銀行顧客の定期預金予測モデル

学部学科：経済学部経済学科

学年・組：2年9組

学籍番号：22106557

名前：河本薫子

提出日：2023/3/13

発表日：2023/3/16

目次

予測モデル制作の目的・手順

目的

低コストで潜在顧客の属性を分析
定期預金の成約率 UP　につなげる

モデルの要件定義

◆ 高精度

正解率　⇒　高く

AUC　⇒　高く

※誤分類・潜在顧客の見逃しを防ぐ
※正解率：予測の的中割合／AUC：ROC曲線の曲線下面積

◆ 低い計算コスト

特徴量　⇒　減らす

※重要度の高い特徴量だけを抽出
※処理コストを減らし、スムーズな稼働を実現

制作手順

1 データ前処理

データセットの尺度を揃える
・カテゴリデータの one-hotエンコーディング
・標準化

2 分類器の実装・性能評価

7種類の分類器の性能を比較
・LogisticRegression　・DecisionTree　・KNeighbor
・SVC　・RandomForest　・AdaBoost　・GradientBoost

3 工夫要素の追加

① オーバーサンプリング
② K分割交差検証
③ パラメータチューニング
④ 特徴量抽出
⑤ 多数決分類器

検証の流れ

検証過程

表1

	評価・検証A	評価・検証B	評価・検証C	評価・検証D
平均正解率	89%	80%	91%	92%
平均再現率	38%	84%	92%	95%
平均AUC	67%	83%	94%	98%
チューニング実装時間			3～12時間	1～2時間
課題	クラスの不均衡分布	分類器の性能を最適化できていない	チューニングの計算コストが大きい	多数決分類器ではチューニングができない
対策	オーバーサンプリング	**パラメータチューニング K分割交差検証**	**特徴量抽出 多数決分類器**	

検証過程1

実装・性能評価 A

分類器名	正解率	再現率
LogisticRegression	90%	35%
DecisionTree	88%	47%
Kneighbor	89%	31%
SVC	90%	31%
RandomForest	90%	37%
AdaBoost	88%	47%
GradientBoost	91%	41%
平均値	89%	38%

表1

問題

正解率/AUCが機能しない

原因

クラスが不均衡

対策Ⅰ：オーバーサンプリング

図1

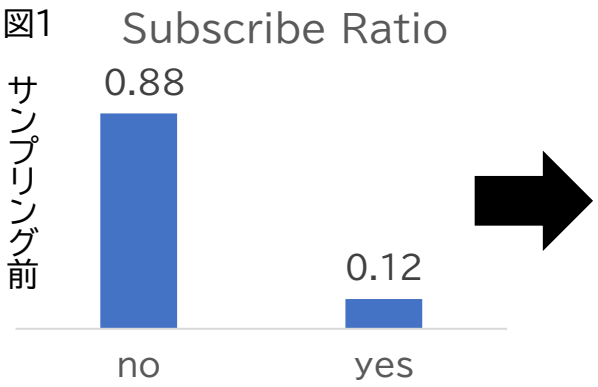
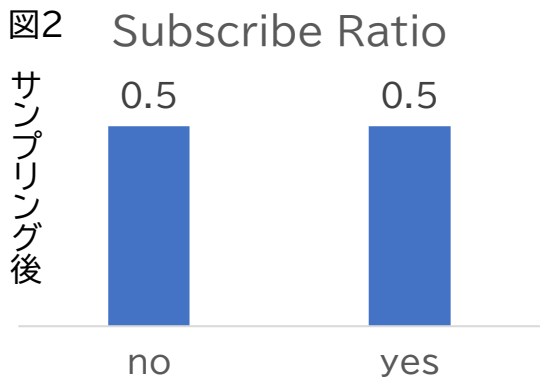


図2

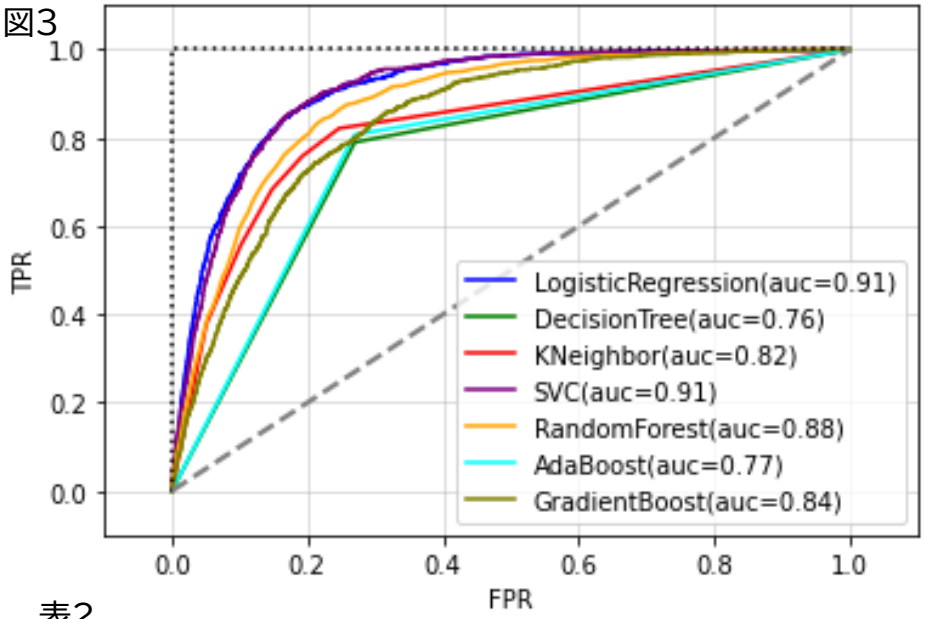


クラスの不均衡を解消

預金者のサンプル数を増やし、2クラスのサンプル数をそろえる

実装・性能評価 B　～対策Ⅰ実施後～

図3



←ROC曲線

表2

分類器名(res)	正解率	AUC
LogisticRegression	85%	91%
DecisionTree	89%	76%
Kneighbor	95%	82%
SVC	91%	91%
RandomForest	91%	88%
AdaBoost	74%	77%
GradientBoost	35%	84%
平均値	80%	84%

問題

正解率の低下

検証過程2

実装・性能評価 B　～課題と対策～

原因

・分類器のパラメータが最適化されていない
・汎化性能を正確に測定できない

対策Ⅱ：

・パラメータチューニング

分類器のパフォーマンスを最大化する組み合わせを見つける

分類器のパラメータの組み合わせを変えて検証

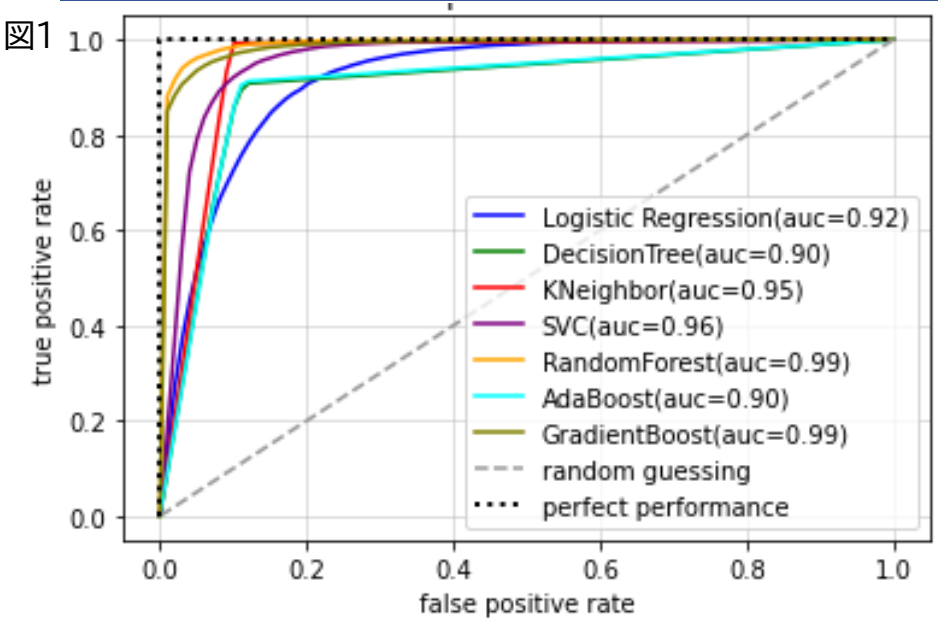
・K分割層化検証（K=5）

モデルの性能をより正確に評価できる

データセットを複数に分割し、テストに使うセットを変えながら検証を繰り返す
計算コスト抑制のため K=5 に設定

実装・性能評価 C　～対策Ⅰ・Ⅱ実施後～

図1



←ROC曲線

表1

分類器名(gs,kf)	正解率	AUC	チューニング計算コスト
LogisticRegression	85%	92%	2～5分
DecisionTree	90%	90%	5～10分
Kneighbor	95%	95%	5～10分
SVC	91%	96%	600分～
RandomForest	95%	99%	110～120分
AdaBoost	90%	90%	10～20分
GradientBoost	93%	99%	120～180分
平均値	**91%**	**94%**	**123分～**

検証過程3

実装･性能評価 C　～課題と対策～

問題

アルゴリズム実行に時間がかかる

原因

計算コスト要因の存在

表1

計算コスト要因	数を減らせるか
サンプル数	×
特徴量の種類	**〇**
K分割交差検証	△
パラメータ候補	×

→特徴量選択後の、モデルの性能低下に対処する必要

対策Ⅲ：

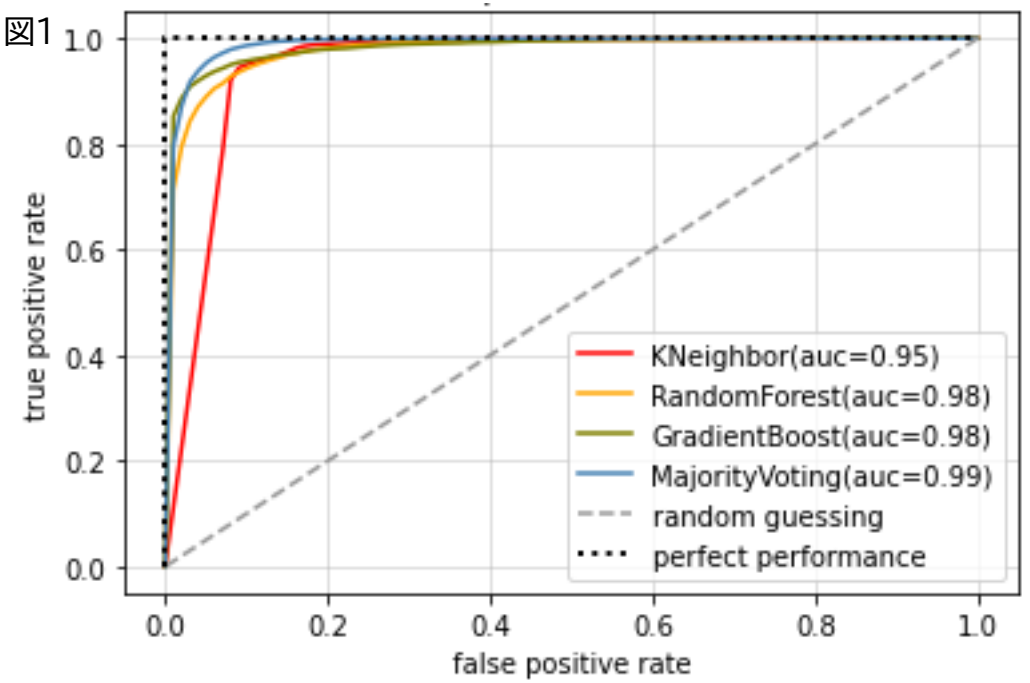
・重要度を基に特徴量抽出→**計算コストを軽減**

・多数決分類器の実装→**誤分類のリスク最小化**

多数決分類器：
分類器の過半数が予測するクラスを採択。誤分類があっても、その影響を抑える。
低性能分類器の影響を受けないよう、スコア上位のもののみを選ぶ。

実装･性能評価 D　～対策Ⅰ・Ⅱ・Ⅲ実施後～

図1



ROC曲線

表2

分類器名	正解率	AUC	チューニングの計算コスト
Kneighbor	90%	95%	3～4分
RandomForest	92%	98%	90分
GradientBoost	91%	98%	70分
Majority Voting	**94%**	**99%**	チューニング不可

特徴量抽出-予測・結果

抽出される特徴量の予測

・預金者と非預金者の間で数値に大差がある特徴量

例:預金者の方が年収が高い

・カテゴリ変数以外の特徴量

カテゴリ変数は名義特徴量。ダミー変数化の際にモデルへの影響力が落ちる可能性

表1

	yes	no
age	41.672140	40.862165
balance	1785.768237	1307.779822
day	15.158816	15.925462
pdays	69.066218	35.653802
duration	532.955585	221.559108
previous	1.159354	0.496456
campaign	2.162853	2.845876

・balance
顧客の年平均口座残高
・duration
銀行員と顧客の通話時間
・previous
過去に銀行と顧客が連絡をとった回数

・定期預金する顧客は預金残高が高め
・銀行と連絡をよく取り合う
→直観に紐付いた予測

抽出結果　～重要度の計算～

表2

DecisionTree	RandomForest	AdaBoost	GradientBoost	best_feat	feat_label
0.271485	0.278747	0.213523	0.430118	0.298468	duration
0.086554	0.083890	0.199720	0.086445	0.114152	campaign
0.047579	0.065666	0.069613	0.077058	0.064979	housing_yes
0.080556	0.069939	0.066806	0.036651	0.063488	balance
0.066826	0.060905	0.063855	0.039003	0.057647	day
0.074405	0.057801	0.065499	0.031779	0.057371	age
0.027398	0.040871	0.039384	0.067068	0.043680	contact_unknown

抽出特徴量　計6つ
・duration　・campaign
・balance　・day
・housing_yes　・age

・campaign
定期預金について
顧客と連絡をとった回数
・day
直近に連絡を取ってからの
経過日数
・housing_yes
顧客の住宅ローン

・durationの重要度 が最も高く圧倒的
・ 抽出後のモデルの**性能は高い**
→予測とは異なる特徴量もあるが
妥当な特徴量を選択できている

特徴量抽出-考察

抽出特徴量と預金者の関係

図1

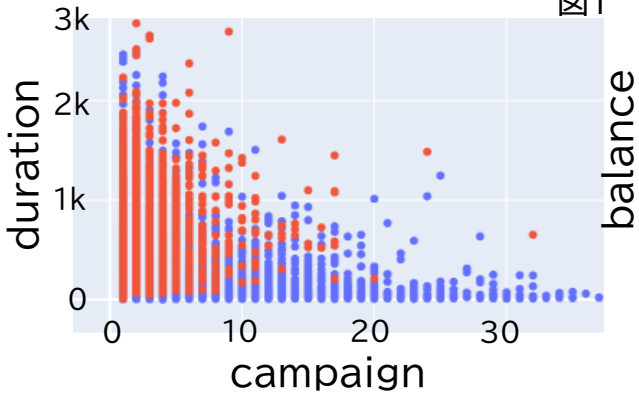
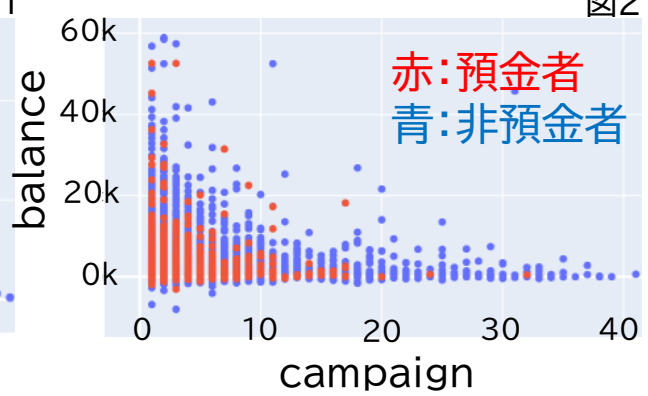


図2

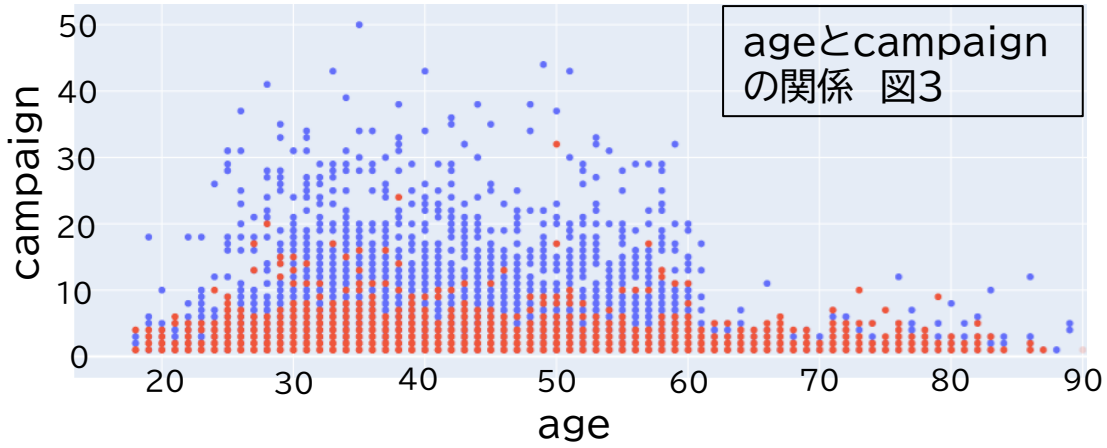


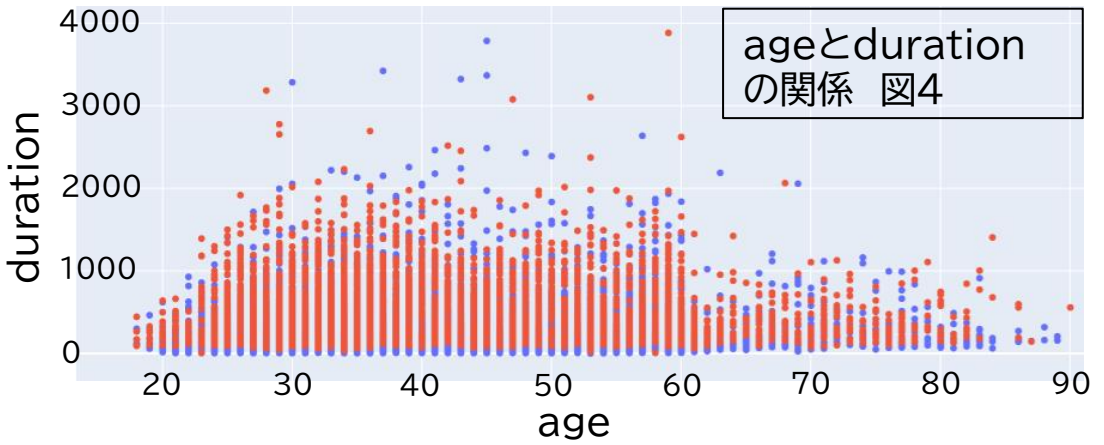
・campaign と duration の関係(左)
campaign は短く、duration は長い方が
成約率は高い

・campaign と balance の関係(右)
balanceの多寡に依らず、campaignが短い方が
成約率は高い

銀行と頻繁に・長く連絡を取っている顧客ほど、
定期預金の成約率は高い(預金残高は関係無い)

赤:預金者　青:非預金者





・60代以上の顧客とは頻繁に連絡を取っていない
・60代以上も20～50代同様、成約が取れている
→60代以上にも中年層同様頻繁に連絡を取るべき

結論

より詳細な分析の提案

景気による成約率変動の可能性

消費者物価指数、雇用率などの景気指数導入
→**成約集中時期に注目し、成約営業のタイミングを絞り込み、営業コストを抑えられる**

年収の成約率への影響

顧客の平均年収などの導入
→**潜在顧客層を事前に絞り込める（年収が極端に低ければ預金しない、など）**

年収データによる順序特徴量の導入

職業別年収で職種をランク付け
→データ前処理の段階で名義特徴量を順序特徴量として扱える
→**特徴量を増やさず、計算コストを減らせる**

実務への導入を見据えて

・実務では分析手法はビジネス要件によって異なる

今回

・多数決分類器などで計算コストが多少かさむ
・成約見込みの高い顧客を確実に拾いたい
→潜在顧客を効率的に見つけ営業コストを抑える

再現率を上げる場合

・見込みのない顧客を誤分類してしまうリスク
・少しでも見込みある潜在顧客は1人残らず見つけたい
→余分な営業コストを負ってでも成約数を増やす

・実務で発生する予算・人材の制約
・実務で求められる目標
に従って分析を変える必要

参考文献

・著者：Sebastian Raschka、Vahid Mirjalili
『[第3版] Python 機械学習プログラミング 達人データサイエンティストによる理論と実践』 株式会社インプレス 2020年10月 発行

・著者：Andreas C. Muller、Sarah Guido
『Pythonではじめる機械学習ーscikit-learnで学ぶ特徴量エンジニアリングと機械学習の基礎ー』 株式会社オライリー・ジャパン 2017年5月発行