	unknown	no
36	technician	married	tertiary	no	1109	no	no	cellular	13	aug	328	2	-1	0	unknown	no
20	student	single	secondary	no	502	no	no	cellular	30	apr	261	1	-1	0	unknown	yes
31	blue-collar	married	secondary	no	360	yes	yes	cellular	29	jan	89	1	241	1	failure	no

説明変数

目的変数

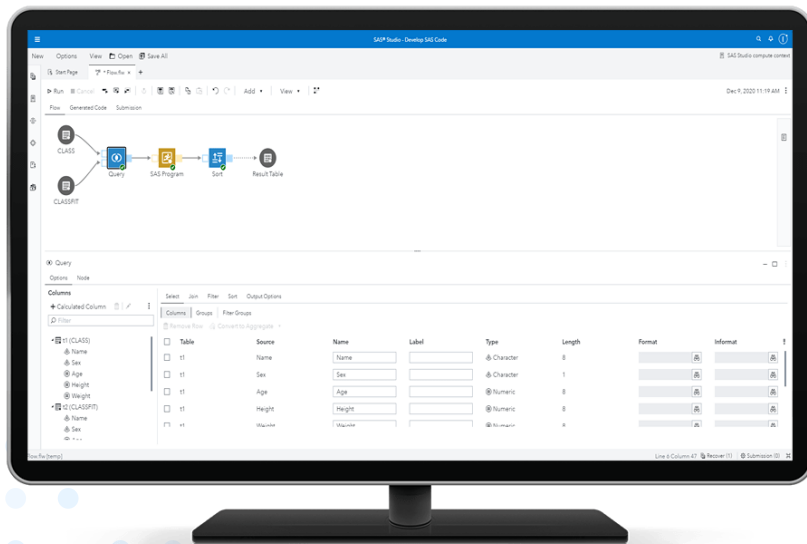
予測(分析)対象を
説明するための変数

予測(分析)
したい対象

SAS Studio について

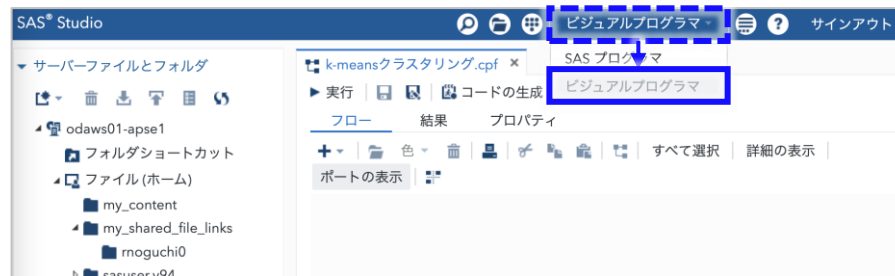
- 今年のウェビナーでは、**SAS Studio** でデモを行います。
- SAS Studio はすべてのSAS製品に付帯しているGUI で、今回は学習用に自宅でもお使い頂けるクラウド型無償版 **SAS OnDemand for Academics** を使っています。
(※無償版の登録については、SAS からの申込完了メールをご参照ください)
- なお、SAS Studio起動時はコード入力画面となっていますが、画面右上の「SASプログラマ」を「**ビジュアルプログラマ**」に変更するとデモと同様の入力画面となります。

▼SAS Studio 画面イメージ



https://www.sas.com/ja_jp/software/studio.html

▼GUI画面への変更方法 (ビジュアルプログラマ)



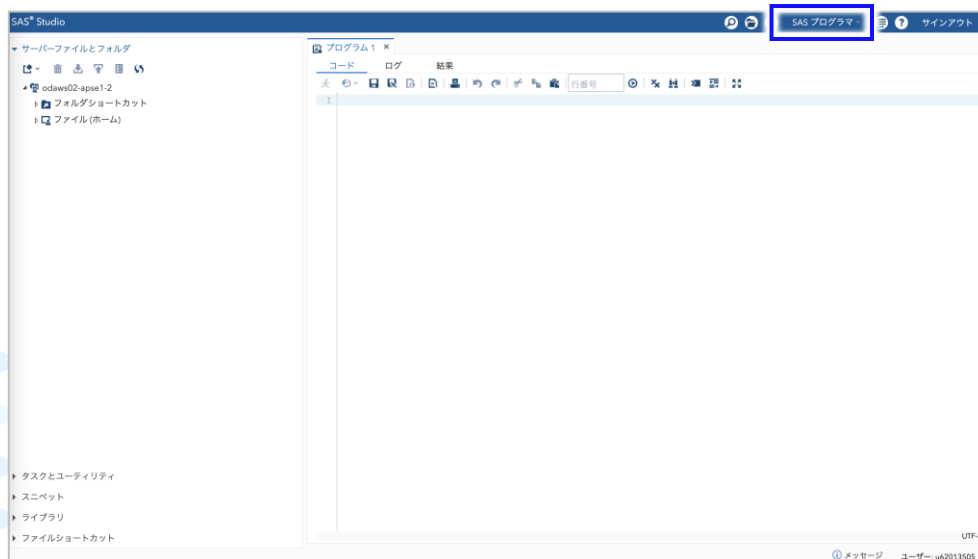
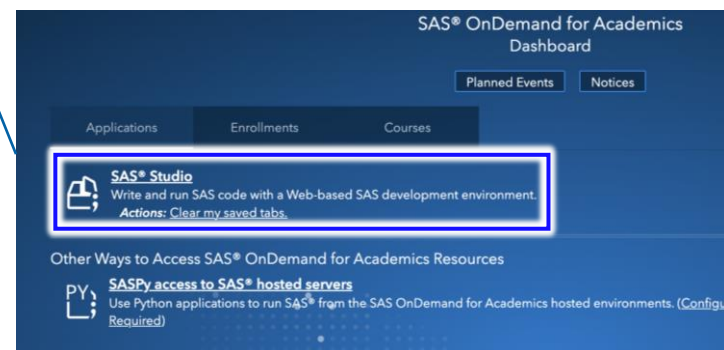
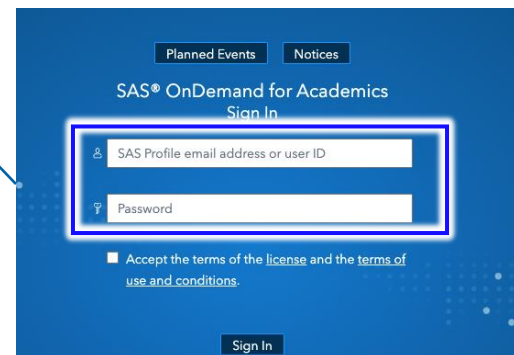
参考 : SAS Studio 起動方法

- SAS OnDemand for Academics にログイン後、Dashboard より SAS Studio を起動
- 起動後、前頁の通り、右上メニューより「ビジュアルプログラマ」を選択

1 SAS OnDemand for Academics にログイン
<https://welcome.oda.sas.com/login>

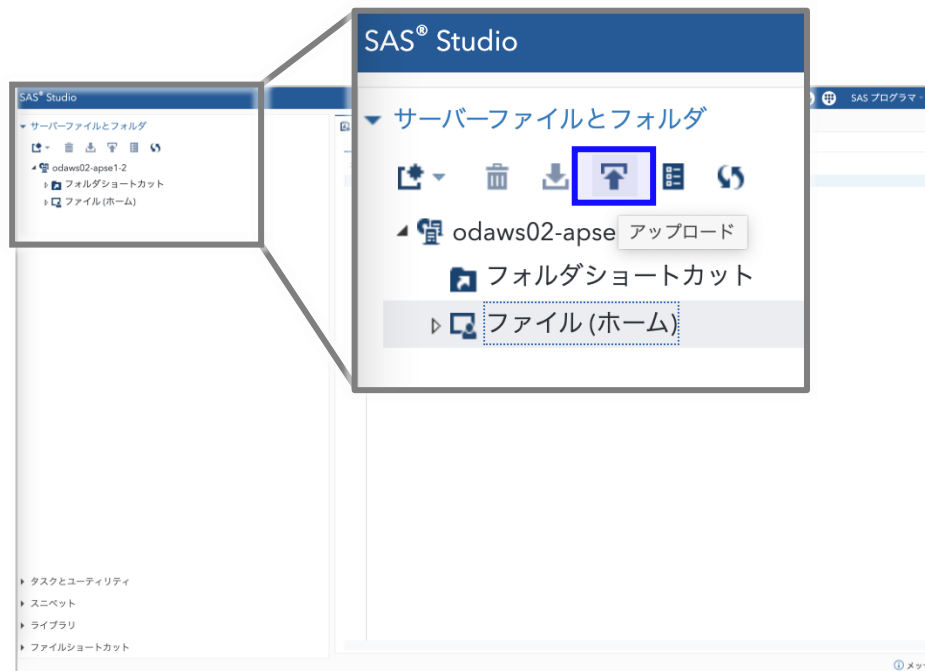
2 SAS OnDemand for Academics Dashboardより
“SAS® Studio” をクリックして起動

3 SAS Studioが起動／右上よりビジュアルプログラマを選択

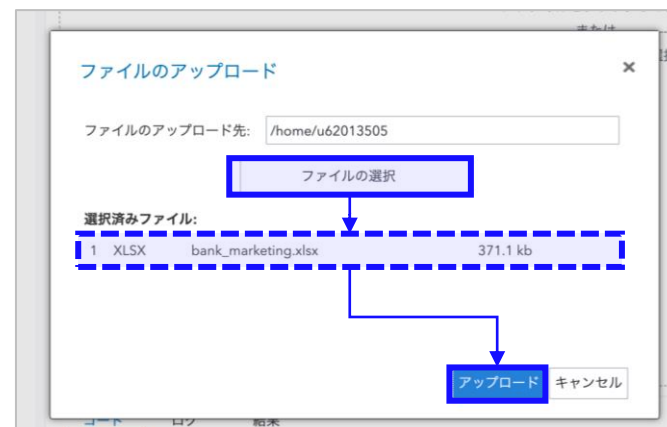


データの読み込み (1/2)

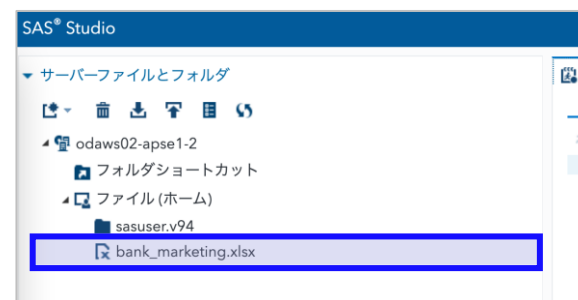
①左パネル内の「アップロード」アイコン をクリック



②「ファイルの選択」ボタンをクリックし、ファイル選択画面で
“bank_marketing.xlsx” を選択し、OKボタン
③「アップロード」ボタンをクリック

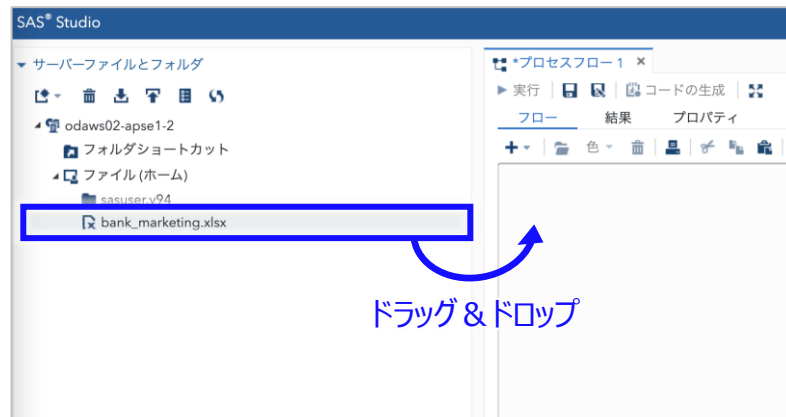


④左パネル内にファイルがアップロードされていることを確認

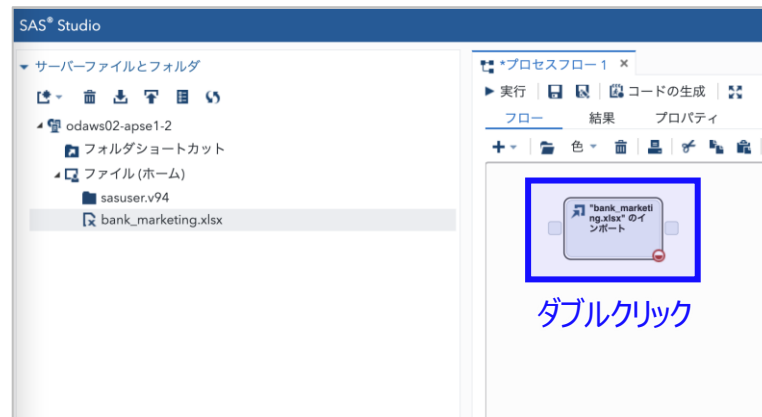


データの読み込み (2/2)

①左パネル内の“bank_marketing.xlsx”を選択し、画面右側のプログラムエリアにドラッグ＆ドロップ



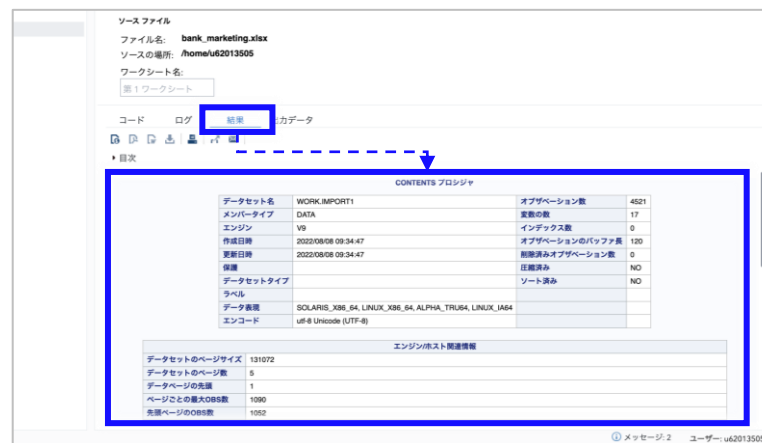
②右側のプロセスフローにノードが生成されるので、当該ノードをダブルクリック



③詳細設定画面が開くので、実行ボタンをクリック (特に各設定は変更不要)



④「結果」のタブ画面に読み込んだデータの概要が出力



読み込んだデータの確認

データ概要の確認

新しいブラウザタブで開く



CONTENTS プロシジャ

データセット名	WORK.IMPORT1	オブザベーション数	4521
メンバータイプ	DATA	変数の数	17
エンジン	V9		
作成日時	2022/08/08 09:34:47		
更新日時	2022/08/08 09:34:47		
保護			
データセットタイプ			
ラベル			
データ表現	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
エンコード	utf-8 Unicode (UTF-8)		

列数と行数を確認！
(ビッグデータ分析の基本)

エンジン/ホスト関連情報

データセットのページサイズ	131072
データセットのページ数	5
データページの先頭	1
ページごとの最大OBS数	1090
先頭ページのOBS数	1052
データセットの修復数	0
ファイル名	/saswork/SAS_work71F80001F3FA_0daws01-apse1-2.oda.sas.com/SAS_workC7860001F3FA_0daws01-apse1-2.oda.sas.com/import1.sas7bdat
作成したリリース	9.0401M6
作成したホスト	Linux
Iノード番号	33850
アクセス権限	rw-r--r--
所有者名	u62013505
ファイルサイズ	768KB
ファイルサイズ (バイト)	

各列のデータ型を確認

#	変数	タイプ	長さ	出力形式	入力形式	ラベル
13	キャンペーン中の連絡	数値	8	BEST.		キャンペーン中の連絡回数
15	キャンペーン前の連絡	数値	8	BEST.		キャンペーン前の連絡回数
5	クレジットカード債務	文字	3	\$3.	\$3.	クレジットカード債務不履行有無
7	住宅ローンの有無	文字	3	\$3.	\$3.	住宅ローンの有無
8	個人ローンの有無	文字	3	\$3.	\$3.	個人ローンの有無
16	前回キャンペーンの結果	文字	7	\$7.	\$7.	前回キャンペーンの結果
4	学歴	文字	9	\$9.	\$9.	学歴
17	定期預金契約有無	文字	3	\$3.	\$3.	定期預金契約有無
6	年間平均残高 (ユーロ)	数値	8	BEST.		年間平均残高 (ユーロ)
1	年齢	数値	8	BEST.		年齢
14	最終連絡からの経過日	数値	8	BEST.		最終連絡からの経過日数
10	最終連絡日	数値	8	BEST.		最終連絡日
12	最終連絡時の会話時間	数値	8	BEST.		最終連絡時の会話時間 (秒)
11	最終連絡月	文字	3	\$3.	\$3.	最終連絡月
3	結婚歴	文字	8	\$8.	\$8.	結婚歴
2	職業	文字	13	\$13.	\$13.	職業
9	連絡手段	文字	9	\$9.	\$9.	連絡手段

生データの確認

プログラム 1 × *bank_marketing ×

設定 | コード/結果 | 分割 | ログ | コード

ファイル情報

ソースファイル

ファイル名: bank_marketing.xlsx

ソースの場所: /home/u62013505

ワークシート名: 第1ワークシート

出力データ

SAS Server: SASApp

データセット名: IMPORT1

ライブラリ: WORK

オプション

ファイルの種類: デフォルト (ファイル拡張子に基づく)

コード ログ 結果 出力データ

テーブル: WORK.IMPORT1 ビュー: 列名 フィルタ: (なし)

列

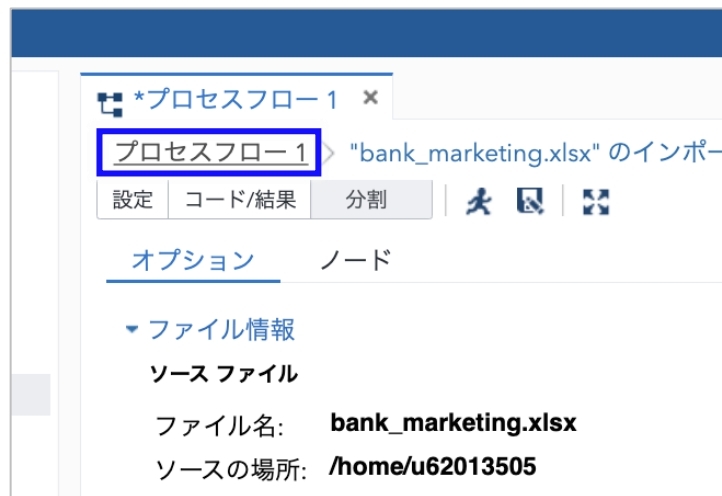
合計行数: 4521 合計変数: 17

	年齢	職業	結婚歴	学歴	クレジット...	年間平均残高 (ユーロ)	住宅C
1	30	unemployed	married	primary	no	1787	no
2	33	services	married	secondary	no	4789	yes
3	35	management	single	tertiary	no	1350	yes
4	30	management	married	tertiary	no	1476	yes
5	59	blue-collar	married	secondary	no	0	yes
6	35	management	single	tertiary	no	747	no
7	36	self-employed	married	tertiary	no	307	yes
8	39	technician	married	secondary	no	147	yes
9	41	entrepreneur	married	tertiary	no	221	yes
10	43	services	married	primary	no	-88	yes
11	39	services	married	secondary	no	9374	yes
12	43	admin.	married	secondary	no	264	yes
13	36	technician	married	tertiary	no	1109	no
14	20	student	single	secondary	no	502	no
15	31	blue-collar	married	secondary	no	360	yes

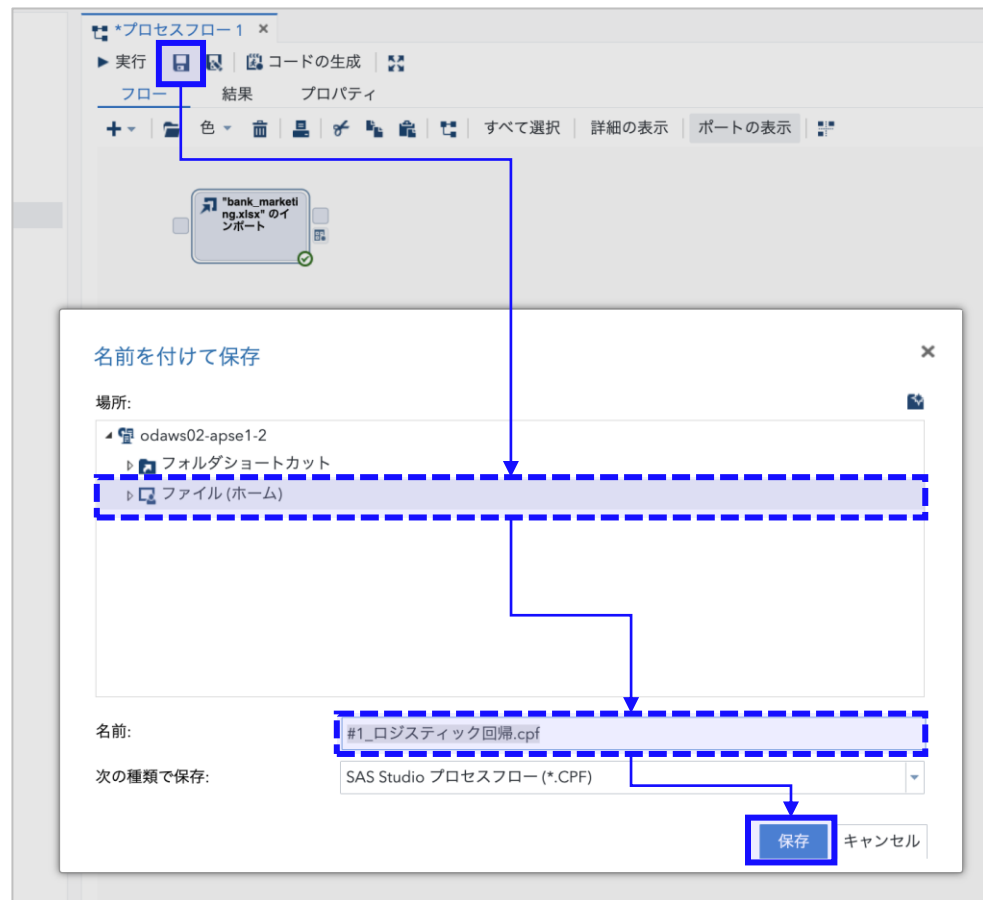
メッセージ: 4 ユーザー: u62013505

作成したプロセスフローの保存

プロセスフローをクリックしてプロセスフロー画面に戻る



保存アイコンをクリックし、保存場所、ファイル名を指定して保存ボタン



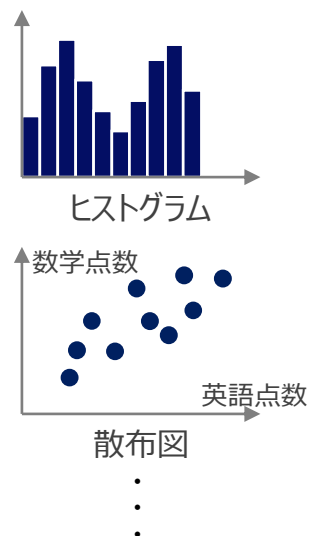
データの特徴の捉え方

- ビッグデータでは個々のデータをくまなく見るのは難しいため、**グラフ**（ヒストグラムや散布図）や**要約統計量**（平均値や標準偏差）を用いて全体傾向を把握する

グラフ化

視覚的にデータの特徴や傾向を把握

ID	英語	数学	...
1	55	48	
2	70	47	
3	66	44	
⋮	⋮	⋮	
18	65	55	
19	57	51	
20	68	43	



数値化（データ要約）

データの特徴を示す値に要約し
比較可能な客観的傾向を掴む
（要約統計量）

ID	英語	数学	...
1	55	48	
2	70	47	
3	66	44	
⋮	⋮	⋮	
18	65	55	
19	57	51	
20	68	43	

- 平均値 = X X X
 - 中央値 = X X X
 - 標準偏差 = X X X
- ⋮

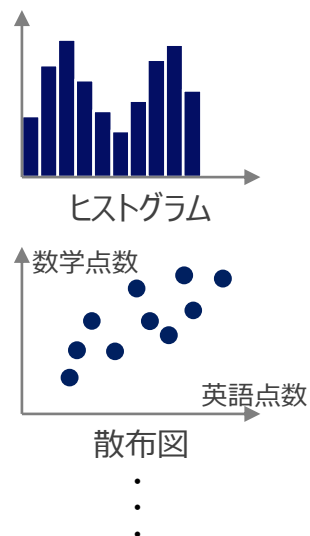
データの特徴の捉え方

- ビッグデータでは個々のデータをくまなく見るのは難しいため、**グラフ**（ヒストグラムや散布図）や**要約統計量**（平均値や標準偏差）を用いて全体傾向を把握する

グラフ化

視覚的にデータの特徴や傾向を把握

ID	英語	数学	...
1	55	48	
2	70	47	
3	66	44	
⋮	⋮	⋮	
18	65	55	
19	57	51	
20	68	43	



数値化（データ要約）

データの特徴を示す値に要約し
比較可能な客観的傾向を掴む
（要約統計量）

ID	英語	数学	...
1	55	48	
2	70	47	
3	66	44	
⋮	⋮	⋮	
18	65	55	
19	57	51	
20	68	43	

- 平均値 = X X X
 - 中央値 = X X X
 - 標準偏差 = X X X
- ⋮

ヒストグラム（度数分布図）

- ヒストグラムは、データを一定間隔（**階級**）ごとに頻度集計（**度数**）をとり、横軸に階級、縦軸に度数をとって柱状（**ビン**）の集合で表したグラフ
- データの分布状況を可視化して、直感的にデータの特徴を捉えやすくする目的がある

元データ

レストランAの
レビューアごとの評点データ

#	評点	#	評点
1	3.5	11	3.0
2	3.5	12	3.5
3	4.5	13	3.0
4	3.5	14	4.0
5	3.5	15	4.0
6	2.5	16	4.0
7	3.0	17	2.5
8	3.5	18	4.5
9	4.0	19	1.0
10	3.0	20	5.0

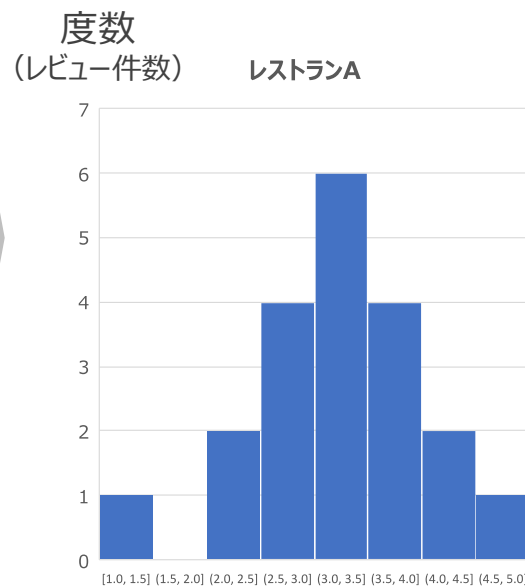
度数分布表

一定間隔（階級）ごとに
レビュー件数（度数）を集計

階級	度数
0.0~0.5	0
0.5~1.0	0
1.0~1.5	1
1.5~2.0	0
2.0~2.5	2
2.5~3.0	4
3.0~3.5	6
3.5~4.0	4
4.0~4.5	2
4.5~5.0	1

ヒストグラム（度数分布図）

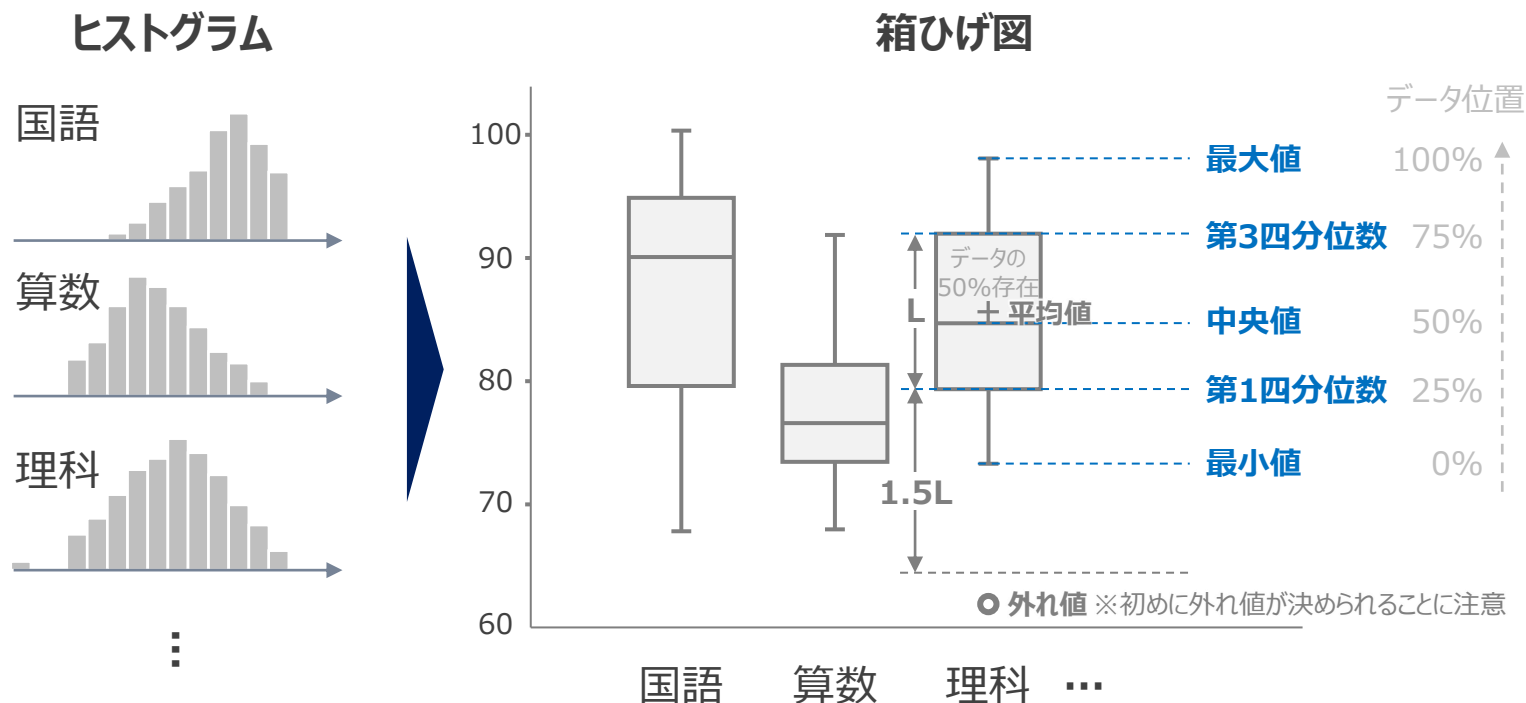
横軸に階級、縦軸に度数を取り、
柱（ビン）の集合で可視化



※境界値は小さい側の階級に含める

箱ひげ図

- 分布形状の把握は極めて重要であるが、主にヒストグラムは1変数、散布図は2変数での分布可視化に適している。一方、箱ひげ図は、分布の概形を**多変数間で比較**するのに適している

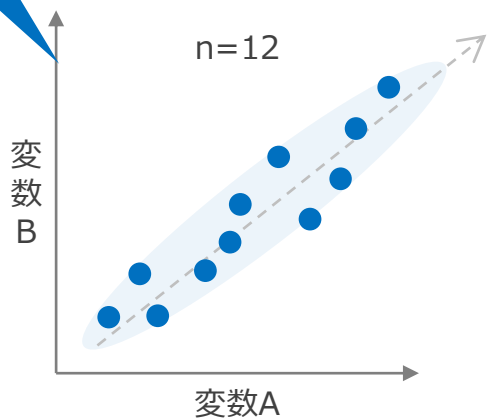


散布図 と 相関

- 二つの値（**変数**）間において、一方が上がれば他方も上がる（or 下がる）ような関係性のことを「**相関**」と呼ぶ
- 各変数を各軸にとってグラフ化した「**散布図**」を描くことで、相関関係の視覚的な把握が可能

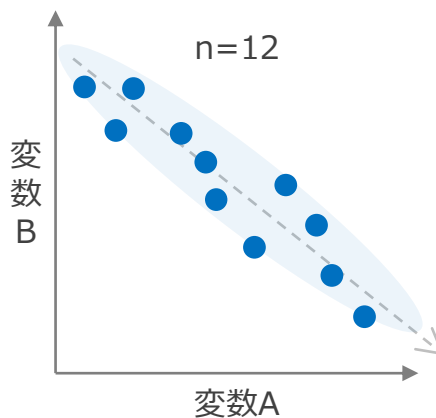
散布図

正の相関



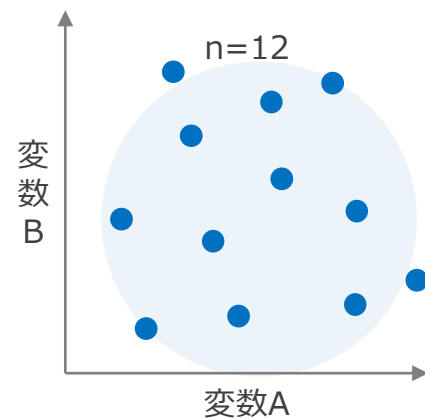
一方 (A) が増加すると
他方 (B) も増加
(右上がりの楕円状)

負の相関



一方 (A) が増加すると
他方 (B) は減少
(右下がりの楕円状)

無相関



双方の変数間に
関連性なし
(まばらな円状)

例) A・・・(夏の)平均気温
B・・・エアコンの販売台数

例) A・・・(冬の)平均気温
B・・・ストーブの販売台数

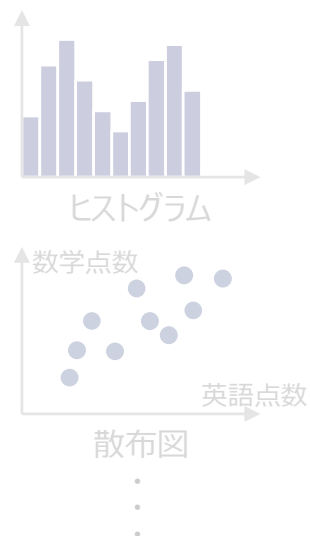
データの特徴の捉え方

- ビッグデータでは個々のデータをくまなく見るのは難しいため、**グラフ**（ヒストグラムや散布図）や**要約統計量**（平均値や標準偏差）を用いて全体傾向を把握する

グラフ化

視覚的にデータの特徴や傾向を把握

ID	英語	数学	...
1	55	48	
2	70	47	
3	66	44	
⋮	⋮	⋮	
18	65	55	
19	57	51	
20	68	43	



数値化（データ要約）

データの特徴を示す値に要約し
比較可能な客観的傾向を掴む
(要約統計量)

ID	英語	数学	...
1	55	48	
2	70	47	
3	66	44	
⋮	⋮	⋮	
18	65	55	
19	57	51	
20	68	43	

- 平均値 = X X X
 - 中央値 = X X X
 - 標準偏差 = X X X
- ⋮

数値化（データ要約）

- ビッグデータでは全データをくまなく見るのは難しいため、「要約した値（＝要約統計量）」を用いて全体傾向を把握するのが一般的

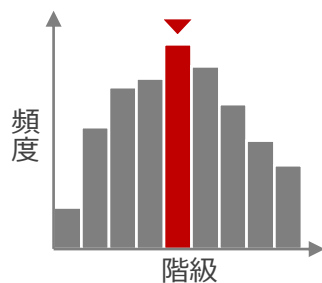
要約統計量

統計量の
種類

代表値

データの中心的な値

分布上での
イメージ

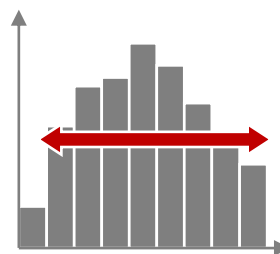


例

平均値、中央値、最頻値

ばらつき指標 (散布度)

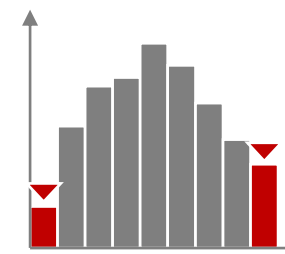
データのばらつき具合



分散、標準偏差

順序統計量

順序関係を表す値



最大値、最小値、パーセンタイル

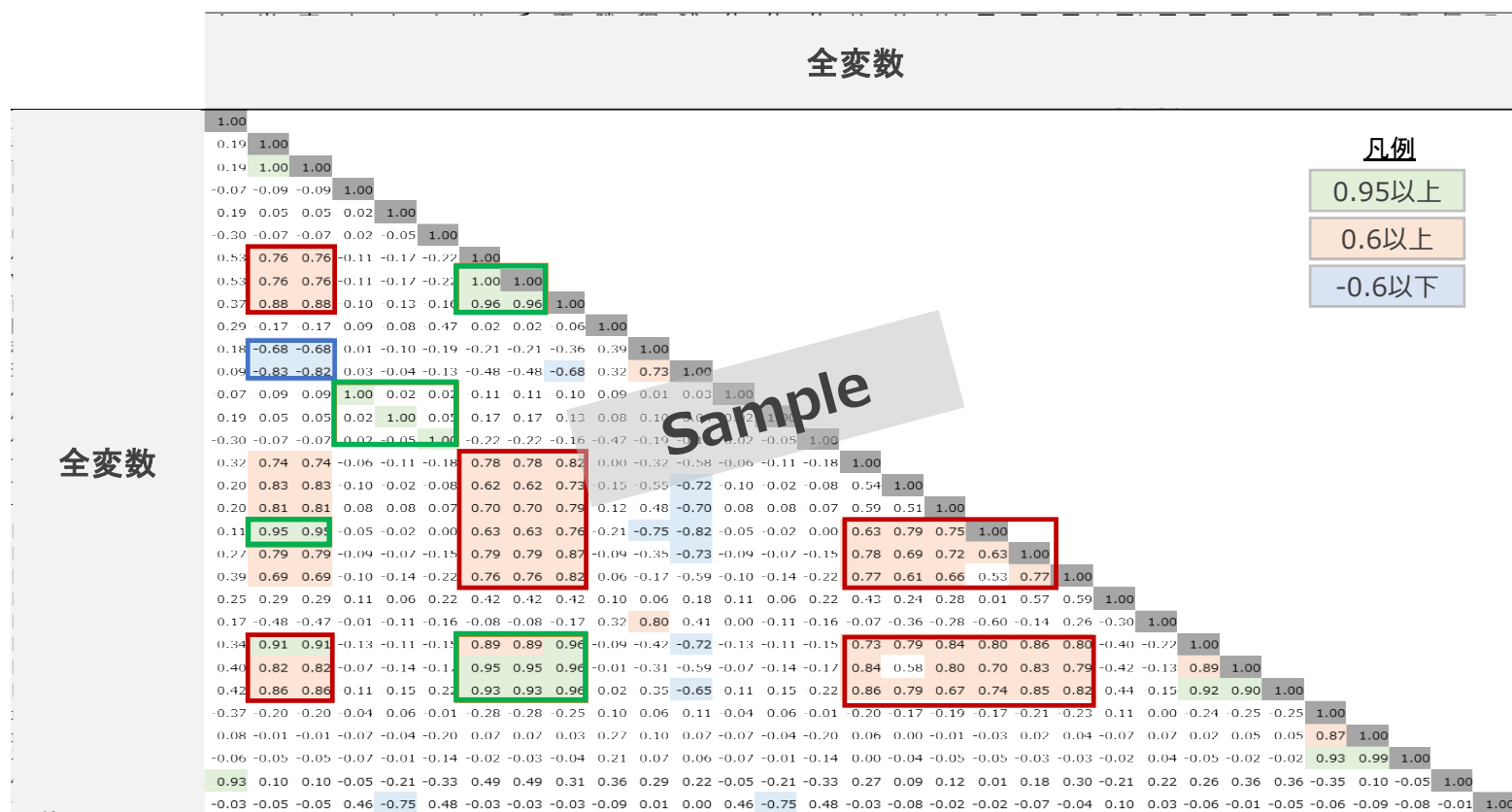
その他

歪度・尖度などの
高次モーメント

相関行列

※第2回で取り扱う予定

- 事前に各変数間の相関係数を総当たりで調べておくと、後々の結果解釈に役立つ（相関行列）
- また、**共線性が高い変数**（相関の高い）が複数混ざっていると、その変数の影響を強く受け、偏った分析結果になることがある。この場合、共線性が高い変数は除外することが有効



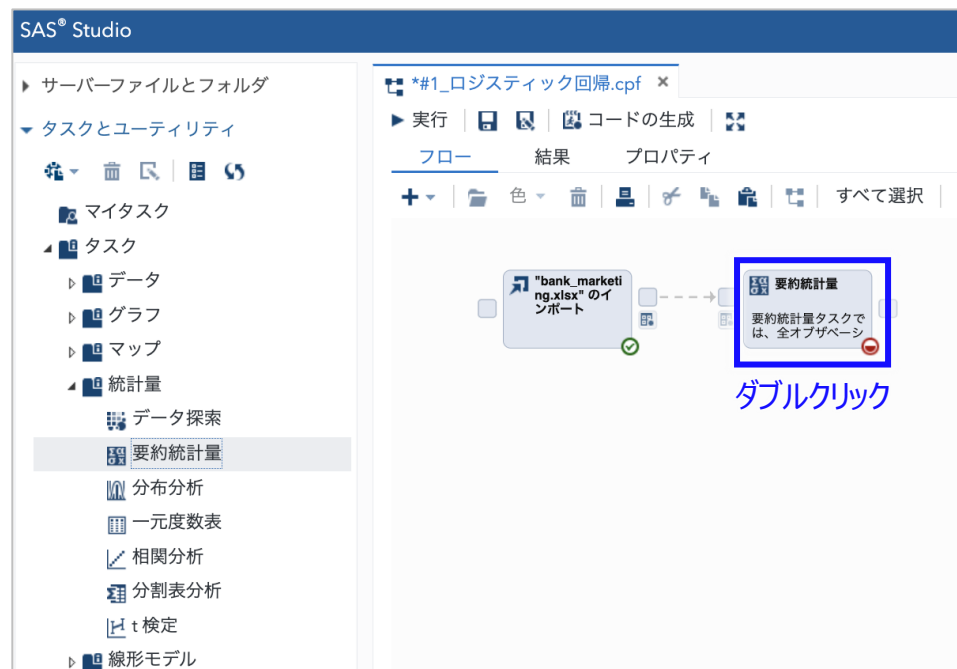
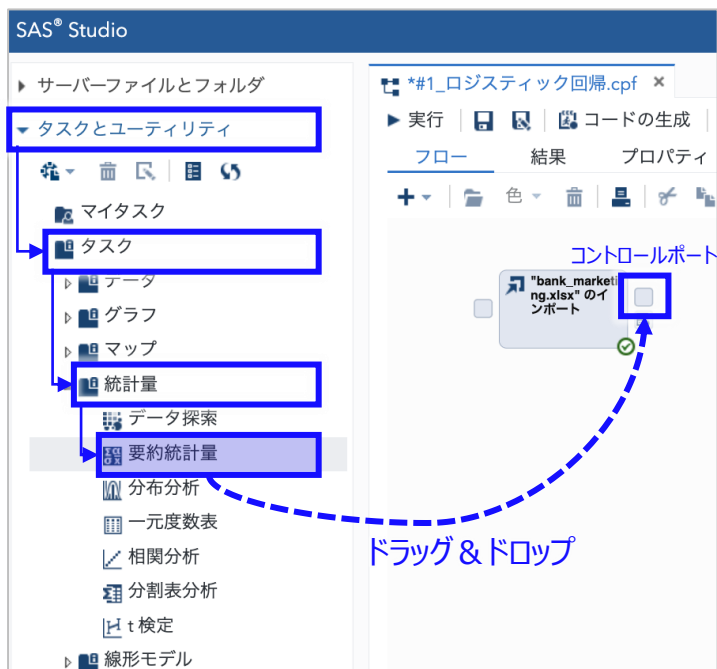
SAS Studio での実装方法

- 要約統計量
- ヒストグラム／箱ひげ図
- 散布図／層別散布図

要約統計量／ヒストグラム／箱ひげ図 – 実行方法 (1/3)

- ①左パネルより、[タスクとユーティリティ]→[タスク]
→[統計量]→[要約統計量]を選択
- ②右側のプロセスフロー内のインポートノードの
右端の四角 □ (コントロールポート) の上ヘドラッグ&ドロップ

- ③プロセスフロー上に要約統計量ノードが生成されるので
ダブルクリックして詳細設定画面を開く



要約統計量／ヒストグラム／箱ひげ図 – 実行方法 (2/3)

④[データ]の画面で、各種設定



データソースの設定 (マイライブラリー-WORK-IMPORT)

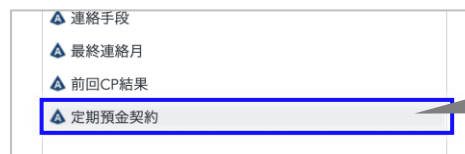


分析変数 (描画対象の変数=数値型の変数) の設定



- Shift (もしくはCtrl) を押しながらかlickで複数選択可能
- 数値型の変数を全て選択し、OK

分類変数 (比較用の分類に用いる変数=目的変数) の設定



- 定期預金契約有無のみを選択し、OKボタン

要約統計量／ヒストグラム／箱ひげ図 – 実行方法 (3/3)

④[オプション]の画面で、出力の設定

The screenshot shows the 'Options' tab in the SAS Studio interface. The 'Statistics' section has checkboxes for 'Average', 'Standard Deviation', 'Minimum', 'Maximum', 'Central Value', 'Number of Outliers', and 'Number of Missing Values'. The 'Plots' section has checkboxes for 'Histogram' and 'Box Plot'. The 'Box Plot' checkbox is highlighted with a blue box. A blue arrow points from the 'Box Plot' checkbox to the text: 「デフォルトでチェック済の項目に加え、[中央値] [欠損値の数] [ヒストグラム] [比較箱ひげ図] のチェックも入れる」.

デフォルトでチェック済の項目に加え、
[中央値] [欠損値の数]
[ヒストグラム] [比較箱ひげ図]
のチェックも入れる

⑤実行ボタンをクリックすると、結果が出力される

The screenshot shows the 'Options' tab with the 'Execute' button (a blue icon with a white 'X') highlighted by a blue box. A blue arrow points from the 'Execute' button to the output table. The output table is titled '新しいブラウザタブ' and contains the following data:

定期預金契約	Obs 数	変数	ラベル	平均	標準偏差
no	4000	年齢	年齢	40.9980000	10.1883977
		年間平均	年間平均	1403.21	3075.35
		キャンペーン前	キャンペーン前	15.9487500	8.2497356
		最終連絡日	最終連絡日	226.3475000	210.3136306
		最終会話時間	最終会話時間	2.8622500	3.2126088
		CP 中連絡回数	CP 中連絡回数	36.0060000	96.2976572
		最終連絡日数	最終連絡日数	0.4712500	1.6273707

要約統計量／ヒストグラム／箱ひげ図 – 実行結果（要約統計量）

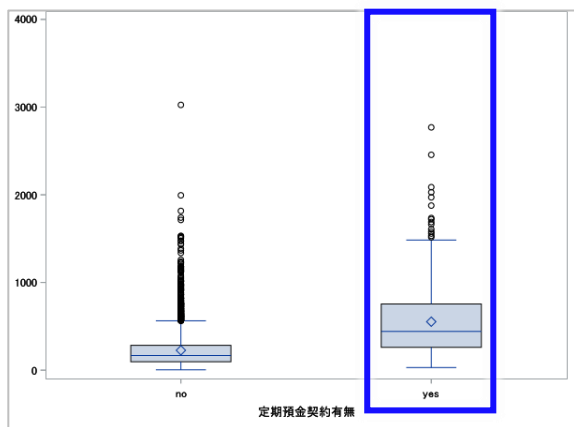
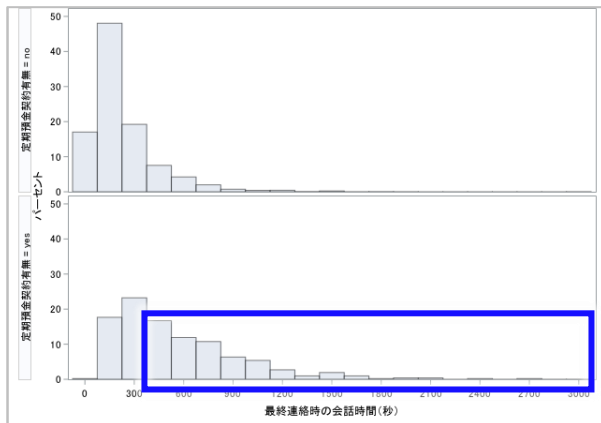
定期預金契約	Obs 数	変数	ラベル	平均	標準偏差	最小値	最大値	中央値	N	欠損値の数
no	4000	年齢	年齢	40.9980000	10.1883977	19.0000000	86.0000000	39.0000000	4000	0
		年間平均残高	年間平均残高	1403.21	3075.35	-3313.00	71188.00	419.5000000	4000	0
		最終連絡日	最終連絡日	15.9487500	8.2497356	1.0000000	31.0000000	16.0000000	4000	0
		最終会話時間	最終会話時間	226.3475000	210.3136306	4.0000000	3025.00	167.0000000	4000	0
		CP中連絡回数	CP中連絡回数	2.8622500	3.2126088	1.0000000	50.0000000	2.0000000	4000	0
		最終連絡日数	最終連絡日数	36.0060000	96.2976572	-1.0000000	871.0000000	-1.0000000	4000	0
		CP前連絡回数	CP前連絡回数	0.4712500	1.6273707	0	25.0000000	0	4000	0
yes	521	年齢	年齢	42.4913628	13.1157723	19.0000000	87.0000000	40.0000000	521	0
		年間平均残高	年間平均残高	1571.96	2444.40	-1206.00	26965.00	710.0000000	521	0
		最終連絡日	最終連絡日	15.6583493	8.2351482	1.0000000	31.0000000	15.0000000	521	0
		最終会話時間	最終会話時間	552.7428023	390.3258046	30.0000000	2769.00	442.0000000	521	0
		CP中連絡回数	CP中連絡回数	2.2667946	2.0920709	1.0000000	24.0000000	2.0000000	521	0
		最終連絡日数	最終連絡日数	68.6391555	121.9630631	-1.0000000	804.0000000	-1.0000000	521	0
		CP前連絡回数	CP前連絡回数	1.0902111	2.0553682	0	14.0000000	0	521	0

目的変数である「定期預金の契約」の有無で、傾向が異なりそうな変数

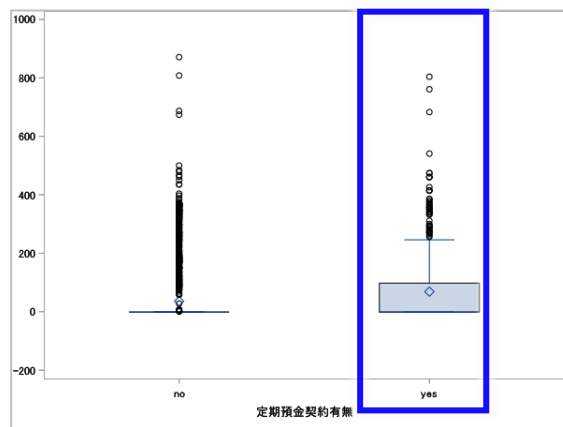
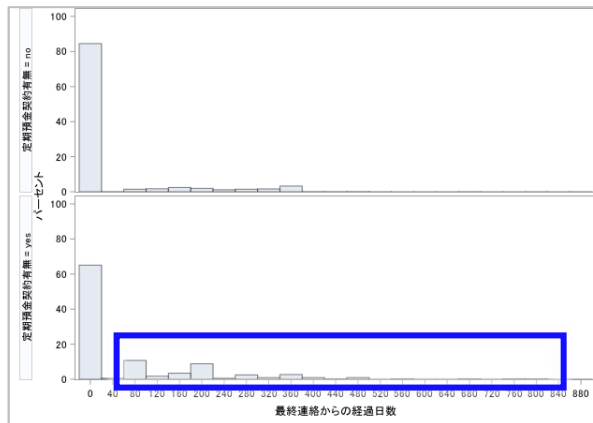
要約統計量／ヒストグラム／箱ひげ図 – 実行結果 (ヒストグラム／箱ひげ図)

※定期預金の契約有無で差が大きい変数

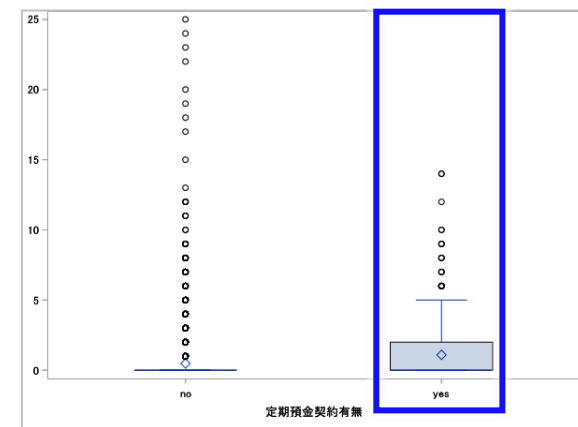
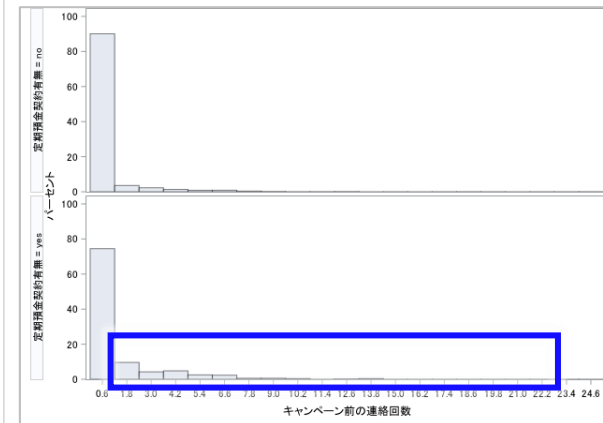
最終連絡時の会話時間 (秒)



最終連絡からの経過日数



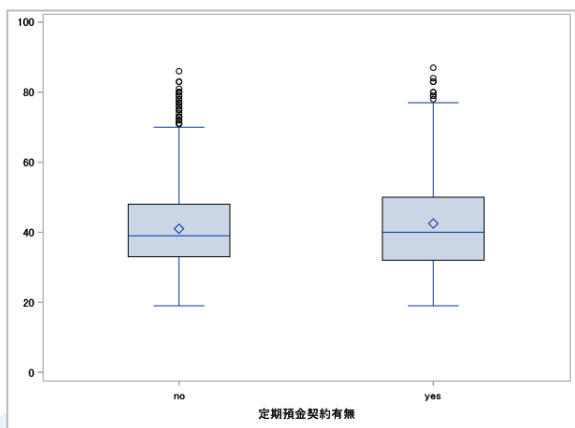
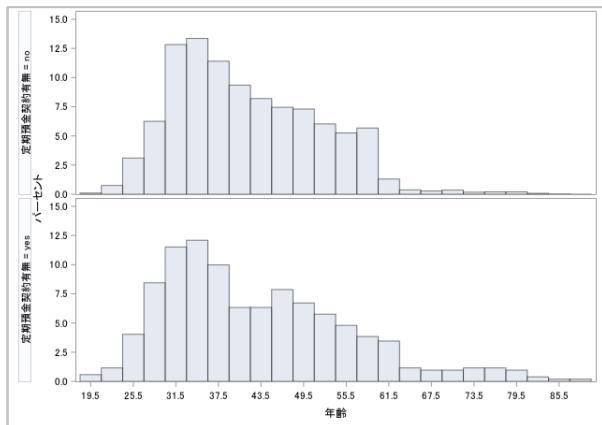
キャンペーン前の連絡回数



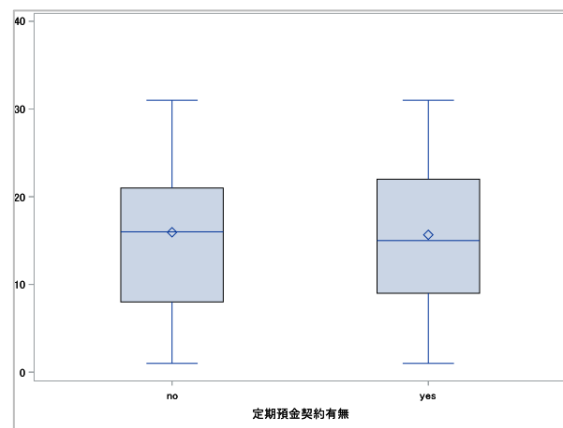
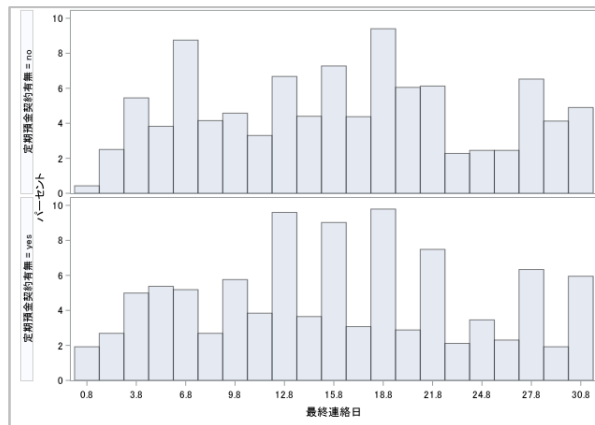
参考：要約統計量／ヒストグラム／箱ひげ図－ 実行結果（ヒストグラム／箱ひげ図）

※定期預金の契約有無であまり差が認められない変数

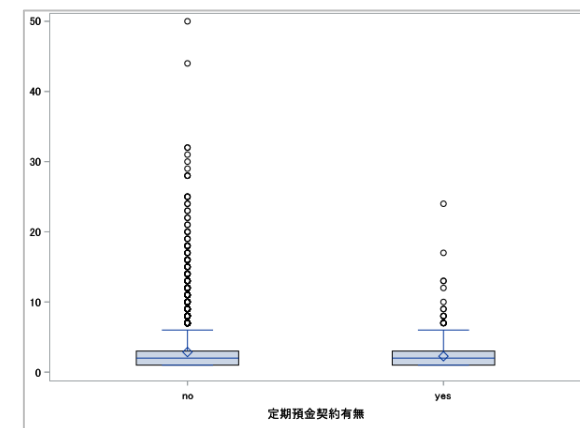
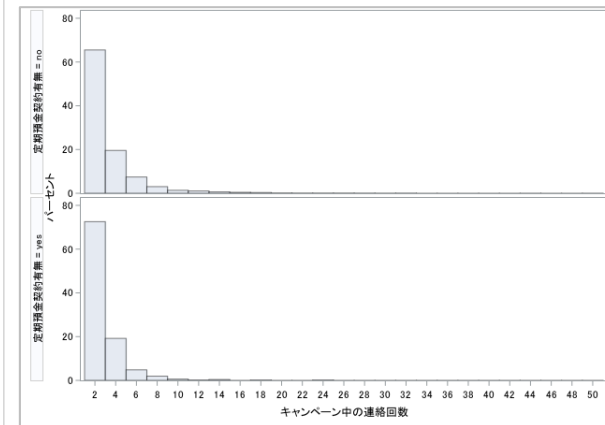
年齢



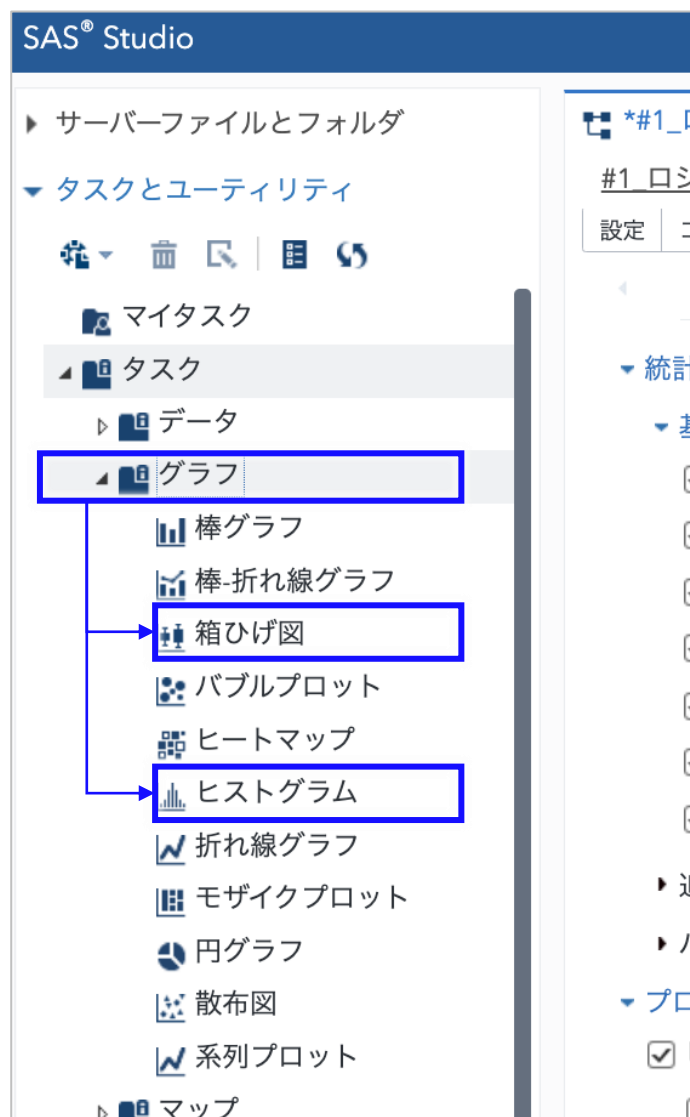
最終連絡日



キャンペーン中の連絡回数



参考：ヒストグラム・箱ひげ図のみ描画したい場合

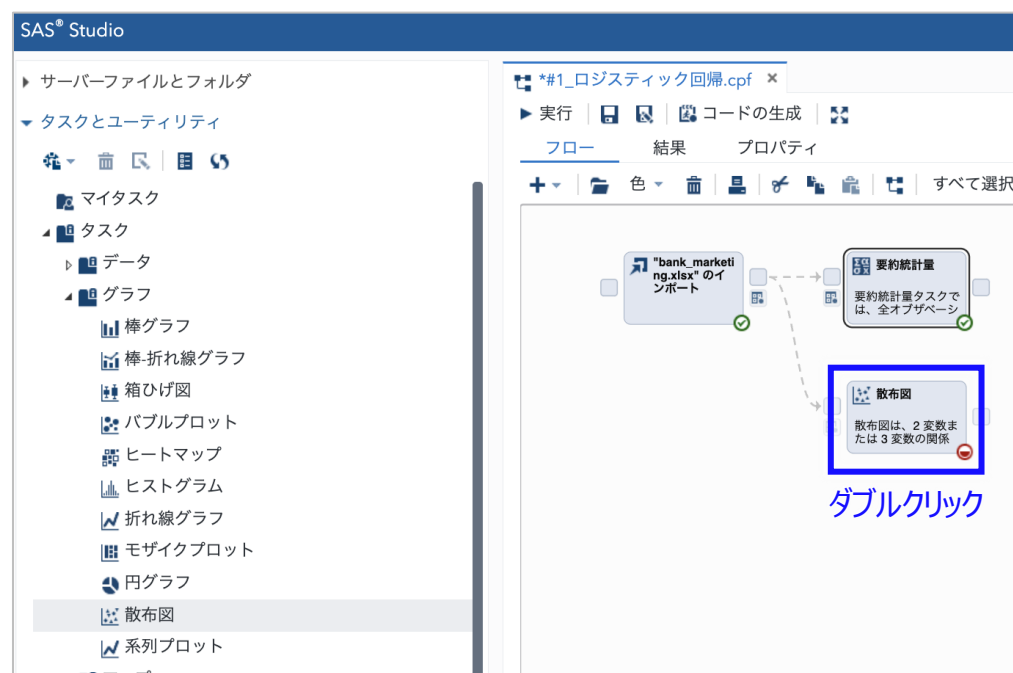
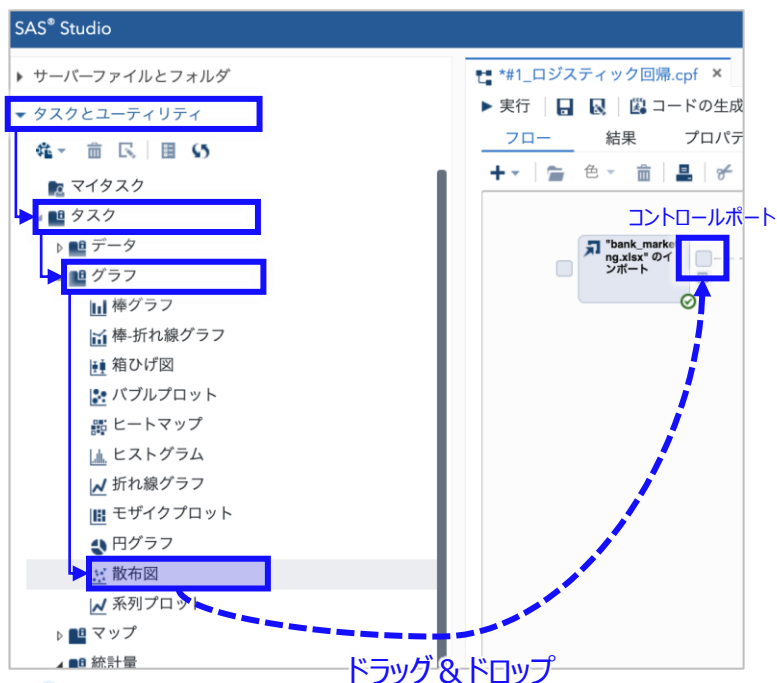


散布図の描画 – 実行方法 (1/2)

①左パネルより、[タスクとユーティリティ]→[タスク]
→[グラフ]→[散布図]を選択

②右側のプロセスフロー内のインポートノードの
右端の四角 □ (コントロールポート) の上ヘドラッグ&ドロップ

③プロセスフロー上に散布図ノードが生成されるので
ダブルクリックして詳細設定画面を開く



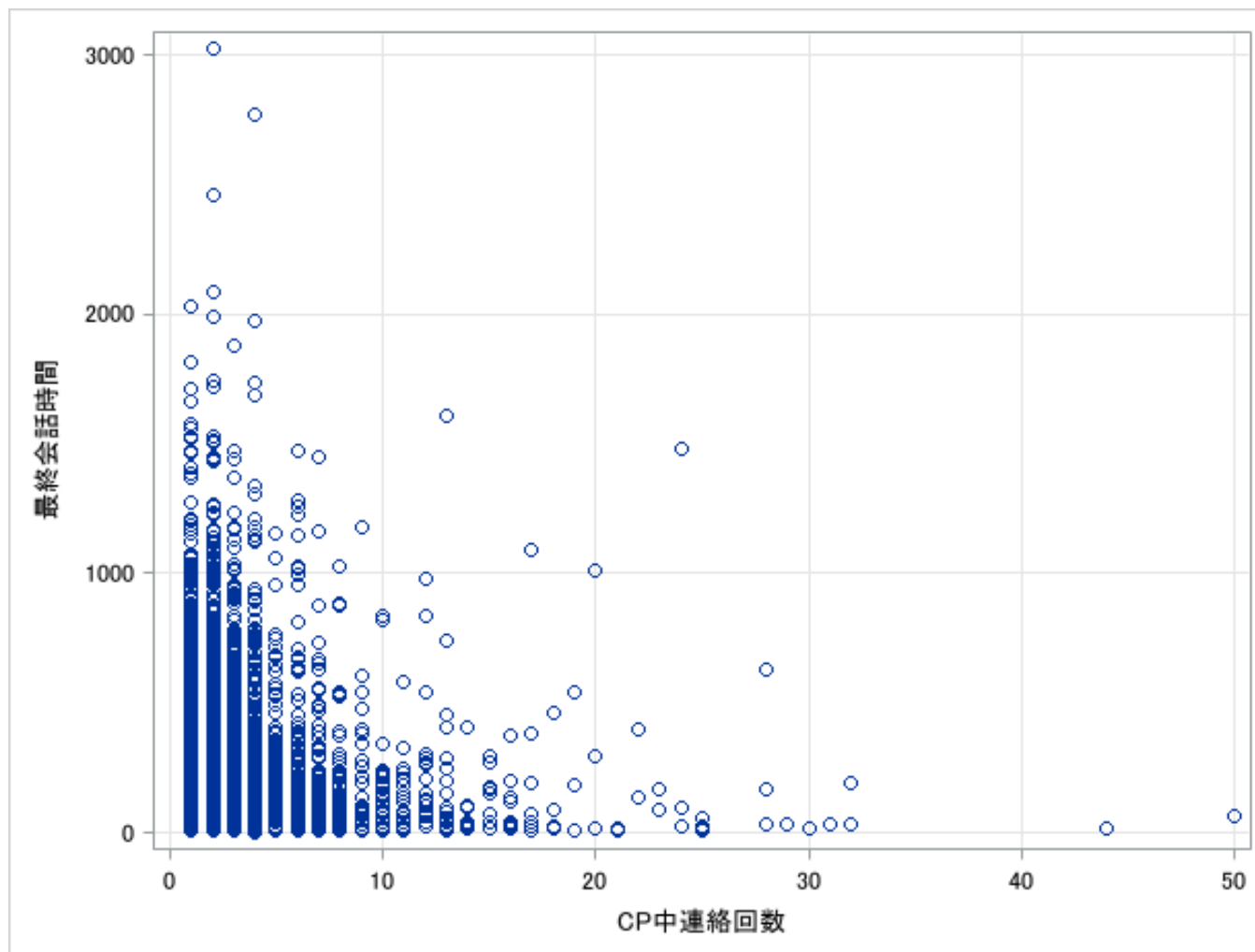
散布図の描画 – 実行方法 (2/2)

④[オプション]の画面で、X軸、Y軸の設定をし、実行ボタン

The screenshot shows the SAS Studio interface with the 'Options' tab selected for a scatter plot. The X-axis is set to 'CP中連絡回数' and the Y-axis is set to '最終会話時間'. The 'Run' button is highlighted. The left sidebar shows the 'Tasks and Utilities' pane with 'Graph' selected. The right pane shows the SAS code editor with the following code:

```
1 /*
2 *
3 * SAS Studio 3.8
4 *
5 * 生成日 '2022/08
6 * 生成者 'u620135
7 * 生成に使用された
8 * 生成に使用された
9 * 生成に使用された
10 * 生成に使用された
11 * 生成に使用された
12 *
13 */
14
15 ods graphics / re
16
17 proc sgplot data=
18     scatter x='CP
19     xaxis grid;
20     yaxis grid;
21 run;
22
23 ods graphics / re
```

散布図の描画 – 実行結果



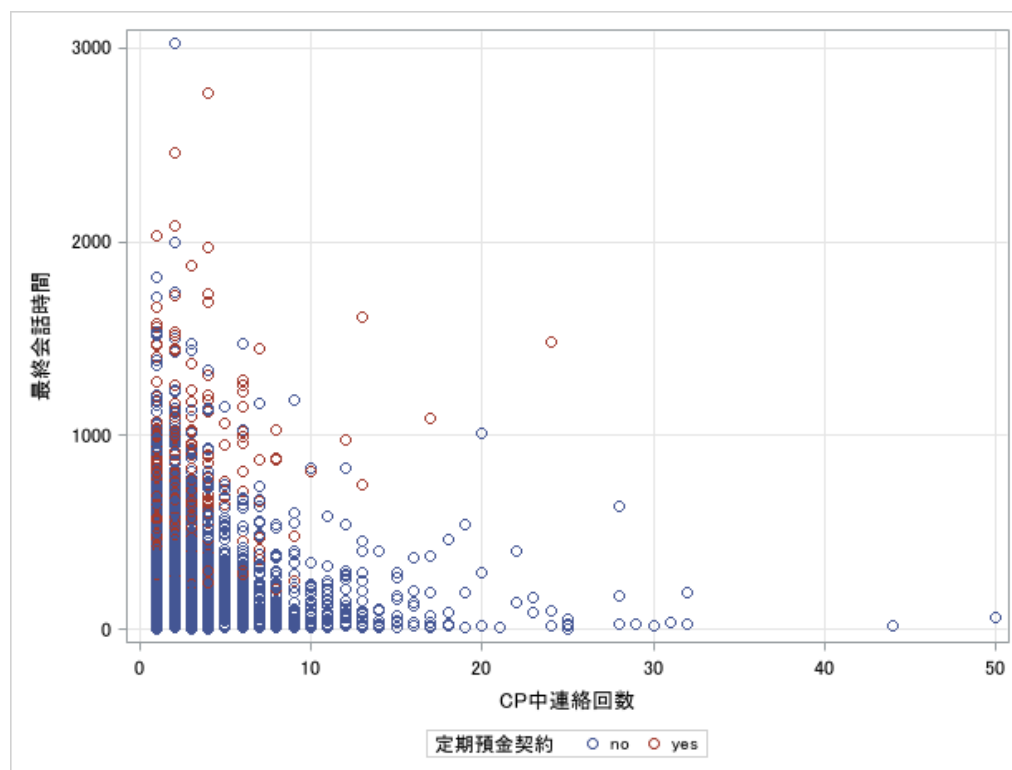
層別散布図 – 実行方法と実行結果

- ・ カテゴリー値で層別化した散布図を描くことで、当初見えなかった関係性が見えてくる可能性がある。分析前における変数間の関係把握、仮説検討を行う上で極めて重要なプロセス

[グループ]に層別したい変数を設定し、再度実行

SAS Studio interface showing the setup for a stratified scatter plot. The 'Run' button is highlighted with a blue box and labeled **実行ボタン**. The 'Data' tab is selected, and 'WORK.IMPORT1' is the data source. The 'X-axis' is 'CP中連絡回数' and the 'Y-axis' is '最終会話時間'. The 'Group' is '定期預金契約'.

「グループ」に層別したい変数を設定
(まずは目的変数を設定してみるのが効果的)



参考：グループごとに散布図を分けて出力

[追加役割]→[グループ分析]
に層別したい変数を設定し、再度実行

*#1_ロジスティック回帰.cpf ×

#1_ロジスティック回帰 実行ボタン

設定 コード/結果 分割

データ 表示 情報 ノード

▼データ

WORK.IMPORT1

フィルタ: (なし)

▼役割

*X軸: (1項目)

CP中連絡回数

*Y軸: (1項目)

最終会話時間

グループ: (1項目)

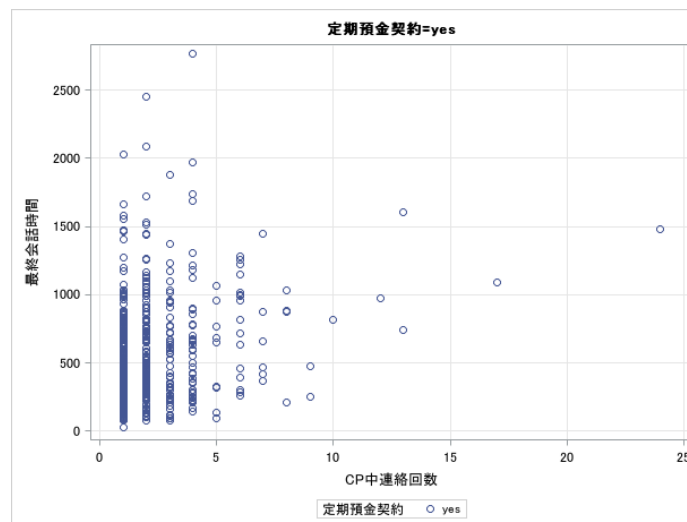
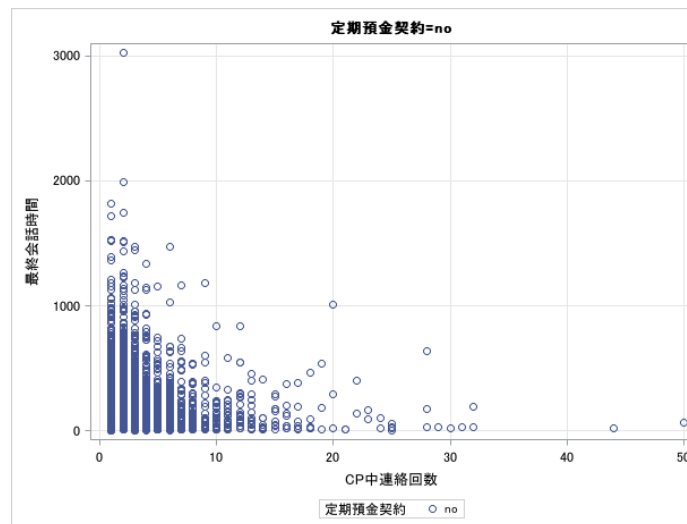
列

▼追加役割

グループ分析: (1項目)

定期預金契約

「グループ分析」に層別したい変数を設定
(まずは目的変数を設定してみるのが効果的)



層別散布図行列 – 実行方法 (1/2)

- ①これまでと同様に、[タスクとユーティリティ]→[タスク]→[統計量]→[データ探索] を選択し、データインポートノードのコントロールポートにドラッグ&ドロップ
- ②生成された [データ探索] ノードをダブルクリックして、詳細設定画面を開く

The screenshot displays the SAS Studio environment. On the left, the 'タスクとユーティリティ' (Tasks and Utilities) pane is expanded to 'タスク' (Tasks), then '統計量' (Statistics), and finally 'データ探索' (Data Exploration), which is highlighted with a blue box. A dashed blue arrow labeled 'ドラッグ & ドロップ' (Drag & Drop) points from this menu item to the 'コントロールポート' (Control Port) of a 'bank_market' data import node in the workflow. The workflow itself is titled '*#1_ロジスティック回帰.cpf'. Below the import node, there are three other nodes: '要約統計量' (Summary Statistics), '散布図' (Scatter Plot), and 'データ探索' (Data Exploration). The 'データ探索' node is highlighted with a blue box, and a dashed blue arrow labeled 'ダブルクリック' (Double Click) points to it. The right pane shows the 'フロー' (Flow) tab with various icons for workflow management.

層別散布図行列 – 実行方法 (2/2)

[データ]の設定

実行ボタン

データ

WORK.IMPORT1

データソースの設定確認

連続変数: (6 項目)

- 年間平均残高
- 最終連絡日
- 最終会話時間
- CP中連絡回数
- 最終連絡日数
- CP前連絡回数

分類変数: (2 項目)

- 定期預金契約

可視化する変数の設定 (数値型) ※上限6変数まで

層別化する変数の設定 ※散布図同様、まずは目的変数を設定

追加役割

[プロット]の設定

実行ボタン

プロット

散布図行列

☒ 散布図行列

☒ ヒストグラムを追加する

☐ 正規分布の密度曲線を追加する

☐ カーネル密度推定を追加する

☐ 予測精度を追加する

対散布図

☐ 対散布図

回帰散布図

☐ 回帰散布図

ヒストグラム

☐ ヒストグラム

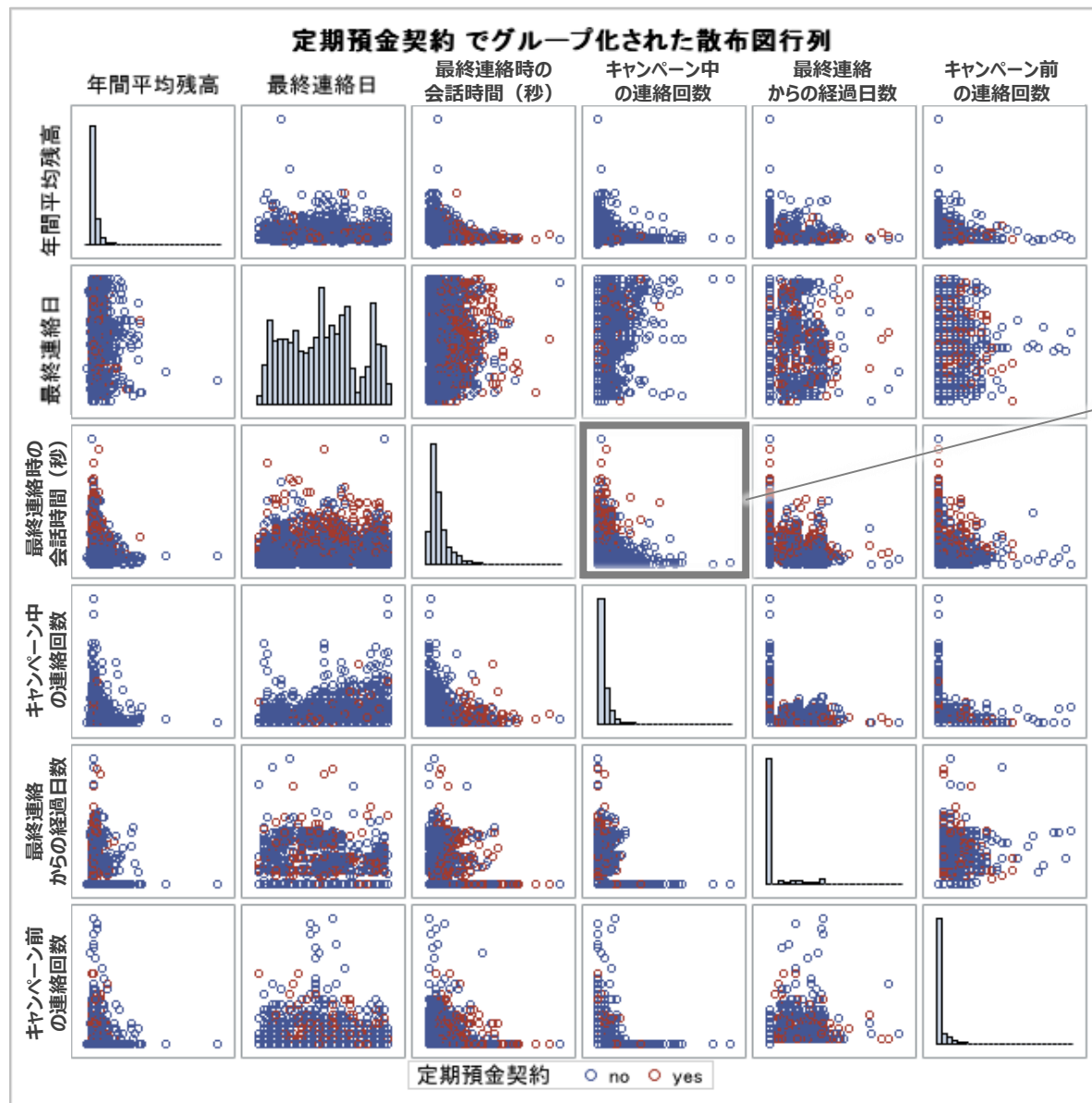
箱ひげ図

分類変数の指定時に使用できます。

☐ 比較箱ひげ図

このノードでも追加で個別のヒストグラム、箱ひげ図を出力可能

層別散布図行列 – 実行結果

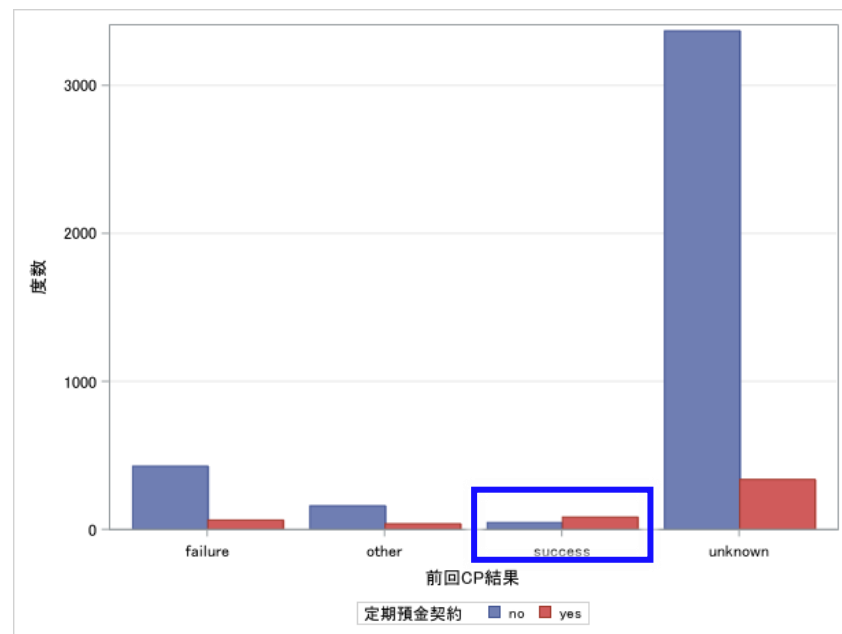
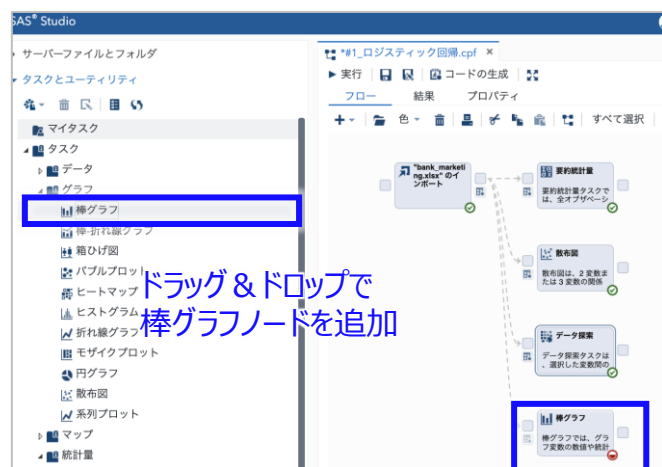


先ほど単一の散布図で確認した2変数

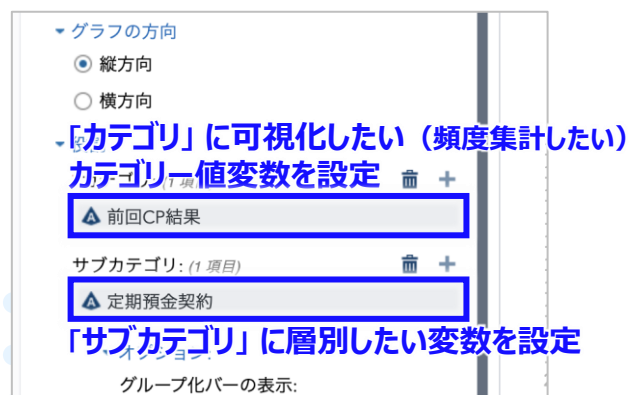
参考：カテゴリー変数の可視化方法

- カテゴリー変数における可視化の基本は、「**頻度集計**」もしくは、前頁までで活用した「**層別化変数**」としての活用である
- ここでは、棒グラフを用いた頻度集計の可視化方法について紹介する

▼棒グラフノードの追加



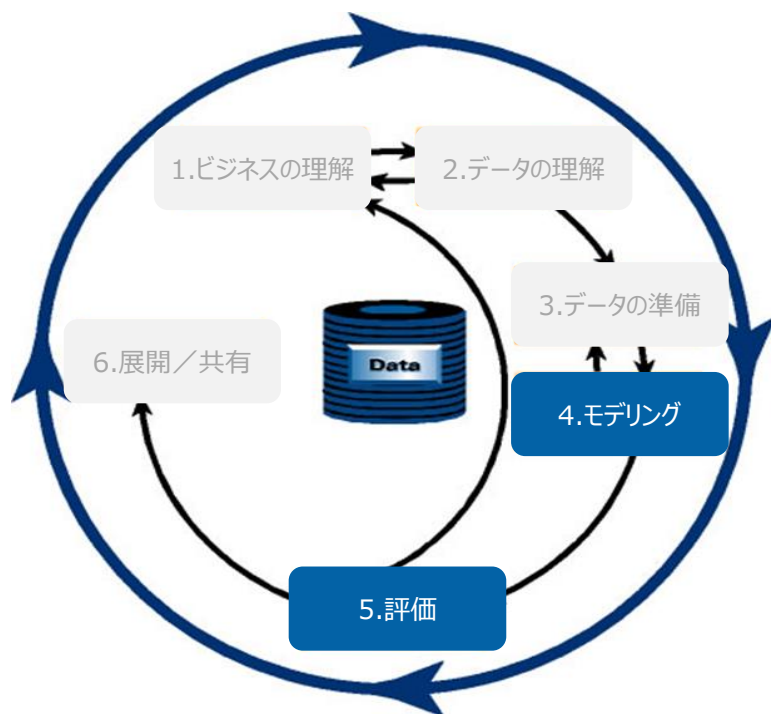
▼ノードの詳細設定



ビッグデータ分析の進め方

- データマイニングの進め方に関する方法論「**CRISP-DM**」に基づいて、分析と評価を繰り返して試行錯誤しながら進めるのが一般的である

CRISP-DM: データマイニング方法論



(CRoss Industry Standard Process for Data Mining)

1. ビジネスの理解

- ビジネス、データマイニング目標の決定
- プロジェクトの立ち上げ

2. データの理解

- データの収集
- データの調査
- データ品質の検証

3. データの準備

- データの選択や除外
- データのクリーニング
- データの構築や統合

4. モデル作成

- モデリング手法の選択
- モデルの作成
- モデルの評価

5. 評価

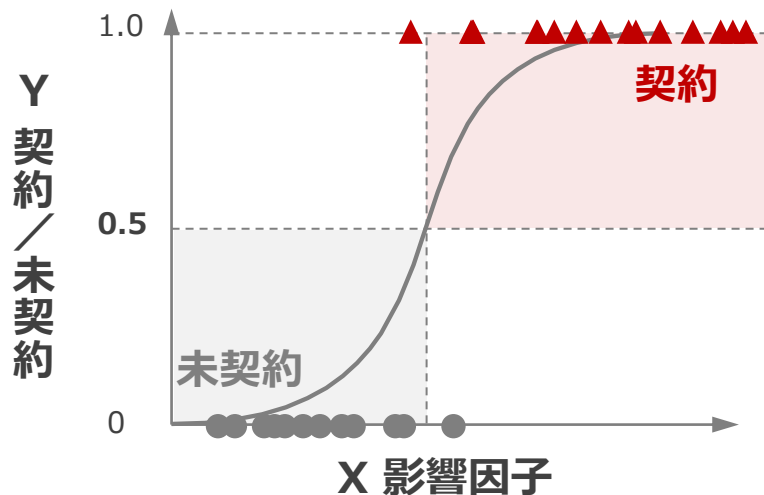
- データマイニングの結果の評価
- プロセスの見直し
- 実行可能なアクションリストの作成

6. 展開／共有

- 業務への導入計画
- モニタリング、メンテナンスの計画

ロジスティック回帰分析の基本

- ロジスティック回帰は、回帰させる関数にS字カーブ状のロジット関数を用いることで、**二クラス分類**（購入する／しない、良品／不良品、生／死 など）を可能とする回帰



オッズ = 発生しない確率 (1-p) に対する発生する確率 (p) の比

$$\ln \left[\frac{P}{1-P} \right] = a_1 X_1 + a_2 X_2 + \dots + a_n X_n + b$$

ロジット

P: ある事象の発生確率

※ 目的変数にロジット関数を用いる以外は重回帰と同様の式

ロジスティック回帰のバリエーション

順序ロジスティック回帰：順序尺度（1,2,3 / 大,中,小 など）の予測に用いる

多項ロジスティック回帰：3クラス以上の名義尺度（犬,猫,人 / 病気の種類 など）の予測に用いる

分類モデルの評価指標

- モデル評価は、分析対象や目的に応じて最適な指標を選択したり、複数の指標を用いて行う

混同行列

※予測対象が「契約」の場合

予測結果			
		契約	未契約
実データ	契約	1 正答 (True Positive)	3 誤り：見逃し (False Negative)
	未契約	2 誤り：誤報 (False Positive)	4 正答 (True Negative)

再現率

特異度

適合率
(陽性的中率)

陰性的中率

予測

実際



誤報



見逃し

モデル評価指標

正解率
(Accuracy)

$$= \frac{1 + 4}{1 + 2 + 3 + 4}$$

適合率, 精度, 陽性的中率
(Precision, PPV*)
*Positive Predictive Value

$$= \frac{1}{1 + 2 + 3 + 4}$$

誤報の少なさ

$$\left[\begin{array}{l} \leftrightarrow \text{陰性的中率 (NPV*)} \\ \text{*Negative Predictive Value} \end{array} \right] = \frac{4}{1 + 2 + 3 + 4}$$

負例の誤報の少なさ

再現率, 感度, 真陽性率
(Recall, Sensitivity)

$$= \frac{1}{1 + 2 + 3 + 4}$$

見逃し (取りこぼし) の少なさ

$$\left[\begin{array}{l} \leftrightarrow \text{特異度, 真陰性率} \\ \text{(Specificity)} \end{array} \right] = \frac{4}{1 + 2 + 3 + 4}$$

負例の見逃しの少なさ

F値, F1スコア
(F-measure)

$$= \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

誤報と見逃しの少なさのバランス
Precision と Recall の調和平均

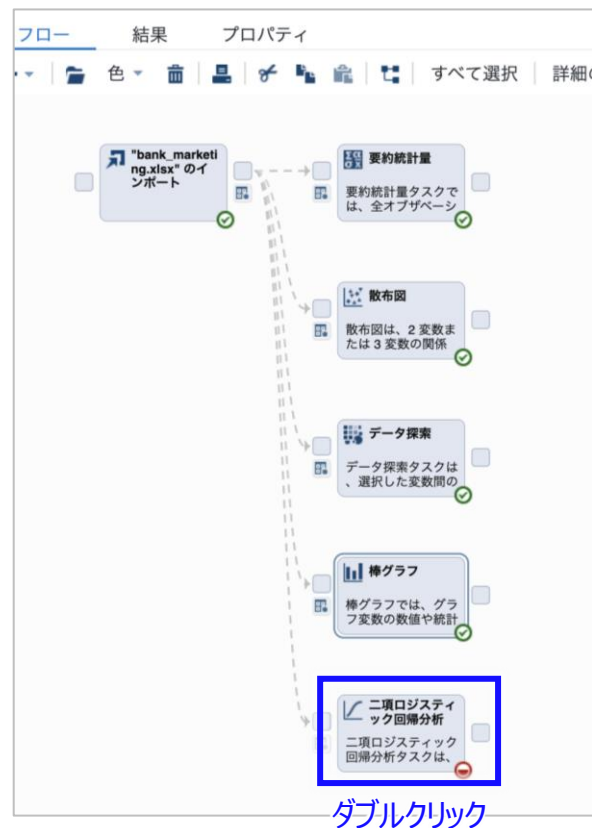
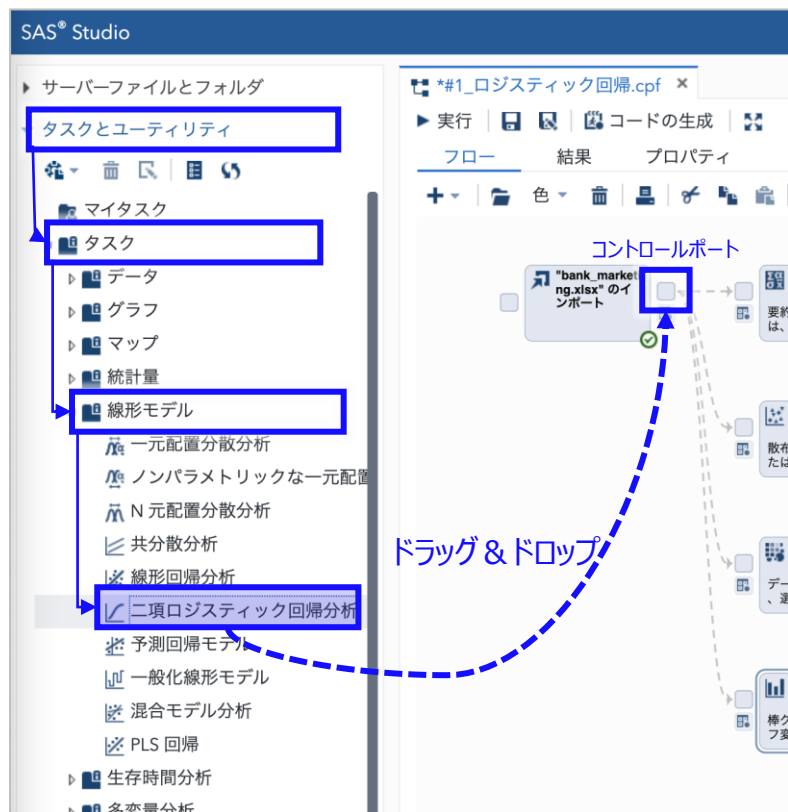
SAS Studio での実装方法

- ロジスティック回帰

ロジスティック回帰 – 実行方法 (1/2) ノードの設置

- ①左パネルより、[タスクとユーティリティ]→[タスク]
→[線形モデル]→[二項ロジスティック回帰分析]を選択
- ②右側のプロセスフロー内のインポートノードの
右端の四角  (コントロールポート) の上ヘドラッグ&ドロップ

- ③プロセスフロー上に二項ロジスティック回帰分析ノードが
生成されるのでダブルクリックして詳細設定画面を開く



ロジスティック回帰 – 実行方法 (2/2) 説明/目的変数・オプションの設定

[データ]の設定 (説明変数・目的変数)

#1 ロジスティック回帰 > 二項ロジスティック回帰分析

設定 コード/結果 分割

データ モデル 選択

データ
WORK.IMPORT1

フィルタ: (なし)

データソースの設定確認

役割

応答
☐ 応答データはイベント数と試行数を含む

* 応答: (1 項目)

☒ 定期預金契約

* 関心のあるイベント:

yes

リンク関数:
ロジット

説明変数

分類変数:

☒ 職業
☒ 結婚歴
☒ 学歴
☒ クレカ債務
☒ 住宅ローン
☒ 個人ローン

▶ 効果のパラメータ化
▶ 欠損値の処理

連続変数:

☒ 年齢
☒ 年間平均残高
☒ 最終連絡日
☒ 最終会話時間
☒ CP中連絡回数

目的変数の設定

目的変数のクラスの中で
発見したいクラスを設定
※yes=1と捉え、yesを予測する
モデル構築を行う

説明変数の設定① : カテゴリー値型変数

[Tips]先に連続変数から
設定の方が選択が楽!

説明変数の設定② : 数値型変数

[オプション]の設定 (各種出力)

#1 ロジスティック回帰 > 二項ロジスティック回帰分析

設定 コード/結果 分割 オプション 出力

フィル 選択 オプション 出力

統計量

表示する統計量の選択:

デフォルトおよび追加統計量

☒ 分類テーブル
☐ 偏相関
☐ 一般化寄与率

▶ 適合度と過分散
▶ 多重比較
▶ 正確検定

パラメータ推定値

☒ 標準化された推定値
☐ 指数化された推定値
☐ パラメータ推定値の相関
☐ パラメータ推定値の共分散

パラメータ推定値の信頼区間:
なし

オッズ比の信頼区間:
なし

信頼水準:
95%

診断統計量

プロット

表示するプロットの選択:

デフォルトおよび追加プロット

☐ 標準化された DFBETA とオブザベーション番号
☐ 影響度に関する統計量とオブザベーション番号
☐ モデルの当てはめへの影響度とパラメータ推定値
☐ 予測確率プロット
☐ 効果プロット

☒ オッズ比プロット
☐ 0 から 2 までのオッズ比を表示する

☒ ROC プロット

表示形式:

[表示する統計量の選択] を
「デフォルトおよび追加統計量」
に変更

[分類テーブル] と
[標準化された推定値] に
チェックを入れる

[表示するプロットの選択] を
「デフォルトおよび追加プロット」
に変更

[オッズ比プロット] と
[ROCプロット] に
チェックを入れる

ロジスティック回帰 – 実行結果（1/4）基本情報

- 分析結果画面の一番初めには、読み込んだデータの件数、目的変数における各クラスの件数内訳が出力される
- 本データは、未契約:4,000件 vs 契約:521件 の不均衡性の高いデータであることがわかる

#1 ロジスティック回帰 実行ボタン

設定 | コード/結果 | 分割 | **実行** | 出力

フィル | 選択 | オプション | 出力

▼ 統計量

表示する統計量の選択:
デフォルトおよび追加統計量

☒ 分類テーブル
☐ 偏相関
☐ 一般化寄与率

▶ 適合度と過分散
▶ 多重比較
▶ 正確検定

▼ パラメータ推定値

☒ 標準化された推定値
☐ 指数化された推定値
☐ パラメータ推定値の相関
☐ パラメータ推定値の共分散

パラメータ推定値の信頼区間:
なし

オッズ比の信頼区間:
なし

信頼水準:
95%

▶ 診断統計量

モデルの情報		
データセット	WORK.IMPORT1	
応答変数	定期預金契約	定期預金契約
応答の水準数	2	
モデル	binary logit	
最適化の手法	Fisher's scoring	

読み込んだオブザベーション数	4521
使用されたオブザベーション数	4521

反応プロファイル		
順番	定期預金契約	度数の合計
1	no	4000
2	yes	521

未契約 : 88.5%

契約 : 11.5%

ロジスティック回帰 – 実行結果 (2/4) 精度

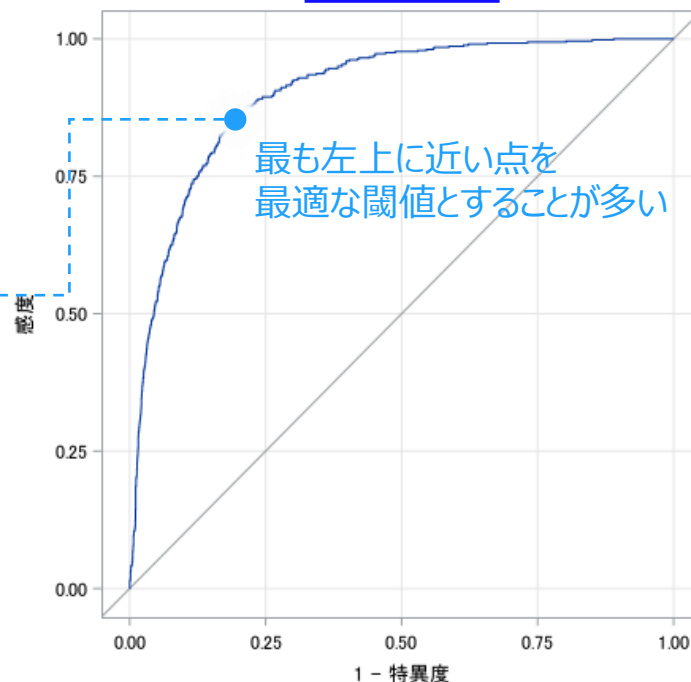
- ROC曲線では、左上に凸形状をしており、曲線下面積（AUC）も0.9程度とまずまずである
- 最もバランスの良い閾値は0.1前後であり、感度（再現率）81.8%，特異度80.8% である

閾値を増加

分類表										
確率水準	正解		不正解		パーセンテージ					
	事象	非事象	事象	非事象	正解	感度	特異度	正の予測	負の予測	
	0.000	521	0	4000	0	11.5	100.0	0.0	11.5	.
	0.020	517	1073	2927	4	35.2	99.2	26.8	15.0	99.6
	0.040	502	2076	1924	19	57.0	96.4	51.9	20.7	99.1
	0.060	477	2623	1377	44	68.6	91.6	65.6	25.7	98.4
	0.080	454	3016	984	67	76.8	87.1	75.4	31.6	97.8
	0.100	426	3232	768	95	80.9	81.8	80.8	35.7	97.1
	0.120	401	3366	634	120	83.3	77.0	84.2	38.7	96.6
	0.140	383	3477	523	138	85.4	73.5	86.9	42.3	96.2
0.160	362	3560	440	159	86.8	69.5	89.0	45.1	95.7	
0.180	344	3607	393	177	87.4	66.0	90.2	46.7	95.3	
0.200	326	3649	351	195	87.9	62.6	91.2	48.2	94.9	
0.220	316	3682	318	205	88.4	60.7	92.1	49.8	94.7	
0.240	308	3714	286	213	89.0	59.1	92.9	51.9	94.6	
0.260	292	3745	255	229	89.3	56.0	93.6	53.4	94.2	
0.280	282	3760	240	239	89.4	54.1	94.0	54.0	94.0	

モデルに対する ROC 曲線

曲線下面積 = 0.9026



▼参考：閾値=0.5の時の精度

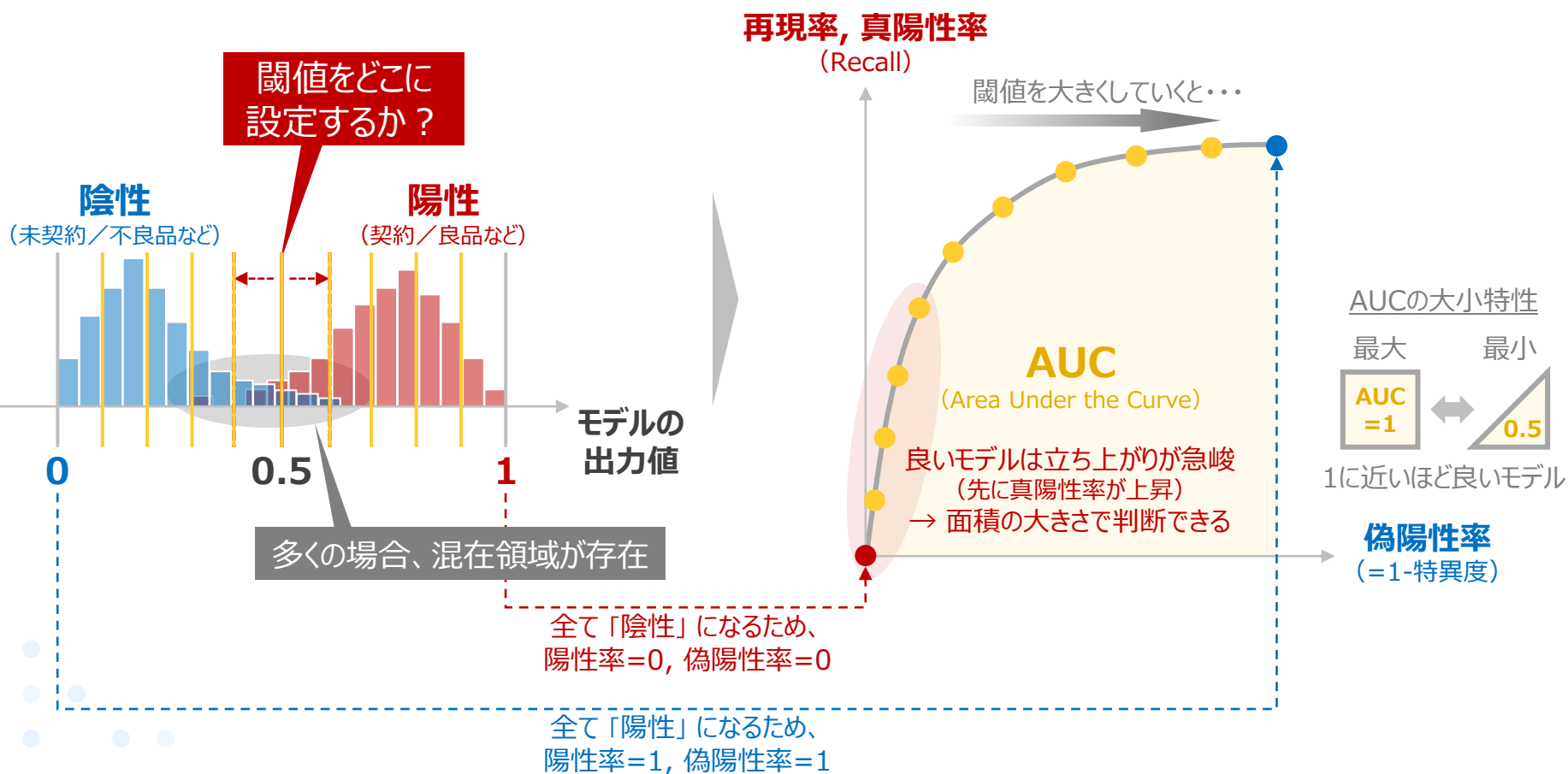
分類表									
確率水準	正解		不正解		パーセンテージ			正の予測	負の予測
	事象	非事象	事象	非事象	正解	感度	特異度		
0.500	166	3907	93	355	90.1	31.9	97.7	64.1	91.7



適合率
(陽性的中率)

参考：ROC曲線によるモデル評価

- **ROC曲線***は、再現率（真陽性率）と偽陽性率のバランスの観点から「良いモデル」を評価する指標 *Receiver Operating Characteristic curve; 受信者動作特性曲線・・・元々は軍事のレーダー性能評価に用いられた指標
- **AUC** (Area Under the Curve; 曲線下面積) の算出と併せて用いる



ロジスティック回帰 – 実行結果 (3/4) 回帰係数

- 標準化回帰係数を見ると、**学歴=tertiary** (高等教育) や、**個人ローン=無**、**連絡手段=電話**、**前回CPの結果=success**、**最終会話時間**で、有意に高い値を示し、関係性が示唆される
- 一方で、最終連絡月が夏 (7, 8月) や5月・11月になると契約率が下がる可能性が示唆される

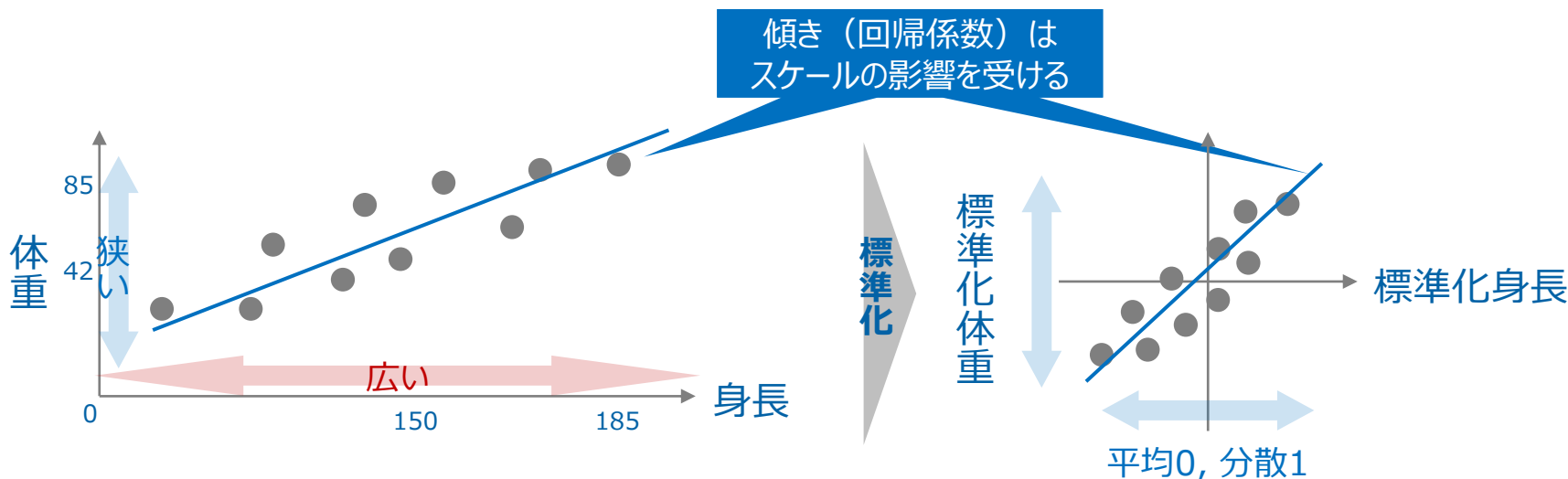
▼各説明変数の回帰係数

パラメータ		自由度	推定値	標準誤差	Wald カイ 2 乗値	Pr > ChiSq	標準化した推定値
Intercept		1	-3.8928	0.9183	17.9716	<.0001	
職業	admin.	1	-0.5206	0.5853	0.7914	0.3737	-0.0883
職業	blue-collar	1	-0.9131	0.5810	2.4695	0.1161	-0.2048
職業	entrepreneur	1	-0.7704	0.6460	1.4223	0.2330	-0.0804
職業	housemaid	1	-0.8736	0.6636	1.7333	0.1880	-0.0749
職業	management	1	-0.5937	0.5707	1.0822	0.2982	-0.1343
職業	retired	1	0.1109	0.5985	0.0343	0.8531	0.0134
職業	self-employed	1	-0.7018	0.6294	1.2433	0.2648	-0.0763
職業	services	1	-0.6663	0.5986	1.2388	0.2657	-0.1063
職業	student	1	-0.1422	0.6464	0.0484	0.8259	-0.0106
職業	technician	1	-0.7133	0.5743	1.5427	0.2142	-0.1477
職業	unemployed	1	-1.1602	0.6711	2.9891	0.0838	-0.1061
職業	unknown	0	0	.	.	.	.
結婚歴	divorced	1	0.3051	0.2038	2.2414	0.1344	0.0540
結婚歴	married	1	-0.1644	0.1478	1.2385	0.2658	-0.0440
結婚歴	single	0	0	.	.	.	.
学歴	primary	1	0.4210	0.3572	1.3890	0.2386	0.0829
学歴	secondary	1	0.5011	0.3245	2.3846	0.1225	0.1381
学歴	tertiary	1	0.7418	0.3361	4.8711	0.0273	0.1872
学歴	unknown	0	0	.	.	.	.
クレカ債務	no	1	-0.5446	0.4315	1.5935	0.2068	-0.0386
クレカ債務	yes	0	0	.	.	.	.
住宅ローン	no	1	0.2600	0.1381	3.5463	0.0597	0.0710
住宅ローン	yes	0	0	.	.	.	.
個人ローン	no	1	0.6296	0.2000	9.9137	0.0016	0.1249
個人ローン	yes	0	0	.	.	.	.

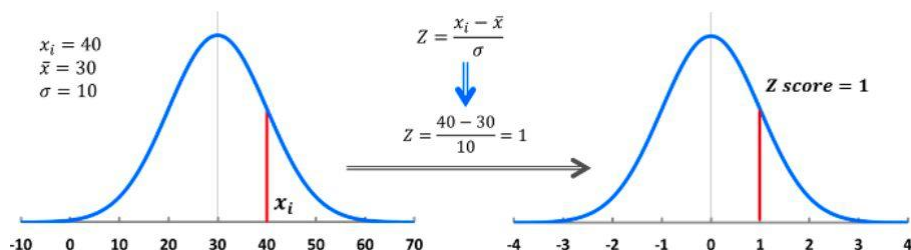
パラメータ		自由度	推定値	標準誤差	Wald カイ 2 乗値	Pr > ChiSq	標準化した推定値
連絡手段	cellular	1	1.4161	0.2277	38.6781	<.0001	0.3747
連絡手段	telephone	1	1.3459	0.3119	18.6232	<.0001	0.1850
連絡手段	unknown	0	0	.	.	.	.
最終連絡月	apr	1	-0.6572	0.4115	2.5504	0.1103	-0.0892
最終連絡月	aug	1	-0.9653	0.3995	5.8382	0.0157	-0.1847
最終連絡月	dec	1	-0.5428	0.7216	0.5657	0.4520	-0.0199
最終連絡月	feb	1	-0.4550	0.4233	1.1552	0.2825	-0.0542
最終連絡月	jan	1	-1.7805	0.4998	12.6885	0.0004	-0.1747
最終連絡月	jul	1	-1.4087	0.4113	11.7294	0.0006	-0.2820
最終連絡月	jun	1	-0.1030	0.4252	0.0587	0.8086	-0.0183
最終連絡月	mar	1	0.8413	0.4959	2.8783	0.0898	0.0480
最終連絡月	may	1	-1.1472	0.3981	8.3049	0.0040	-0.2924
最終連絡月	nov	1	-1.5002	0.4228	12.5876	0.0004	-0.2320
最終連絡月	oct	1	0.7038	0.4559	2.3832	0.1226	0.0512
最終連絡月	sep	0	0	.	.	.	.
前回CP結果	failure	1	0.1216	0.3199	0.1445	0.7038	0.0208
前回CP結果	other	1	0.6128	0.3495	3.0749	0.0795	0.0690
前回CP結果	success	1	2.5665	0.3148	66.4610	<.0001	0.2356
前回CP結果	unknown	0	0	.	.	.	.
年齢		1	-0.00423	0.00713	0.3528	0.5525	-0.0247
年間平均残高		1	-3.91E-6	0.000017	0.0500	0.8230	-0.00649
最終連絡日		1	0.0164	0.00816	4.0427	0.0444	0.0746
最終会話時間		1	0.00423	0.000202	437.3055	<.0001	0.6053
CP中連絡回数		1	-0.0704	0.0282	6.2314	0.0126	-0.1207
最終連絡日数		1	-0.00010	0.000996	0.0097	0.9217	-0.00540
CP前連絡回数		1	-0.00551	0.0382	0.0208	0.8853	-0.00514

参考：回帰係数とデータの標準化

- 機械学習では、各変数間でスケール（値範囲）が大きく異なると、計算に時間がかかったり、回帰係数などのパラメータの直接比較が困難になるため、**スケールを揃える**ことが有効
- 特に、各変数を**平均0, 分散1に変換する「標準化」**を用いることが多い



▼標準化の方法（Zスコア）



$$Z \text{ score} = \frac{x_i - \mu}{\sigma}$$

x_i = 母集団を構成する要素iの値
 μ = 母平均
 σ = 標準偏差

※偏差値は、「平均が50、標準偏差が10」
になるように変換したもの： $50 + 10 \times Z \text{スコア}$

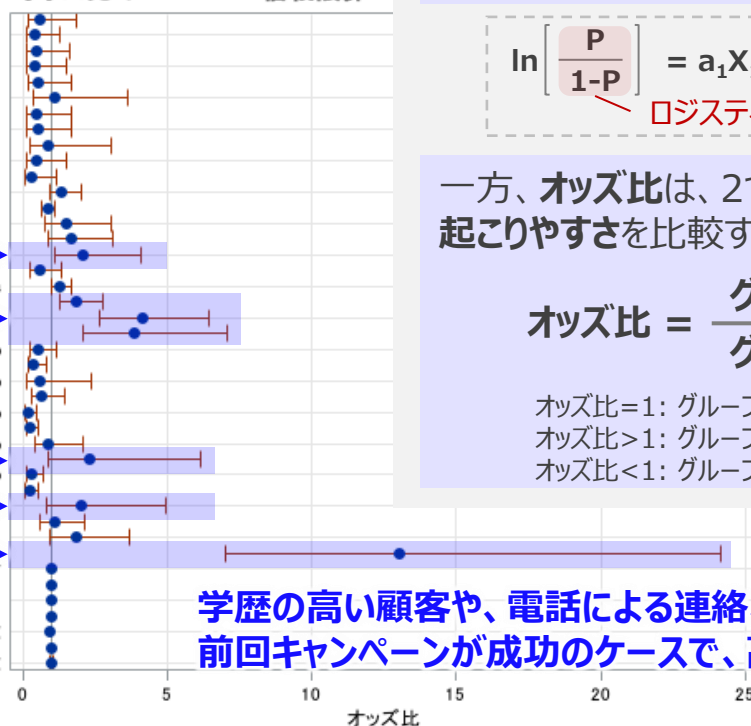
ロジスティック回帰 – 実行結果 (4/4) オッズ比

- 2群間の事象の起こりやすさを表す「**オッズ比**」では、関係性が明白である「前回キャンペーン成功」以外にも、学歴や電話連絡、3月の連絡など、意外な影響因子が浮かび上がった

効果	点推定	95% Wald 信頼限界
職業 admin. vs unknown	0.594	0.189 1.871
職業 blue-collar vs unknown	0.401	0.128 1.253
職業 entrepreneur vs unknown	0.463	0.130 1.642
職業 housemaid vs unknown	0.417	0.114 1.533
職業 management vs unknown	0.552	0.180 1.690
職業 retired vs unknown	1.117	0.346 3.611
職業 self-employed vs unknown	0.496	0.144 1.702
職業 services vs unknown	0.514	0.159 1.660
職業 student vs unknown	0.867	0.244 3.079
職業 technician vs unknown	0.490	0.159 1.510
職業 unemployed vs unknown	0.313	0.084 1.168
結婚歴 divorced vs single	1.357	0.910 2.023
結婚歴 married vs single	0.848	0.635 1.133
学歴 primary vs unknown	1.524	0.756 3.068
学歴 secondary vs unknown	1.651	0.874 3.118
学歴 tertiary vs unknown	2.100	1.087 4.057
クレカ債務 no vs yes	0.580	0.249 1.351
住宅ローン no vs yes	1.297	0.989 1.700
個人ローン no vs yes	1.877	1.268 2.778
連絡手段 cellular vs unknown	4.121	2.637 6.436
連絡手段 telephone vs unknown	3.842	2.085 7.079
最終連絡月 apr vs sep	0.518	0.231 1.161
最終連絡月 aug vs sep	0.381	0.174 0.833
最終連絡月 dec vs sep	0.581	0.141 2.391
最終連絡月 feb vs sep	0.634	0.277 1.455
最終連絡月 jan vs sep	0.169	0.063 0.449
最終連絡月 jul vs sep	0.244	0.109 0.547
最終連絡月 jun vs sep	0.902	0.392 2.076
最終連絡月 mar vs sep	2.319	0.878 6.130
最終連絡月 may vs sep	0.318	0.146 0.693
最終連絡月 nov vs sep	0.223	0.097 0.511
最終連絡月 oct vs sep	2.021	0.827 4.940
前回CP結果 failure vs unknown	1.129	0.603 2.114
前回CP結果 other vs unknown	1.846	0.930 3.661
前回CP結果 success vs unknown	13.020	7.025 24.132
年齢	0.996	0.982 1.010
年間平均残高	1.000	1.000 1.000
最終連絡日	1.017	1.000 1.033
最終会話時間	1.004	1.004 1.005
CP中連絡回数	0.932	0.882 0.985
最終連絡日数	1.000	0.998 1.002
CP前連絡回数	0.995	0.923 1.072

オッズ比の 95% Wald 信頼限界

職業 blue-collar vs unknown
 職業 housemaid vs unknown
 職業 retired vs unknown
 職業 services vs unknown
 職業 technician vs unknown
 結婚歴 divorced vs single
 学歴 primary vs unknown
 住宅ローン no vs yes
 最終連絡月 apr vs sep
 最終連絡月 dec vs sep
 最終連絡月 jan vs sep
 最終連絡月 jun vs sep
 最終連絡月 mar vs sep
 最終連絡月 may vs sep
 最終連絡月 nov vs sep
 最終連絡月 oct vs sep
 前回CP結果 other vs unknown
 年齢
 最終連絡日
 CP中連絡回数
 CP前連絡回数



学歴の高い顧客や、電話による連絡、3月の連絡、
 前回キャンペーンが成功のケースで、高いオッズ比

オッズは、ある事象が起こる「**見込み**」を表す指標

$$\text{オッズ} = \frac{P}{1 - P} \quad \begin{array}{l} \cdots \text{ある事象が起こる確率} \\ \cdots \text{ある事象が起こらない確率} \end{array}$$

$$\ln \left[\frac{P}{1 - P} \right] = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n + b$$

ロジスティック回帰式のこの部分！

一方、オッズ比は、2つの群間における、ある事象の
 起こりやすさを比較するための指標

$$\text{オッズ比} = \frac{\text{グループ2のオッズ}}{\text{グループ1のオッズ}}$$

オッズ比=1: グループ間で起こりやすさは同じ
 オッズ比>1: グループ2の方が起こりやすい
 オッズ比<1: グループ1の方が起こりやすい

まとめ

- グラフによるデータ観察

- 散布図、層別散布図、散布図行列を用いて、**変数間の関係性を俯瞰的に把握**した
- ヒストグラム、箱ひげ図を用いて、**各変数の分布や外れ値を確認**した

- ロジスティック回帰分析

- 教師あり学習の分類手法には、ロジスティック回帰の他、決定木やサポートベクターマシンなどがある
- ロジスティック回帰の特徴は、**決定境界が直線的**であること、0/1の予測だけでなく**発生確率も出力**できることである
- ロジスティック回帰を用いて、**契約確度の高い顧客を予測**し、精度の評価を行なった
- 回帰係数の比較や、オッズ比の確認を行うことで、**影響因子の探索**を行った



End of File