

SASお客様向けwebinar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

2021年12月1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
 - 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

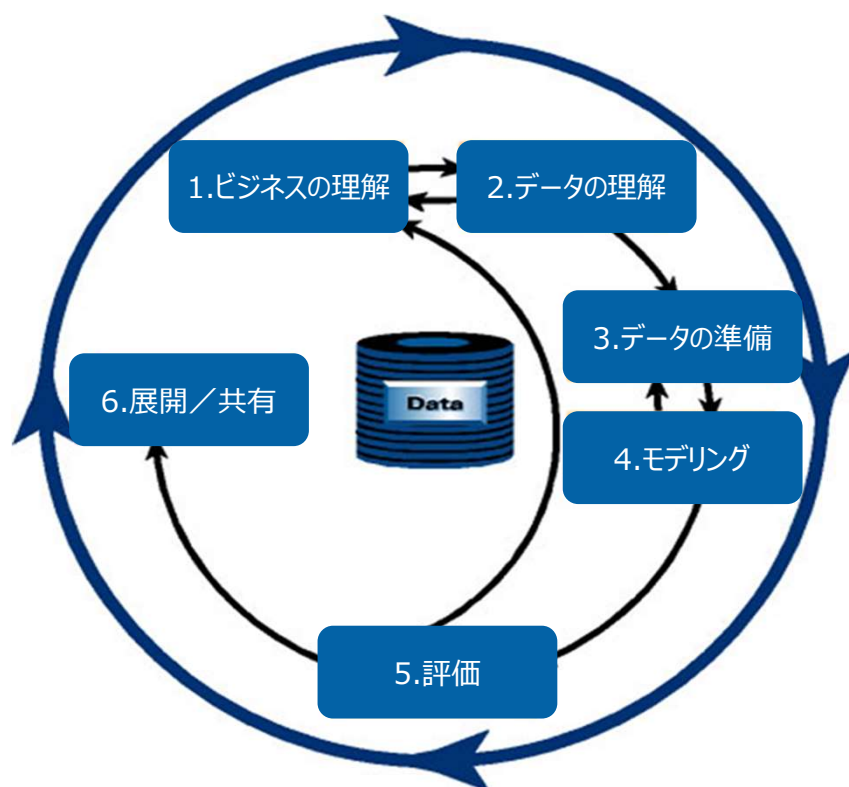
Veracity

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

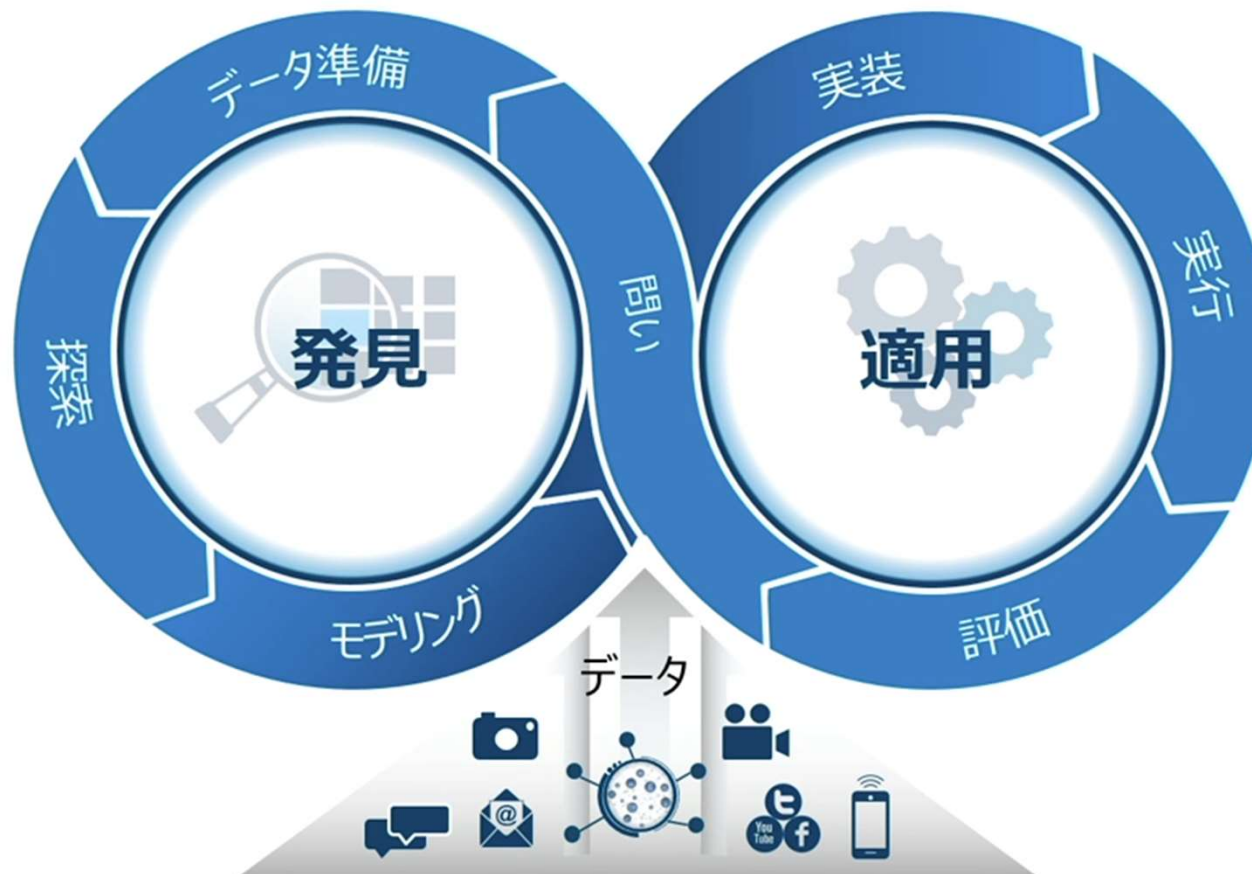
6. 展開／共有

- 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



ビッグデータからの新知識獲得＝「ビッグデータ分析」

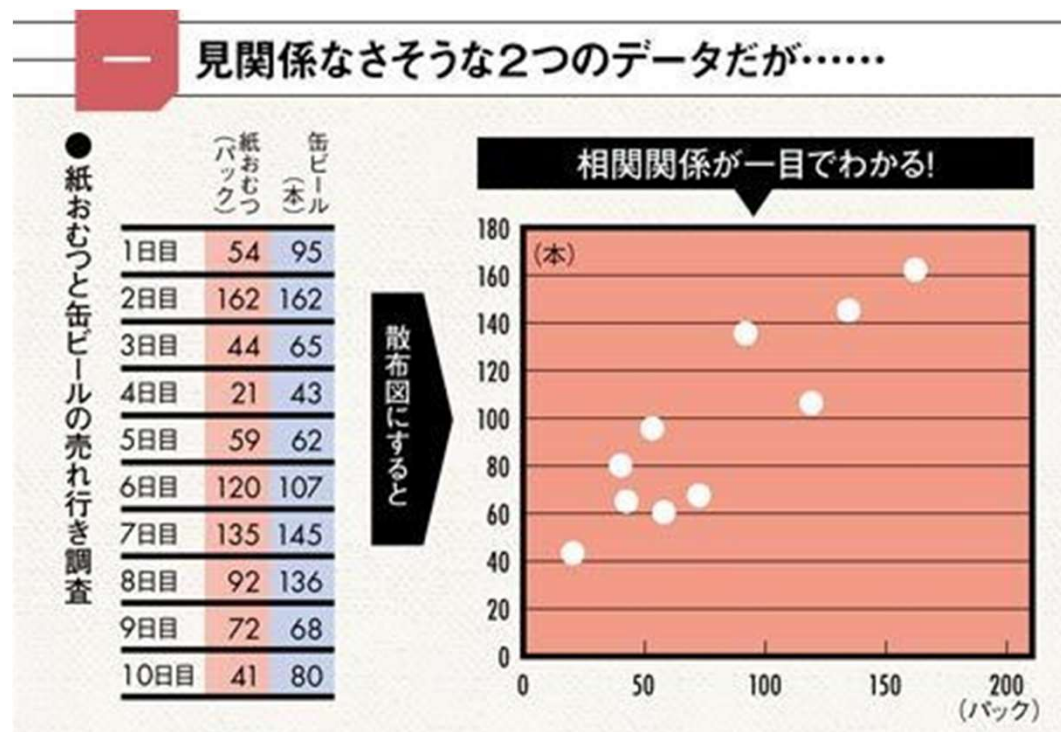
- ビッグデータ分析は、膨大なデータ（＝事実、結果）に潜むデータ間の関係性や特徴（＝新知識）をデータドリブンで抽出する試みである
- データから「**一見関連のなさそうな新たな洞察を得る**」手法と言える

ビール



X

オムツ



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)

参考：AIと機械学習とディープラーニング

- AIは最も広い概念で、基本的には手法に依らないというのが従来の考え方
- 近年では、AI≒ディープラーニング になりつつある

AI（人工知能）

= 人間の知的活動をコンピュータのプログラムで表現

機械学習（ML）

ニューラルネットワーク （NN）

ディープラーニング （深層学習, DL）

DLはニューラルネットの進化版！

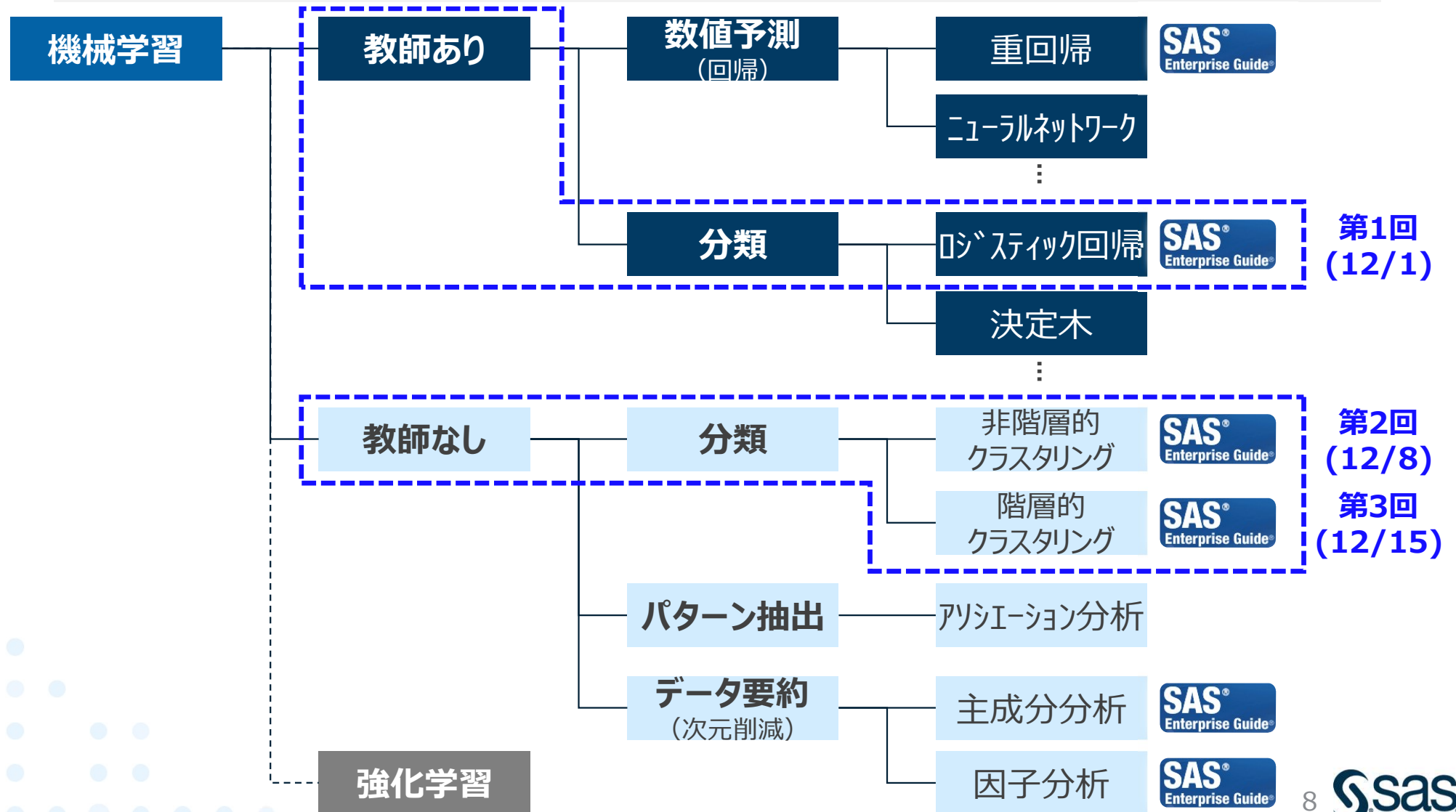
1950

1980

2010

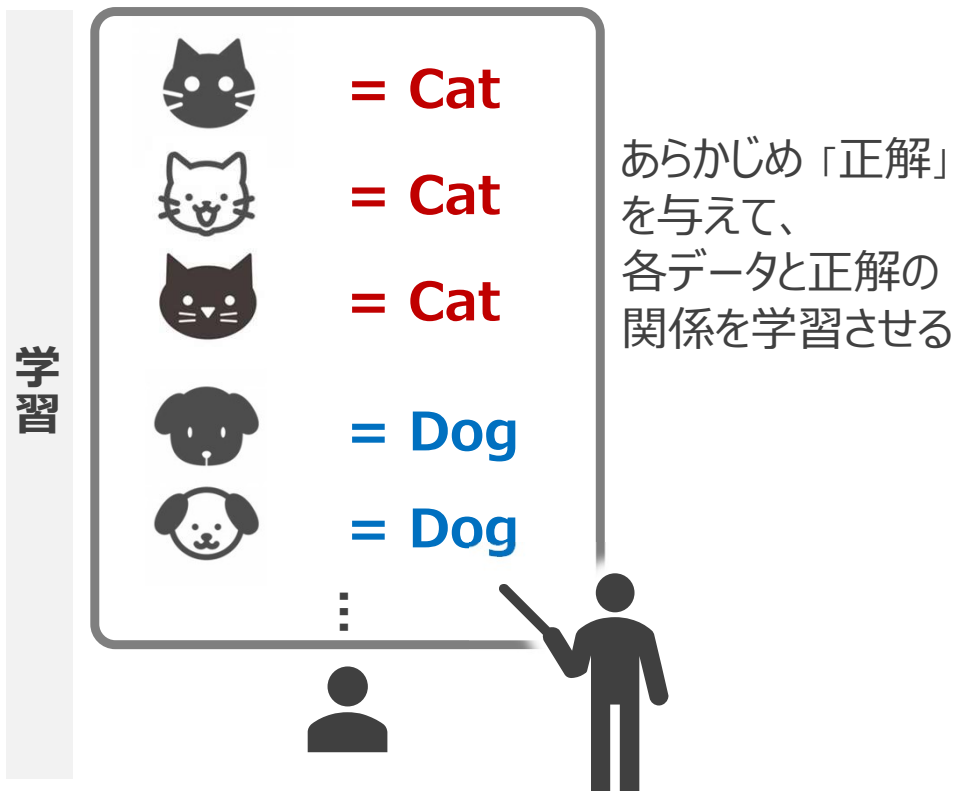
代表的な機械学習手法

- 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

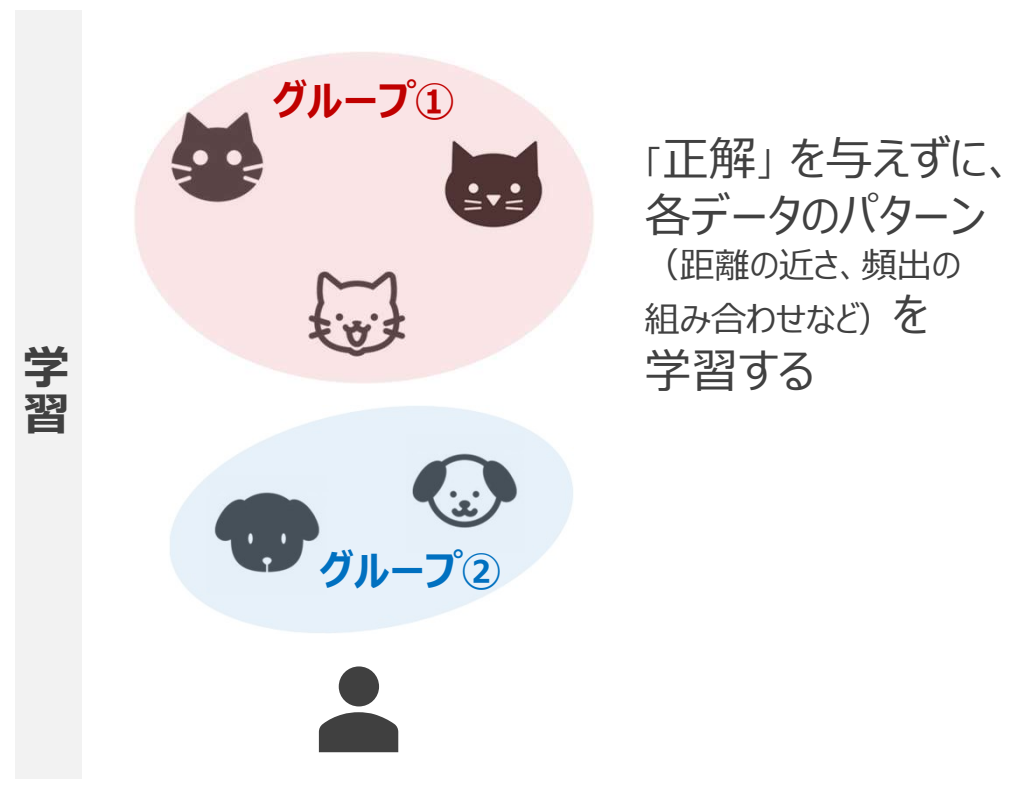


推論



Dog!

教師なし学習



推論



グループ②!

※分類されたグループの
意味づけは人が行う

教師あり学習のイメージ（数値予測と分類）

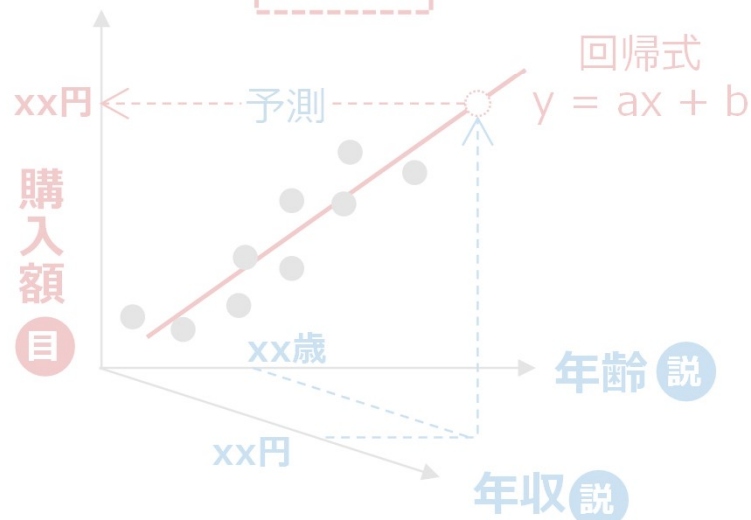
- 各顧客レコードに対して数値 or カテゴリー値（クラス）の解答を与え関係性を学習

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

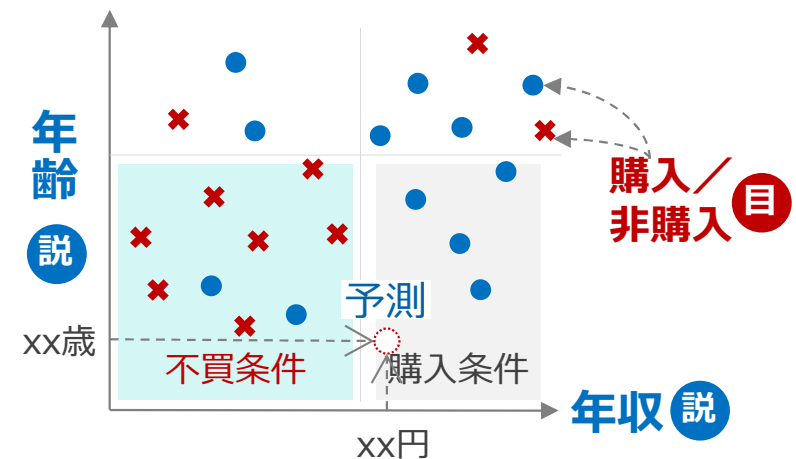
クラス分類

（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

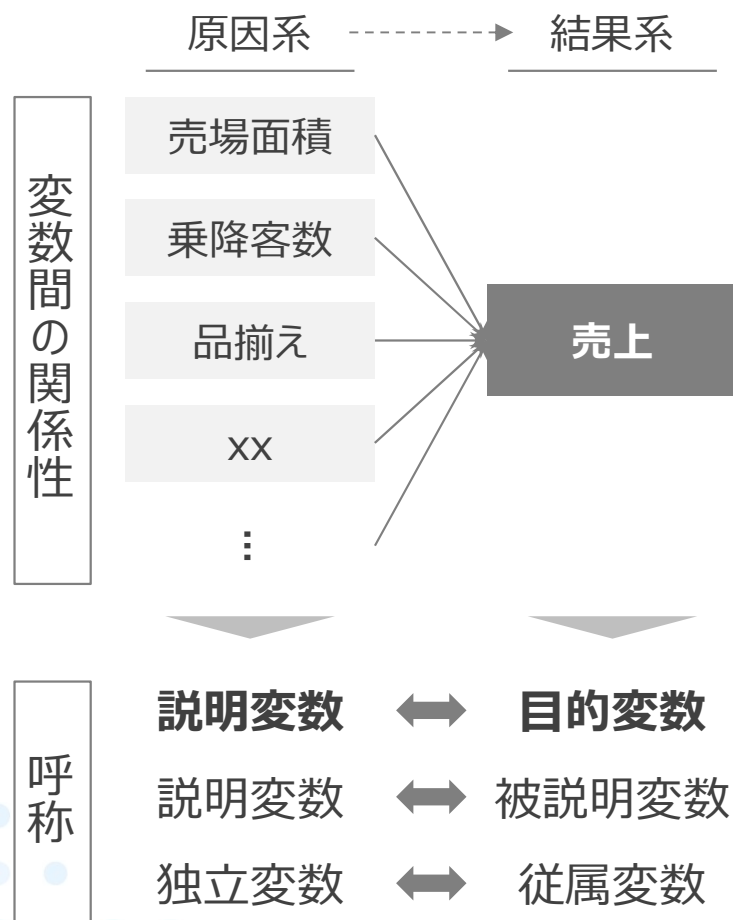
データ例

イメージ

分析例

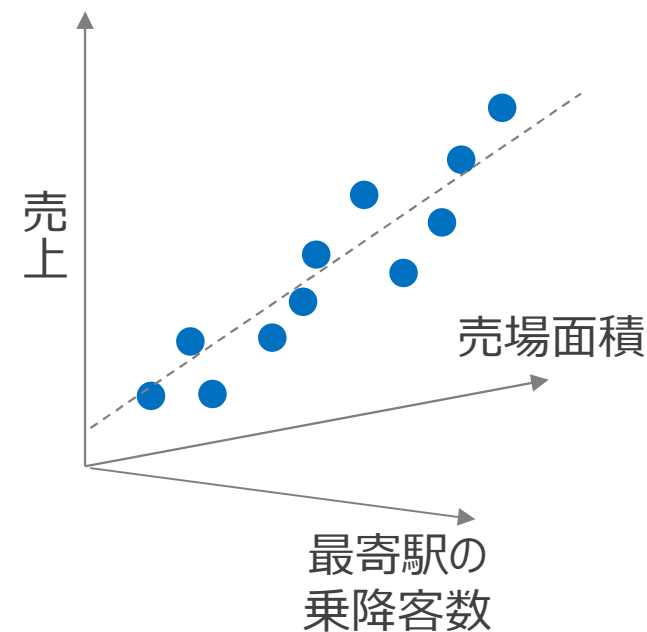
参考：説明変数と目的変数

- 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



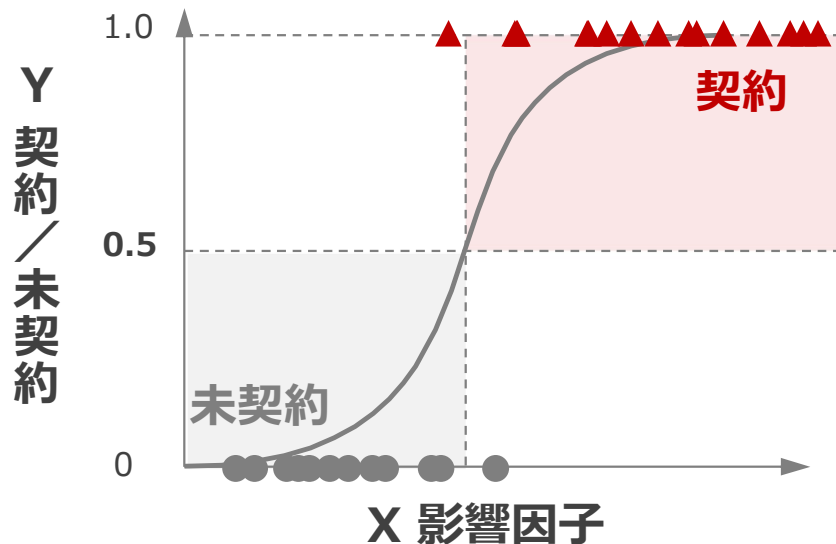
重回帰の場合：

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



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

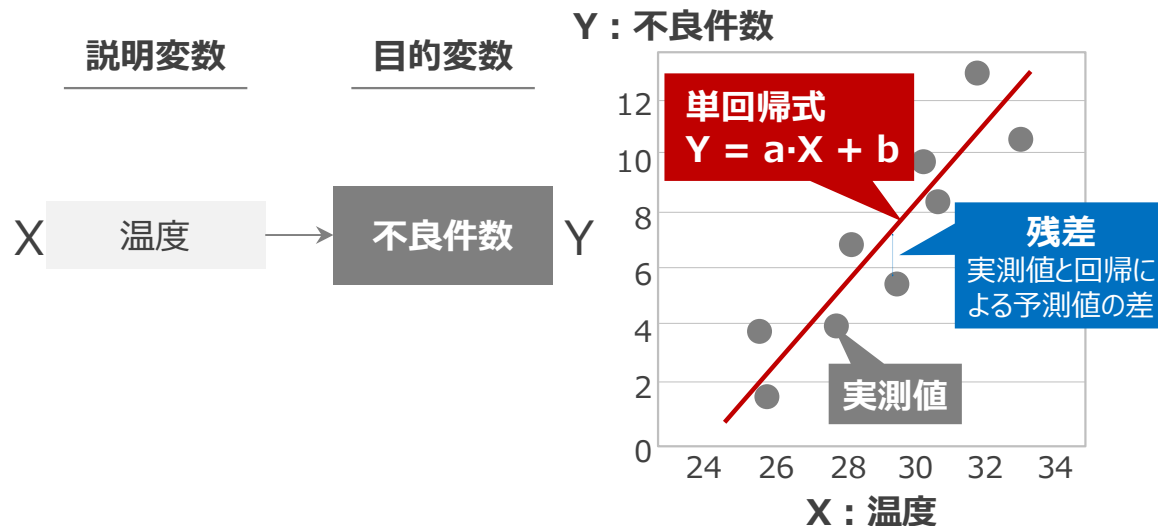
順序ロジスティック回帰：順序尺度（1,2,3 / 大,中,小 など）の予測に用いる

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

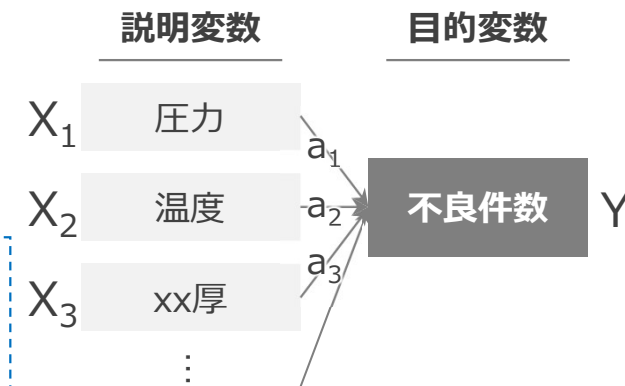


残差の2乗和が最小になるように
係数a, 定数項bを決める (最小二乗法)

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

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

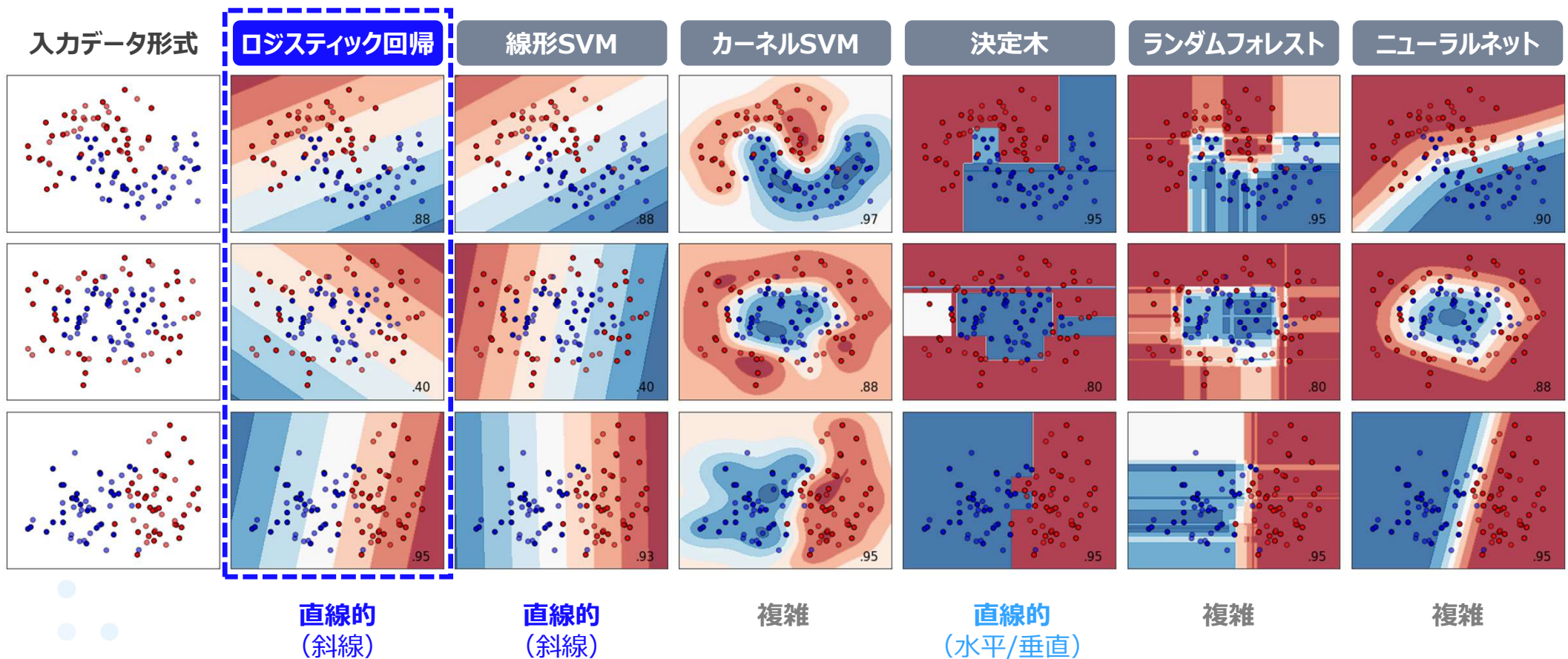


$$Y = a_1 X_1 + a_2 X_2 + \dots + a_n X_n + b$$

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**



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 required ('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

予測(分析)対象を
説明するための変数

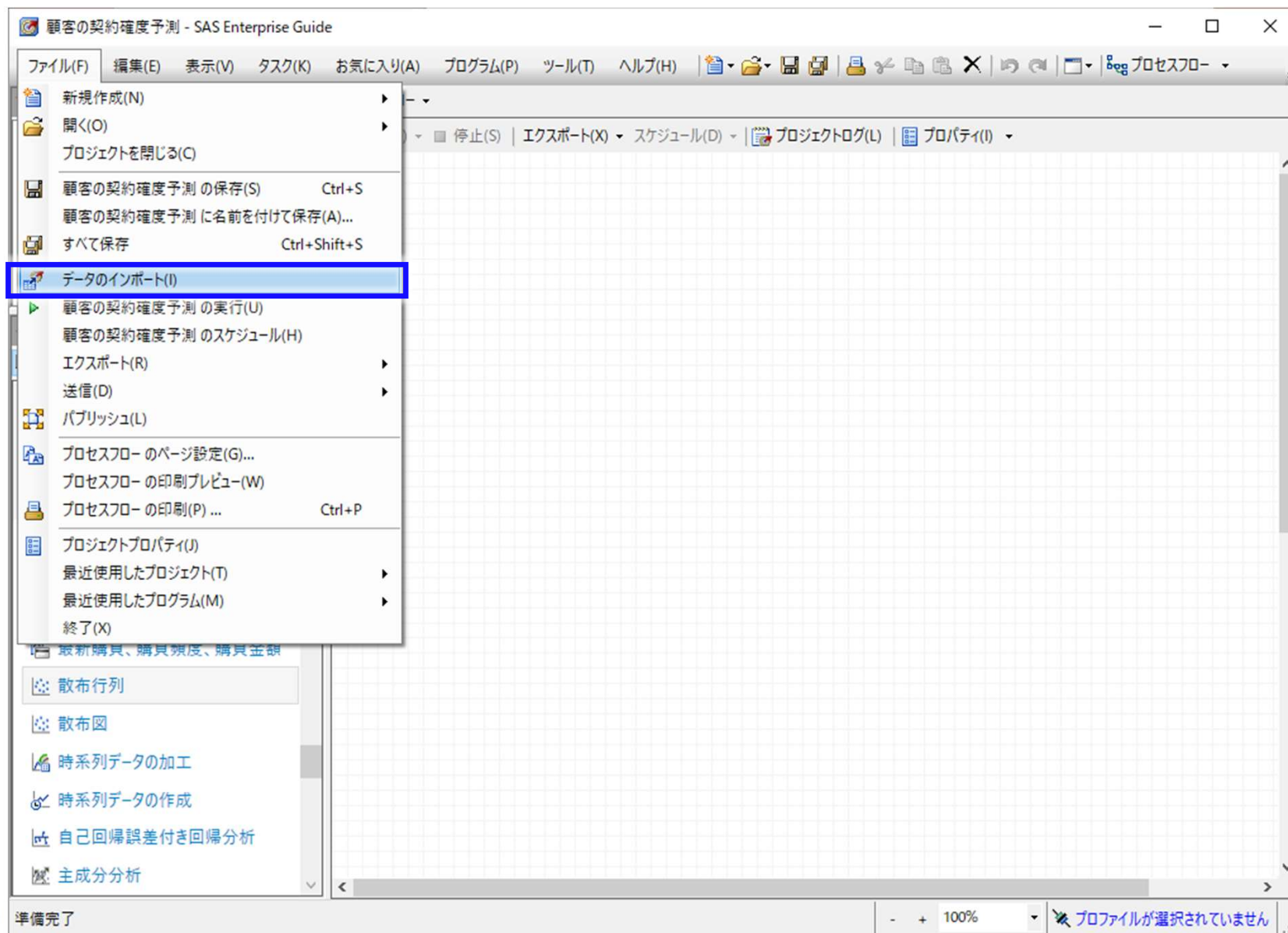
予測(分析)
したい対象

年齢	職業	結婚歴	学歴	クレジットカード債務不履行有無	年間平均残高 (ユーロ)	住宅ローンの有無	個人ローンの有無	連絡手段	最終連絡日	最終連絡月	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	定期預金契約有無
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	no	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
40	management	married	tertiary	no	194	no	yes	cellular	29	aug	189	2	-1	0	unknown	no
56	technician	married	secondary	no	4073	no	no	cellular	27	aug	239	5	-1	0	unknown	no
37	admin.	single	tertiary	no	2317	yes	no	cellular	20	apr	114	1	152	2	failure	no
25	blue-collar	single	primary	no	-221	yes	no	unknown	23	may	250	1	-1	0	unknown	no
31	services	married	secondary	no	132	no	no	cellular	7	jul	148	1	152	1	other	no
38	management	divorced	unknown	no	0	yes	no	cellular	18	nov	96	2	-1	0	unknown	no
42	management	divorced	tertiary	no	16	no	no	cellular	19	nov	140	3	-1	0	unknown	no
44	services	single	secondary	no	106	no	no	unknown	12	jun	109	2	-1	0	unknown	no
44	entrepreneur	married	secondary	no	93	no	no	cellular	7	jul	125	2	-1	0	unknown	no
26	housemaid	married	tertiary	no	543	no	no	cellular	30	jan	160	3	-1	0	unknown	no

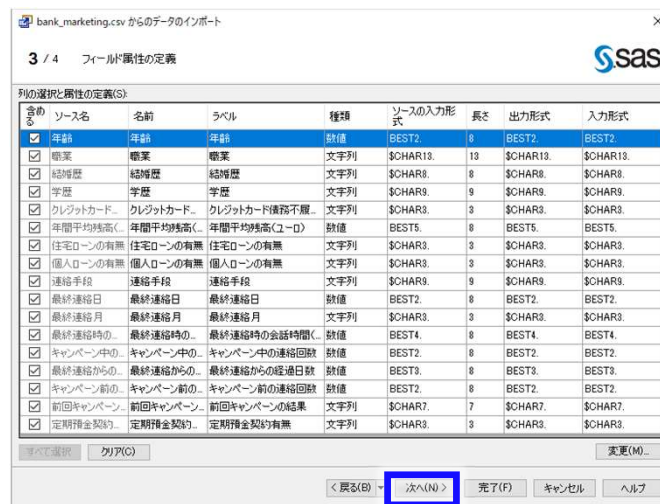
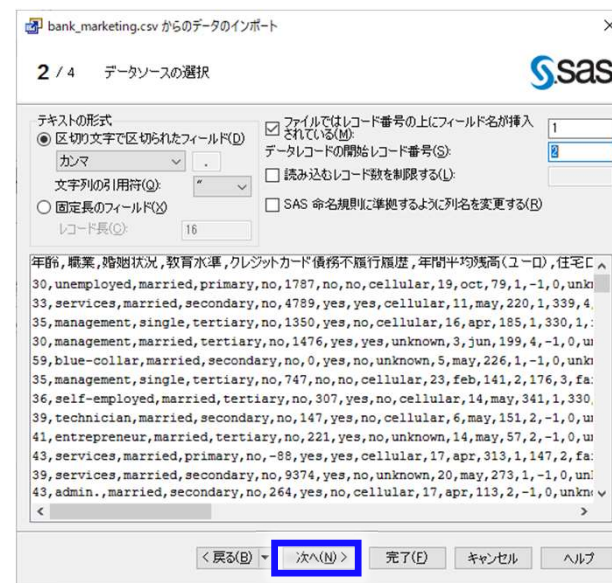
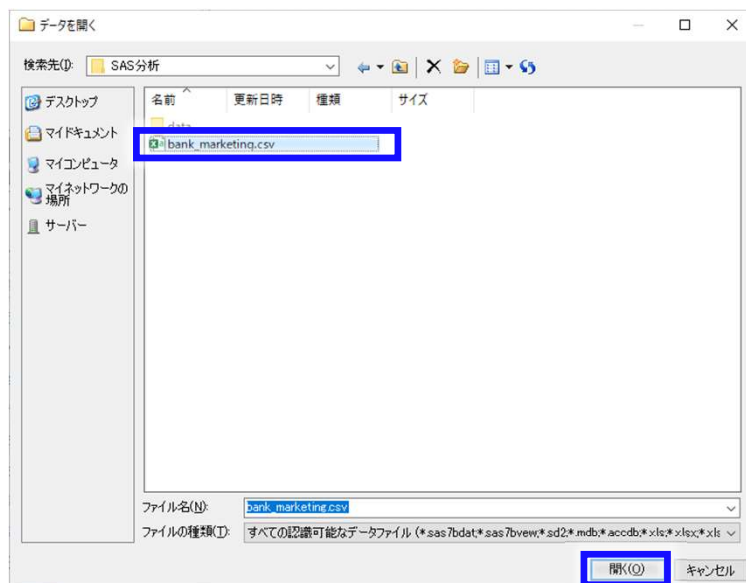
説明変数

目的変数

データの読み込み (1/2)



データの読み込み (2/2)



読み込んだデータの確認

prise Guide

タスク(K) お気に入り(A) プログラム(P) ツール(T) ヘルプ(H) | | log プロセスフロー | [現在のプロジェクトの検索]

データ (bank_marketing.csv) のインポート

コード ログ 出力データ

タスクの変更(Y) | フィルタと並べ替え(L) | クエリビルダ(Q) | Where 式(W) | データ(D) | 記述統計(E) | グラフ(G) | 分析(Z) | エクスポート(X) | 送信(N) | プロパティ

読み込まれたデータ

	年齢	職業	結婚歴	学歴	クレジットカードの債務不履行	年間平均所得高(ユーロ)	住宅ローンの有無	個人ローンの有無	連絡手段	最終更新日時
1	30	unemployed	married	primary	no	1787	no	no	cellular	1
2	33	services	married	secondary	no	4789	yes	yes	cellular	1
3	35	management	single	tertiary	no	1350	yes	no	cellular	1
4	30	management	married	tertiary	no	1476	yes	yes	unknown	1
5	59	blue-collar	married	secondary	no	0	yes	no	unknown	1
6	35	management	single	tertiary	no	747	no	no	cellular	2
7	36	self-employed	married	tertiary	no	307	yes	no	cellular	1
8	39	technician	married	secondary	no	147	yes	no	cellular	1
9	41	entrepreneur	married	tertiary	no	221	yes	no	unknown	1
10	43	services	married	primary	no	-88	yes	yes	cellular	1
11	39	services	married	secondary	no	9374	yes	no	unknown	2
12	43	admin.	married	secondary	no	264	yes	no	cellular	1
13	36	technician	married	tertiary	no	1109	no	no	cellular	1
14	20	student	single	secondary	no	502	no	no	cellular	3
15	31	blue-collar	married	secondary	no	360	yes	yes	cellular	2
16	40	management	married	tertiary	no	194	no	yes	cellular	2
17	56	technician	married	secondary	no	4073	no	no	cellular	2
18	37	admin.	single	tertiary	no	2317	yes	no	cellular	2
19	25	blue-collar	single	primary	no	-20	yes	no	unknown	2
20	31	services	married	secondary	no	0	yes	no	cellular	1
21	38	management	divorced	unknown	no	0	yes	no	cellular	1
22	42	management	divorced	tertiary	no	16	no	no	cellular	1
23	44	services	single	secondary	no	106	no	no	unknown	1
24	44	entrepreneur	married	secondary	no	93	no	no	cellular	1
25	26	housemaid	married	tertiary	no	543	no	no	cellular	3
26	41	management	married	tertiary	no	5883	no	no	cellular	2
27	55	blue-collar	married	primary	no	627	yes	no	unknown	1
28	67	retired	married	unknown	no	696	no	no	telephone	1
29	56	self-employed	married	secondary	no	784	no	yes	cellular	3
30	53	admin.	married	secondary	no	105	no	yes	cellular	2
31	68	retired	divorced	secondary	no	4189	no	no	telephone	1
32	31	technician	married	secondary	no	171	no	no	cellular	2
33	59	management	married	secondary	no	42	no	no	cellular	2
34	32	management	single	tertiary	no	2536	yes	no	cellular	2
35	49	technician	married	tertiary	no	1235	no	no	cellular	1
36	42	admin.	divorced	secondary	no	1811	yes	no	unknown	1
37	78	retired	divorced	primary	no	229	no	no	telephone	2
38	32	blue-collar	married	secondary	no	2089	yes	no	cellular	1
39	33	management	married	secondary	no	3935	yes	no	cellular	1
40	23	services	single	tertiary	no	363	yes	no	unknown	3
41	38	management	single	tertiary	no	11971	yes	no	unknown	1
42	26	management	single	tertiary	no	559	no	no	cellular	1

プロファイルが選択されていません

プロパティ

bank_marketing.csv からインポートされたデータのプロパティ

一般
列
詳細
概要

名前	種類	長さ	出力形式	入力形式	ラベル
年齢	数値	8	BEST2	BEST2	
職業	文字	13	\$CHAR13.	\$CHAR13.	
結婚歴	文字	8	\$CHAR8.	\$CHAR8.	
学歴	文字	9	\$CHAR9.	\$CHAR9.	
クレジットカード債務不履行有無	文字	3	\$CHAR3.	\$CHAR3.	
年間平均所得高(ユーロ)	数値	8	BEST5.	BEST5.	
住宅ローンの有無	文字	3	\$CHAR3.	\$CHAR3.	
個人ローンの有無	文字	3	\$CHAR3.	\$CHAR3.	
連絡手段	文字	9	\$CHAR9.	\$CHAR9.	
最終連絡日	数値	8	BEST2.	BEST2.	
最終連絡月	文字	3	\$CHAR3.	\$CHAR3.	
最終連絡時の会話時間(秒)	数値	8	BEST4.	BEST4.	
キャンペーン中の連絡回数	数値	8	BEST2.	BEST2.	
最終連絡からの経過日数	数値	8	BEST3.	BEST3.	
キャンペーン前の連絡回数	数値	8	BEST2.	BEST2.	
前回キャンペーンの結果	文字	7	\$CHAR7.	\$CHAR7.	
定期預金契約の有無	文字	3	\$CHAR3.	\$CHAR3.	

データセットを識別するラベル。LABEL= オプションで指定されます。

OK キャンセル

各列のデータ型を確認

bank_marketing.csv からインポートされたデータのプロパティ

一般
列
詳細
概要

データセットのプロパティ

サーバー: ローカル

SAS 参照名: WORK.BANK_MARKETING_0001

列: 17

行: 4521

インデックス: いいえ

圧縮: いいえ

選択ペインを使用して、表示するオプションのカテゴリを選択します。

OK キャンセル

列数と行数を確認！
(ビッグデータ分析の基本)

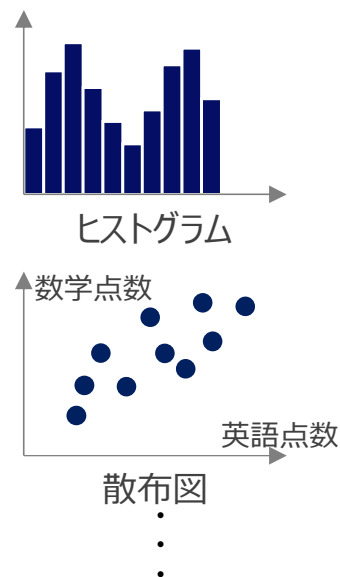
データの特徴の捉え方

- 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
- ⋮

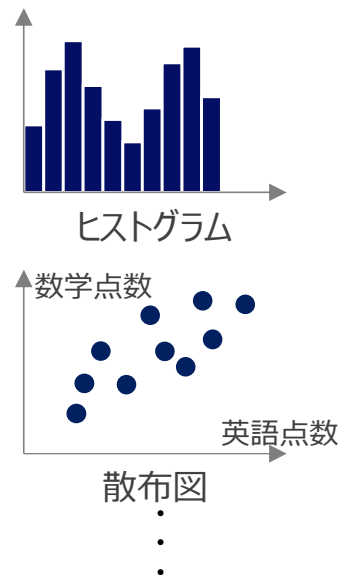
データの特徴の捉え方

- 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

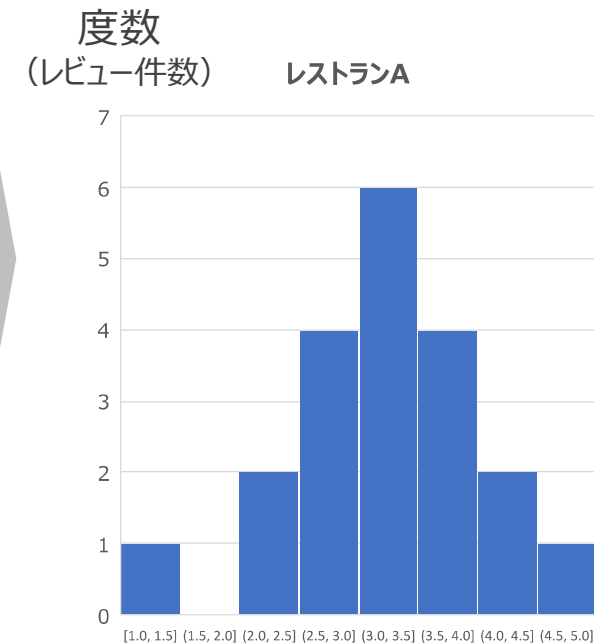
度数分布表

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

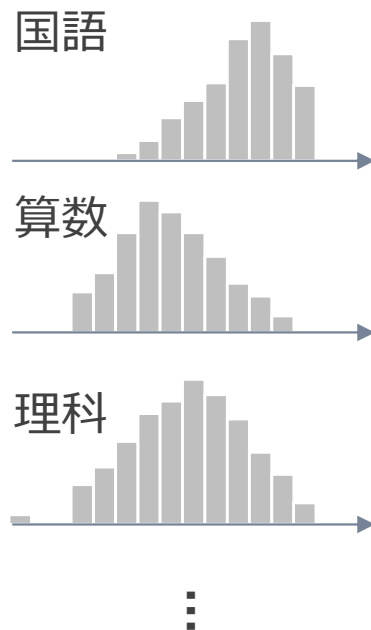


※境界値は小さい側の階級に含める

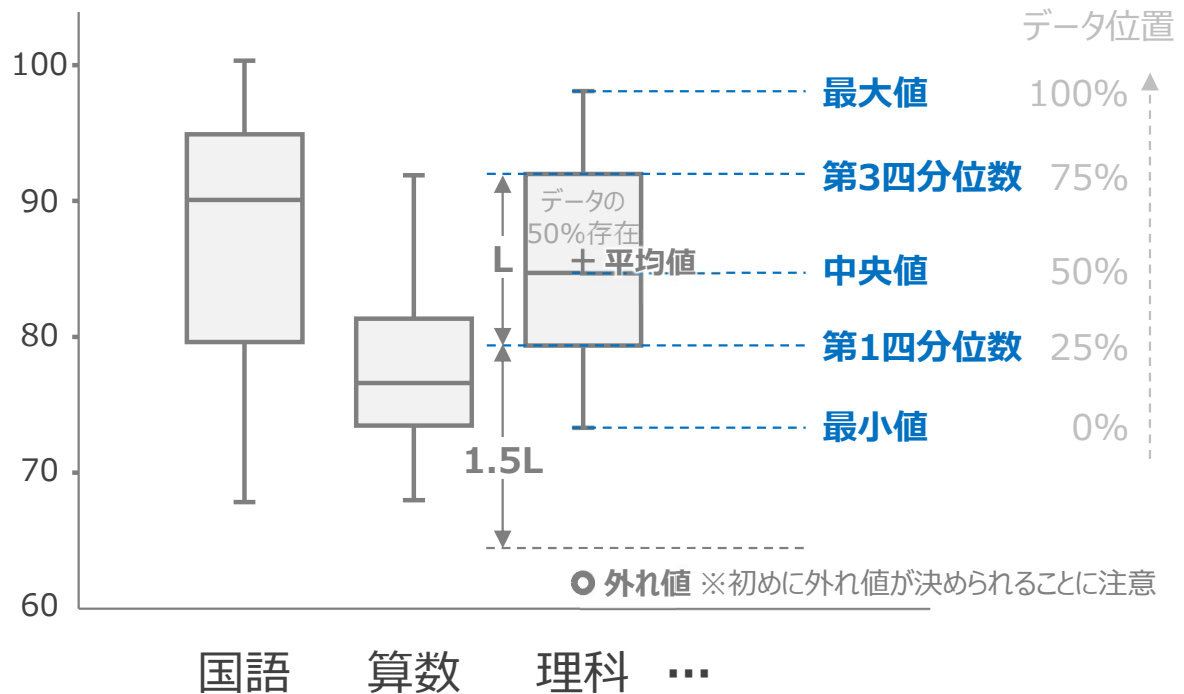
箱ひげ図

- 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

ヒストグラム



箱ひげ図

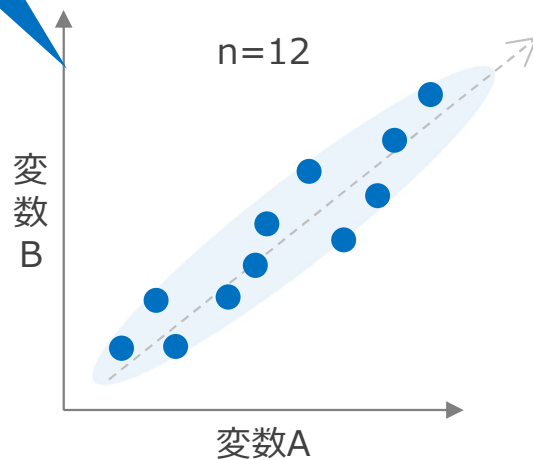


散布図 と 相関

- 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

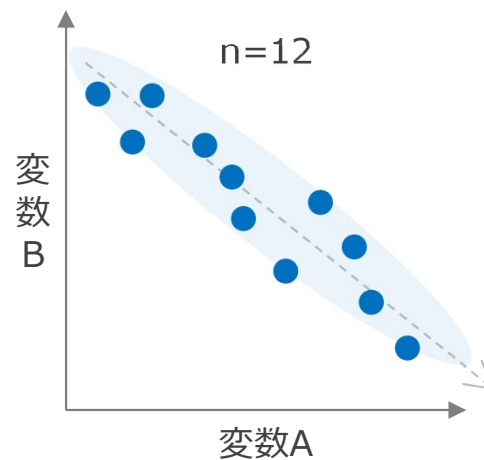
正の相関



一方 (A) が増加すると
他方 (B) も増加
(右上がりの楕円状)

例) A・・・(夏の)平均気温
B・・・エアコンの販売台数

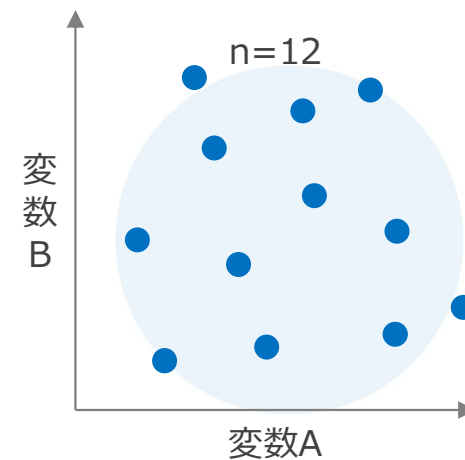
負の相関



一方 (A) が増加すると
他方 (B) は減少
(右下がりの楕円状)

例) A・・・(冬の)平均気温
B・・・ストーブの販売台数

無相関



双方の変数間に
関連性なし
(まばらな円状)

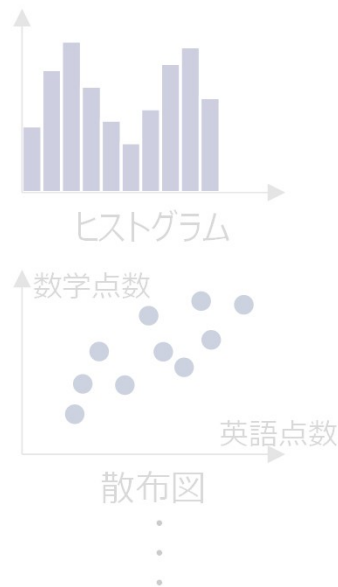
データの特徴の捉え方

- 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）

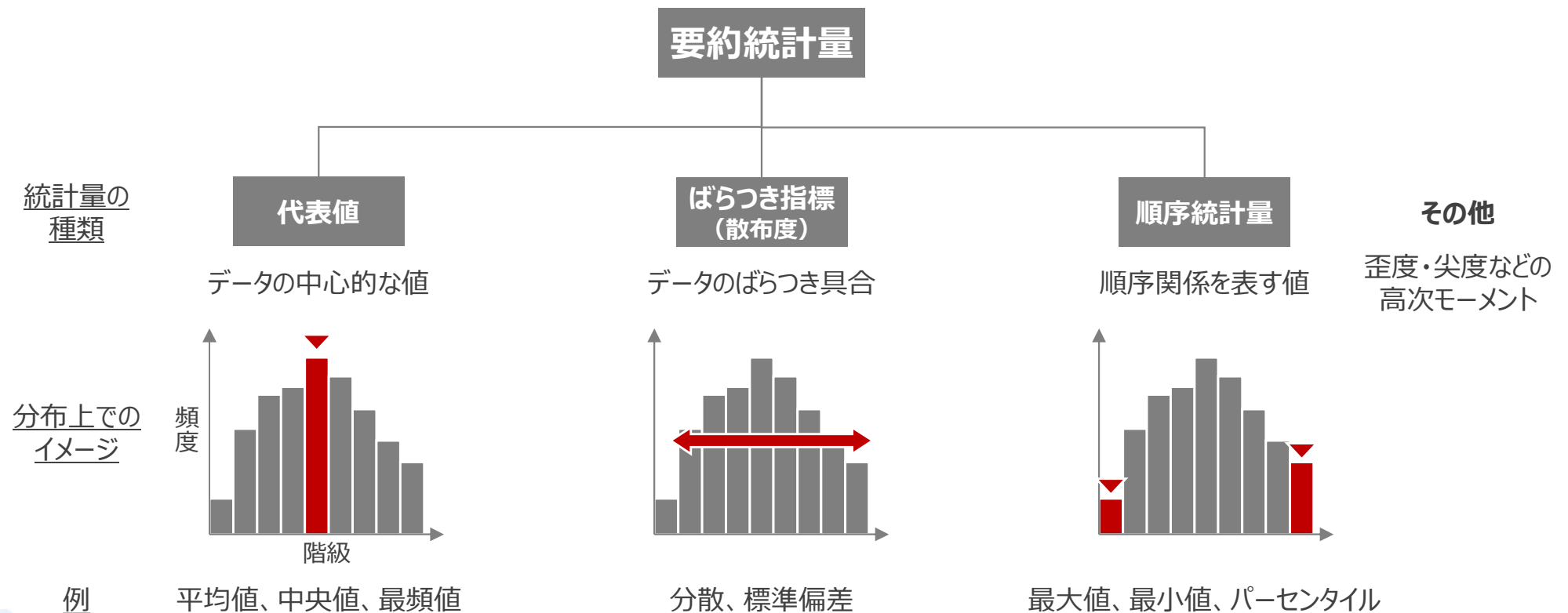
データの特徴を示す値に要約し
比較可能な客観的傾向を掴む
(要約統計量)

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
- ⋮

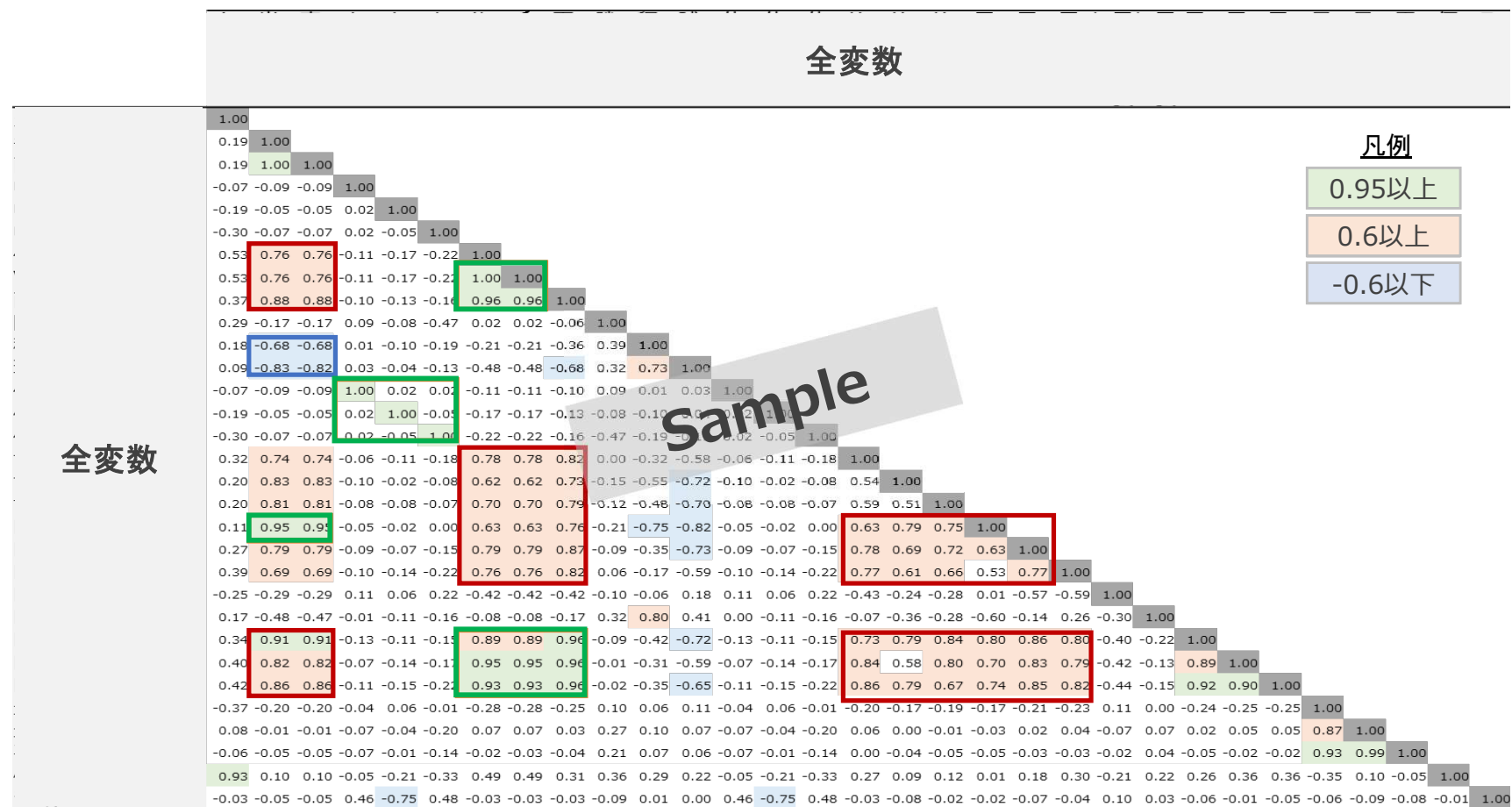
数値化（データ要約）

- 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



相関行列 ※第2回 (12/8) 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 Enterprise Guide での実装方法

- 要約統計量
- 箱ひげ図
- 散布図／層別散布図／散布図行列

要約統計量の算出 – 実行方法

Local:WORK.BANK_MARKETING_0001 の要約統計量

2 / 4 変数を役割に割り当てます。

使用可能な変数(V):

- 年齢
- 職業
- 結婚歴
- 学歴
- クレジットカード債務不履行有無
- 年間平均残高(ユーロ)
- 住宅ローンの有無
- 個人ローンの有無
- 連絡手段
- 最終連絡日
- 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
- 定期預金契約有無

描画対象の変数 (数値型) を設定

要約統計量を生成する変数 (分析変数):

- 年齢
- 年間平均残高(ユーロ)
- 最終連絡日
- 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

分類に使用する変数 (分類変数):

- 定期預金契約有無

「分類に使用する変数」に目的変数を設定

値ごとに別々の表を作成する変数 (グループ変数):

(オプション)
ここに変数をドロップしてください。

詳細(A)...

< 戻る(B) 次へ(N) > 完了(F) キャンセル ヘルプ

Local:WORK.BANK_MARKETING_0001 の要約統計量

3 / 4 統計量と結果を指定します。

統計量: MEAN, STD, MIN, MAX, N

編集(E)...

結果:

- ☒ 統計量を表示する(T)
- ☒ ヒストグラム(H)
- ☒ 箱ひげ図(W)
- ☐ 統計量をデータセットに保存する(D)

SAS Server: Local
ライブラリ: WORK
データセット: MEANSummaryStats

参照(R)...

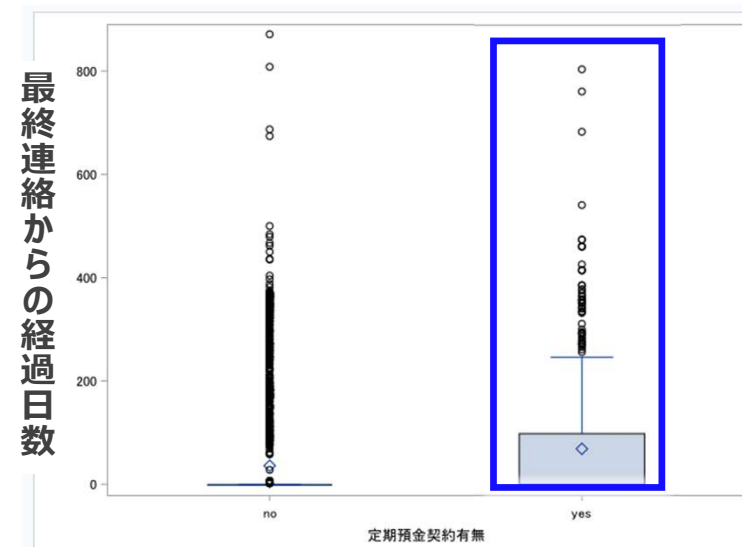
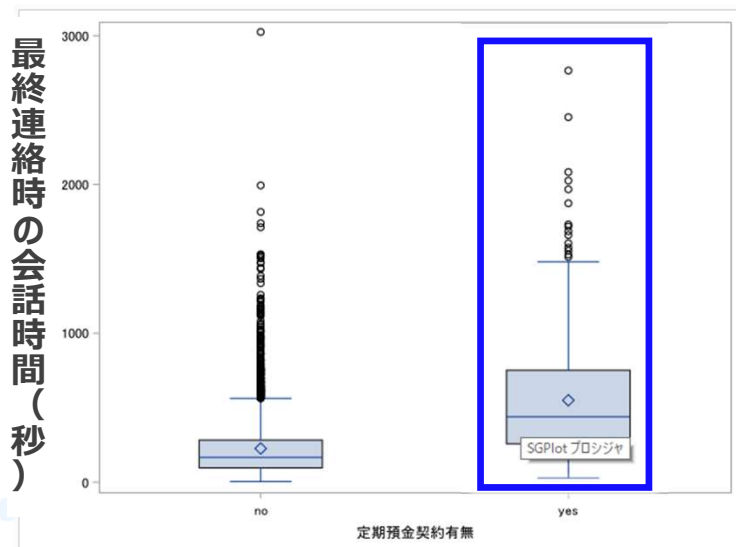
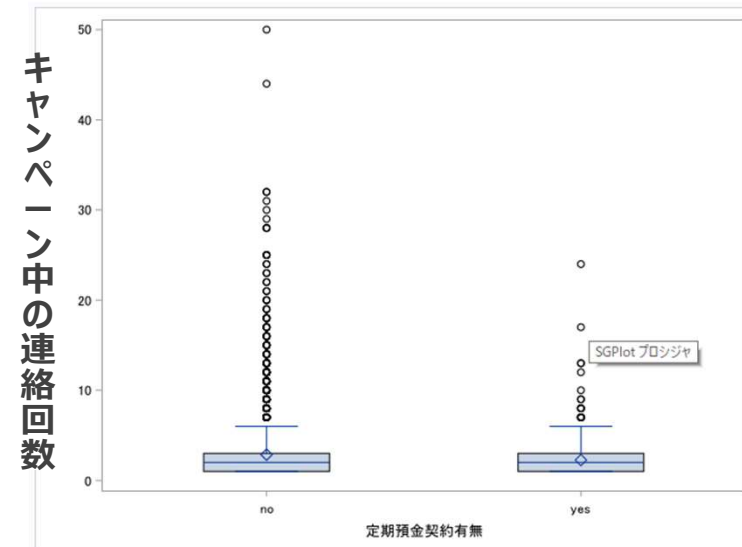
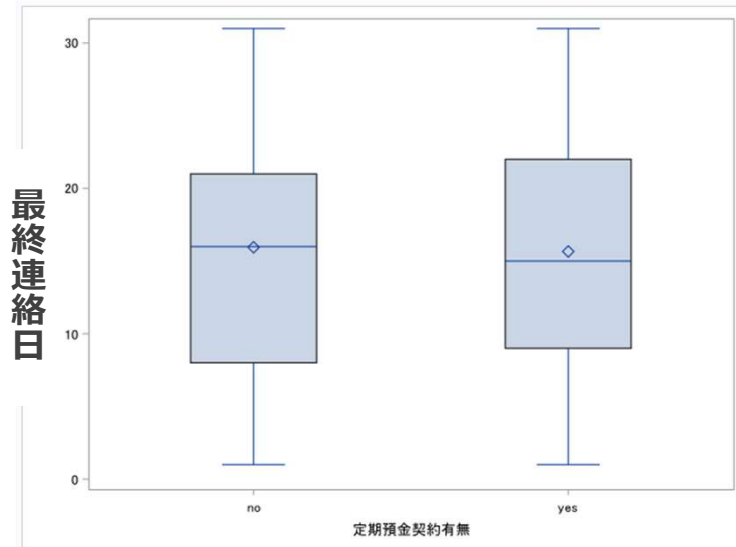
< 戻る(B) 次へ(N) > 完了(F) キャンセル ヘルプ

同時にヒストグラム、
箱ひげ図も出力可能

要約統計量の算出 – 実行結果 (要約統計量)

定期預金契約有無	Obs 数	変数	平均	標準偏差	最小値	最大値	N
no	4000	年齢	40.9980000	10.1883977	19.0000000	86.0000000	4000
		年間平均残高(ユーロ)	1403.21	3075.35	-3313.00	71188.00	4000
		最終連絡日	15.9487500	8.2497356	1.0000000	31.0000000	4000
		最終連絡時の会話時間(秒)	226.3475000	210.3136306	4.0000000	3025.00	4000
		キャンペーン中の連絡回数	2.8622500	3.2126088	1.0000000	50.0000000	4000
		最終連絡からの経過日数	36.0060000	96.2976572	-1.0000000	871.0000000	4000
		キャンペーン前の連絡回数	0.4712500	1.6273707	0	25.0000000	4000
yes	521	年齢	42.4913628	13.1157723	19.0000000	87.0000000	521
		年間平均残高(ユーロ)	1571.96	2444.40	-1206.00	26965.00	521
		最終連絡日	15.6583493	8.2351482	1.0000000	31.0000000	521
		最終連絡時の会話時間(秒)	552.7428023	390.3258046	30.0000000	2769.00	521
		キャンペーン中の連絡回数	2.2667946	2.0920709	1.0000000	24.0000000	521
		最終連絡からの経過日数	68.6391555	121.9630631	-1.0000000	804.0000000	521
		キャンペーン前の連絡回数	1.0902111	2.0553682	0	14.0000000	521

要約統計量の算出 – 実行結果 (箱ひげ図)



参考：箱ひげ図のみ描画したい場合

顧客の契約確度予測 - SAS Enterprise Guide

ファイル(F) 編集(E) 表示(V) タスク(K) お気に入り(A) プログラム(P) ツール(T) ヘルプ(H) | 現在プロジェクトの検索

プロジェクトツリー

- プロセスフロー
 - bank_marketing.csv
 - データ (bank_marketing.csv) のインポート
 - bank_marketing.csv からインポートされたデータ
 - 散布行列
 - 散布図

サーバー

- 最新の情報に更新(R) 接続解除(D)
- サーバー
 - ローカル
 - ライブラリ
 - 複数のファイル
 - 個人用 OLAP Server

散布図

入力データ コード ログ 結果

フィルタと並べ替え(L) クエリビルダ(Q) Where 式(W) データ(D) 記述統計(E) グラフ(G) 分析(Z) エクスポート(X) 送信(N)

年齢 職業 結婚歴 学歴

1 30 unemployed married primary no
2 33 services married secondary no
3 35 management single tertiary no
4 30 management married tertiary no
5 59 blue-collar married secondary no
6 35 management single tertiary no
7 36 self-employed married tertiary no
8 39 technician married secondary no
9 41 entrepreneur married tertiary no
10 43 services married primary no
11 39 services married secondary no
12 43 admin. married secondary no
13 36 technician married tertiary no
14 20 student single secondary no
15 31 blue-collar married secondary no
16 40 management married tertiary no
17 56 technician married secondary no
18 37 admin. single tertiary no
19 25 blue-collar single primary no
20 31 services married secondary no
21 38 management divorced unknown no
22 42 management divorced tertiary no
23 44 services single secondary no
24 44 entrepreneur married secondary no
25 26 housemaid married tertiary no
26 41 management married tertiary no
27 55 blue-collar married primary no
28 67 retired married unknown no
29 56 self-employed married secondary no
30 53 admin. married secondary no
31 68 retired divorced secondary no
32 31 technician married secondary no
33 59 management married secondary no
34 32 management single tertiary no
35 49 technician married tertiary no
36 42 admin. divorced secondary no
37 78 retired divorced primary no
38 32 blue-collar married secondary no
39 33 management married secondary no
40 23 services single tertiary no
41 38 management single tertiary no
42 26 management single tertiary no

箱ひげ図(O)...

準備完了

100% プロファイルが選択されていません

散布図の描画 – 実行方法 (1/2)

顧客の契約確度予測 - SAS Enterprise Guide

ファイル(F) 編集(E) 表示(V) タスク(K) お気に入り(A) プログラム(P) ツール(T) ヘルプ(H) | 100% | プロセスフロー | 現在のプロジェクトの検索

プロジェクトツリー

- プロセスフロー
 - bank_marketing.csv
 - データ (bank_marketing.csv) のインポート
 - bank_marketing.csv からインポートされたデータ
 - 散布行列

タスク

- 曲面プロット
- 個々の測定チャート
- 混合モデル
- 最新購買、購買頻度、購買金額
- 散布行列
- 散布図
- 時系列データの加工
- 時系列データの作成
- 自己回帰誤差付き回帰分析
- 主成分分析
- 集計表
- 集計表ウィザード
- 出力形式の作成
- 正準相関分析
- 折れ線グラフ
- 折れ線グラフウィザード
- 線形モデル
- 線形回帰分析

散布行列

入力データ | コード | ログ | 結果

フィルタと並べ替え(L) クエリビルダー(Q) Where 式(W) データ(D) 記述統計(E) グラフ(G) 分析(Z) エクスポート(X) 送信(N)

年齢 職業 婚姻状況 教育水準 クレジットカード債務不履行

1 30 unemployed married primary no

2 33 services married secondary no

3 35 management single tertiary no

4 30 management married tertiary no

5 59 blue-collar married secondary no

6 35 management single tertiary no

7 36 self-employed married tertiary no

8 39 technician married secondary no

9 41 entrepreneur married tertiary no

10 43 services married primary no

11 39 services married secondary no

12 43 admin. married secondary no

13 36 technician married tertiary no

14 20 student single secondary no

15 31 blue-collar married secondary no

16 40 management married tertiary no

17 56 technician married secondary no

18 37 admin. single tertiary no

19 25 blue-collar single primary no

20 31 services married secondary no

21 38 management divorced unknown no

22 42 management divorced tertiary no

23 44 services single secondary no

24 44 entrepreneur married secondary no

25 26 housemaid married tertiary no

26 41 management married tertiary no

27 55 blue-collar married primary no

28 67 retired married unknown no

29 56 self-employed married secondary no

30 53 admin. married secondary no

31 68 retired divorced secondary no

32 31 technician married secondary no

33 59 management married secondary no

34 32 management single tertiary no

35 49 technician married tertiary no

36 42 admin. divorced secondary no

37 78 retired divorced primary no

38 32 blue-collar married secondary no

39 33 management married secondary no

40 23 services single tertiary no

41 38 management single tertiary no

42 26 management single tertiary no

106 no no unknown 12

93 no no cellular 7

543 no no cellular 30

5883 no no cellular 20

627 yes no unknown 5

696 no no telephone 17

784 no yes cellular 30

105 no yes cellular 21

4189 no no telephone 14

171 no no cellular 27

42 no no cellular 21

2536 yes no cellular 26

1235 no no cellular 13

1811 yes no unknown 14

229 no no telephone 22

2089 yes no cellular 14

3935 yes no cellular 6

363 yes no unknown 30

11971 yes no unknown 17

559 no no cellular 11

最終接続の日

19

11

16

3

5

23

14

6

14

17

20

17

13

30

29

29

27

20

23

7

18

19

12

7

30

20

5

17

14

22

14

6

30

11

17

30

21

14

27

21

26

13

14

22

14

6

30

17

11

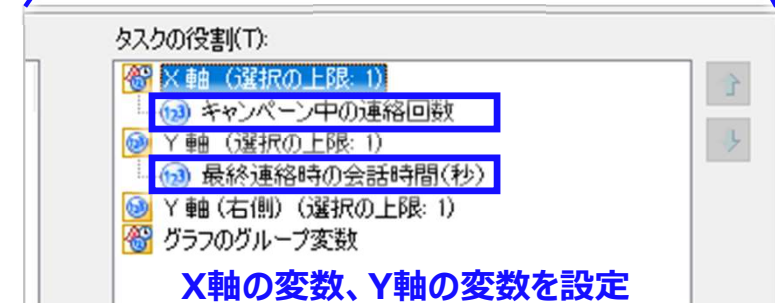
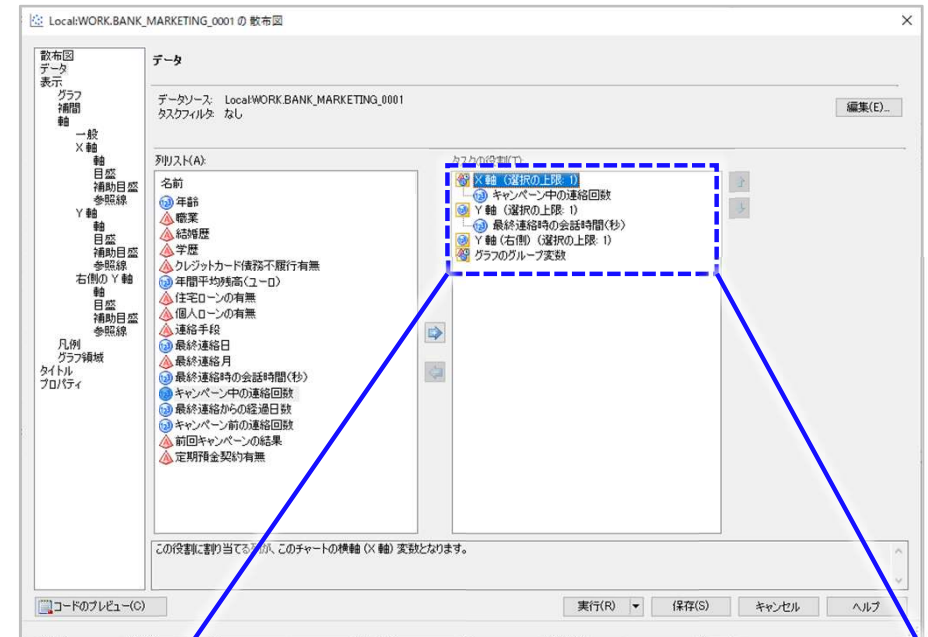
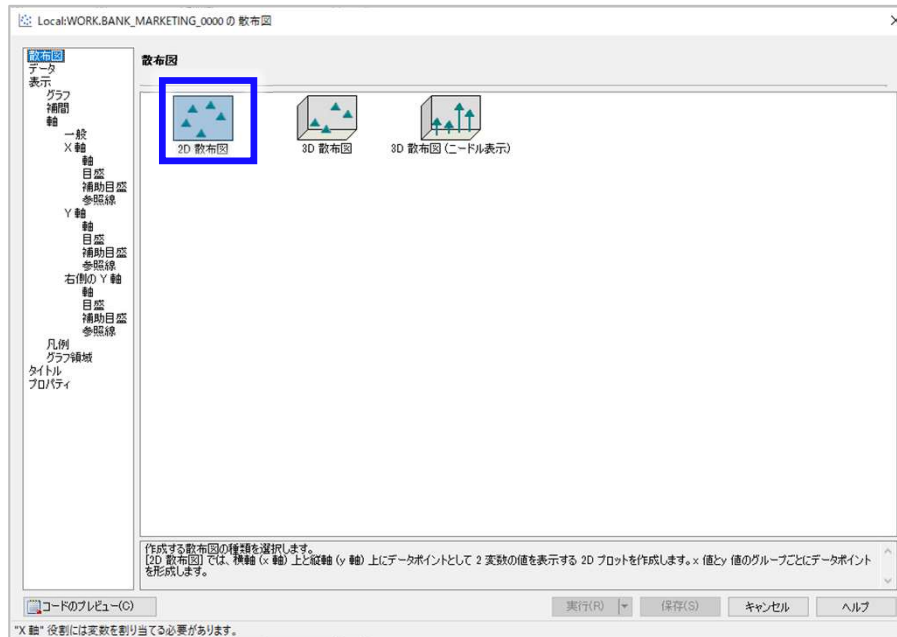
11

準備完了

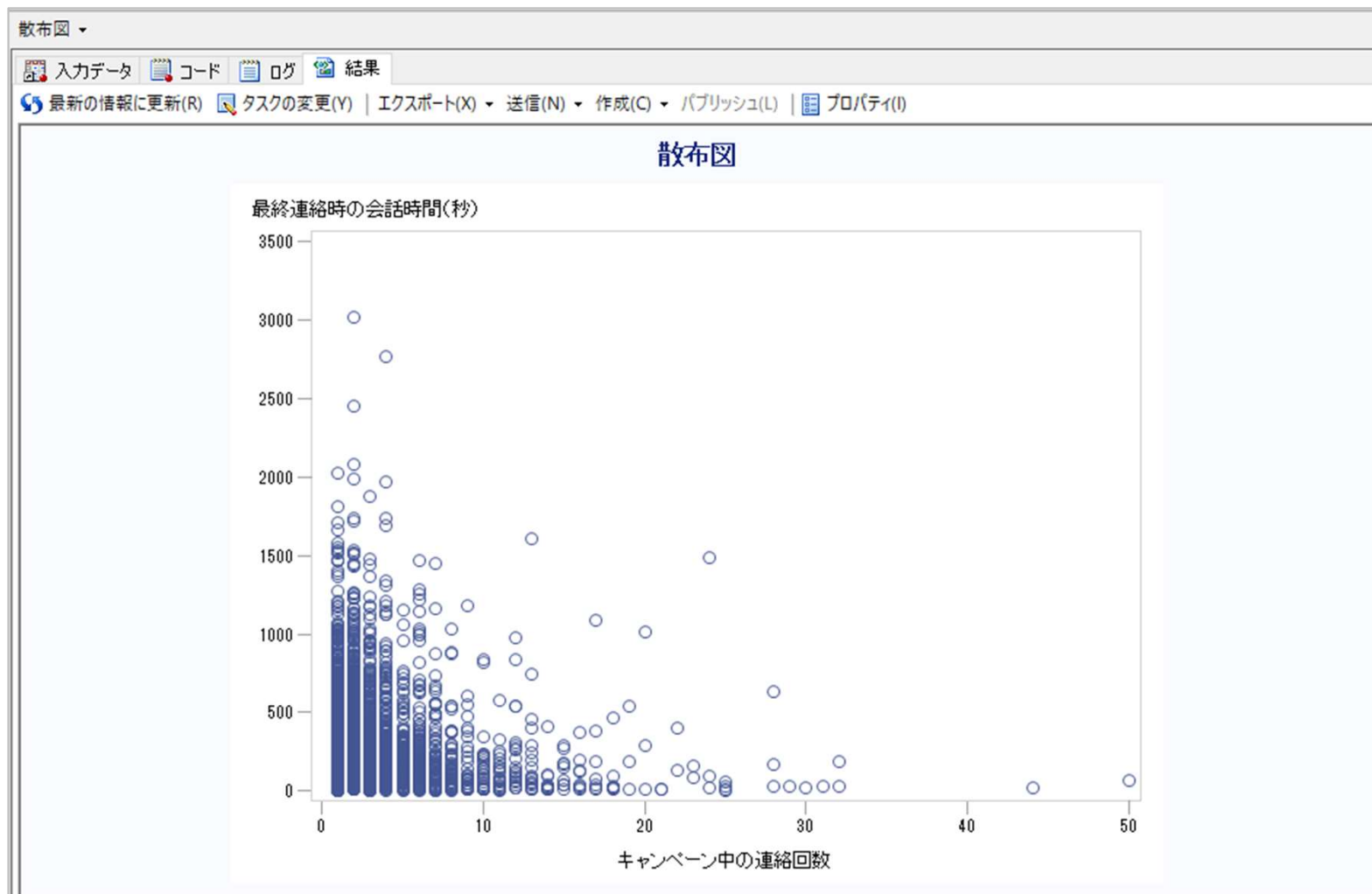
100%

プロファイルが選択されていません

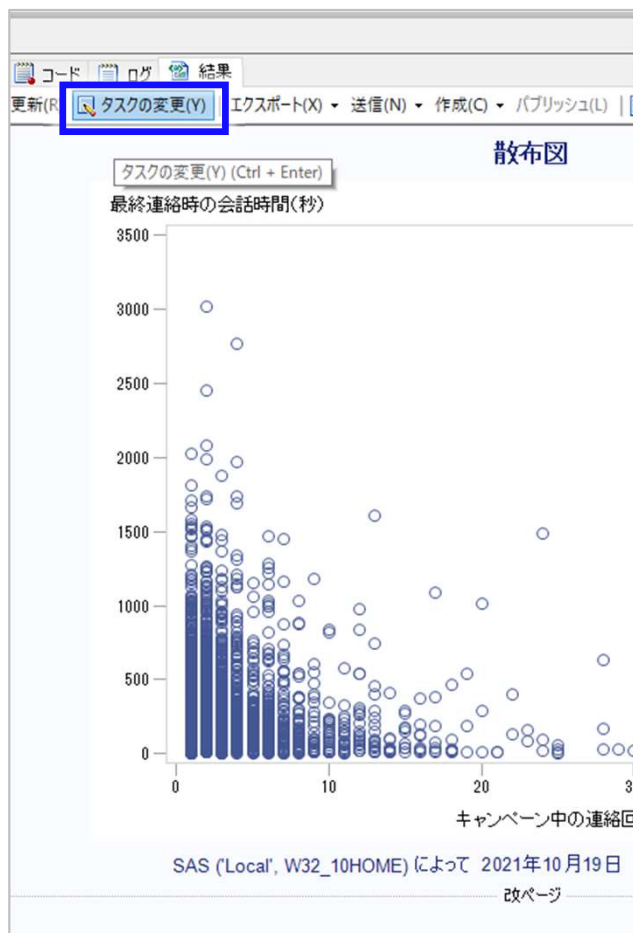
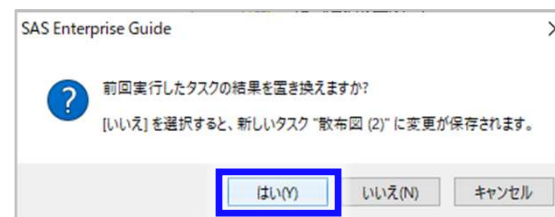
散布図の描画 - 実行方法 (2/2)



散布図の描画 – 実行結果



グラフの変更 と 層別散布図 – 実行方法

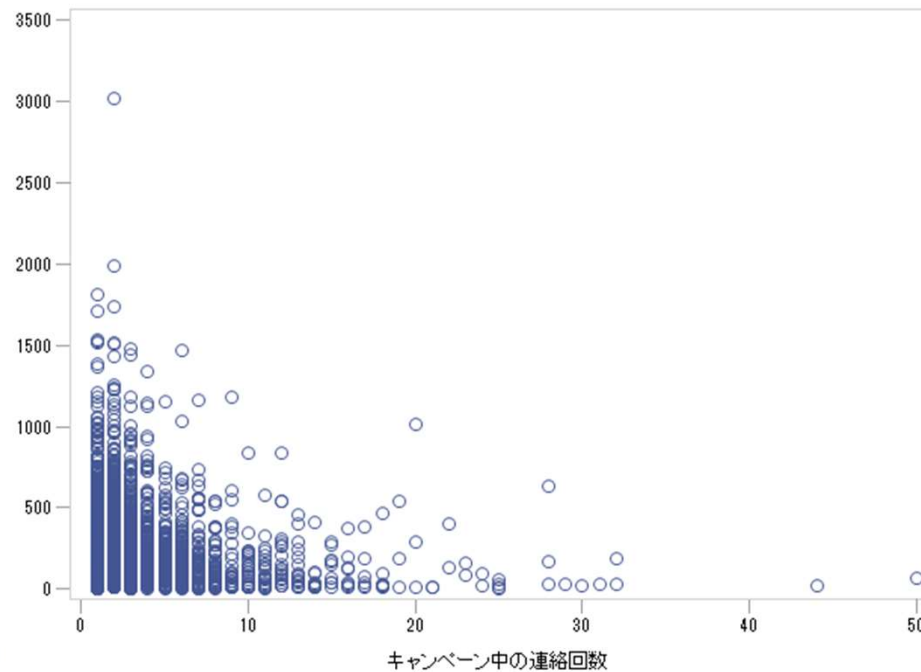



グラフの変更 と 層別散布図 – 実行結果

- ・ 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

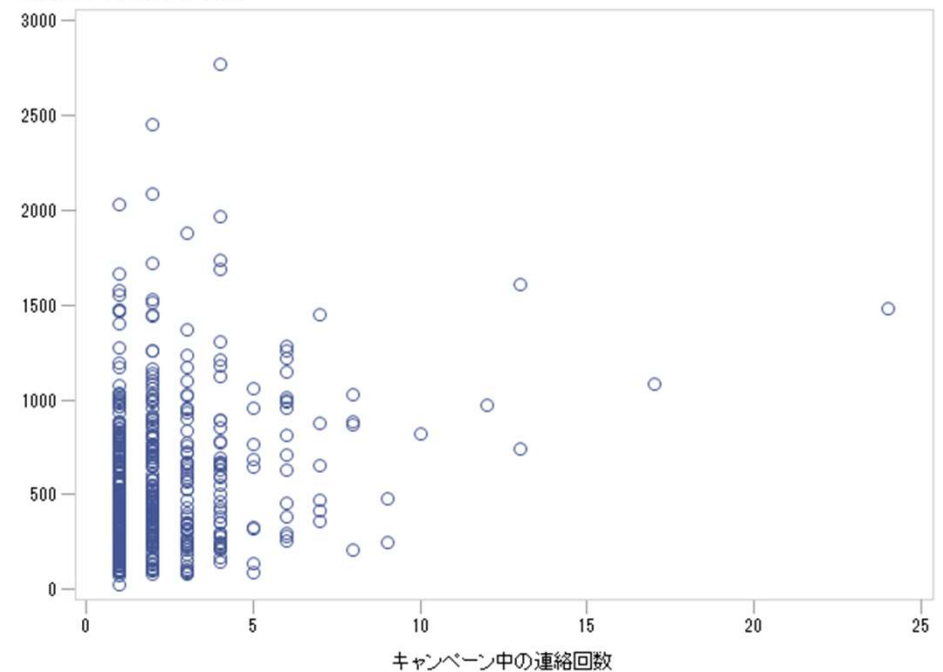
散布図
定期預金契約有無=no

最終連絡時の会話時間(秒)



散布図
定期預金契約有無=yes

最終連絡時の会話時間(秒)



散布図行列（層別） – 実行方法（1/2）

顧客の契約確度予測 - SAS Enterprise Guide

ファイル(F) 編集(E) 表示(V) タスク(K) お気に入り(A) プログラム(P) ツール(T) ヘルプ(H) | プロセッサー

プロジェクトツリー

- プロセスフロー
 - bank_marketing.csv
 - データ (bank_marketing.csv) のイ

タスク

- 曲面プロット
- 個々の測定チャート
- 混合モデル
- 最新購買、購買頻度、購買金額
- 散布行列**
- 散布図
- 時系列データの加工
- 時系列データの作成
- 自己回帰誤差付き回帰分析
- 主成分分析

データ (bank_marketing.csv) のインポート

コード ログ 出力データ

タスクの変更(V) フィルタと並べ替え(L) クエリビルダ(Q) Where 式(W) データ(D) 記述統計(E) **グラフ(G)** 分析(Z)

年齢 職業 婚姻状況 教育水準 クレジットカード債務... 年間平均残高(ユーロ)

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
16	40	management	married	tertiary	no	194	no
17	56	technician	married	secondary	no	4073	no
18	37	admin.	single	tertiary	no	2317	yes
19	25	blue-collar	single	primary	no	-221	yes
20	31	services	married	secondary	no	132	no
21	38	management	divorced	unknown	no	0	yes
22	42	management	divorced	tertiary	no	16	no
23	44	services	single	secondary	no	106	no
24	44	entrepreneur	married	secondary	no	93	no
25	26	housemaid	married	tertiary	no	543	no
26	41	management	married	tertiary	no	5883	no

準備完了

100% プロファイルが選択されていません

グラフ(G) 分析(Z)

- 棒グラフウィザード(H)...
- 棒グラフ(B)...
- 円グラフウィザード(E)...
- 円グラフ(P)...
- 折れ線グラフウィザード(W)...
- 折れ線グラフ(L)...
- 散布図(S)...**
- 散布行列(X)...**
- 領域プロット(A)...
- 棒-折れ線グラフ(I)...
- バブルプロット(U)...
- ドーナツチャート(D)...
- 等高線プロット(N)...
- 箱ひげ図(O)...
- レーダーチャート(R)...
- 曲面プロット(F)...
- タイルグラフ(T)...
- マップチャート(M)...
- ODS Graphics Designer
- ODS 統計グラフの表示(C)

散布図行列（層別） – 実行方法（2/2）

散布図行列 Local:WORK.BANK_MARKETING_0001

ラベル(L): 散布行列

データソース: Local:WORK.BANKMARKETING_0001

タスクフィルタ(F): なし

行列変数(M): 年齢, 年間平均残高(ユーロ), 最終連絡日, 最終連絡時の会話時間(秒), キャンペーン

グループ変数(G): グループなし

グラフタイトル(T): 散布図行列

対角方向(D): 変数名の表示, 右下がり

楕円(P): 予測, 95% 信頼

凡例(N): 凡例, 下, 枠なし

グラフ領域(A): デフォルトサイズ, データチップなし

チャートのフット

☐ デフォルトのチャートサイズより優先する(O)

高さ(H): 0

幅(W): 0

☒ カーソルを合わせた際にデータチップを表示する(I)

キャンセル ヘルプ

可視化する変数の設定

列リスト(A):

名前

年齢

職業

結婚歴

学歴

クレジットカード債務不履行有無

年間平均残高(ユーロ)

住宅ローンの有無

個人ローンの有無

連絡手段

最終連絡日

最終連絡月

最終連絡時の会話時間(秒)

列の役割(R):

行列変数

年齢

年間平均残高(ユーロ)

最終連絡日

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

OK キャンセル

数値型の変数をドラッグ & ドロップ

層別化する変数の設定

グループ変数(G): グループなし

グラフタイトル(T): 最終連絡の日

対角方向(D): 変

楕円(P): 予測, 9

凡例(N): 定期預金契約有無

グラフ領域(A): デ

散布図同様、まずは目的変数を設定

対角線上に表示する内容の設定

※見た目の設定（オプション）

対角方向(D): 変数名の表示, 右下がり

☒ ヒストグラム(H)

☐ 核密度推定(K)

☐ 正規密度推定(N)

方向

☒ 上から開始, 右下がり(T)

☐ 下から開始, 右上がり(B)

SYSSOPLY ...

ヒストグラムがおすすめ

マウスオーバー表示の設定

※見た目の設定（オプション）

グラフ領域(A): デフォルトサイズ, データチップなし

チャートのフット

☐ デフォルトのチャートサイズより優先する(O)

高さ(H): 0

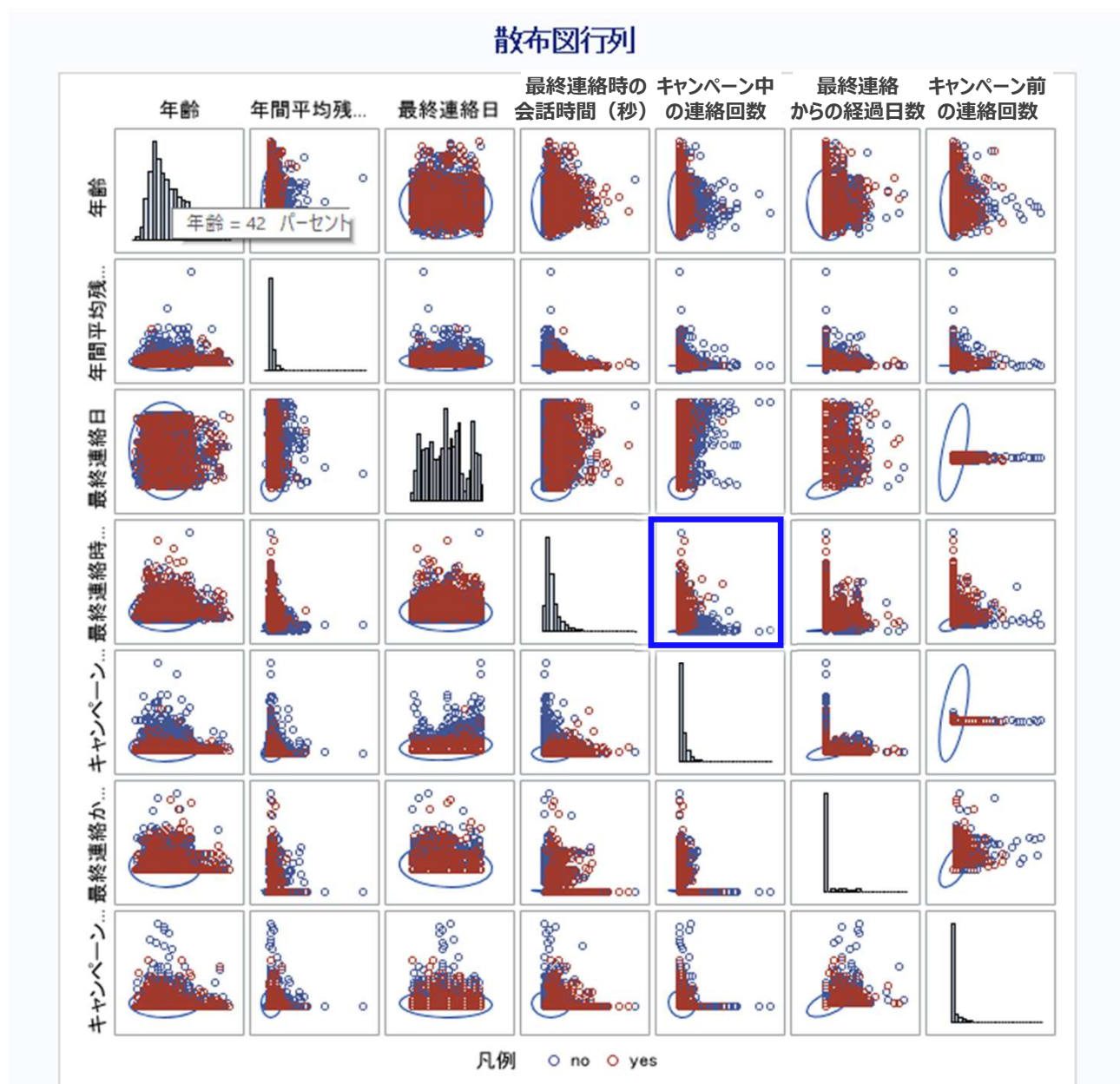
幅(W): 0

☒ カーソルを合わせた際にデータチップを表示する(I)

コードのプレ

データチップ表示（マウスオーバー時に値表示）
ONが便利

散布図行列 (層別) – 実行結果



参考：カテゴリー値の可視化方法

- カテゴリー値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、頻度集計による棒グラフ可視化方法について紹介する

Local:WORK.BANK_MARKETING_0001 の棒グラフ

2 / 4 変数を役割に割り当てます。

棒グラフ(O)

必須
バー(A): 前回キャンペーンの結果

可視化したい（頻度集計したい）カテゴリー値変数を設定

バーの高さ(E): 定期預金契約有無

「グループ変数」に層別したい変数を設定

オプション
グループ化する基準(G): 定期預金契約有無

2 軸(D):

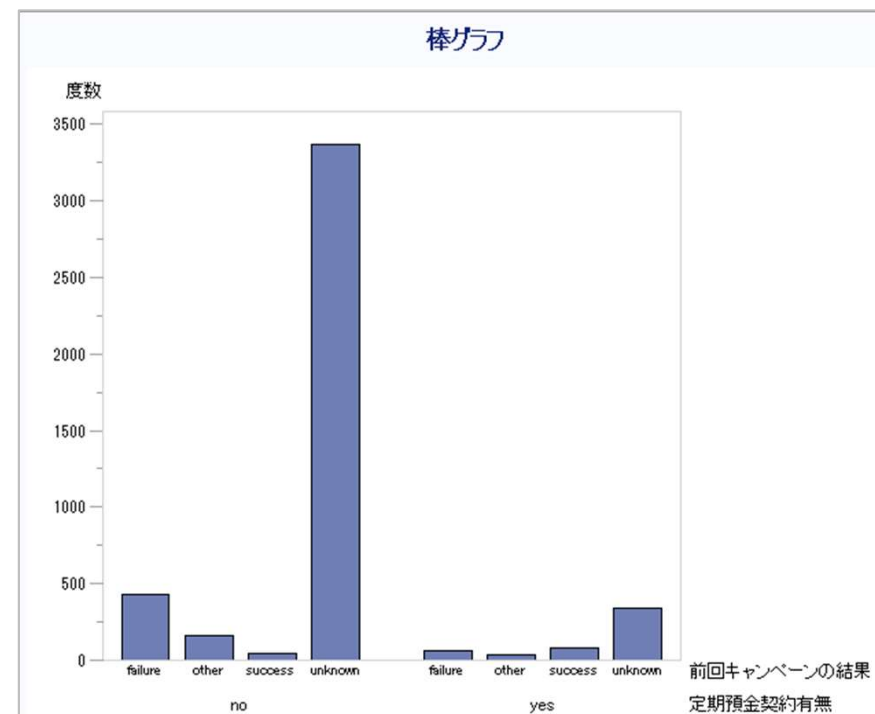
積み上げの基準(T):

チャートの作成基準(C): 年齢

サンプルチャート:

(度数)

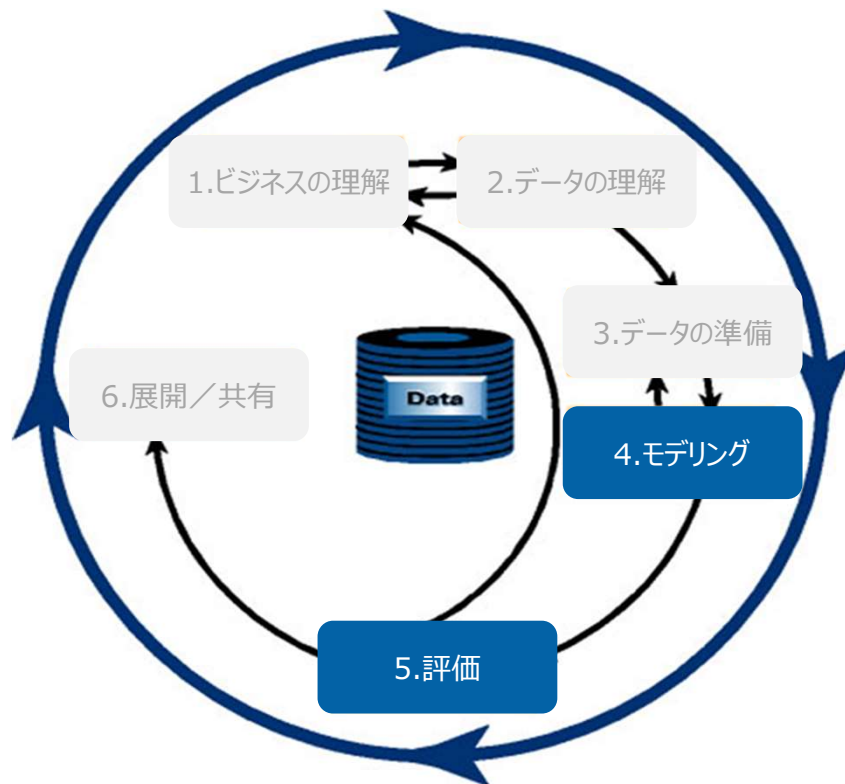
< 戻る(B) 次へ(N) > 完了(F) キャンセル ヘルプ



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

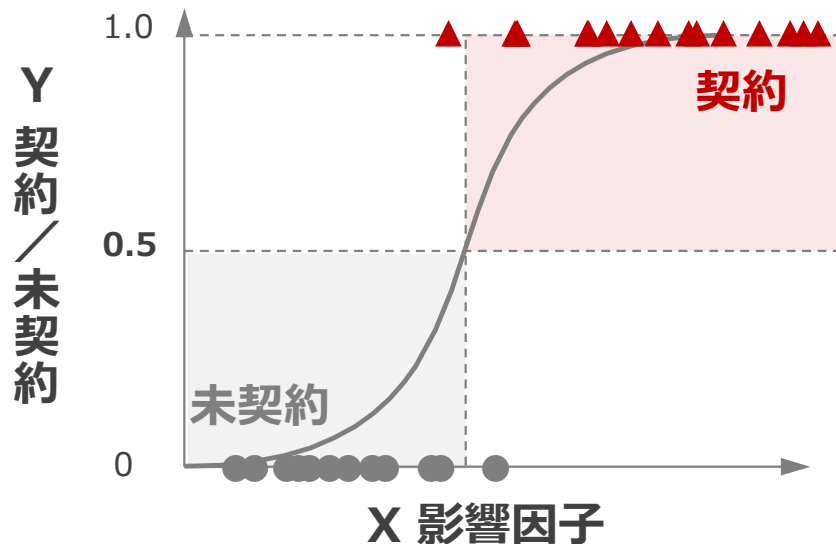
- データマイニングの結果の評価
- プロセスの見直し
- 実行可能なアクションリストの作成

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

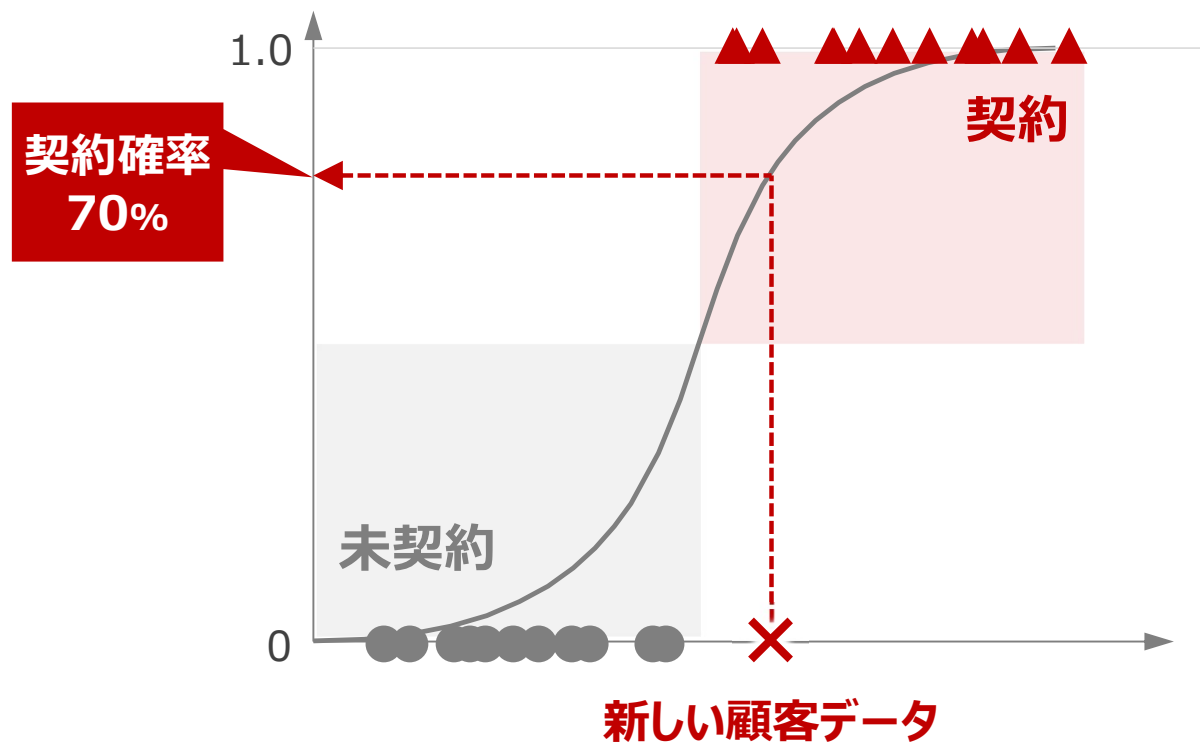
ロジスティック回帰のバリエーション

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

ロジスティック回帰における「発生確率」

- ロジスティック回帰の出力は最終的には 0 or 1 の2クラスであるが、内部的には「発生確率」の値を持ち、**閾値（多くは0.5）を境界として2クラスの予測結果を出力している**
- つまり、0 or 1 の予測結果に加えて、その**発生確率も同時に活用**することができる



$$\ln \left[\frac{P}{1-P} \right] = a_1X_1 + a_2X_2 + \cdots + a_nX_n + b$$
$$\Leftrightarrow \text{確率}P = \frac{1}{1 + \exp(-(a_1X_1 + a_2X_2 + \cdots + a_nX_n + b))}$$

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)	再現率
	未契約	2 誤り：誤報 (False Positive)	4 正答 (True Negative)	

適合率
(陽性的中率)

陰性的中率

予測

実際



災害
アラート



誤報



アラート
無し



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

⇔ 陰性的中率
(NPV*)
*Negative Predictive Value

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

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

⇔ 特異度, 真陰性率
(Specificity)

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

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 Enterprise Guide 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 (1/5) メニュー立ち上げ

顧客の契約確度予測 - SAS Enterprise Guide

ファイル(F) 編集(E) 表示(V) タスク(K) お気に入り(A) プログラム(P) ツール(T) ヘルプ(H) | 入力データ | コード | ログ | 結果 | プロセスフロー | 現在のプロジェクトの検索

プロジェクトツリー

- プロセスフロー
 - bank_marketing.csv
 - データ (bank_marketing.csv) のインポート
 - bank_marketing.csv からインポートされたデータ
 - 散布行列
 - 散布図
 - 箱ひげ図
 - 要約統計量

サーバー

- 最新の情報に更新(R) 接続解除(D)
- サーバー
 - ローカル
 - ライブラリ
 - 複数のファイル
 - 個人用 OLAP Server

要約統計量

入力データ | コード | ログ | 結果

フィルタと並べ替え(L) クエリビルダ(Q) Where 式(W) データ(D) 記述統計(E) グラフ(G) 分析(Z) エクスポート(X) 送信(N)

分析(Z) のサブメニュー:

- 回帰分析(R)
 - 多変量回帰分析(M)
 - 生存時間分析(S)
 - 工程能力分析(Y)
 - 管理図(C)
 - パレート図(P)...
 - 時系列分析(T)
 - データマイニング(N)
- HP 線形回帰分析(H)...
- 線形回帰分析(L)...
- 非線形回帰分析(N)...
- ロジスティック回帰分析(S)...**
- HP ロジスティック回帰分析(P)...
- 一般化線形モデル(G)...

年齢	職業	結婚歴	学歴	クレジットカード所有	年間高	公営分析(A)	入ローンの有無	連絡手段
30	unemployed	married	primary	no				
33	services	married	secondary	no				
35	management	single	tertiary	no				
30	management	married	tertiary	no				
59	blue-collar	married	secondary	no				
35	management	single	tertiary	no				
36	self-employed	married	tertiary	no				
39	technician	married	secondary	no				
41	entrepreneur	married	tertiary	no				
43	services	married	primary	no	-88	yes	yes	cellular
39	services	married	secondary	no	9374	yes	no	unknown
43	admin.	married	secondary	no	264	yes	no	cellular
36	technician	married	tertiary	no	1109	no	no	cellular
20	student	single	secondary	no	502	no	no	cellular
31	blue-collar	married	secondary	no	360	yes	yes	cellular
40	management	married	tertiary	no	194	no	yes	cellular
56	technician	married	secondary	no	4073	no	no	cellular
37	admin.	single	tertiary	no	2317	yes	no	cellular
25	blue-collar	single	primary	no	-221	yes	no	unknown
31	services	married	secondary	no	132	no	no	cellular
38	management	divorced	unknown	no	0	yes	no	cellular
42	management	divorced	tertiary	no	16	no	no	cellular
44	services	single	secondary	no	106	no	no	unknown
44	entrepreneur	married	secondary	no	93	no	no	cellular
26	housemaid	married	tertiary	no	543	no	no	cellular
41	management	married	tertiary	no	5883	no	no	cellular
55	blue-collar	married	primary	no	627	yes	no	unknown
67	retired	married	unknown	no	696	no	no	telephone
56	self-employed	married	secondary	no	784	no	yes	cellular
53	admin.	married	secondary	no	105	no	yes	cellular
68	retired	divorced	secondary	no	4189	no	no	telephone
31	technician	married	secondary	no	171	no	no	cellular
59	management	married	secondary	no	42	no	no	cellular
32	management	single	tertiary	no	2536	yes	no	cellular
49	technician	married	tertiary	no	1235	no	no	cellular
42	admin.	divorced	secondary	no	1811	yes	no	unknown
78	retired	divorced	primary	no	229	no	no	telephone
32	blue-collar	married	secondary	no	2089	yes	no	cellular
33	management	married	secondary	no	3935	yes	no	cellular
23	services	single	tertiary	no	363	yes	no	unknown
38	management	single	tertiary	no	11971	yes	no	unknown
26	management	single	tertiary	no	559	no	no	cellular

準備完了

100% | プロファイルが選択されていません

ロジスティック回帰 – 実行方法 (2/5) 説明変数・目的変数の設定

Local:WORK.BANK_MARKETING_0001 のロジスティック回帰分析

データ
モデル
応答
効果
選択
オプション
グラフ
予測値
タイトル
プロパティ

データ
データソース: Local:WORK.BANK_MARKETING_0001
タスクフィルタ: なし

変数リスト(A)

名前

- 年齢
- 職業
- 結婚歴
- 学歴
- クレジットカード債務不履行有無
- 年間平均残高(ユーロ)
- 住宅ローンの有無
- 個人ローンの有無
- 連絡手段
- 最終連絡日
- 最終連絡月
- 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
- 定期預金契約有無

タスクの役割(T):

従属変数 (選択の上限: 1)

- 定期預金契約有無

説明変数

- 年齢
- 年間平均残高(ユーロ)
- 最終連絡日
- 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

分類変数

- 職業
- 結婚歴
- 学歴
- クレジットカード債務不履行有無
- 住宅ローンの有無
- 個人ローンの有無
- 連絡手段
- 最終連絡月
- 前回キャンペーンの結果

グループ分析

- 度数カウント (選択の上限: 1)
- 重み (選択の上限: 1)

目的変数の設定

説明変数の設定①
: 数値型変数

説明変数の設定②
: カテゴリー型変数

タスクの役割に割り当てることができる変数のリスト。

ドラッグ & ドロップ

コードのプレビュー(C)

実行(R) 保存(S) キャンセル ヘルプ

ロジスティック回帰 – 実行方法 (3/5) 主効果の設定

Local:WORK.BANK_MARKETING_0001 のロジスティック回帰分析

モデル > 効果

主効果の設定

分類変数と量的変数(V):

- 年齢
- 年間平均残高(ユーロ)
- 最終連絡日
- 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
- 職業
- 結婚歴
- 学歴
- クレジットカード債務不履行有無
- 住宅ローンの有無
- 個人ローンの有無
- 連絡手段
- 最終連絡月
- 前回キャンペーンの結果

主効果(M)

交互作用効果(O)

ネスト効果(N)

組み合わせ効果(F)

次数(D): 16

多項式効果(P)

次数(G): 2

効果(E):

- '年齢'
- '年間平均残高(ユーロ)'
- '最終連絡日'
- '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'
- '職業'
- '結婚歴'
- '学歴'
- 'クレジットカード債務不履行有無'
- '住宅ローンの有無'
- '個人ローンの有無'
- '連絡手段'
- '最終連絡月'
- '前回キャンペーンの結果'

☒ 切片を含める(I)

効果の削除(T)

コードのプレビュー(C)

実行(R) 保存(S) キャンセル ヘルプ

ロジスティック回帰 – 実行方法 (4/5) 正例の指定

Local:WORK.BANK_MARKETING_0001 のロジスティック回帰分析

データ
モデル
応答
効果
選択
オプション
グラフ
予測値
タイトル
プロパティ

モデル > 応答

応答の種類(E): 2 値

モデルの種類:

- ☒ ロジット(O)
- ☐ プロビット(B)
- ☐ 補対数-対数(M)
- ☐ 一般化ロジット(G)

定期預金契約有無 の応答水準(L):
no
yes

モデルを当てはめる水準を「yes」に設定する
※yes=1と捉え、yesを予測するモデル構築を行う

モデルを当てはめる水準(F): no
no
yes

コードのプレビュー(C)

実行(R) 保存(S) キャンセル ヘルプ

ロジスティック回帰 – 実行方法 (5/5) オプション／出力内容の設定

必要に応じて、画面出力する内容にチェックを入れる

Local:WORK.BANK_MARKETING_0001 のロジスティック回帰分析

データ
モデル
応答
効果
プロパティ

モデル > オプション

推定値の詳細

☐ 推定値相関行列(O)

☐ 推定値の共分散行列(V)

分類テーブル

☒ 分類テーブルを表示する(T)

モデルの当てはまりに関する評価値

☐ 影響度に関する統計量(I)

☐ Hosmer - Lemeshow の適合度検定(H)

☐ デビアンسおよび Pearson の χ^2 乗統計量(D)

☐ 一般化 R2 乗(G)

信頼限界

パラメータ

☐ Wald(W)

☐ プロファイル尤度に基づく(F)

条件付きオッズ比

☐ Wald(A)

☐ プロファイル尤度に基づく(K)

信頼水準(E):

モデルの当てはめ手法

当てはめ手法(H): 自動 (選択なし)

☐ Firth の調整済み尤度(A)

コードのプレビュー(C)

実行(R) 保存(S) キャンセル ヘルプ

Local:WORK.BANK_MARKETING_0001 のロジスティック回帰分析

データ
モデル
応答
効果
選択
オプション
プロパティ

予測値

予測データ

☒ 元のサンプル(O)

☐ 追加データ(A)

参照(B)

出力データを保存する

☒ 予測値(P)

☐ 診断統計量(G)

Local:WORK.PREDLogRegPri

参照(E)

追加統計量

☐ 残差(U)

☐ 予測限界(L)

☒ 出力とプロットを表示する(D)

☐ 予測値を表示する(W)

選択ペインで、タスクのオプションの異なる組み合わせを選択できます。

コードのプレビュー(C)

実行(R) 保存(S) キャンセル ヘルプ

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

ロジスティック回帰分析の結果

LOGISTIC プロシジャ

モデルの情報	
データセット	WORK.SORTTEMPTABLESORTED
応答変数	定期預金契約有無
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	定期預金契約有無	度数の合計
1	no	4000
2	yes	521

未契約 : 88.5%

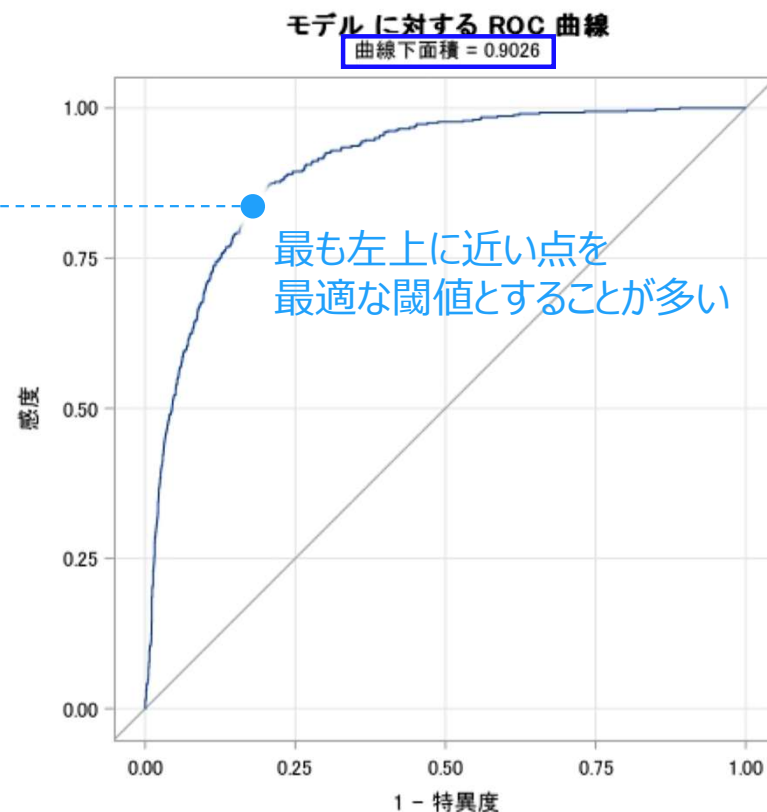
契約 : 11.5%

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

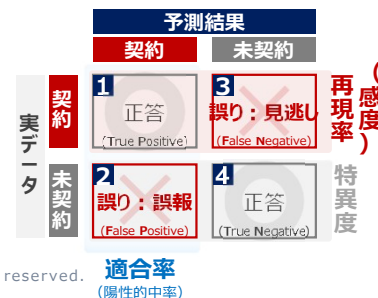
確率水準	分類表				パーセンテージ				
	正解		不正解		正解	感度	特異度	正の予測	負の予測
	事象	非事象	事象	非事象					
0.000	521	0	4000	0	11.5	100.0	0.0	11.5	.
0.020	517	1073	2927	4	35.2	99.2	26.8	15.0	99.6
0.040	502	2076	1924	19	57.0	96.4	51.9	20.7	99.1
0.060	477	2623	1377	44	68.6	91.6	65.6	25.7	98.4
0.080	454	3016	984	67	76.8	87.1	75.4	31.6	97.8
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
0.300	269	3786	214	252	89.7	51.6	94.7	55.7	93.8
0.320	259	3799	201	262	89.8	49.7	95.0	56.3	93.5
0.340	249	3814	186	272	89.9	47.8	95.4	57.2	93.3
0.360	238	3836	164	283	90.1	45.7	95.9	59.2	93.1

閾値を増加



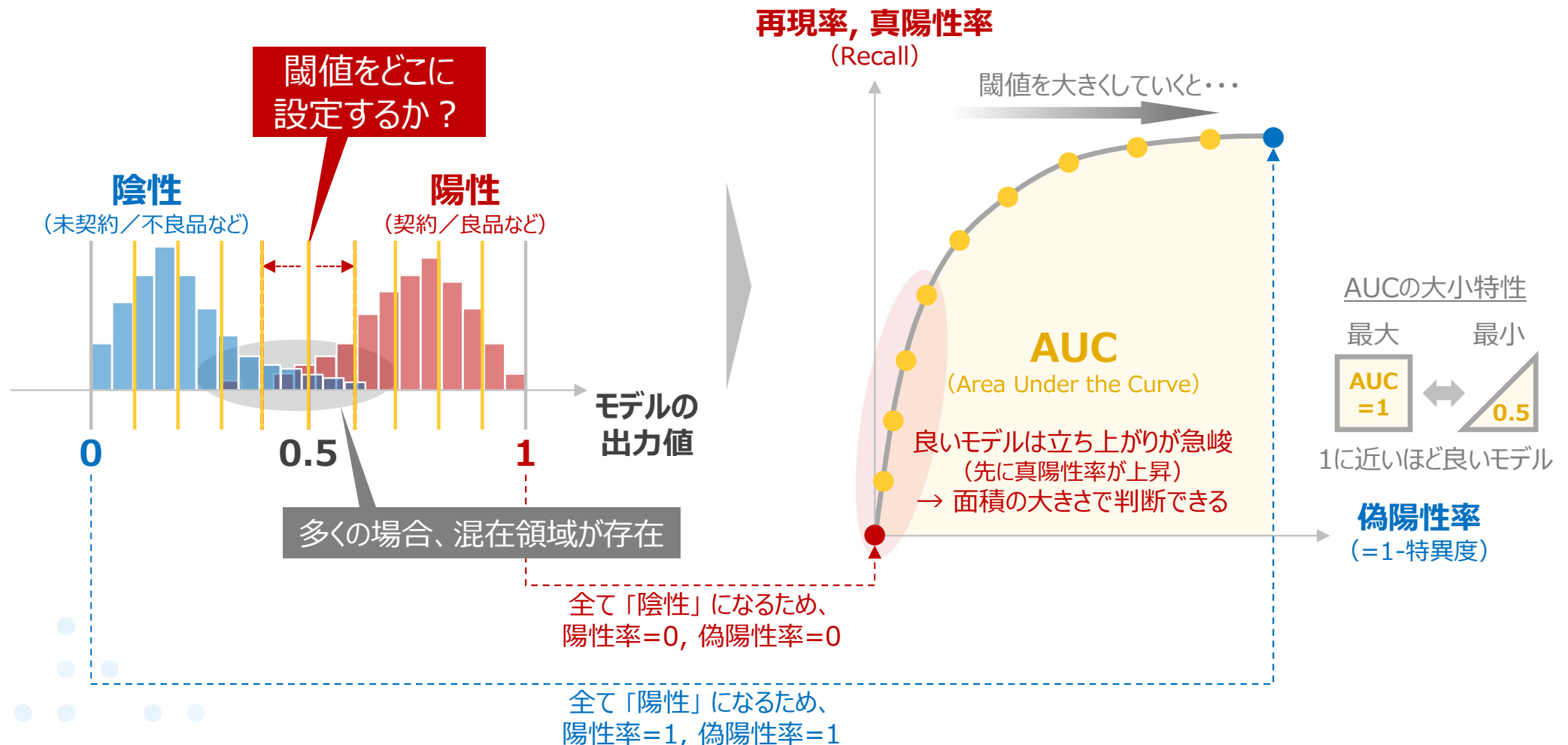
▼参考：閾値=0.5の時の精度

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



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

- 回帰係数を見ると、職業=retired や、最終連絡月=3月、前回キャンペーンの結果=success など、有意に高い値を示し、関係性が示唆される
- 一方で、最終連絡月が年末年始にかかると、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

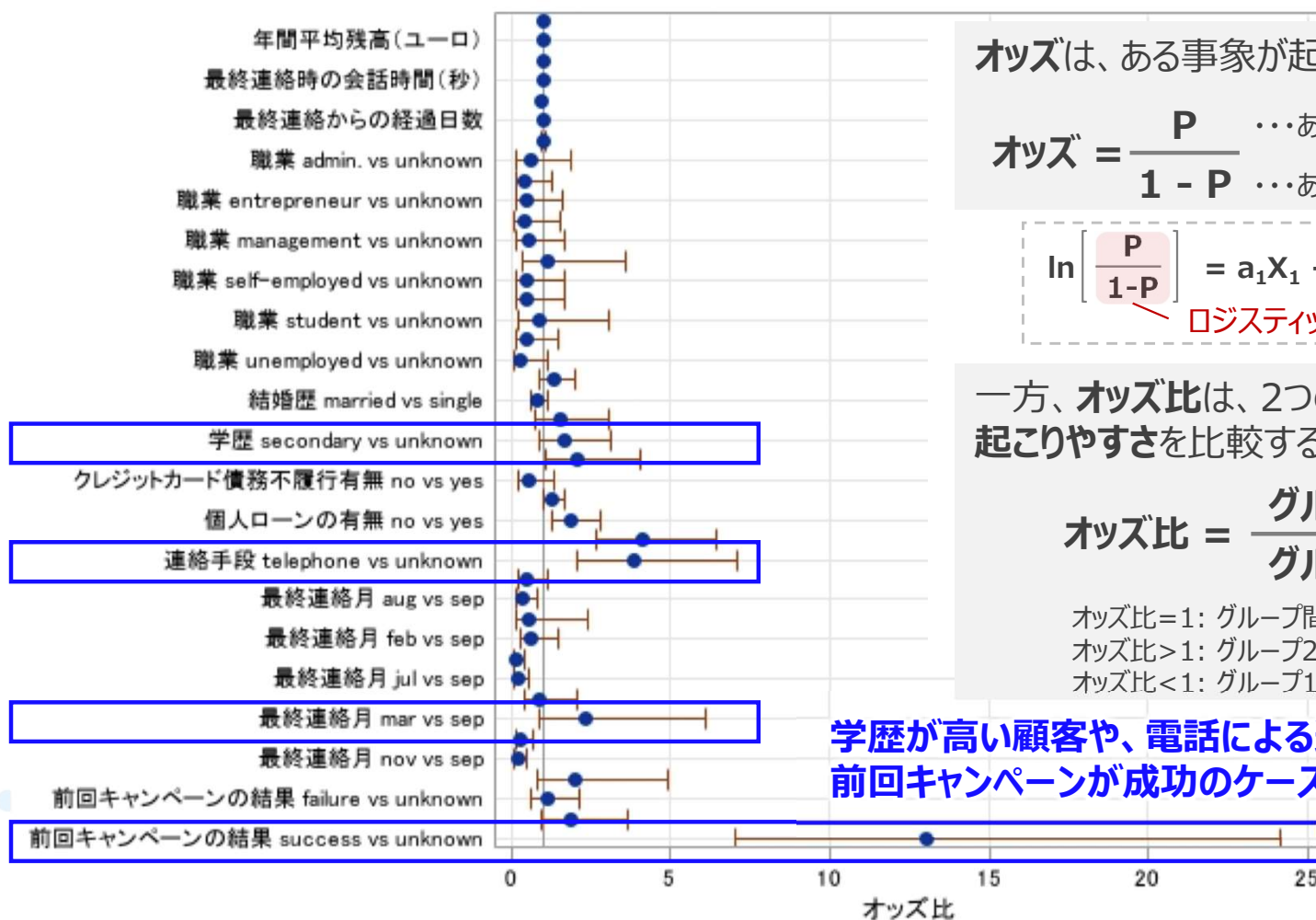
パラメータ		自由度	推定値	標準誤差	Wald カイ 2 乗値	Pr > ChiSq
年齢		1	-0.00423	0.00713	0.3528	0.5525
年間平均残高 (ユーロ)		1	-3.91E-6	0.000017	0.0500	0.8230
最終連絡日		1	0.0164	0.00816	4.0427	0.0444
最終連絡時の会話時間		1	0.00423	0.000202	437.3055	<.0001
キャンペーン中の連絡		1	-0.0704	0.0282	6.2314	0.0126
最終連絡からの経過日		1	-0.00010	0.000996	0.0097	0.9217
キャンペーン前の連絡		1	-0.00551	0.0382	0.0208	0.8853
職業	admin.	1	0.0580	0.1895	0.0938	0.7594
職業	blue-collar	1	-0.3344	0.1747	3.6627	0.0556
職業	entrepreneur	1	-0.1917	0.3147	0.3712	0.5424
職業	housemaid	1	-0.2949	0.3469	0.7230	0.3952
職業	management	1	-0.0150	0.1583	0.0089	0.9247
職業	retired	1	0.6896	0.2434	8.0236	0.0046
職業	self-employed	1	-0.1231	0.2852	0.1863	0.6660
職業	services	1	-0.0876	0.2165	0.1638	0.6857
職業	student	1	0.4365	0.3189	1.8734	0.1711
職業	technician	1	-0.1346	0.1597	0.7100	0.3995
職業	unemployed	1	-0.5815	0.3554	2.6773	0.1018
結婚歴	divorced	1	0.2582	0.1164	4.9195	0.0266
結婚歴	married	1	-0.2113	0.0836	6.3911	0.0115
学歴	primary	1	0.00504	0.1628	0.0010	0.9753
学歴	secondary	1	0.0851	0.1183	0.5178	0.4718
学歴	tertiary	1	0.3258	0.1391	5.4859	0.0192

パラメータ		自由度	推定値	標準誤差	Wald カイ 2 乗値	Pr > ChiSq
クレジットカード債務	no	1	-0.2723	0.2157	1.5935	0.2068
住宅ローンの有無	no	1	0.1300	0.0690	3.5463	0.0597
個人ローンの有無	no	1	0.3148	0.1000	9.9137	0.0016
連絡手段	cellular	1	0.4954	0.1129	19.2564	<.0001
連絡手段	telephone	1	0.4252	0.1670	6.4841	0.0109
最終連絡月	apr	1	-0.0726	0.1908	0.1449	0.7034
最終連絡月	aug	1	-0.3808	0.1690	5.0757	0.0243
最終連絡月	dec	1	0.0418	0.5804	0.0052	0.9426
最終連絡月	feb	1	0.1295	0.2216	0.3417	0.5588
最終連絡月	jan	1	-1.1959	0.3240	13.6227	0.0002
最終連絡月	jul	1	-0.8241	0.1826	20.3663	<.0001
最終連絡月	jun	1	0.4815	0.2241	4.6172	0.0317
最終連絡月	mar	1	1.4259	0.3214	19.6793	<.0001
最終連絡月	may	1	-0.5627	0.1671	11.3422	0.0008
最終連絡月	nov	1	-0.9156	0.2088	19.2357	<.0001
最終連絡月	oct	1	1.2884	0.2638	23.8489	<.0001
前回キャンペーンの結	failure	1	-0.7036	0.1656	18.0635	<.0001
前回キャンペーンの結	other	1	-0.2124	0.1922	1.2221	0.2689
前回キャンペーンの結	success	1	1.7413	0.1797	93.9424	<.0001

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

オッズ比の 95% Wald 信頼限界



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

学歴が高い顧客や、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

参考：モデルチューニング（変数選択）

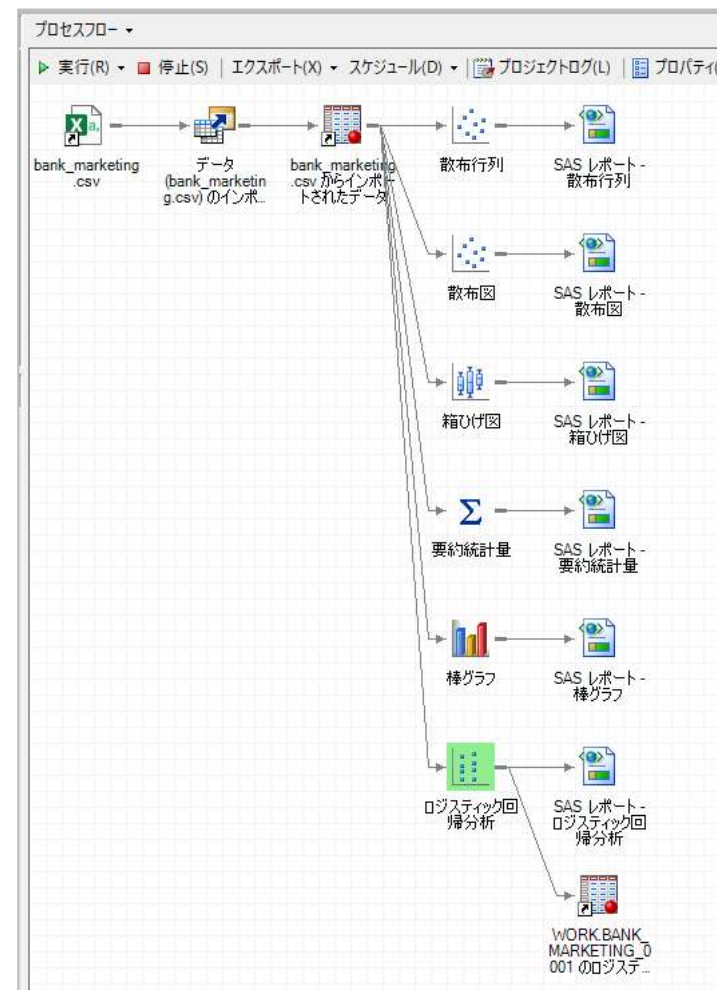
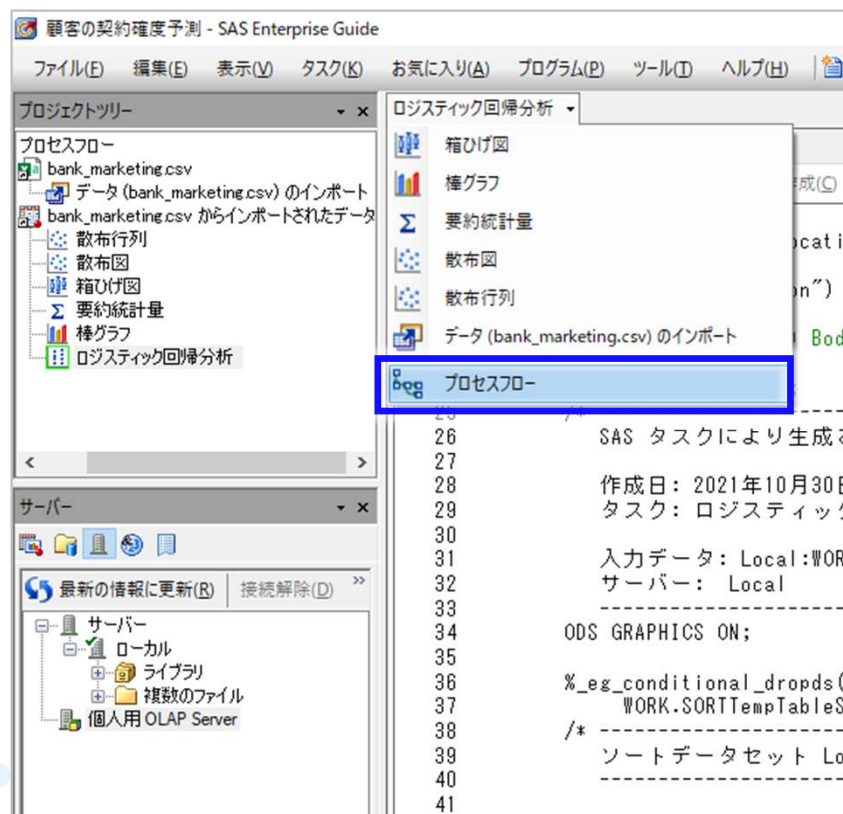
- 今回は、全ての説明変数を用いてモデル構築を行ったが、説明変数の追加／除外による精度変化を見ることが最適な変数選択を行うこともある
- ただし、予めドメイン知識を活用して変数を絞り込んでおくことも重要である



- **変数選択なし（強制投入法）**：全ての説明変数を強制的に入力として与える
- **変数増加法**：まず1変数（単回帰）でモデル構築し、精度を確認しながら順次変数を追加する
- **変数減少法**：まず全説明変数を入れた重回帰モデルを構築し、から出発し、精度を確認しながら順次変数を減らす
- **変数増減法（ステップワイズ法）**：精度変化を見ながら追加と除外を判定する。最もよく用いられる

参考：プロセスフローでの確認

- SAS Enterprise Guideでは、「プロセスフロー」によりデータ加工やグラフ観察、モデル構築などの一連の分析プロセスを俯瞰的／反復的に確認可能



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