銀行顧客データを用いた
定期預金・NISA開設予測モデル構築

機械学習による提案方針最適化シミュレーション
(詳細版)

Created by T. Yoshino

GitHub:https://github.com/tetsuyayoshino777-coder/customer-targeting-portfolio

本レポートの構成

- イントロダクション　3p
- メイン　11p
- まとめ　41p
- 付録　45p

イントロダクション

イントロダクションは以下の内容について記載する。

- 分析が必要になった背景と目的
- 分析で使用するデータ
- 分析で設定された課題と分析の結果

分析が必要になった背景

Yさんは、ある銀行のマーケティング担当者として働いている。

Yさんの働く銀行では、商材ごとに異なる部署が組成されており、それぞれの部署が自部署の商材の拡販を最善に業務を遂行している。

Yさんの働くマーケティング部は、自社の各部署から依頼を受け、自社が有する顧客名簿に向けたメールの配信や、Web広告の出稿など、各種マーケティング施策を遂行している。

今回、会計年度末が近いこともあり、「定額預金申し込み」を扱う部署Aと、「NISA口座開設申し込み」を扱う部署Bから同時期に、自社が有する顧客名簿に向けた特別キャンペーンメールの配信依頼の依頼を受けている。

ただし、特別キャンペーンメールの乱用は会社のブランド価値を棄損することから、１顧客に対して月に１回のみ実施が許可されているため、同じ顧客に対して定期預金とNISA口座開設の両方をメール配信することができない状態である。

なお、部署A・Bともに目標達成が厳しいことから、それぞれ、できるだけ多くの顧客に対するメール配信を希望している。

Yさんはメールマーケティングの担当者として、会社全体としての売上の最大化を目指す立場であり、根拠をもって施策を考案し、適切な意思決定のもと、部署A・部署Bともに納得できるよう説明する義務を負っている。

分析の目的

分析方針：

自社保有のデータから「定額預金申し込み」、および「NISA口座開設申し込み」に寄与している変数を見つけ出し、顧客別に「定額預金申し込み」と「NISA口座開設申し込み」のどちらを優先して訴求するべきかを明らかにする。

分析の目的：

部署Aと部署Bのマネージャーに対して、客観的な根拠に基づく提案を行うことを目的とする。提案に際して、本分析レポートを提出する。
最終的には、商材別成功確率を比較し、顧客単位で配信商材を最適化する。

分析で使用するデータ

分析には、以下のデータを使用する。

▼ *データセット名：銀行の顧客ターゲティング(学習用データ)*

このデータはSIGNATEの下記URLにおける「学習データ」をベースとして、問題設定に対応するよう加工を加えたダミーデータである。

（サンプルサイズ 27,128、カラム数 18含まれるデータ）

（ベースとなる分析データの取得元）

https://user.competition.signate.jp/ja/competition/detail/?competition=092375ab3c4a43c18c8277e1fd264aa9&task=f3c678327db64f3b988fb85d8a49e5ed&tab=dataset

分析で使用するデータ

今回分析全体の目的変数として「定期預金申し込み有無」および「NISA口座開設申し込み有無」を使用し、その他の変数は説明変数として扱う。

なお、「NISA口座開設申し込み有無」については付録記載のP40によって付与したダミーデータであり、その他の項目はすべてSIGNATEのサンプルデータをそのまま採用している。

カラム	ヘッダ名称	データ型	説明
0	id	int	行の通し番号
1	age	int	年齢
2	job	varchar	職種
3	marital	varchar	未婚/既婚
4	education	varchar	教育水準
5	default	varchar	債務不履行があるか（yes, no）
6	balance	int	年間平均残高（€）
7	housing	varchar	住宅ローン（yes, no）
8	loan	varchar	個人ローン（yes, no）
9	contact	varchar	連絡方法
10	day	int	最終接触日
11	month	char	最終接触月
12	duration	int	最終接触時間（秒）
13	compaign	int	現キャンペーンにおける接触回数
14	pdays	int	経過日数：前キャンペーン接触後の日数
15	previous	int	接触実績：現キャンペーン以前までに顧客に接触した回数
16	postcome	varchar	前回のキャンペーンの成果
17	y	boolean	定額預金申し込み有無（1:有り, 0:無し）※目的変数
18	x	boolean	NISA口座開設申し込み有無（1:有り、0:無し)※目的変数

ダミーデータの付与要件

以下①～③を条件として、NISA口座開設申し込み有無を示す変数xを付与する

①NISA口座を開設するペルソナ像として下記を想定。

- 年齢（age）：30歳以上の人は、資産運用への関心が高いと仮定。
- 銀行口座残高（balance）：2,000€以上の人は十分な資産を持っている可能性が高い
- 職業（job）：下記の職業は比較的収入が安定し、投資への意欲が高いと仮定
  （management（管理職）/entrepreneur（起業家）/technician（技術職））

②上記すべての条件を見ていたとしても、成功率は70%とする

③定額預金申し込み有無の総数と同程度、NISA口座開設申し込み有無も発生するものとする

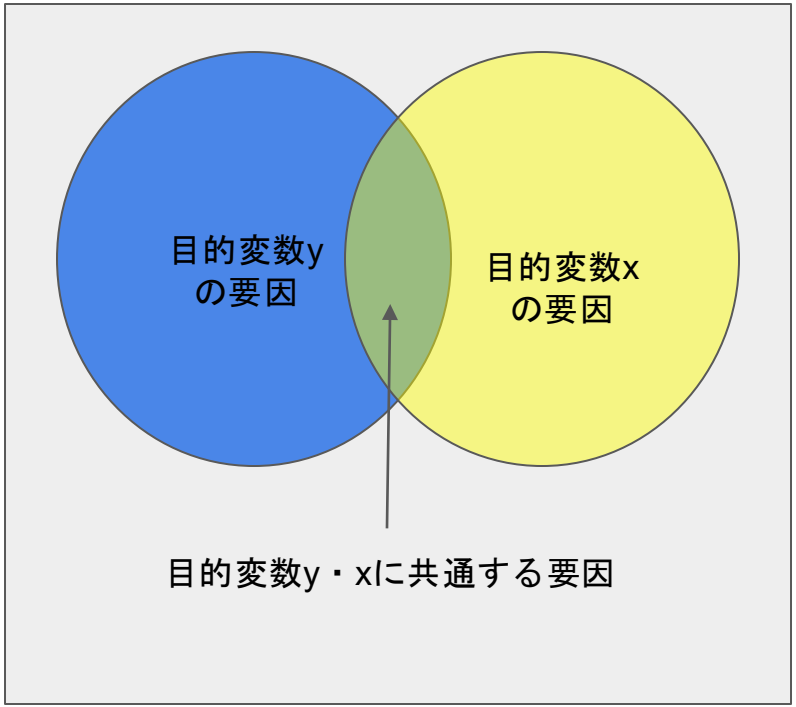
※詳細条件は付録記載のp41を参照

分析で設定された課題と分析方針

本分析では、analytics_report_sample.csvを使用して、商材ごとに目的達成（口座開設）に寄与する要因を見つけ出し、顧客ごとにより可能性の高い商材を優先して紹介することを目的として分析を行う。

なお、商材ごとに目的達成に寄与する要因が重複する場合、多変量解析もしくは機械学習により優先すべき商材を統計的に分析を行し、提案する。

参考：要因分類のイメージ図



分析結果（概要）

データ分析の結論として以下の実施を提案する。

- 「NISA口座」においては、前回の定額預金向けキャンペーンに成功していない方、予算残高が多い方、職種が管理職や技術者、起業家といった属性の顧客を優先してメール配信する

- 「定額預金」においては、過去のキャンペーン実績に成功している、過去の接点において通話時間が長い、職種が学生や定年退職者といった属性の顧客を優先してメール配信する

なお、詳細は次ページ以降に記述する。

メインパート

メインパートの構成

（例）

メインパートでは、以下について記載する。

代表値の比較

特徴量重要度分析

- ランダムフォレスト（特徴量重要度分析）
- ロジスティック回帰分析（特徴量の方向性分析）

クラスタリング分析

商材別の成功確率の試算に基づく最適化提案

代表値の比較

本パートではデータセット全件およびy=1、x=1の3つの分類について、各変数ごとの分布状況を確認する。

結論として、分布の傾向から、下記の特徴があると考えられる。

（次ページ以降、一部変数のヒストグラムを抜粋して記載）

xとyの両方に重要な変数

- **数値型**
  - `balance`: yでは中程度残高、xでは高額残高が成功率に寄与。
- **オブジェクト型**
  - `job`: **管理職（management）や技術者（technician）**多く、x=1でより顕著。

yにとって重要な変数

- **数値型**
  - `duration`: **長い通話時間**（500秒以上）が成功率に影響。
  - `pdays`: 過去に**接触経験**がある顧客（100日以内）が多い。
- **オブジェクト型**
  - `postcome`: **過去成功**（success）の顧客が多い。

xにとって重要な変数

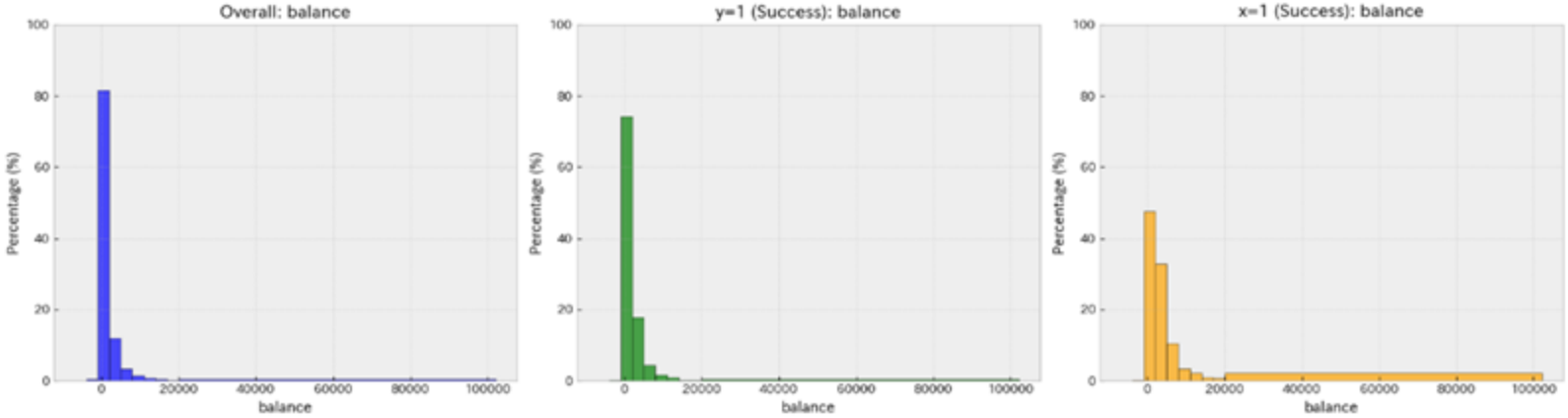
- **数値型**
  - `balance`: **高額口座残高**（5,000ユーロ以上）が成功率に影響。
- **オブジェクト型**
  - `education`: **高等教育（tertiary）**が大半を占める。

代表値の比較：balance（口座残高）

全体: 残高が0より低い分布が最も多く、y=1、x=1の傾向と異なる。

y=1: 全体より残高が高い傾向があり、特に1,000～3,000ユーロの顧客が多い。

x=1: y=1と同様に1,000～3,000ユーロの顧客が多いが、より高額な割合が他と比して多い。



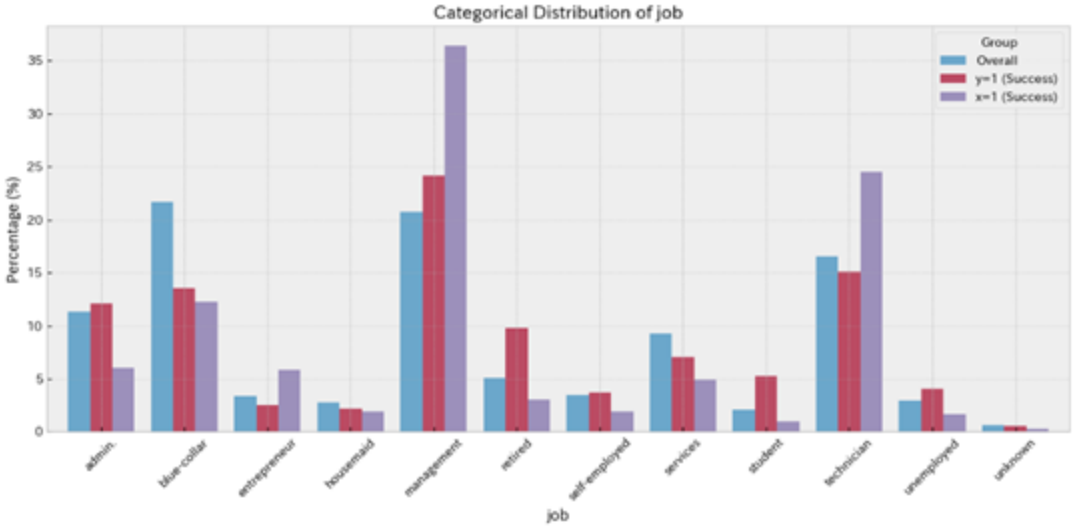
代表値の比較：job（職業）

全体: blue-collar（労働者）が最多で、次にmanagement（管理職）やtechnician（技術職）が多い。

y=1: management（管理職）とtechnician（技術職）の割合が増加、blue-collarの割合が減少。

x=1: y=1と同様にmanagement（管理職）とtechnician（技術職）の割合が増加、blue-collarの割合が減少。

傾向: 職業別で成功傾向が異なり、特にmanagementがyとxの両方で高い成功率を示す。また、よりxのほうが分布の偏りが顕著に表れている。

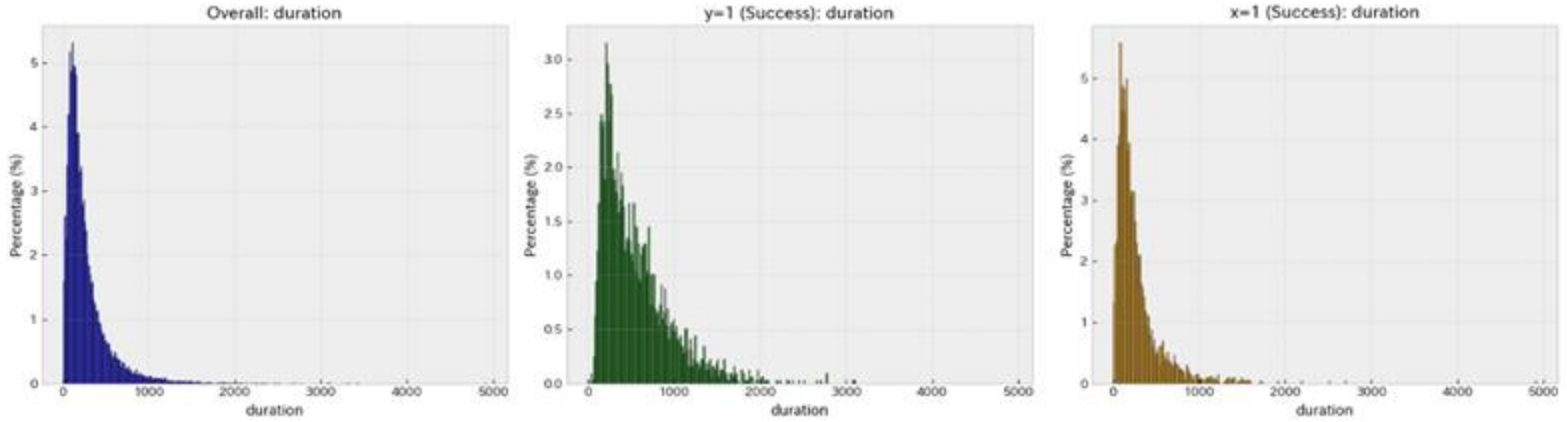


代表値の比較：duration（最終接触時間（秒））

全体では通話時間は100～300秒が中心で、一部で長い通話が存在。

y=1では通話時間が長いほど成功率が高く、500秒以上の割合が顕著に高い。

x=1も全体と同様の傾向。

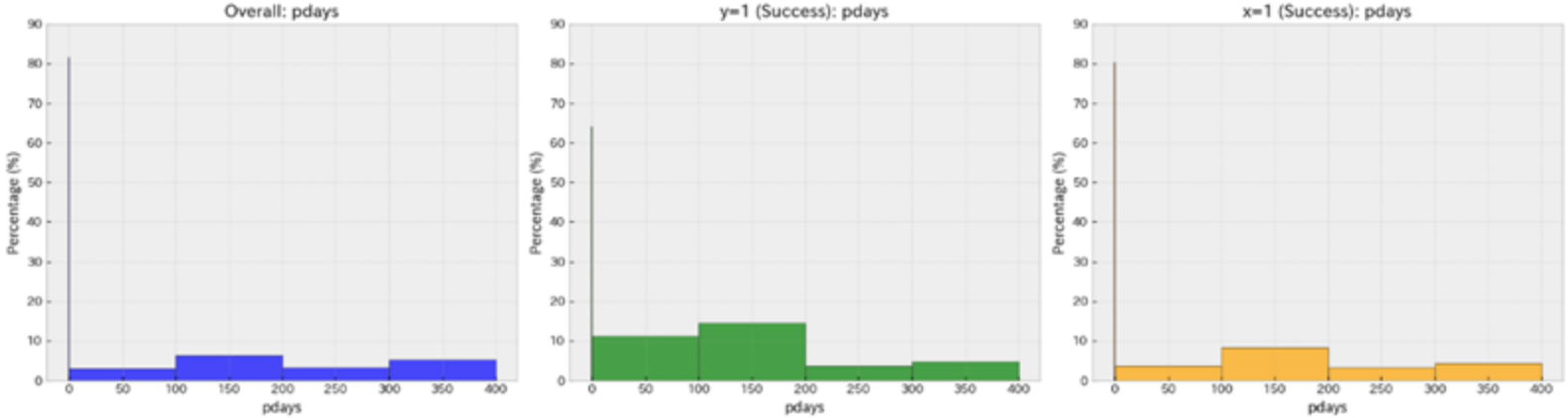


代表値の比較：pdays（最後のキャンペーンからの日数）

いずれの分類においても、-1(推定：未接触）が最も多い。

全体およびx-1は、-1（未接触）の割合が80%程度に対し、**y=1では65%程度。**

y=1では100日以内に接触経験のある顧客が比較的多い。

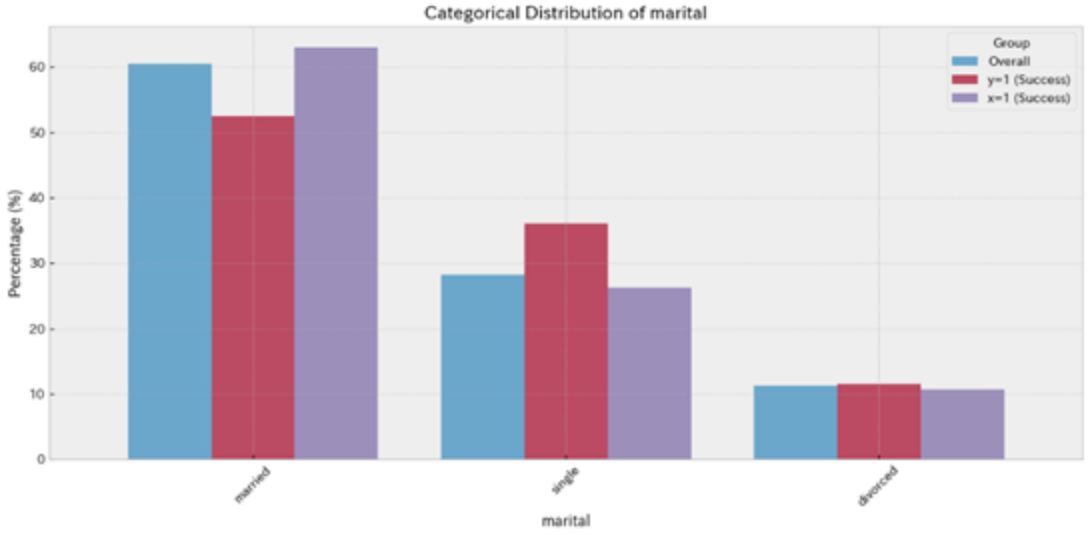


代表値の比較：marital（婚姻状況）

全体: married（既婚）が過半数を占め、次にsingle（未婚）、divorced（離婚）が続く。

y=1: single（未婚）の割合が増加し、married（既婚）の割合が減少。

x=1: 分布が全体とほぼ一致し、婚姻状況による顕著な偏りは少ない。

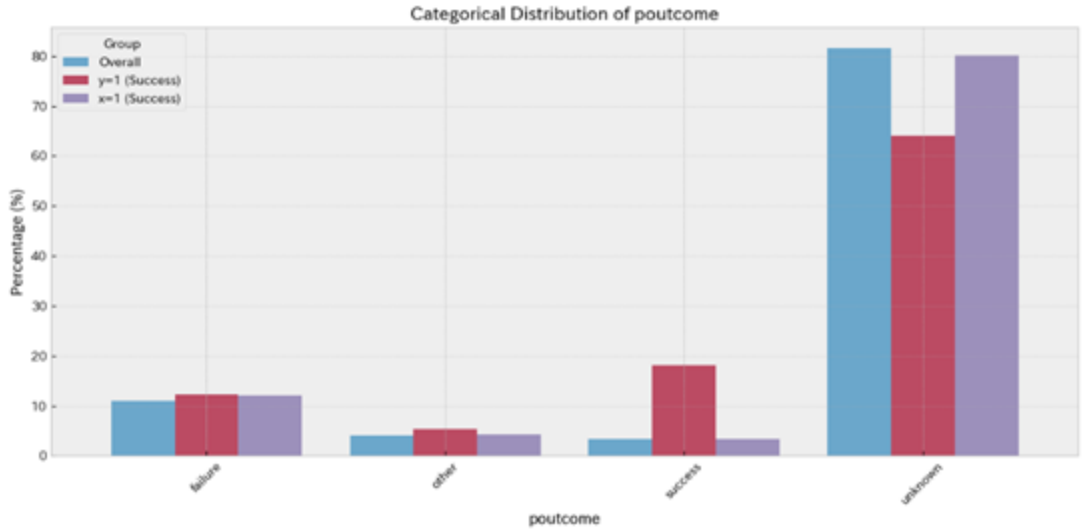


代表値の比較：postcome（過去のキャンペーン結果）

全体: unknownが大半を占めるが、failureとotherも一定割合を占める。

y=1: success（過去成功）が増加し、unknownの割合が減少。

x=1: 全体の傾向と類似



代表値の比較：数値型変数の傾向まとめ

変数	全体	y=1 の傾向	x=1 Tの傾向	Comment
age	年齢（顧客の年齢）。数値範囲は広いが、多くは中高年層に集中。	◎ 中高年層（30代後半以上）が成功率が高い。	○ 同様に高年齢層が成功率が高いが、yよりも年齢層に広がりがある	yとxの両方で重要だが、yでの成功傾向がより顕著。
balance	低～中程度の残高が多い	◎ 中程度の口座残高（1,000～3,000ユーロ）が多い	○ 高額口座残高（5,000ユーロ以上）が多い	y=1は中程度の残高が多く、x=1は高額残高への偏りがやや強い
duration	中程度（300秒前後）が多い	◎ 長い通話時間（500秒以上）が多い	△ 短い通話時間（100～300秒）が多い	y=1は長時間通話が多いが、x=1は短い時間帯の分布に近い
campaign	1～3回が中心	△ 少ない接触回数（1～2回）がやや多い	△ 接触回数の多さは成功率に影響しない	両者とも接触回数に大きな偏りは見られない
pdays	未接触（-1）が大半	◎ 過去に接触経験がある顧客（pdaysが100日以内）が多い	△ 過去の接触がなくても成功しやすい	y=1は過去接触が重要だが、x=1は未接触でも成功可能
previous	過去接触なしが多い	○ 過去の接触回数が1～2回の顧客が多い	△ 過去接触回数の影響は小さい	y=1は過去接触がやや影響、x=1は影響が少ない
day	最後に連絡を取った日（その月の1日～31日）。均等に分布している。	△ 特定の日付では成功率がやや高いが、強い影響はない。	△ 同様に弱い影響が見られる。	日付そのものより、通話内容やタイミングの方が重要と考えられる。

代表値の比較：オブジェクト型変数の傾向まとめ

変数	全体	y=1 の傾向	x=1 Tの傾向	Comment
job	労働者（blue-collar）が最多	◎ 管理職（management）、技術職（technician）が多い	◎ 管理職（management）が圧倒的に多い	x=1は管理職に顕著な偏りがあり、y=1は技術職も多い
marital	既婚（married）が多数派	△ 未婚（single）の割合がやや高い	△ 婚姻状況による顕著な影響は少ない	y=1は未婚に若干偏りがあり、x=1はほぼ全体分布に近い
education	中等教育（secondary）が最多	○ 高等教育（tertiary）の割合が高い	◎ 高等教育（tertiary）が大半を占める	x=1の方が高等教育に偏る度合いが強い
default	債務不履行なし（no）が圧倒的に多い	× 債務不履行なし（no）がほとんどを占める	× 債務不履行の有無による影響はほぼない	両者とも全体分布とほぼ一致し、影響は見られない
housing	住宅ローンあり（yes）が多数	△ 住宅ローンなし（no）の割合がやや高い	△ 住宅ローンの有無による顕著な影響はない	y=1はローンなしがやや有利、x=1はほぼ全体分布に近い
loan	個人ローンなし（no）が圧倒的に多い	◎ 個人ローンなし（no）が大半を占める	◎ 個人ローンなし（no）が中心	両者とも個人ローンなしが重要
contact	携帯電話（cellular）が主流	◎ 携帯電話（cellular）が圧倒的に多い	◎ 携帯電話（cellular）が圧倒的に多い	両者とも携帯電話での連絡が中心
month	5月（may）が最多	○ 8月（aug）の成功率がやや高い	△ 月別の顕著な偏りはない	y=1は8月にやや偏りがあり、x=1はほぼ全体分布に近い
postcome	不明（unknown）が大半	◎ 過去成功（success）の顧客が多い	△不明（unknown）が大半	y=1は過去成功の影響が特に顕著

特徴量重要度分析

目的:

各商材（定額預金/NISA口座）における成功率を高めるため、どの変数が重要かを特定する。

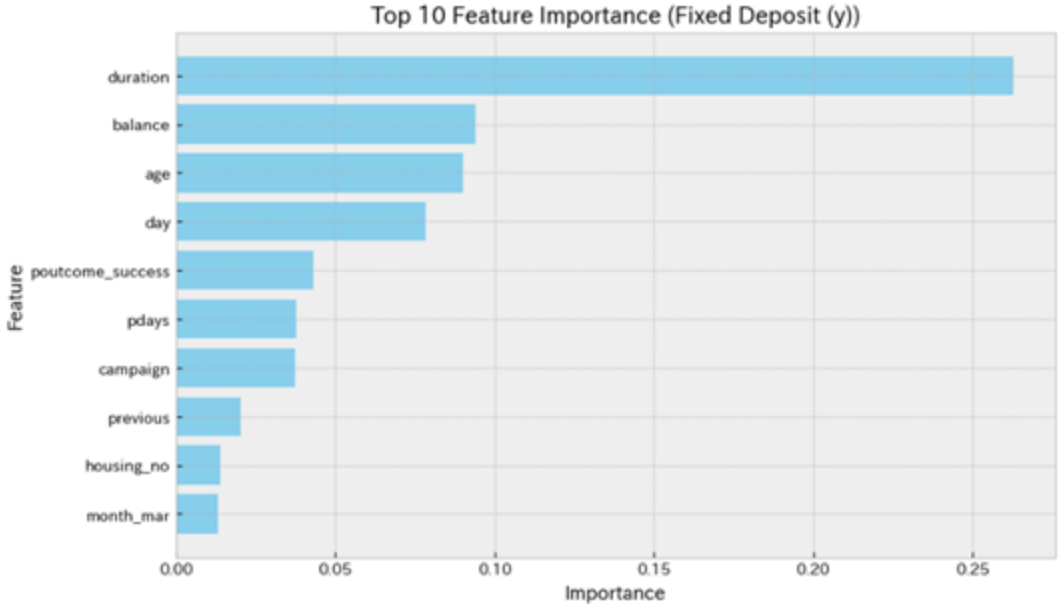
手法:

ランダムフォレストやロジスティック回帰モデルを使用して、変数の重要度を可視化する。

a. ランダムフォレストの 特徴量重要度（Importance） は、モデル全体の予測力に対する各変数の寄与割合を確認する。
b. ロジスティック回帰モデルでは、特徴量の「方向性」を理解する。

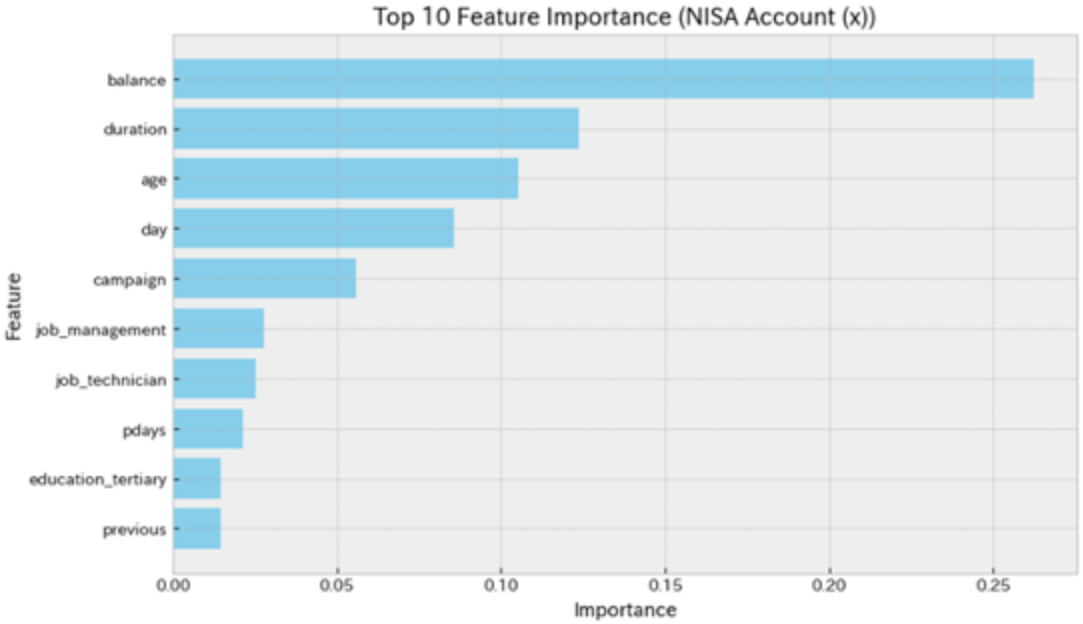
ランダムフォレストによる特徴量重要度：定期預金(y=1)

- **上位10個の特徴量の重要度：69.02%**
- **特徴的な傾向**:
  - 通話時間（duration）が最も重要で、長時間のコミュニケーションが成功に直結。
  - 銀行残高（balance）や顧客の年齢（age）も成功に大きく影響。
  - 過去の成功経験（postcome_success）や最近の接触（pdays）が重要。
- **戦略的示唆**:
  - 長い通話時間を確保できるリソース配分を行う。
  - 過去に成功したキャンペーン対象顧客を優先的にアプローチ。



ランダムフォレストによる特徴量重要度：NISA口座(x=1)

- **上位10個の特徴量の重要度：73.68%**
- **特徴的な傾向**:
  - 銀行残高（`balance`）が圧倒的に重要。高額残高の顧客が開設しやすい。
  - 年齢（`age`）や職業（`job_management`、`job_technician`）が関連。
  - 高等教育（`education_tertiary`）も影響を持つ。
- **戦略的示唆**:
  - 高額残高の顧客や特定の職業層をターゲットにした配信を行う。
  - 教育水準が高い顧客に特化したメッセージを作成。



ランダムフォレストによる特徴量重要度：その他の示唆

- **共通点**:
  - 通話時間（duration）や口座残高（balance）、年齢（age）が両商材で重要。
- **相違点**:
  - 定額預金は過去の成功経験や住宅ローンの有無が重要。
  - NISA口座は職業や教育水準が成功率に影響。
- **参考**
  - 年齢（age）が重要という結果は、変数間の相関性が影響していると想定し、相関状況を確認したものの、特に強い相関は確認できなかった

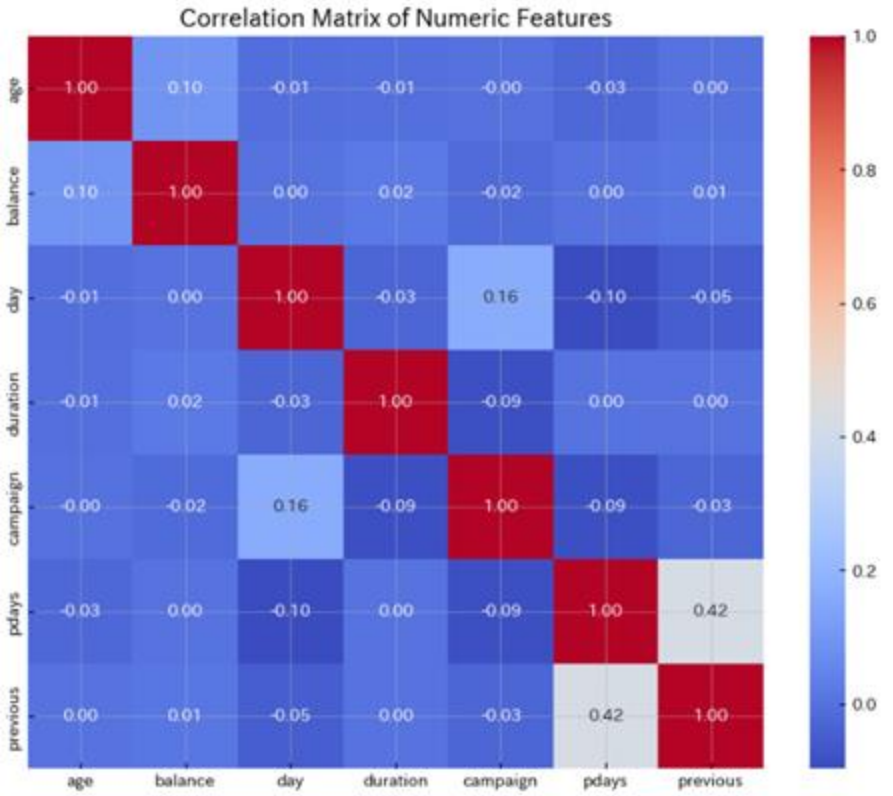
参考：変数間の相関分析

他の変数との相関が低い:

- age と他の変数との相関はすべて |0.1| 未満であり、ほぼ独立している。
- この結果から、age のモデルでの重要度は、他の変数の影響を受けず、純粋に年齢自体が重要な特徴量として機能していると考えられる。

可能性の示唆:

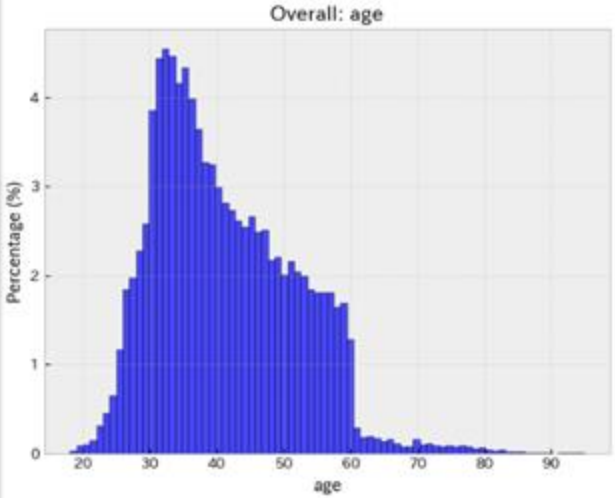
- age の重要度が高い理由は、顧客の年齢そのものがターゲット変数（y または x）に直接影響しているためと推測できる。
- 例えば、特定の年齢層が定額預金やNISA口座の申し込みに積極的である可能性がある。

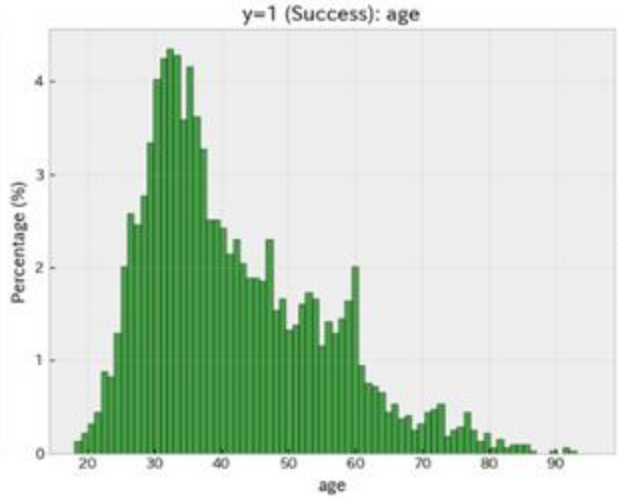


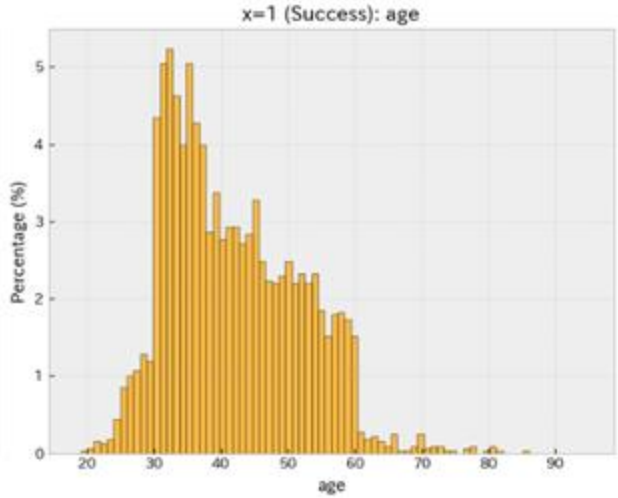
参考：age（年齢）

年齢層は広く分布しており、30代から50代が中心。

y=1(定額預金)は比較的年齢層の幅が他の分類と比して広い。







ロジスティック回帰分析（変数の寄与度の確認）

目的

- ランダムフォレストでは、特徴量の重要度（`importance`）が提供されますが、その影響が「正の方向（成功確率を増加させる）」か「負の方向（成功確率を減少させる）」かは不明。
- ロジスティック回帰の **係数（`coefficients`）** を用いることで、各特徴量がターゲットに対して持つ「方向性」と「強度」を解釈する。

ロジスティック回帰の結果から得られる洞察

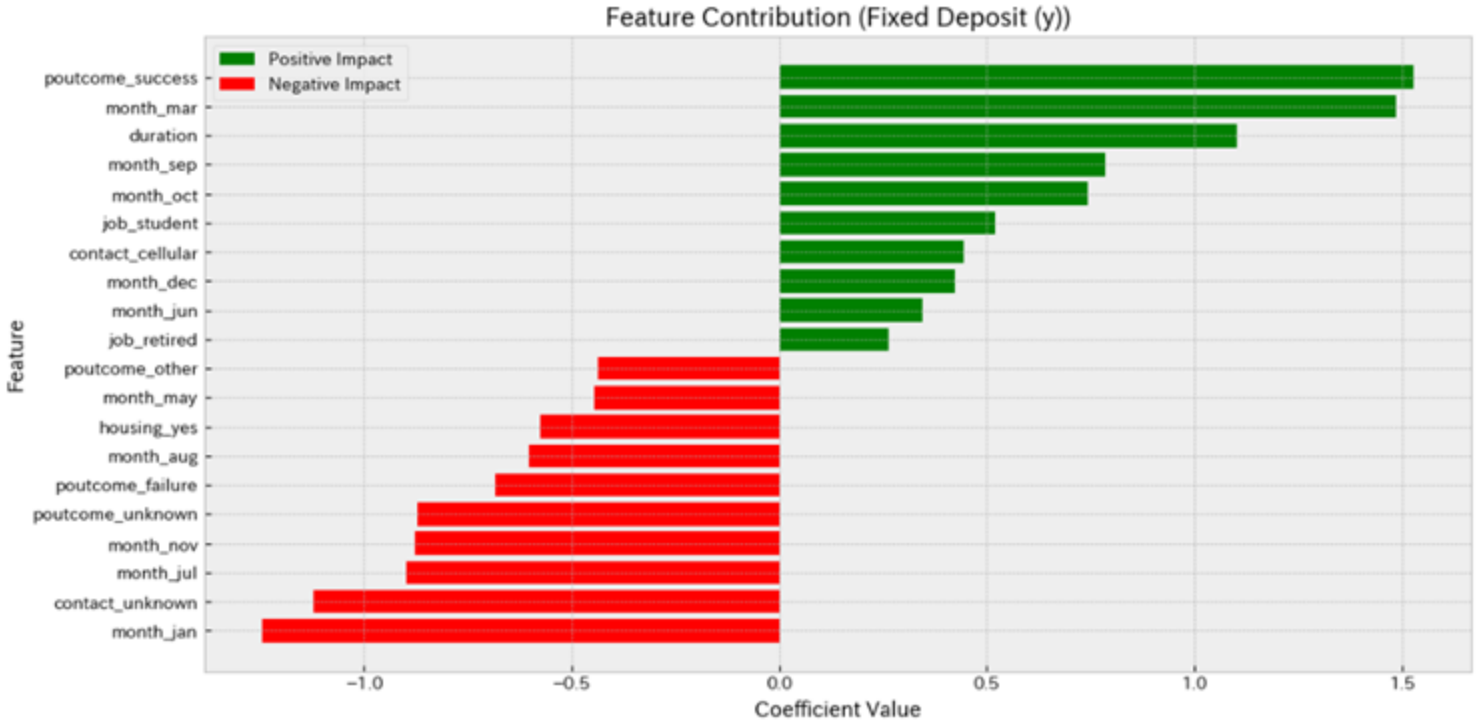
定期預金 (y=1) の特徴

- 成功確率に強い影響を与える特徴は、**時期（`month_*`）** と **通話内容（`duration`, `postcome_success`）**。
- 特定の職業層（学生や定年退職者）が高い成功率を示す。

NISA 口座 (x=1) の特徴

- **職業層（`job_*`）** と **資産状況（`balance`）** が大きな影響を与える。
  a. **職業層（`job_*`）においては、定期預金 (y=1) と異なる傾向がみられる。**
  b. **時期（month_*）**の影響もみられるが、**定期預金 (y=1) と比べて影響が小さい**

ロジスティック回帰分析：定期預金（y=1)の場合



ロジスティック回帰分析：定期預金（y=1)の場合

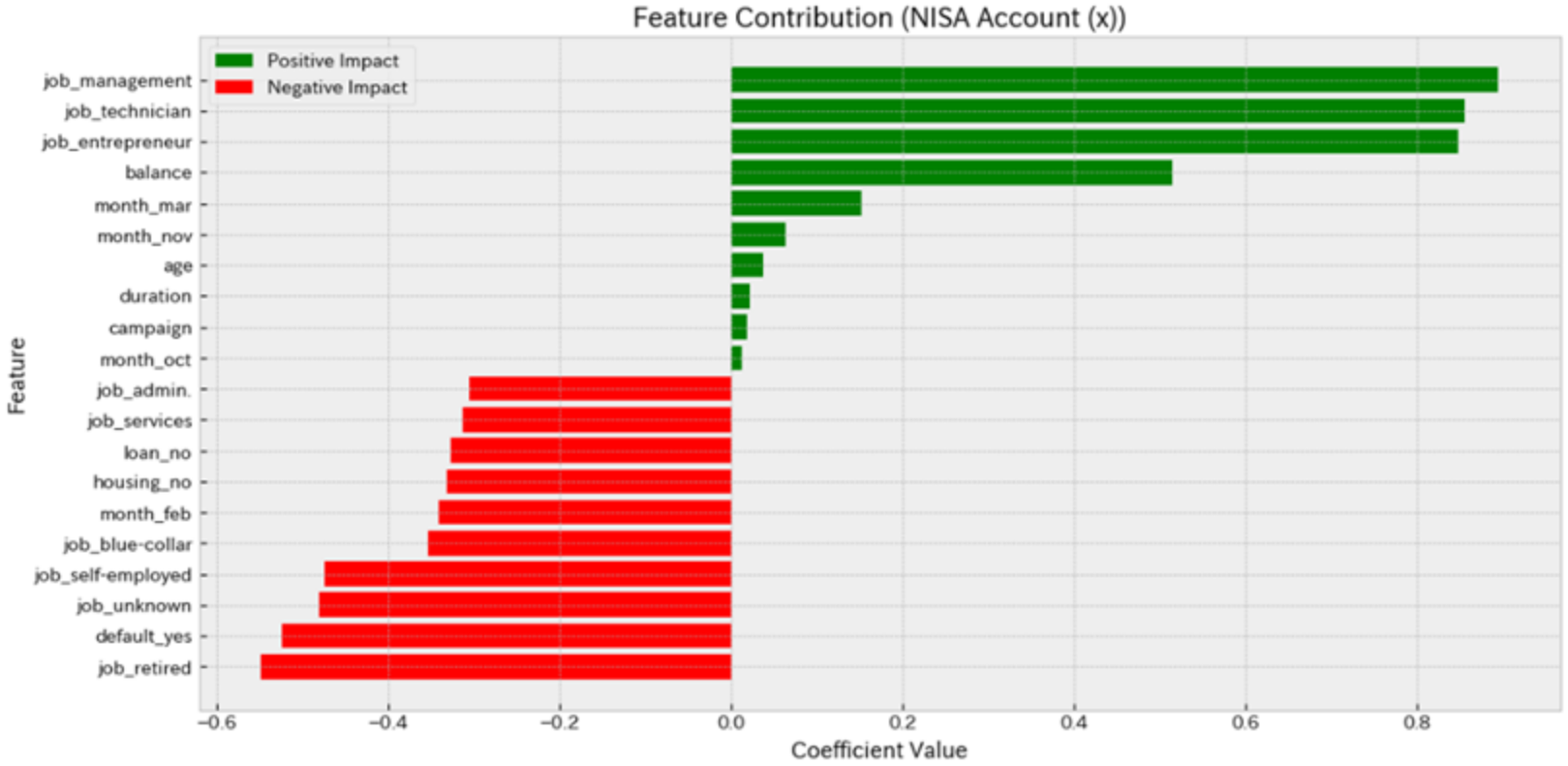
ポジティブな特徴（成功確率を高める特徴）

1. `postcome_success`（係数: 1.527970）:
   - 過去のキャンペーンで成功した顧客は、定額預金キャンペーンでも成功率が非常に高い。
   - 最も強い正の影響を持つ特徴量。
2. `month_mar`（係数: 1.485630）:
   - 3月に連絡を取った顧客は成功率が高い。
   - 時期的な要因が影響している可能性。
3. `duration`（係数: 1.102351）:
   - 通話時間が長いほど成功確率が高い。
   - 顧客との丁寧なコミュニケーションが重要。
4. **職業関連（例: `job_student`, `job_retired`）:
   - `job_student`（**係数: 0.521242**）: 学生が成功率で顕著に高い。
   - `job_retired`（**係数: 0.265190**）: 定年退職者も成功率が高い。
   - 顧客層のターゲティングとして、特定の職業が効果的。

ネガティブな特徴（成功確率を下げる特徴）

1. `month_jan`（係数: -1.245002）:
   - 1月に連絡を取った顧客の成功確率が特に低い。
   - 季節性の影響が大きい可能性。
2. `contact_unknown`（係数: -1.121665）:
   - 連絡手段が不明な顧客の成功確率が低い。
   - コンタクト情報の質が重要。
3. `postcome_unknown`（係数: -0.870679）:
   - 過去のキャンペーン履歴が不明な顧客も成功確率が低い。
4. `month_may`（係数: -0.445282）:
   - 5月に連絡した場合の成功確率も低い。
   - 時期に応じた施策の調整が必要。

ロジスティック回帰分析：NISA口座（x=1)の場合



ロジスティック回帰分析：NISA口座（x=1)の場合

ポジティブな特徴（成功確率を高める特徴）

1. `job_management`（係数: 0.895079）:
   - 管理職が成功率で最も高い。
   - 職業による成功傾向が顕著。
2. `job_technician`（係数: 0.855078）:
   - 技術職も高い成功率を示している。
3. `job_entrepreneur`（係数: 0.848420）:
   - 起業家も高い成功率を示す。
   - 特定の職業層へのターゲティングが有効。
4. `balance`（係数: 0.515032）:
   - 高額残高の顧客が成功率を引き上げる。
   - 金融資産の多い顧客がターゲット。

ネガティブな特徴（成功確率を下げる特徴）

1. `job_retired`（係数: -0.547720）:
   - 定年退職者の成功率が低い。
   - `y`（定額預金）で高かったのと対照的。
2. `default_yes`（係数: -0.523989）:
   - 債務不履行の履歴がある顧客は成功率が低い。
   - 信用リスクが影響。
3. `job_unknown`（係数: -0.480737）:
   - 職業が不明な顧客も成功確率が低い。
4. `loan_no` と `housing_no`（係数: -0.327273, -0.330850）:
   - 住宅ローンや個人ローンを持っていない顧客の成功率が低い。
   - NISA口座に対する金融ニーズが低い可能性。

クラスタリング分析

期待される結果

クラスタリング結果:

- 顧客属性に基づくセグメントの特徴が視覚的に確認可能。
  - エルボー法による事前分析の結果、クラスター数は2つが最適と判断し実行

セグメントごとの成功率:

- 定額預金（y）とNISA口座（x）の成功率をクラスター単位で比較。
- 特定セグメントに対するターゲティング戦略を立案可能。

結論

- クラスター0はNISA口座（x）、クラスター1は定額預金（y）を優先したほうが、成功率が高い

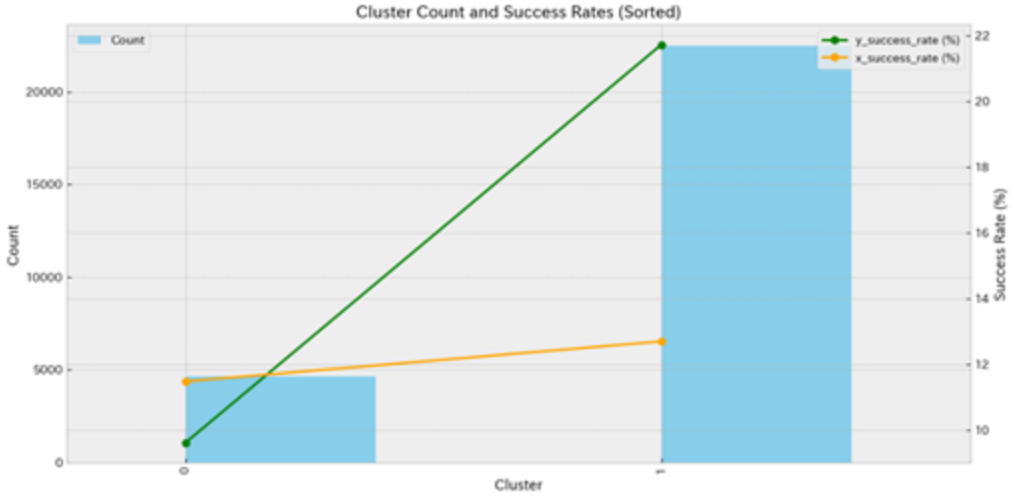
クラスタリング分析：クラスタリングの状況

クラスター 0:

- 顧客の **17.09%** を占める小規模なセグメント。特定の特徴を持つ少数派。
- キャンペーンの成功率
  a. NISA口座（x）：**12.71%**。
  b. 定額預金（y）：**9.63%**
- 定額預金（y）に比べると、NISA口座（x）に対する成功率の差は小さい。

クラスター 1:

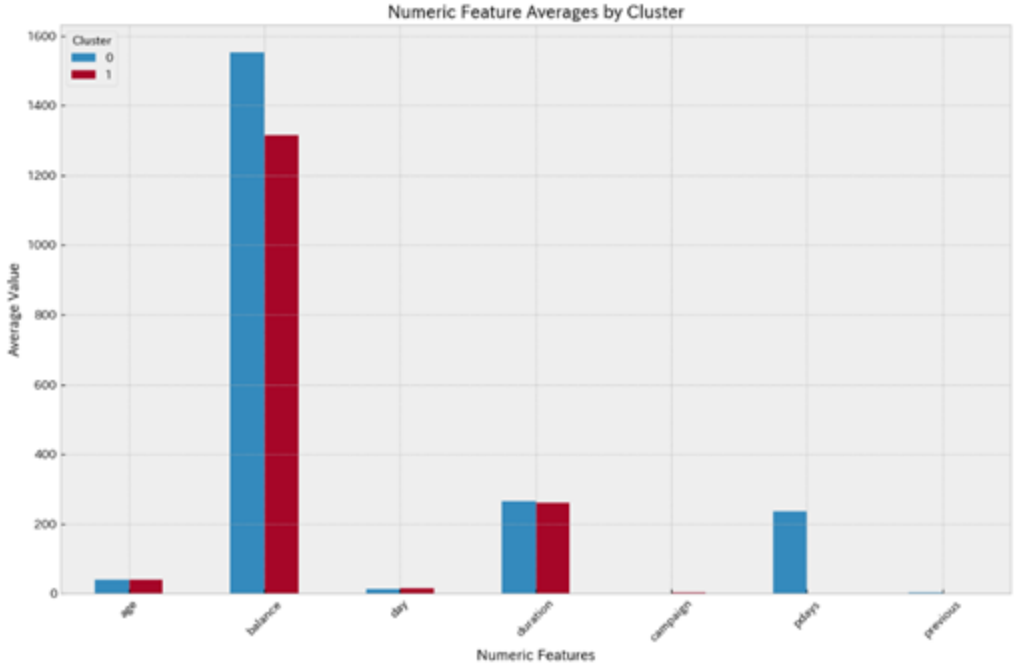
- 顧客の **82.91%** を占める支配的なセグメント。全体の大多数の顧客がこのクラスターに属する。
- キャンペーンの成功率
  a. NISA口座（x）：**11.49%**
  b. 定額預金（y）：**21.73%**
- 定額預金（y）の成功率が非常に高い（約 2倍以上）。



クラスタリング分析：クラスターごとの特徴(数値型変数)

数値型変数の平均値を比較。いずれの傾向も、ランダムフォレストやロジスティック回帰分析の結果とも整合している

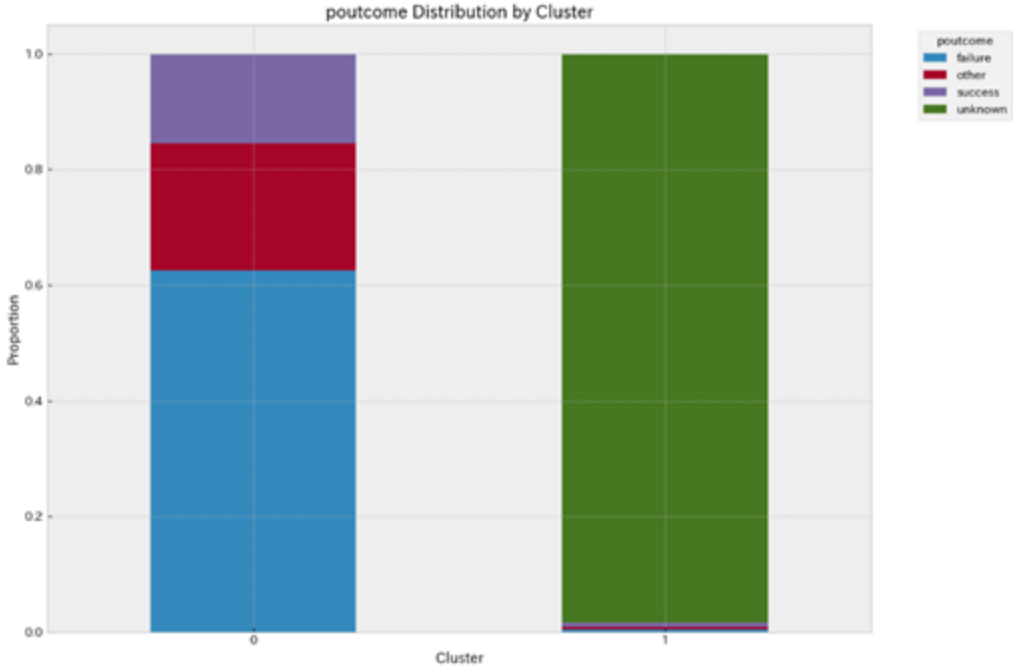
- balance（年間平均残高）がクラスター0のほうが高い
    a. NISA口座のほうがbalance（年間平均残高）が高いほどポジティブな影響を受ける
- pdays(経過日数：前キャンペーン接触後の日数）がクラスター0のほうが高い
    a. クラスター0はpdaysの結果に左右されにくい



クラスタリング分析：クラスターごとの特徴(カテゴリ型変数)

カテゴリ型変数については特徴的な傾向のある変数を抜粋して記載。

- postcome(前回のキャンペーン実績)については、クラスター0がfailureが最も多く6割程度。続くotherが20%程度であり、合わせて83%程度を占める
- クラスター1については、unknownが大半を占める。
- postcomeはyの前回実績であることから、前回の結果の時点でfailureもしくはotherであれば、改めて接触を持った場合においても、成功率は低いと考えられる。



商材別の成功確率の試算に基づく最適化提案

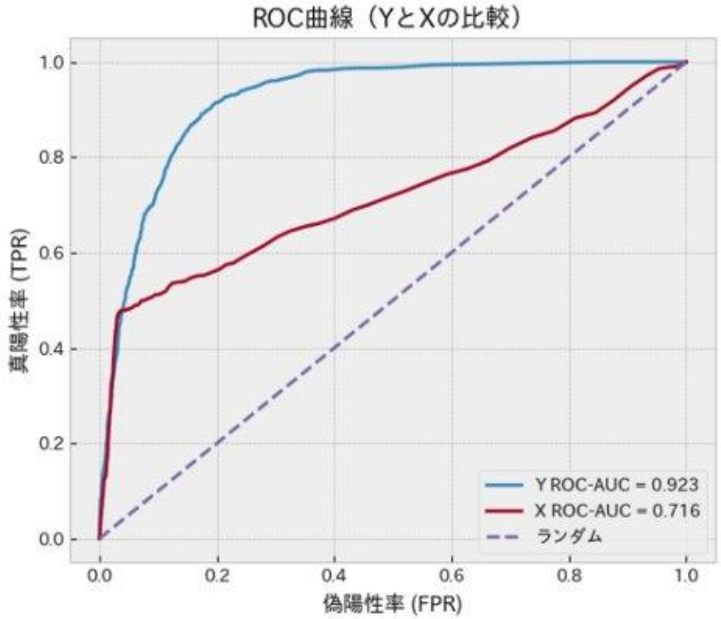
予測モデルの構築方法と性能評価

- 商材別の成功確率を予測するモデルを、定期預金（Y)とNISA口座（X)のそれぞれで構築
- 学習モデル：RandomForest + CalibratedClassifierCV（sigmoid, prefit）
- データ分割：Train 60% / Validation 20% / Test 20%(RANDOM_STATE=42)
  - Test：5426件での性能を確認
- 学習済みモデルから顧客ごとに「定期預金（Y）」と「NISA（X）」の成功確率を推定し、
  それぞれの確率を比較して商材を選択しました。具体的な判定ルールは次の通りです。
  - **P(Yの成功確率) ≥ P(Xの成功確率)** の場合：定期預金を推奨（Y）
  - **P(Xの成功確率) > P(Yの成功確率)** の場合：NISAを推奨（X）
  - 同点の場合は、優先順位ルールにより「定期預金（Y）」を選択
- ベースライン比較では以下の3パターンを用いて、提案手法（Policy）の効果を評価しています。
  - **Always Y**：常に定期預金を配信した場合
  - **Always X**：常にNISAを配信した場合
  - **Random（0.5）**：YとXを半々でランダムに配信した場合

予測モデルの性能評価

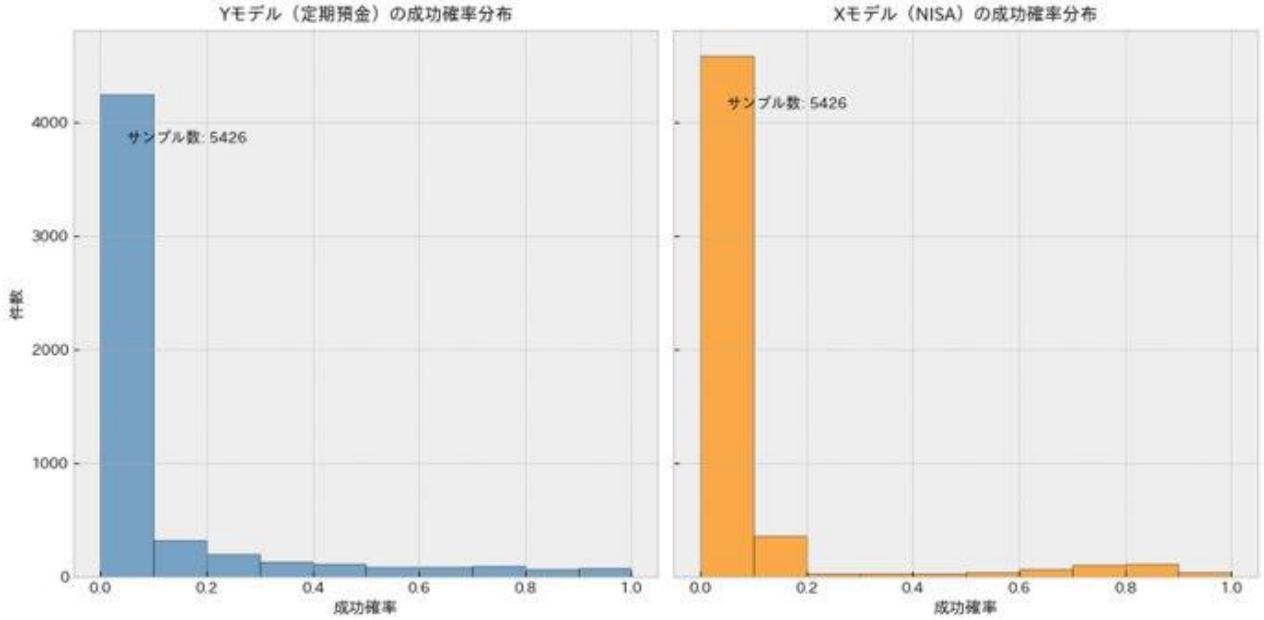
- 定期預金モデルは高精度・安定
- NISAモデルは中精度だが有用

評価指標	定期預金(Y)	NISA(X)
ROC-AUC	0.92	0.72
PR-AUC	0.59	0.42
Brier Score	0.07	0.08



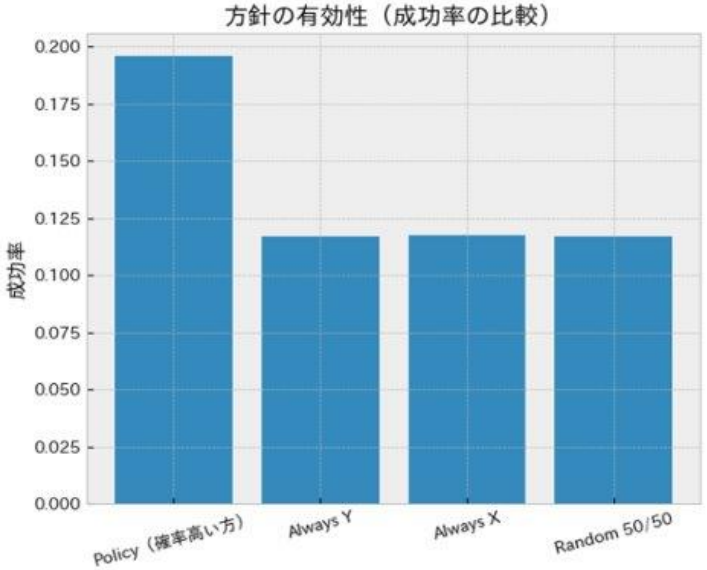
提案方針決定ロジック

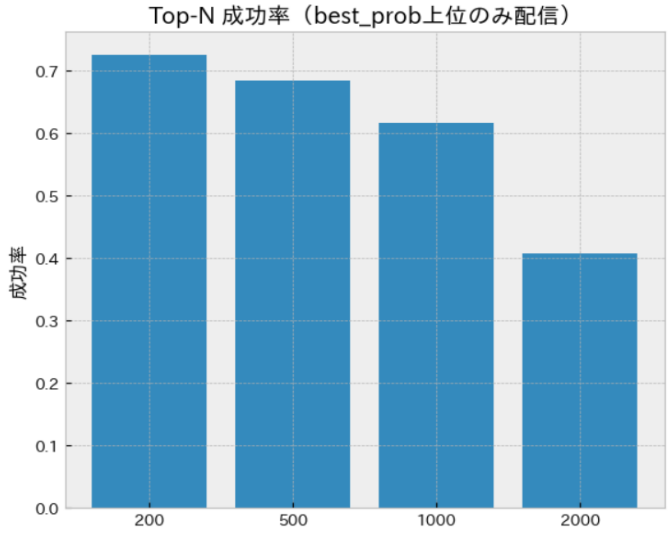
- 各顧客にY成功確率とX成功確率を算出し、より成功率が高い商材を提案する
- 確率分布を確認すると、大半は僅差であるが、明確なY/X優先層も存在する



提案商材最適化による成果

- 成功確率が高い商材を選択した場合の成功確率は**19.6%**。常にX/Yを提案した場合の成功確率と比較し、**成功確率は＋7.9pt（約1.7倍）の改善**
- 特に事前に試算した成功率が高いほど、実際に成功する確率が高い傾向にあり（Top1000顧客成功率：**61.6%**）





まとめパート

まとめ

本分析では、SIGNATE公開データを利用（一部改変）して、「定額預金」「NISA口座」の開設に寄与している要因を見つけ出し、マーケティング施策であるキャンペーンメールを実施する顧客の選定に役立てることを目的に分析を行った。

データ分析の結論として以下の実施を提案する。

- **定額預金：過去の顧客との接点履歴から優先するほうが良い**
  - キャンペーン実績に成功している
  - 過去の接点において通話時間が長い
- **NISA口座：下記の属性を優先してメール配信するほうが良い**
  - 予算残高が多い方
  - 職種が管理職や技術者、起業家
  - 前回の定額預金向けキャンペーンに成功していない
- **機械学習による予測モデルに基づく最適化を図ることにより、キャンペーンの成功確率を改善する可能性がある（試算：約1.7倍）**

反省・今後の展望

＜反省＞

- データの制約
  - 属性情報＋目的変数2つを持つデータセットを用意できず、ダミーデータを用いた
  - ダミーデータの付与により、一定の恣意性が残った
- 商材別の成功確率の改善余地は大きい
  - 特徴量拡充（残高、年収、行動履歴等）
  - 他モデルの検討

＜今後の展望＞

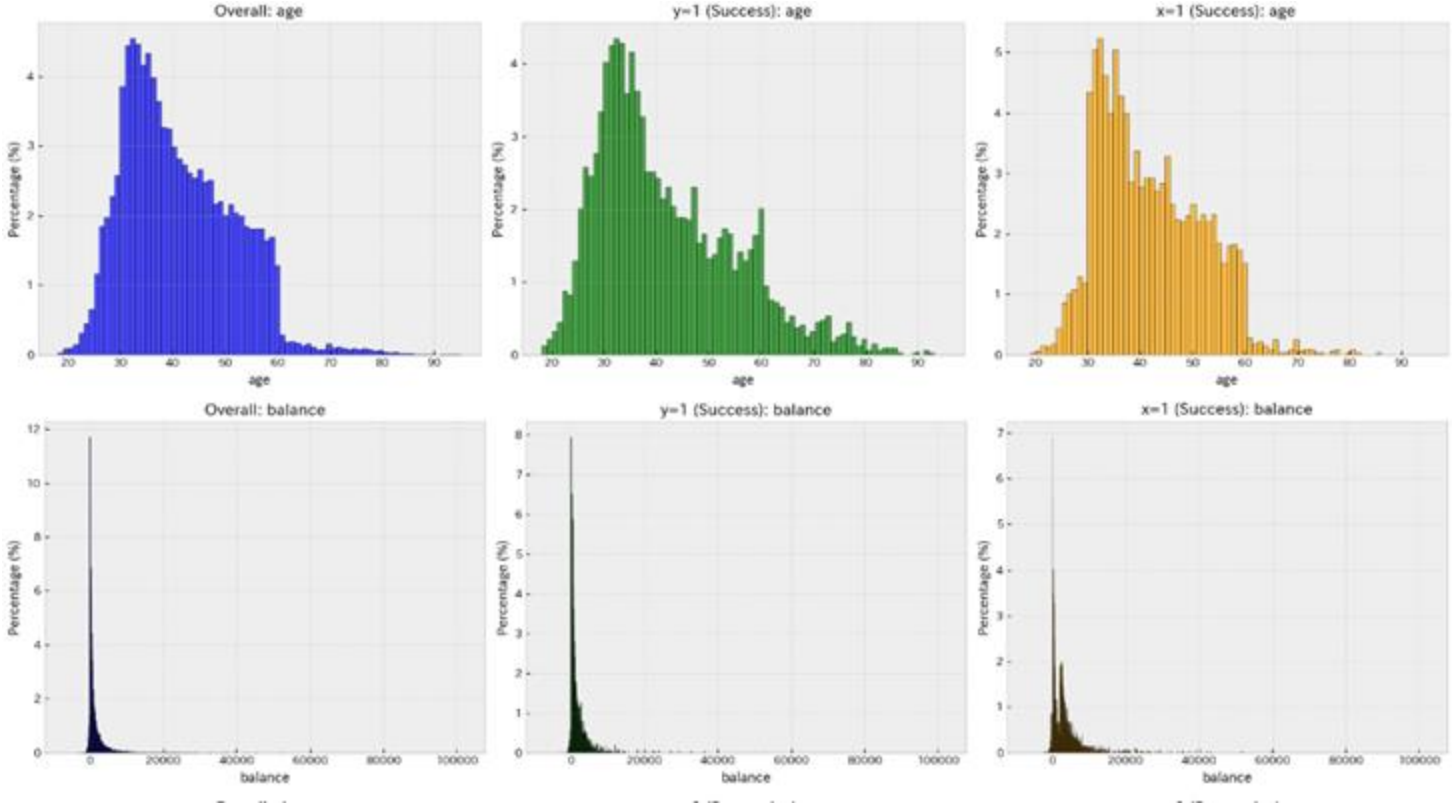
- 複数モデルによる成功確率の試算
- 裏付けの補強（統計的検定による補足：2群間の比較）
- 実データへの応用
- SHAP導入による説明性強化

付録

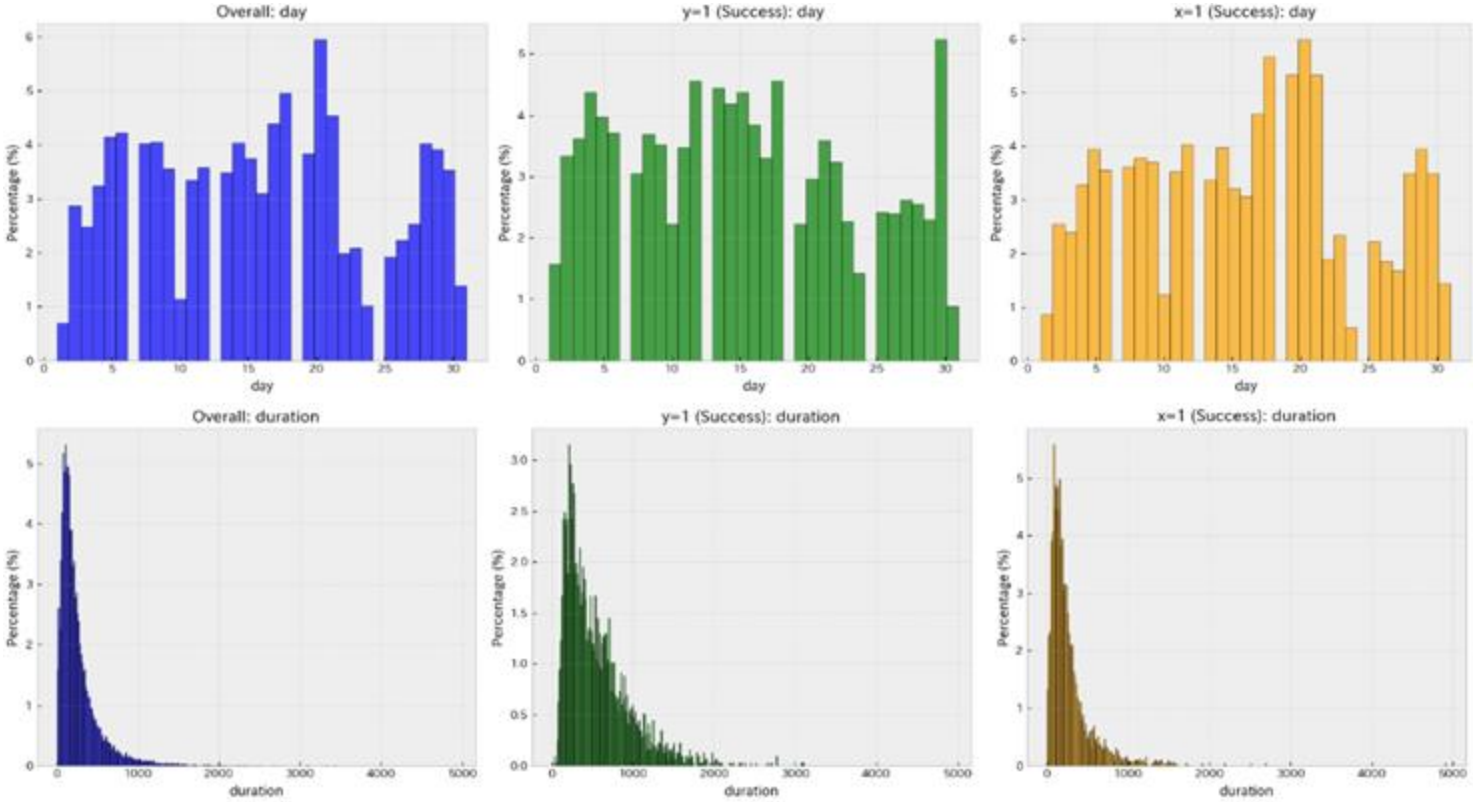
ダミーデータを付与したPythonコード

```python
import numpy as np

# 初期設定
np.random.seed(42)  # 再現性のための乱数シード

# 条件を満たす場合でも成功確率を70%に設定
success_probability = 0.7

# ダミーデータの作成（条件に基づき70%の確率で成功）
df['x'] = np.where(
    (df['age'] >= 30) & (df['balance'] >= 2000) & (df['job'].isin(['management', 'entrepreneur', 'technician'])),
    np.random.choice([1, 0], size=len(df), p=[success_probability, 1 - success_probability]),
    0
)

# y の成功件数に合わせて成功件数を調整
y_success_count = df['y'].sum()
current_x_success_count = df['x'].sum()

if current_x_success_count > 0 and y_success_count > current_x_success_count:
    additional_success_count = y_success_count - current_x_success_count

    # x が 0 のインデックスを取得
    zero_indices = df[df['x'] == 0].index

    # 追加で成功とするインデックスをランダムに選択
    selected_indices = np.random.choice(zero_indices, size=additional_success_count, replace=False)

    # 選択されたインデックスを 1 に変更
    df.loc[selected_indices, 'x'] = 1

# 確認用
df['x'].value_counts()
```

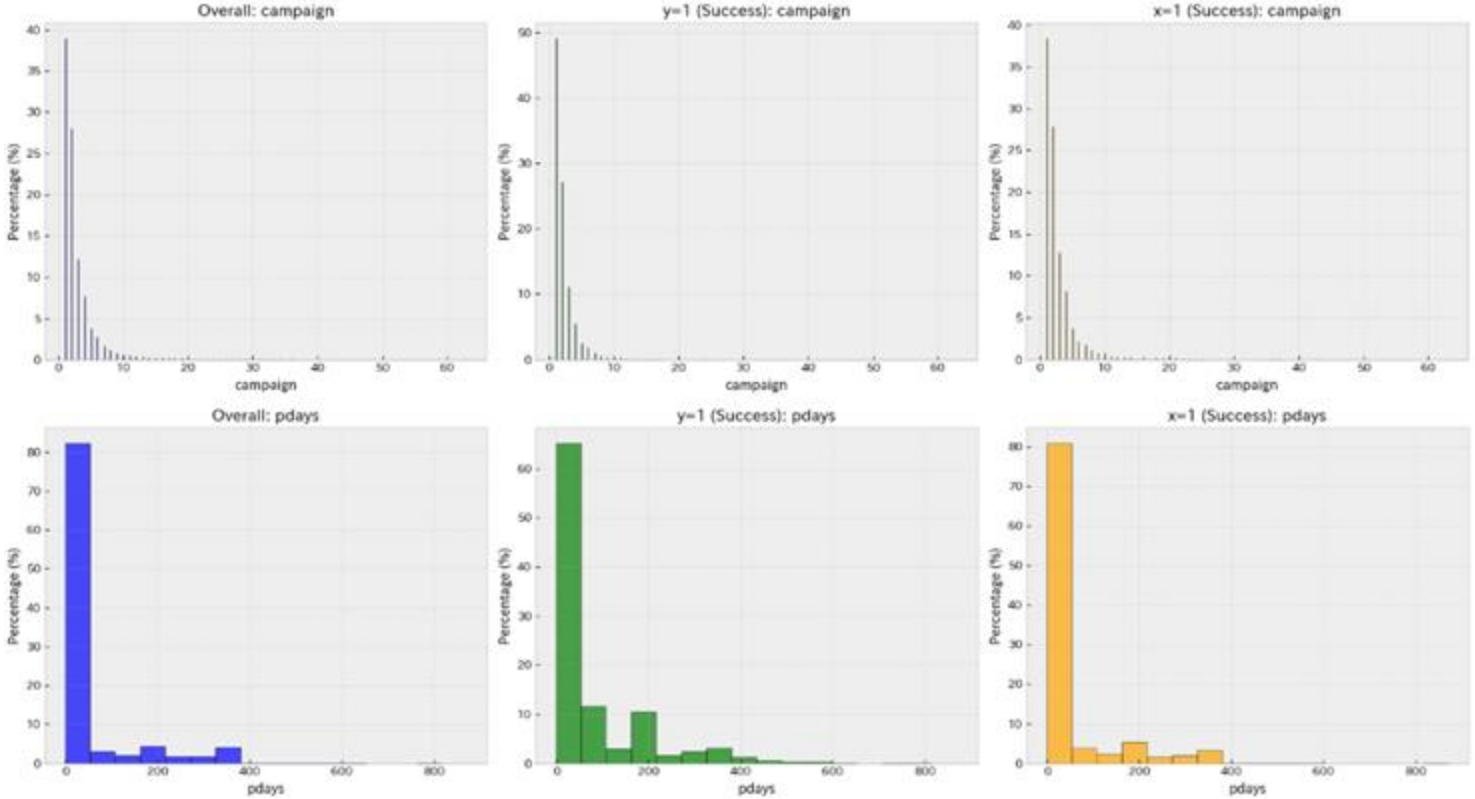
代表値の比較：age、balance



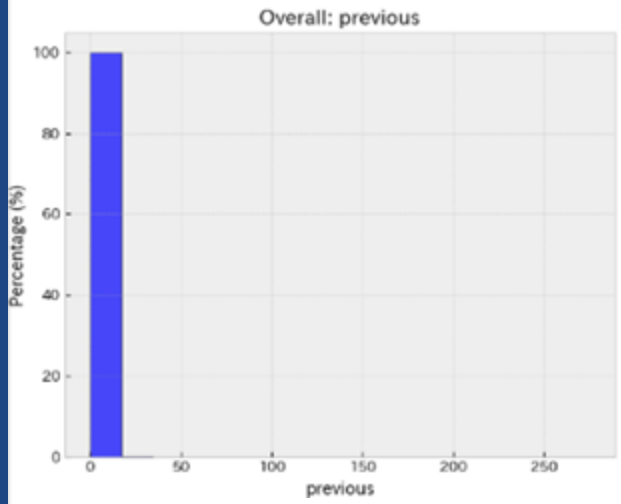
代表値の比較：day、duration

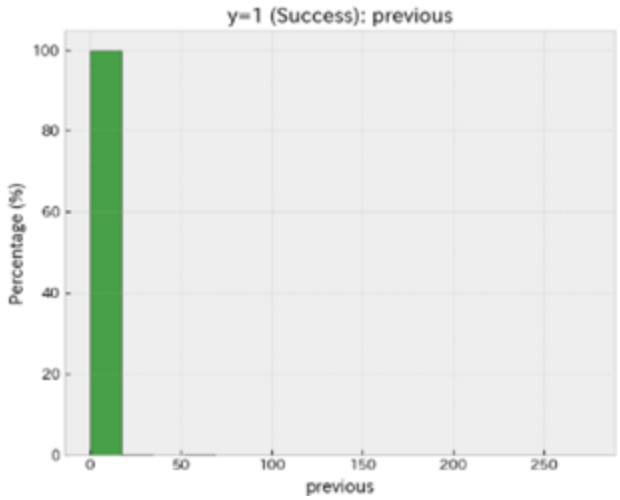


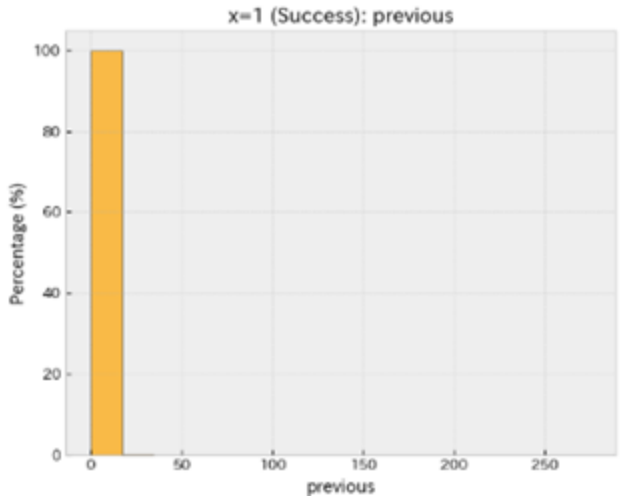
代表値の比較：campaign、pdays



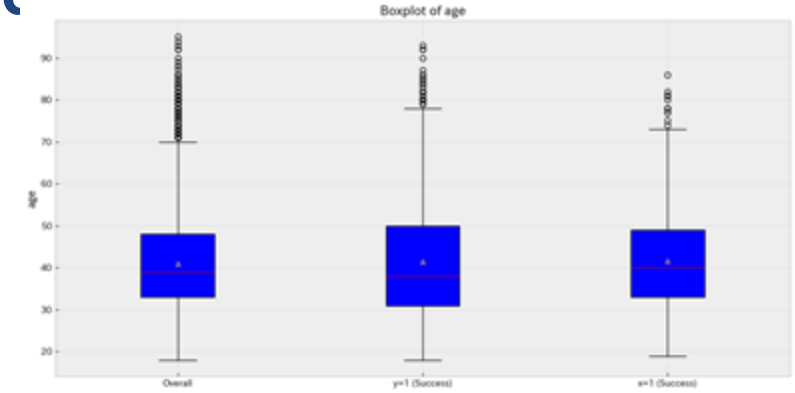
代表値の比較：previous

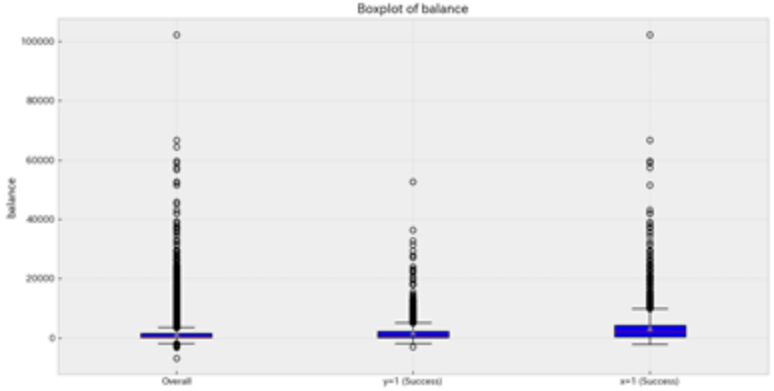


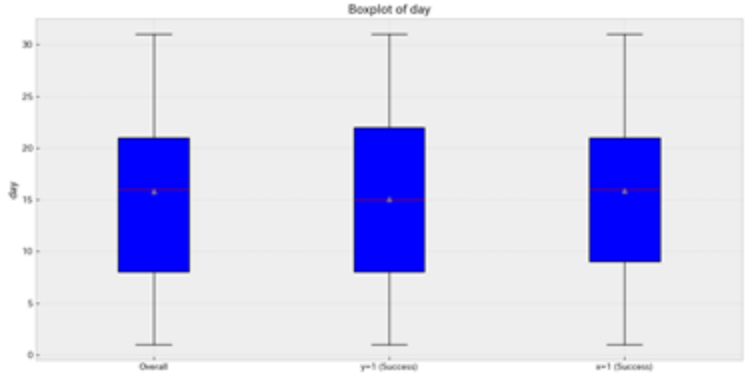


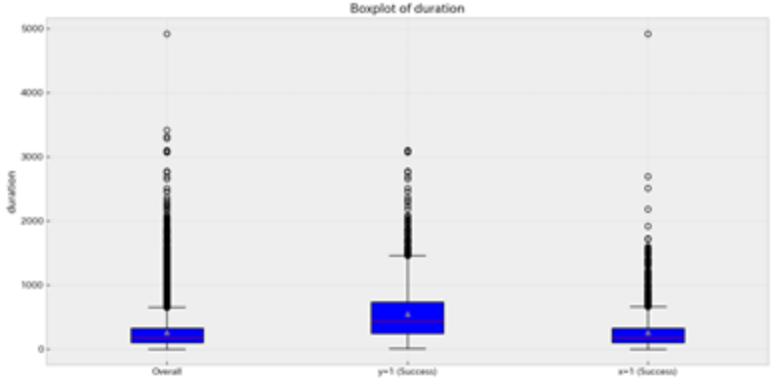


代表値の比較(箱ひげ図）：age、balance、day、duration

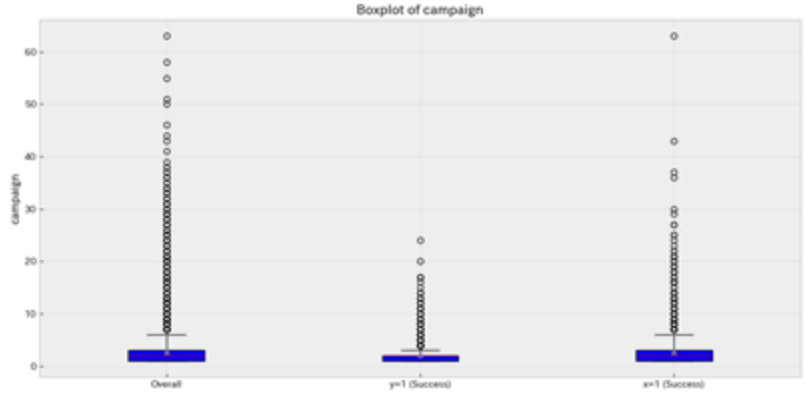


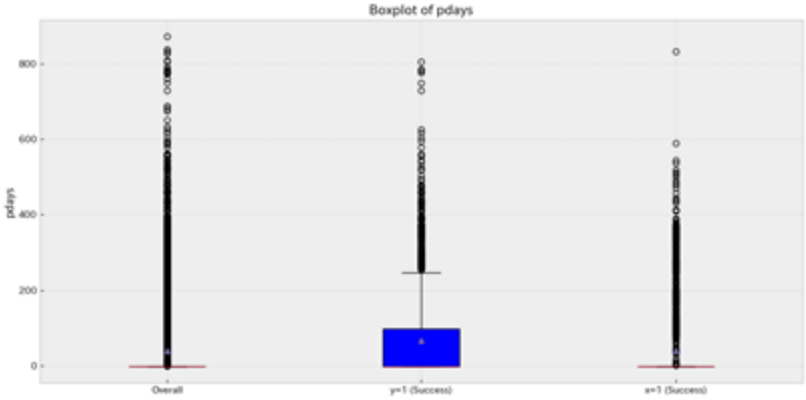


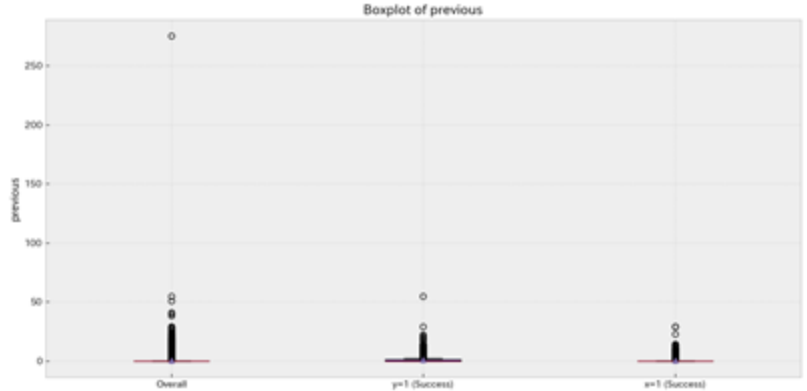




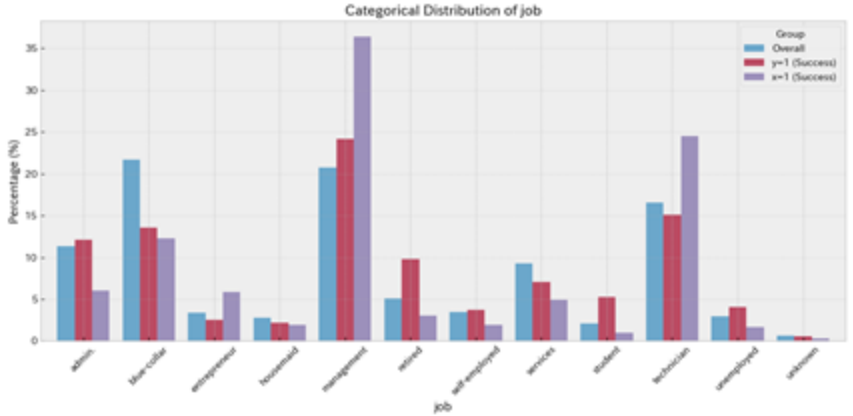
代表値の比較(箱ひげ図）：campaign、pdays、previous

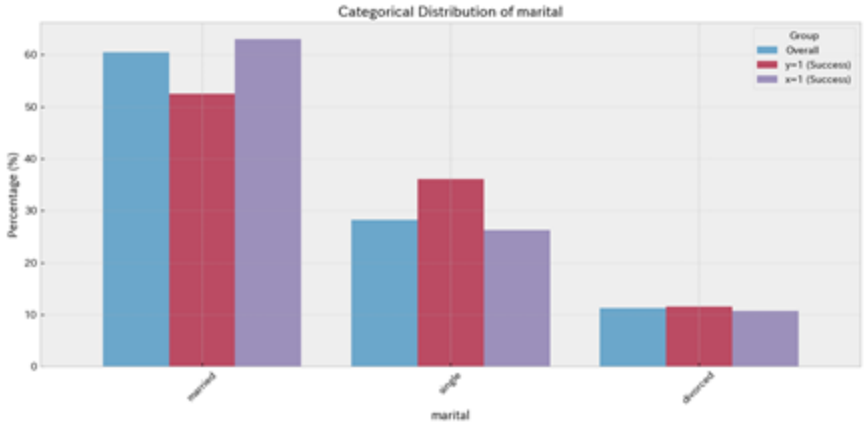


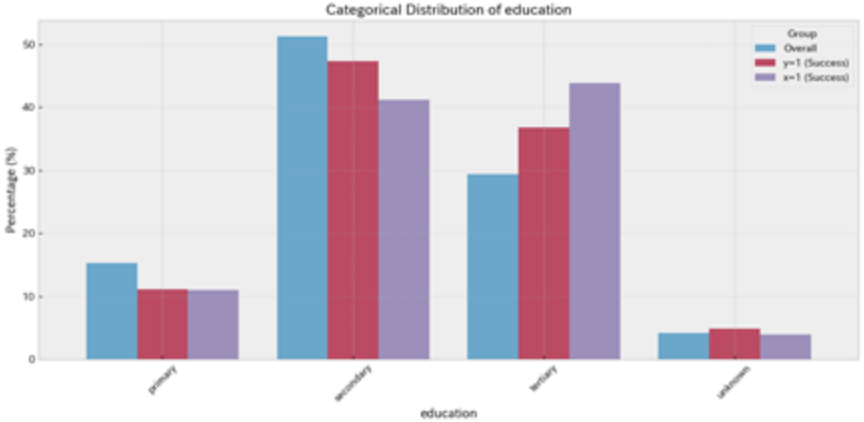


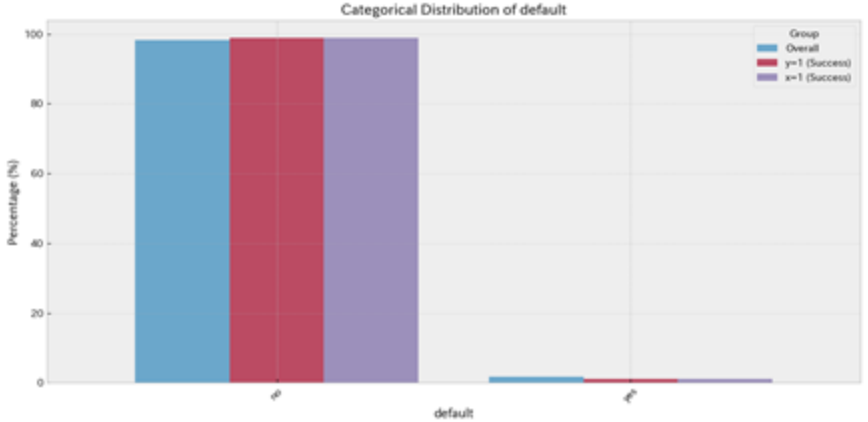


代表値の比較：jobs、marital、education、default

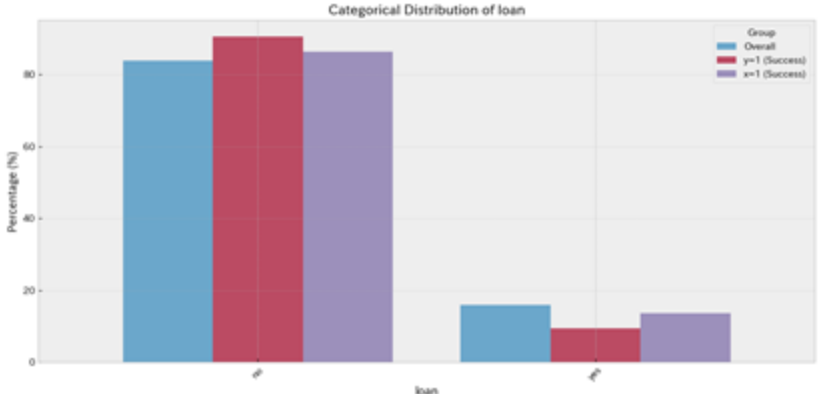


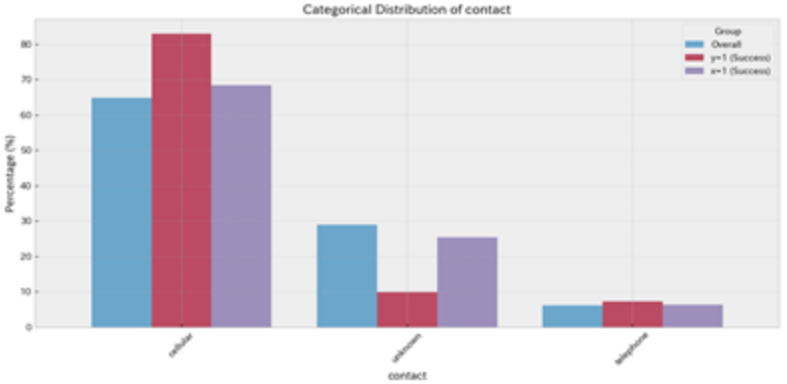


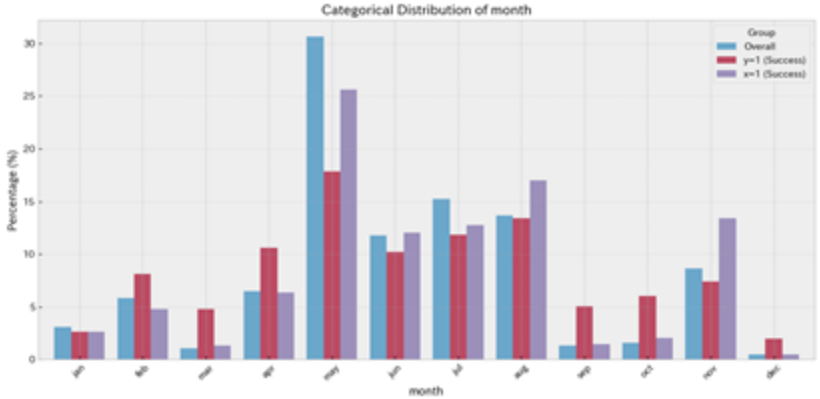




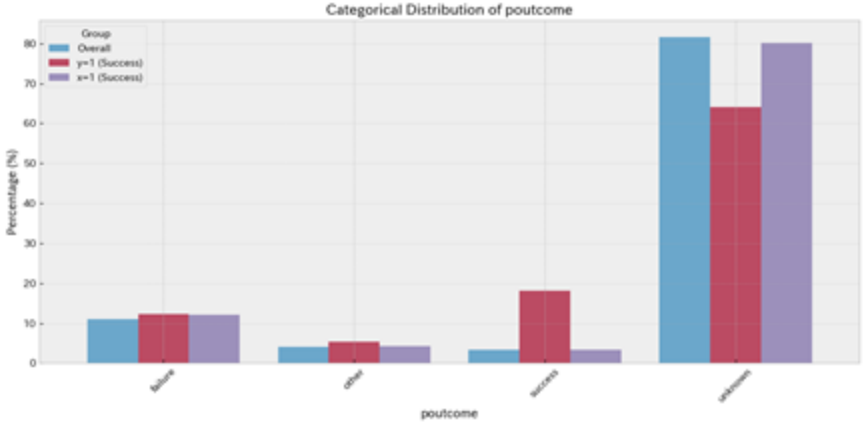
代表値の比較：housing、loan、contact、month



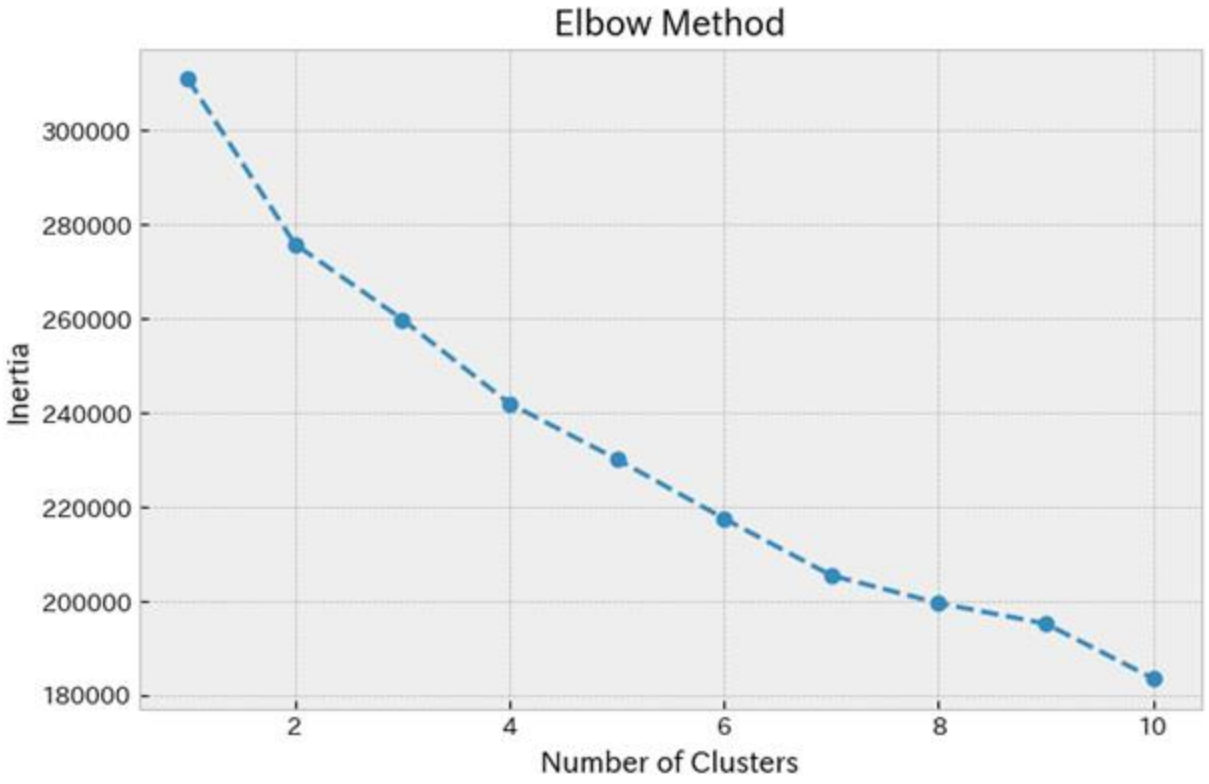




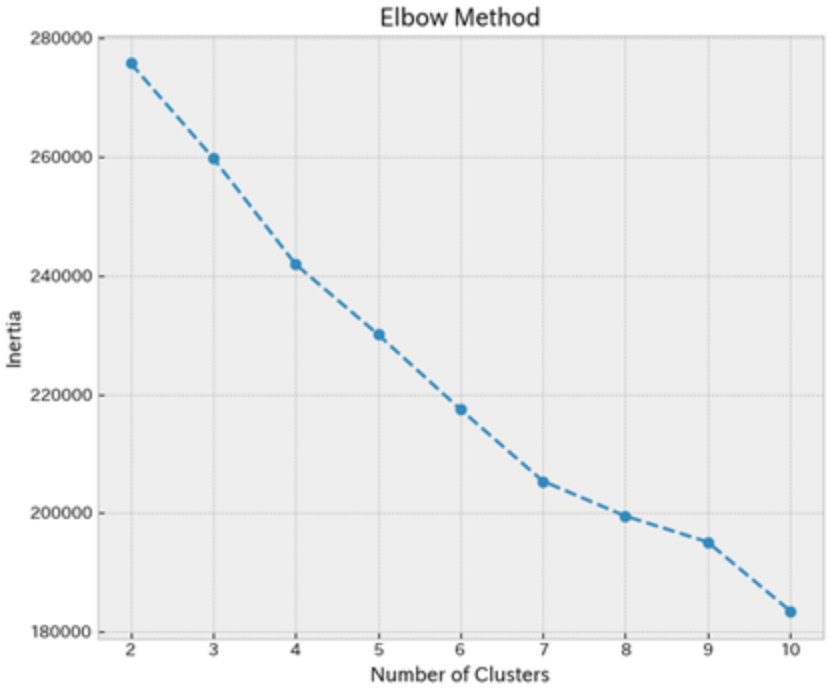
代表値の比較：postcome

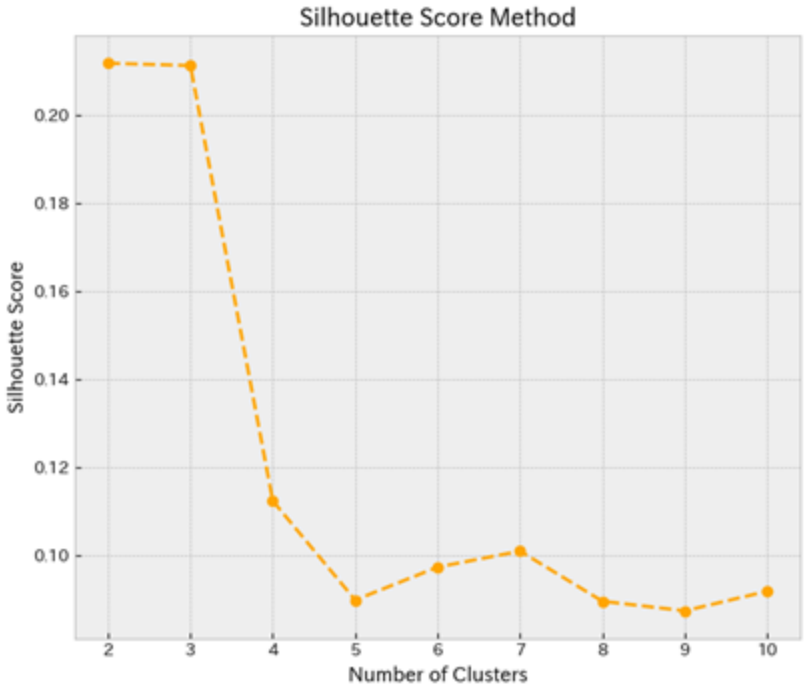


エルボー法による最適なクラスター数の検討

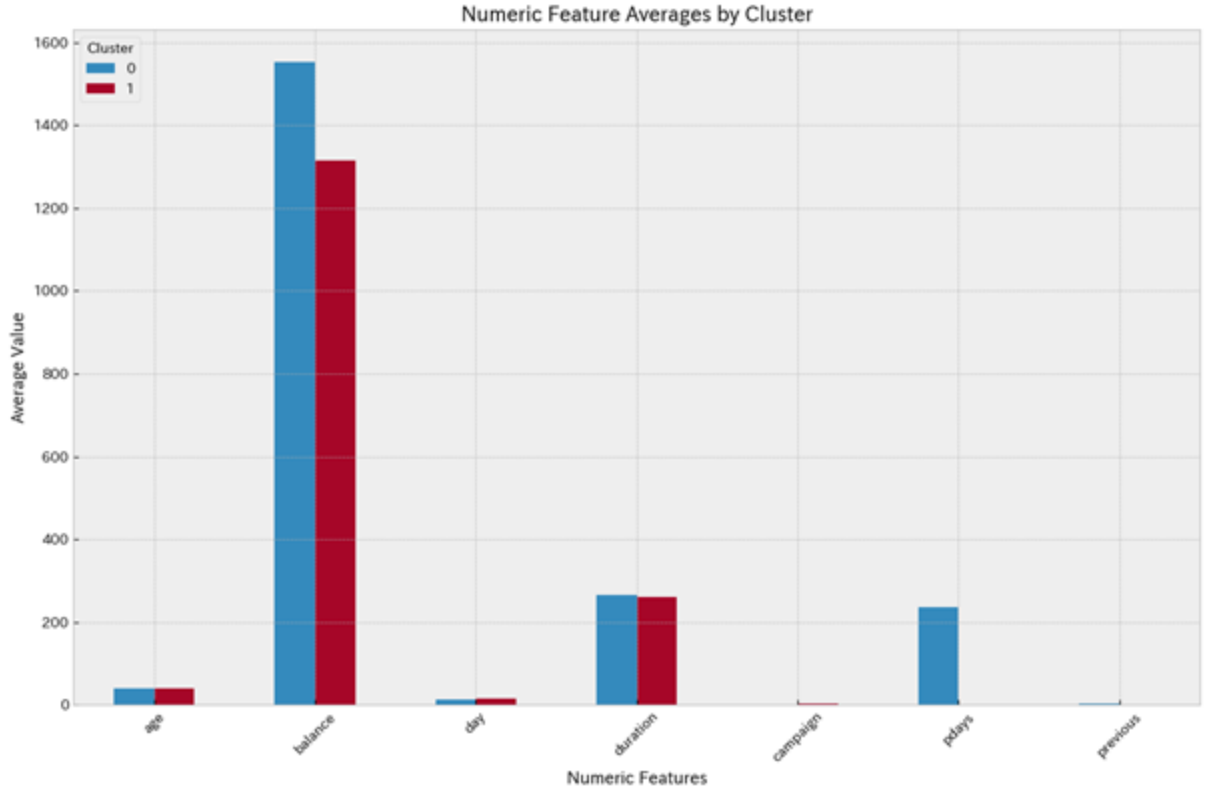


エルボー法による最適なクラスター数の検討

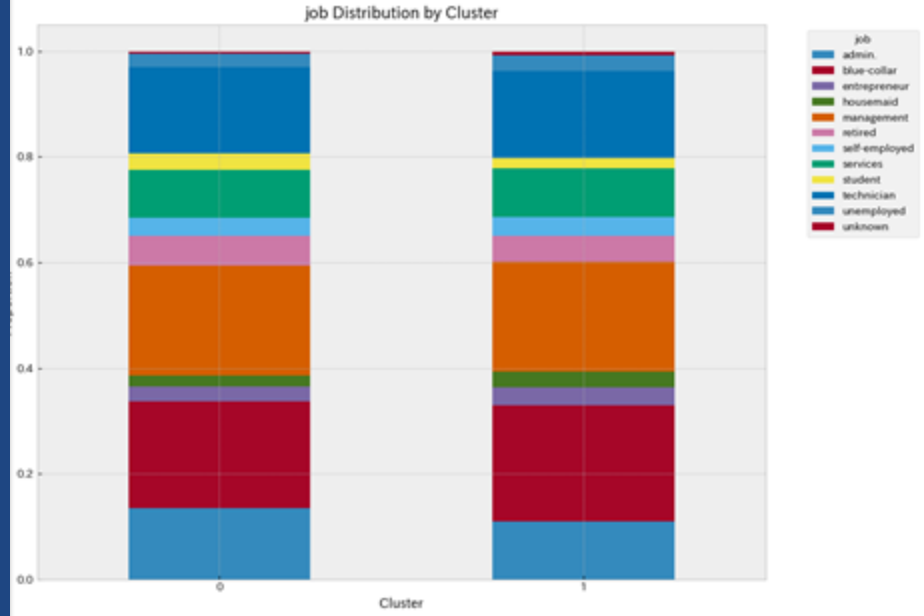


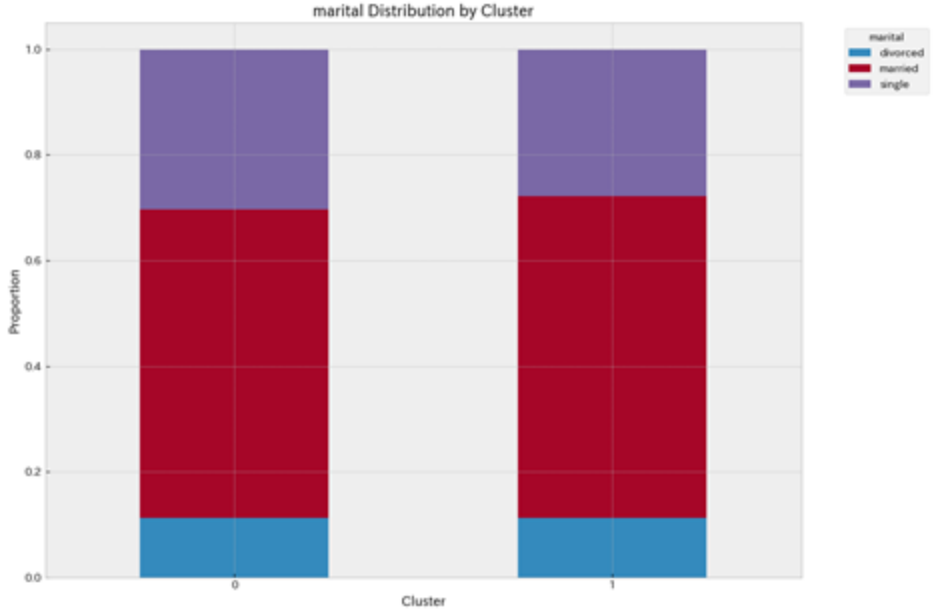


クラスター別の属性情報の分布：数値型変数

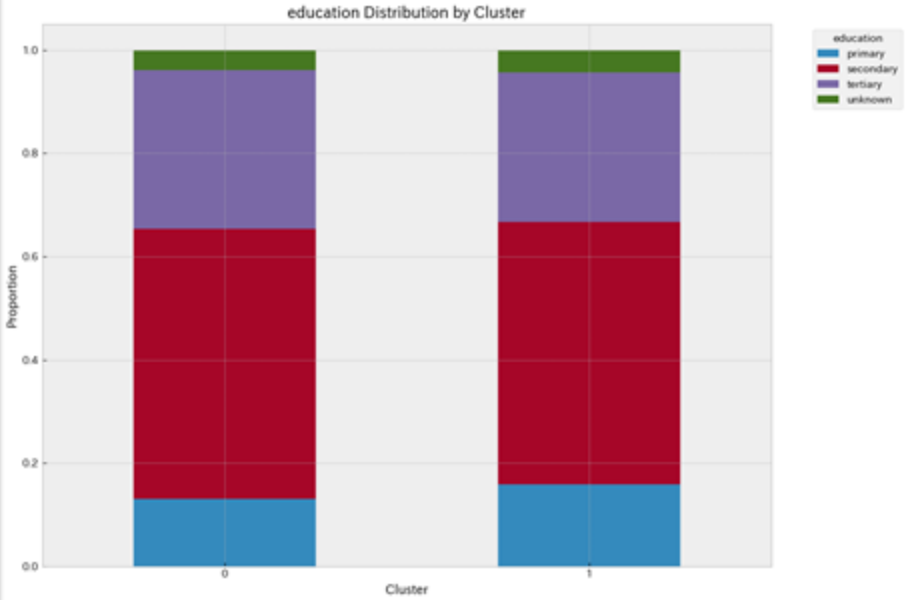


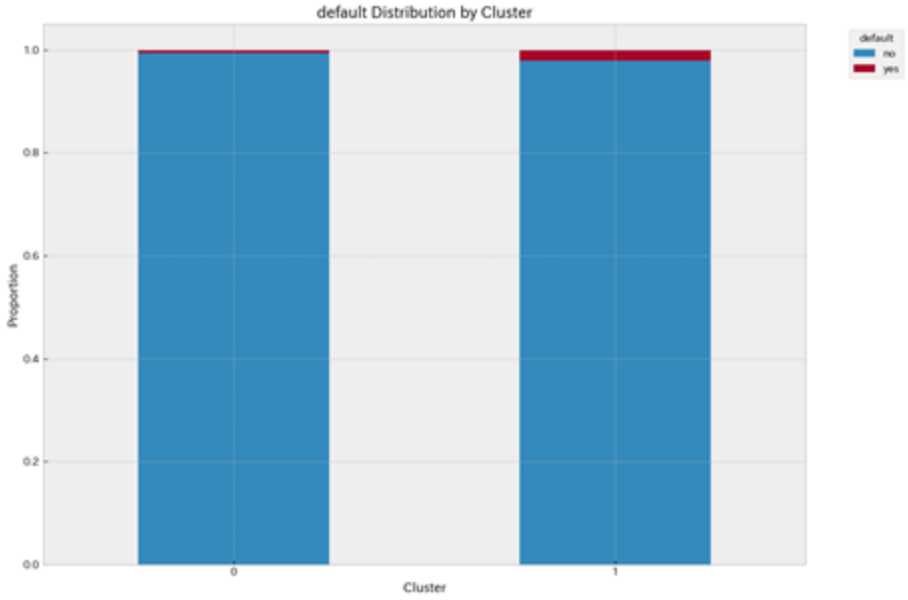
クラスター別の属性情報の分布：job、marital



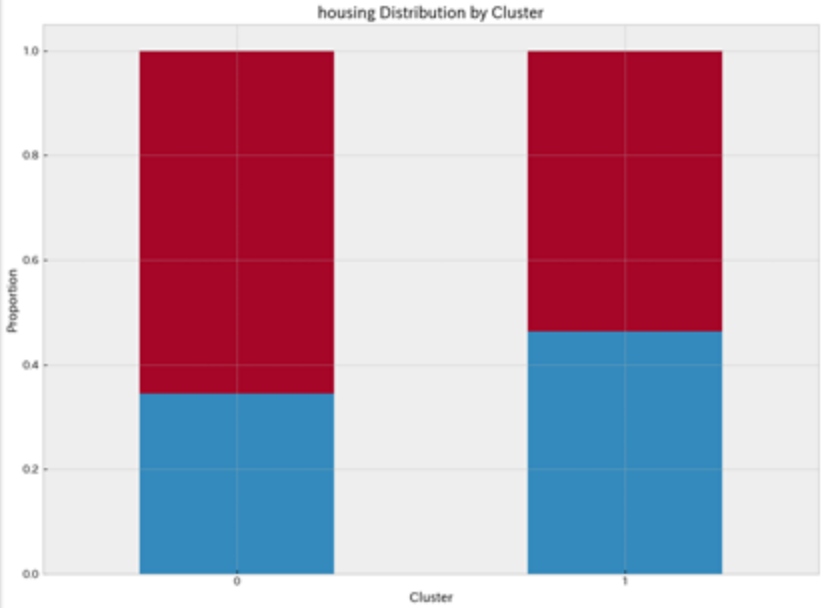


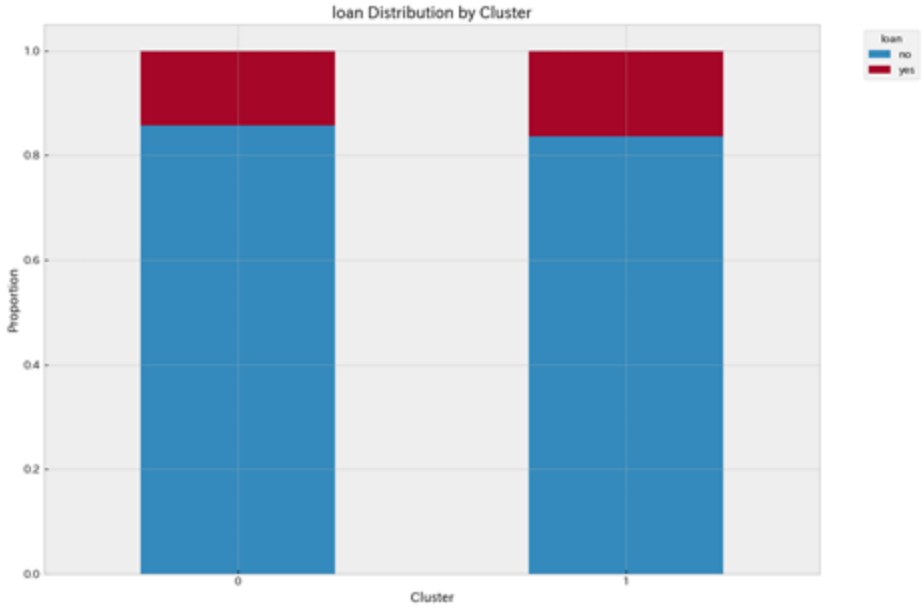
クラスター別の属性情報の分布：education、default



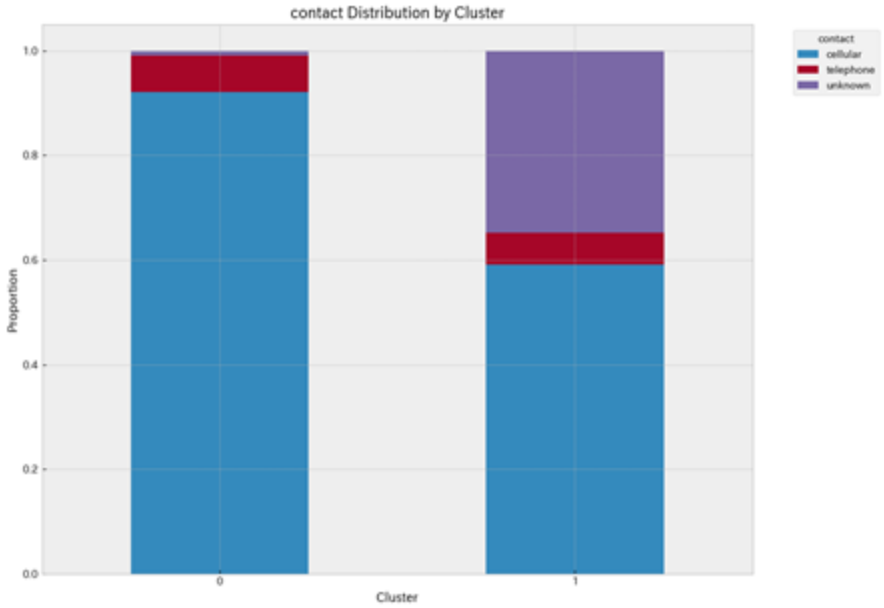


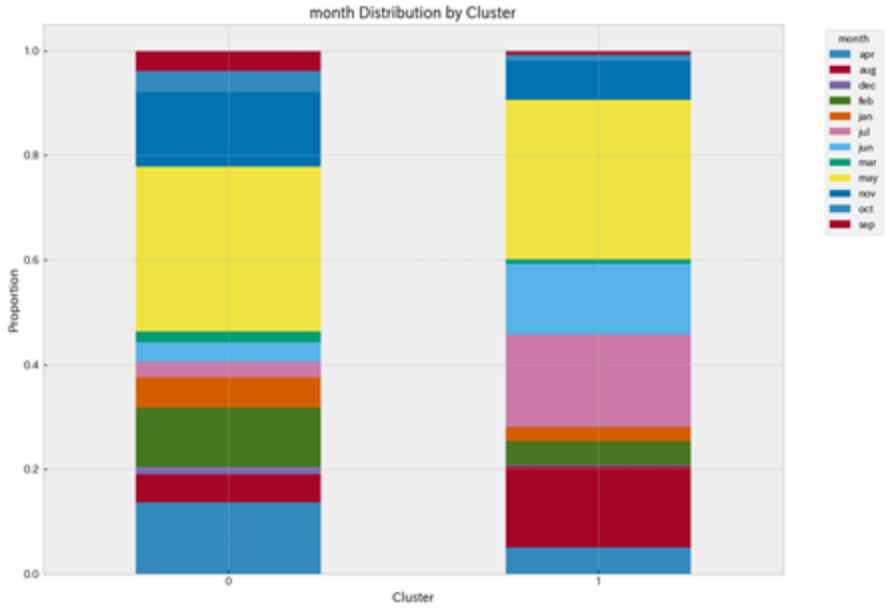
クラスター別の属性情報の分布：housing、loan





クラスター別の属性情報の分布：contact、month





クラスター別の属性情報の分布：postcome