



# AI・機械学習ハンズオン ～実践Kaggle 初級編～

AIや機械学習の技術をKaggleの実践を通して身につけてもらう  
ハンズオン型イベントです

2019年7月2日19時30分から

# 自己紹介

- 2018年6月財務省を退職
- kaggleへの挑戦をメインに生活している（その他研修講師なども実施）
- 2019年4月Kaggleマスターになる
  - ▶ 4/14開催の技術書典6で、kaggleのチュートリアル第3版を公開。
- 2019年6月地震コンペで3位獲得
- youtube始めました。kaggleのことを説明するための動画を公開しています

twitterより



カレー🍛 専業kaggler

@currypurin

2018年6月末に公務員を退職し専業kagglerになり、2019年4月kaggleマスターになる。今は年内にkaggleグランドマスターになることを目指して挑戦中。金メダルあと3つでグランドマスター🏆  
🏆🏆



# 自己紹介 2

Competitions Master

Rank

90

of 112,370

2

2

2

LANL Earthquake Prediction

4 hours ago · Top 1%

3<sup>rd</sup> of 4541

Santander Value Prediction...

9 months ago · Top 1%

8<sup>th</sup> of 4484

PLAsTiCC Astronomical Cl...

6 months ago · Top 2%

16<sup>th</sup> of 1094

🔄 自分がリツイート



カレー@kaggler @currypurin · 4月8日

#技術書典 の原稿入稿しました！

5章を全部書き直して、付録B、付録Gを追加しました。他の部分もKernelの仕様の変更とか、サイトの変更などにあわせて更新するなど手を入れています。

後日、コードや内容の公開とか解説動画の収録など行います。  
よろしくお願いします。

Pythonによる

【第3版】

kaggle のチュートリアル

こんなに楽しく機械学習を学べる  
ネットゲームがあったとは！

本編 付録

1.1 Kaggleとは	6
1.2 Kaggleへログイン	7
1.3 Kaggleトップページの説明	8
1.4 コンペ (Competitions) のページの概要	8
第2章 コンペのページの翻訳など	11
2.1 Overview (概要)	11
2.2 Data (データ)	17
2.3 Kernels (カーネル環境)	20
2.4 Discussion (ディスカッション)	23
2.5 Leaderboard (リーダーボード)	24
2.6 Rules (ルール)	25
2.7 Team (チーム)	27
2.8 My Submissions (自分のサブミッター)	27
2.9 Submit Predictions (予測のサブミット)	27
第3章 まずは、サブミットしてみる	28
3.1 データの作成・サブミット方法	28
3.2 カーネル環境のノートブックを用いる方法	28
3.3 スクリプトでサブミットする方法	32
3.4 ローカルPCで作成したデータをサブミットする方法	34
第2部	
第4章 タイタニックデータの概要	36
4.1 データの概要の確認	36
4.2 Pandas-profilingを用いて各特徴を個別に把握	44
4.3 各特徴とtargetとの関係を可視化	47
4.4 まとめ	65
第5章 LightGBMでのタイタニック	66
5.1 前処理	66
5.2 ホールドアウト法での学習・推論	67
B.2 カテゴリカル変数の指定	83
B.3 重要度の表示	84
B.3 Scikit-learn interfaceを使用しない方法	85
C Santander Value Prediction Challengeで金メダルを獲得しました	87
C.1 経緯	87
C.2 コンペの概要	88
C.3 最終順位	91
C.4 参考 HomeCreditコンペの結果	91
D HomeCreditコンペ 銀メダル獲得するために行ったこと(寄稿)	92
D.1 私のスペック	92
D.2 コンペ参加の準備	92
D.3 コンペ内容と基本的な内容の確認	93
D.4 コンペ参加	94
D.5 結果	102
D.6 感想	103
E Kaggleの称号と用語集	104
E.1 Kaggleの称号の説明	104
E.2 Kaggle用語集	105
F データ分析の勉強方法	107
F.1 楽しく学ぶ	107
F.2 本で学ぶ	107
G kaggleに9ヶ月取り組んで学んだこと	108
G.1 コンペの選び方	108
G.2 信頼できる検証セットの大切さ	108
G.3 実験の重要さと実験の管理	109
H 4 寄稿の書き方	111



1



13



102



# 今日の内容

●Kaggleのチュートリアル第3版の内容をもとにしたハンズオンと解説

▶A pandas-profilingでのEDA

▶5章 LightGBMでのタイタニック(ハンズオン)、B LightGBMの補足説明

▶C Santander Value Prediction Challengeで金メダルを獲得しました（解説）

▶E Kaggleの称号と用語集（解説）

▶F データ分析の勉強方法（解説）

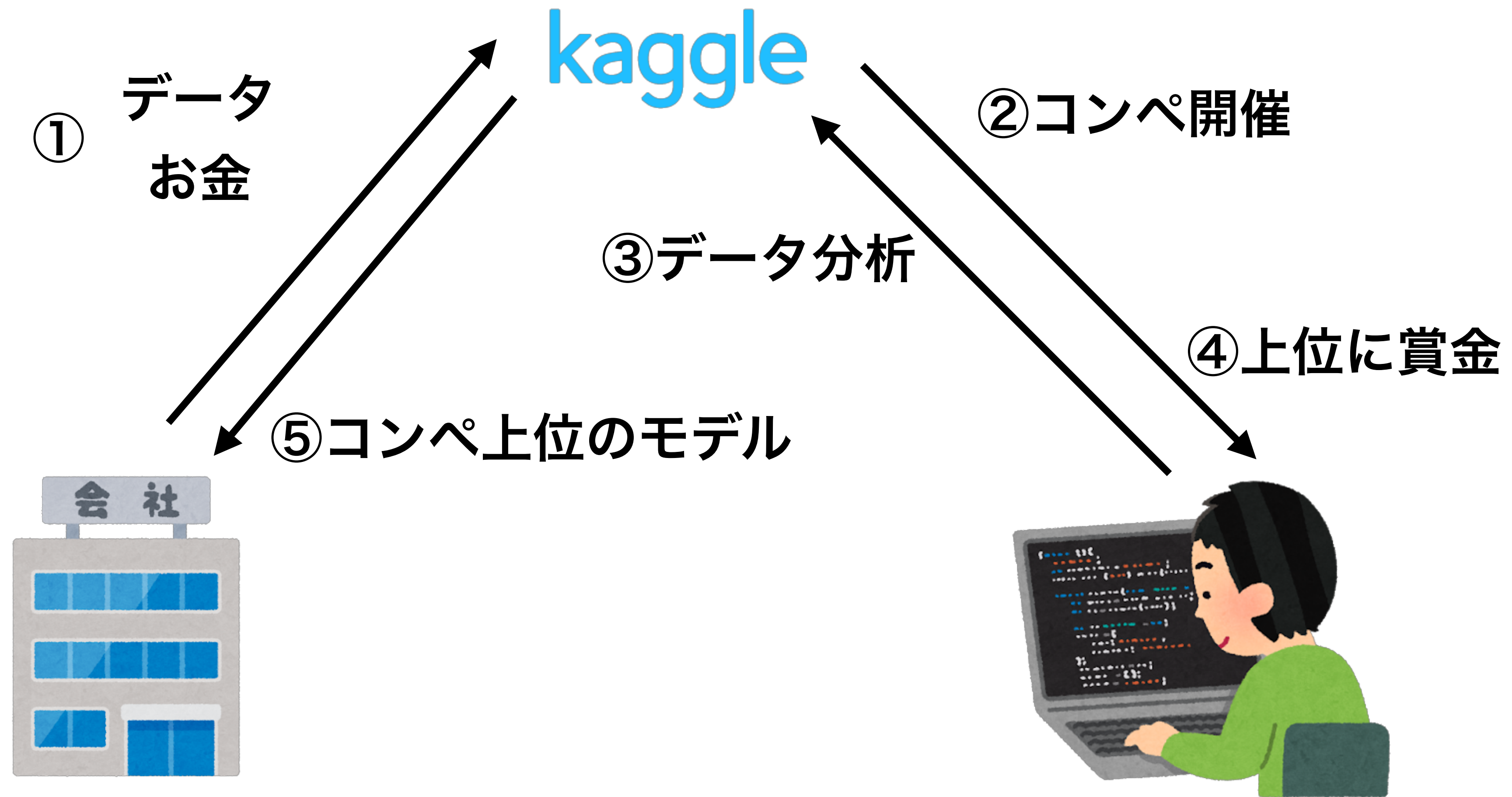
▶D HomeCreditコンペ 銀メダルを獲得するために行ったこと（解説）

▶HomeCreditコンペに挑戦してみる（ハンズオン）

# 今日説明したいこと

- LightGBMでのホールドアウト法と交差検証で学習・推論
- Kaggleのコンペに挑戦する様子
- Kaggleのメダル獲得条件、称号獲得条件                      など

# 1章 Kaggleについて



# 2章 コンペのページの翻訳など

# A PandasProfilingでのEDA

- <https://www.youtube.com/watch?v=8XwORj3Qsjs>
- 1行書くだけでEDAしてくれる便利なライブラリ
- コンペが始まったらとりあえず使うのがオススメ
- 実務でも使える
- <https://www.kaggle.com/currypurin/titanic-data-pandas-profiling>



# A PandasProfilingでのEDA

```
import pandas as pd
import pandas_profiling as pdp

df = pd.read_csv('../input/train.csv')

pdp.ProfileReport(df)
```

# B LightGBMでのタイタニック(ハンズオン)

- LightGBMはマイクロソフトが作成した、決定木ベースの勾配ブースティングを行うライブラリ
- 精度も高く、早いので、コンペでもよく使われている
- 実務でも使われている
- 欠損値の処理やスケーリング（大きさを揃える）をしなくても良いのでとても簡単である

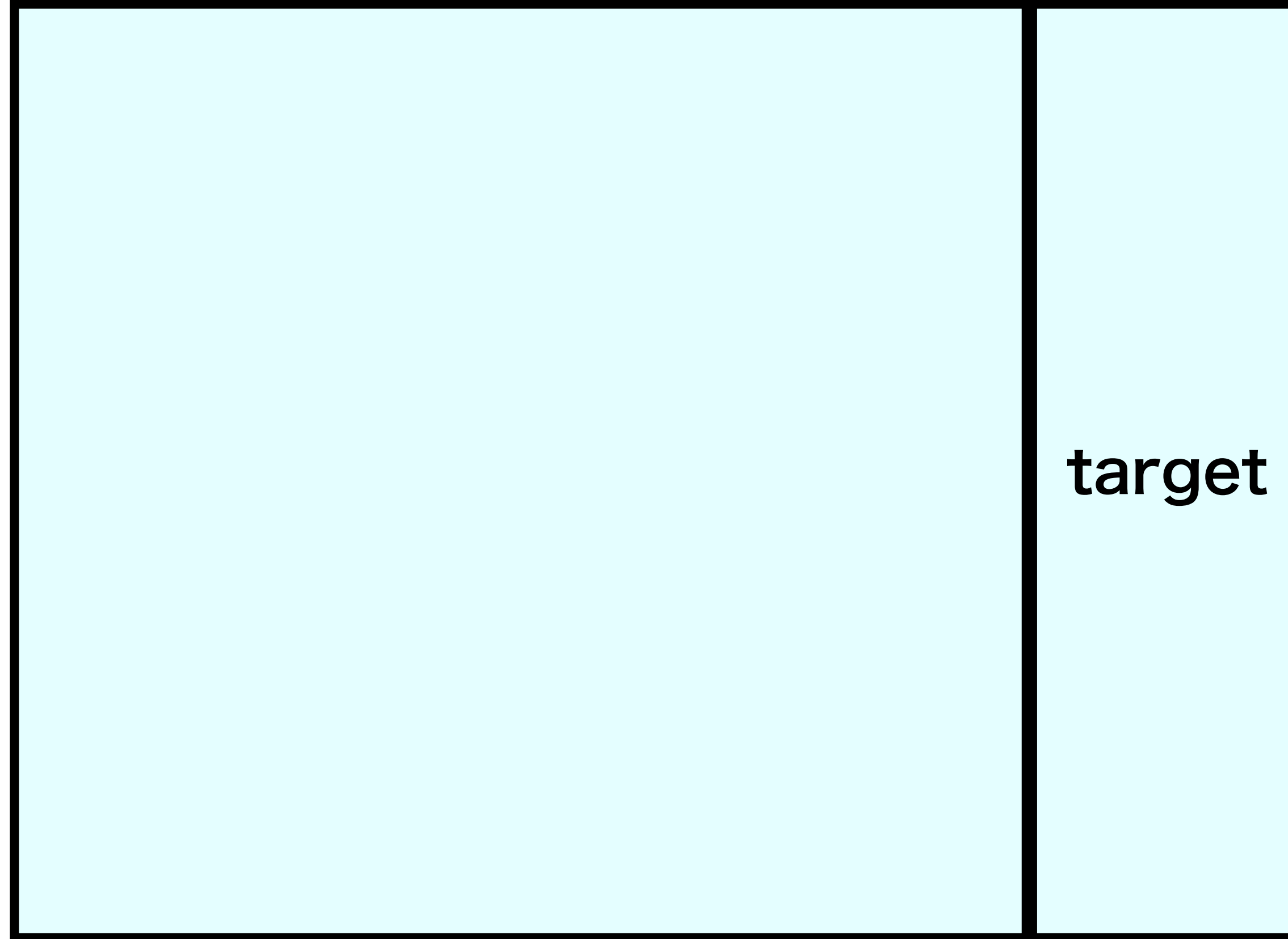
# ハンズオンはノートブックで

- ノートブック
  - その1 : <https://www.kaggle.com/currypurin/titanic-lightgbm>
  - その2 : <https://www.kaggle.com/currypurin/titanic-lightgbm-ex>

-

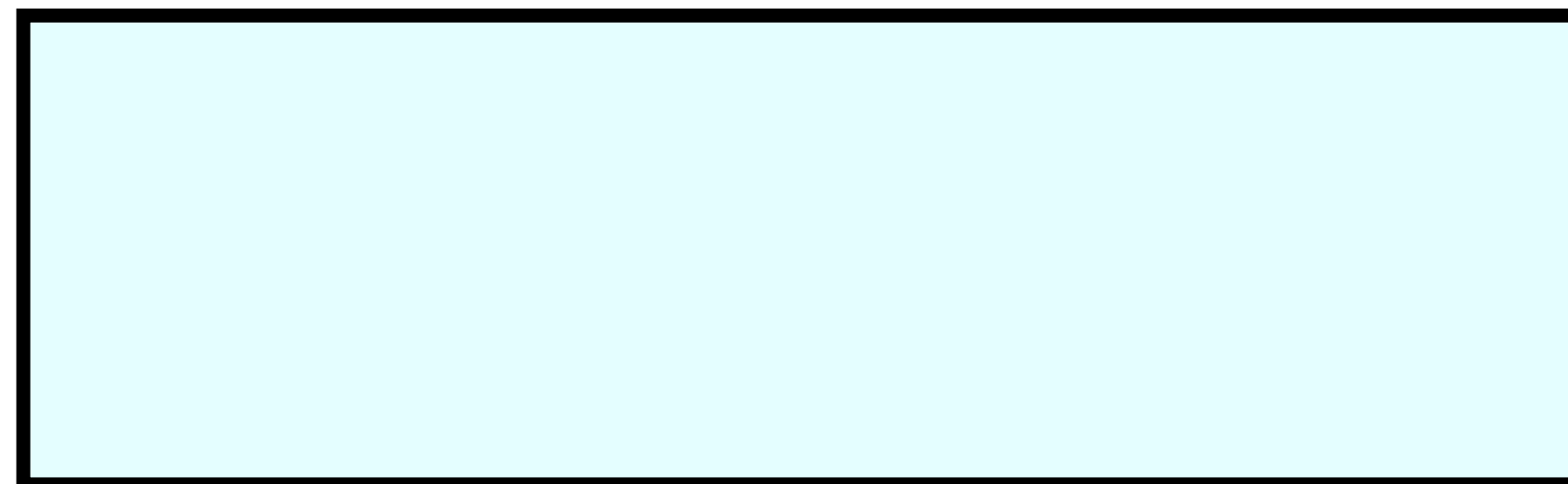
# トレーニングデータとテストデータ

train



target

test





# トレーニングデータとテストデータ

train

X\_train

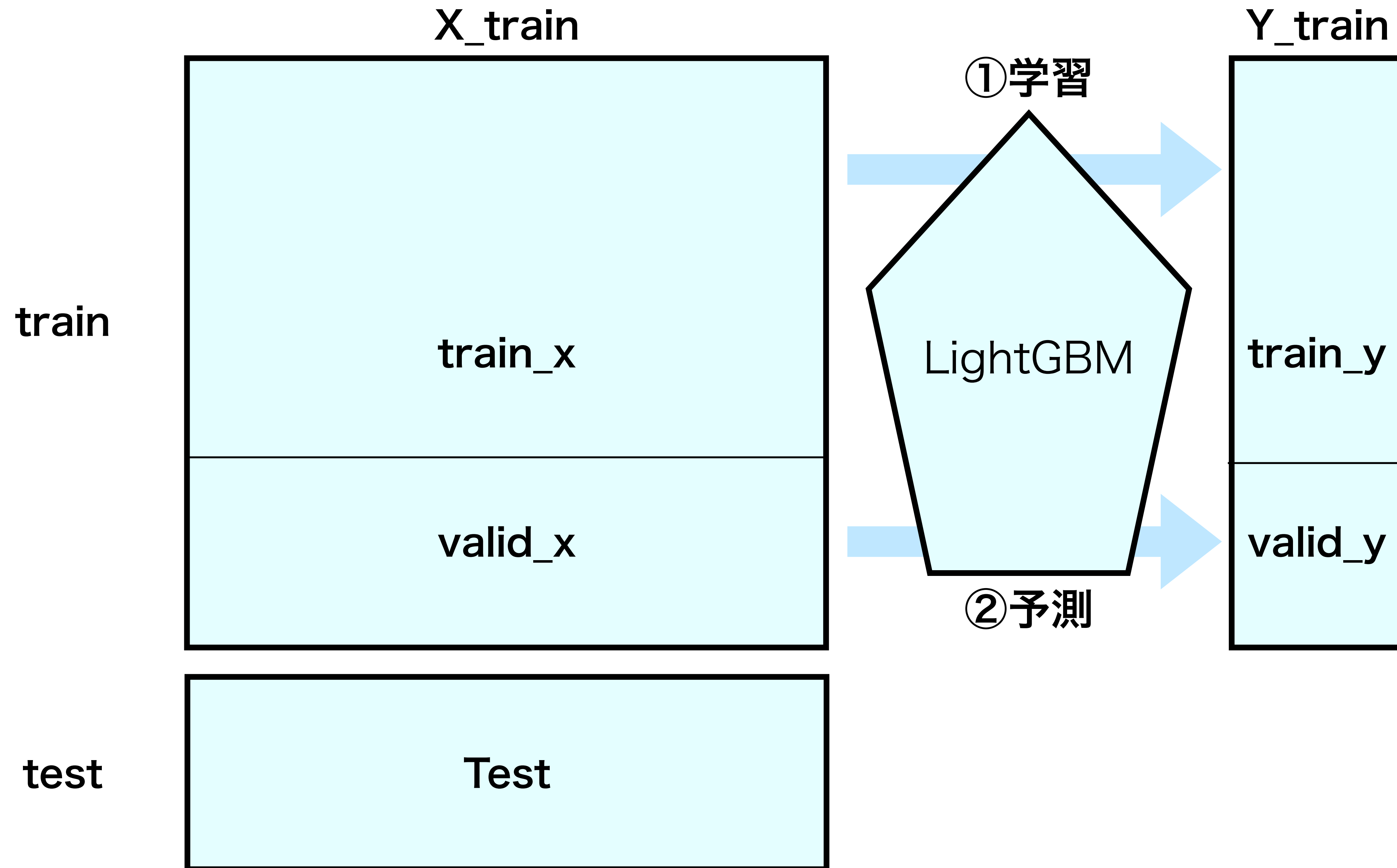
Y\_train

target

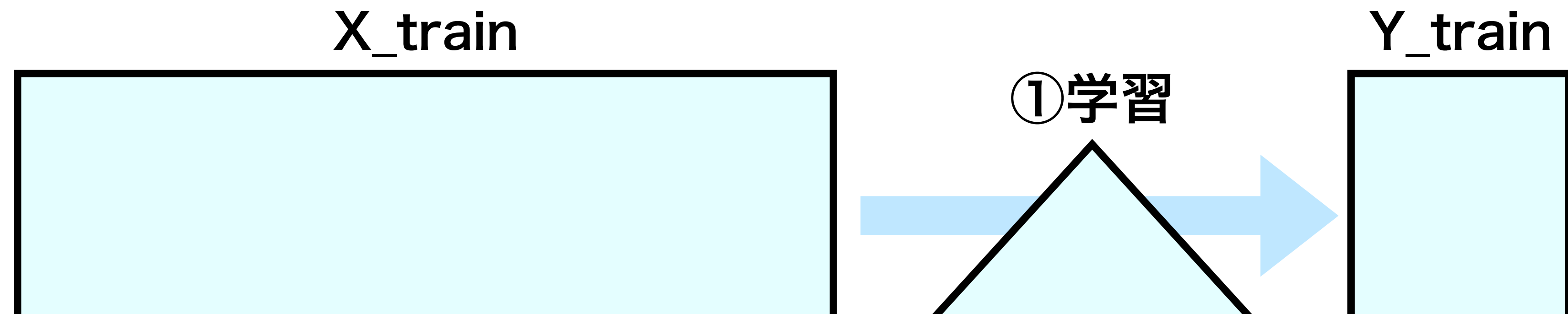
test

# ホールドアウト法

trainとvalidを完全に分離

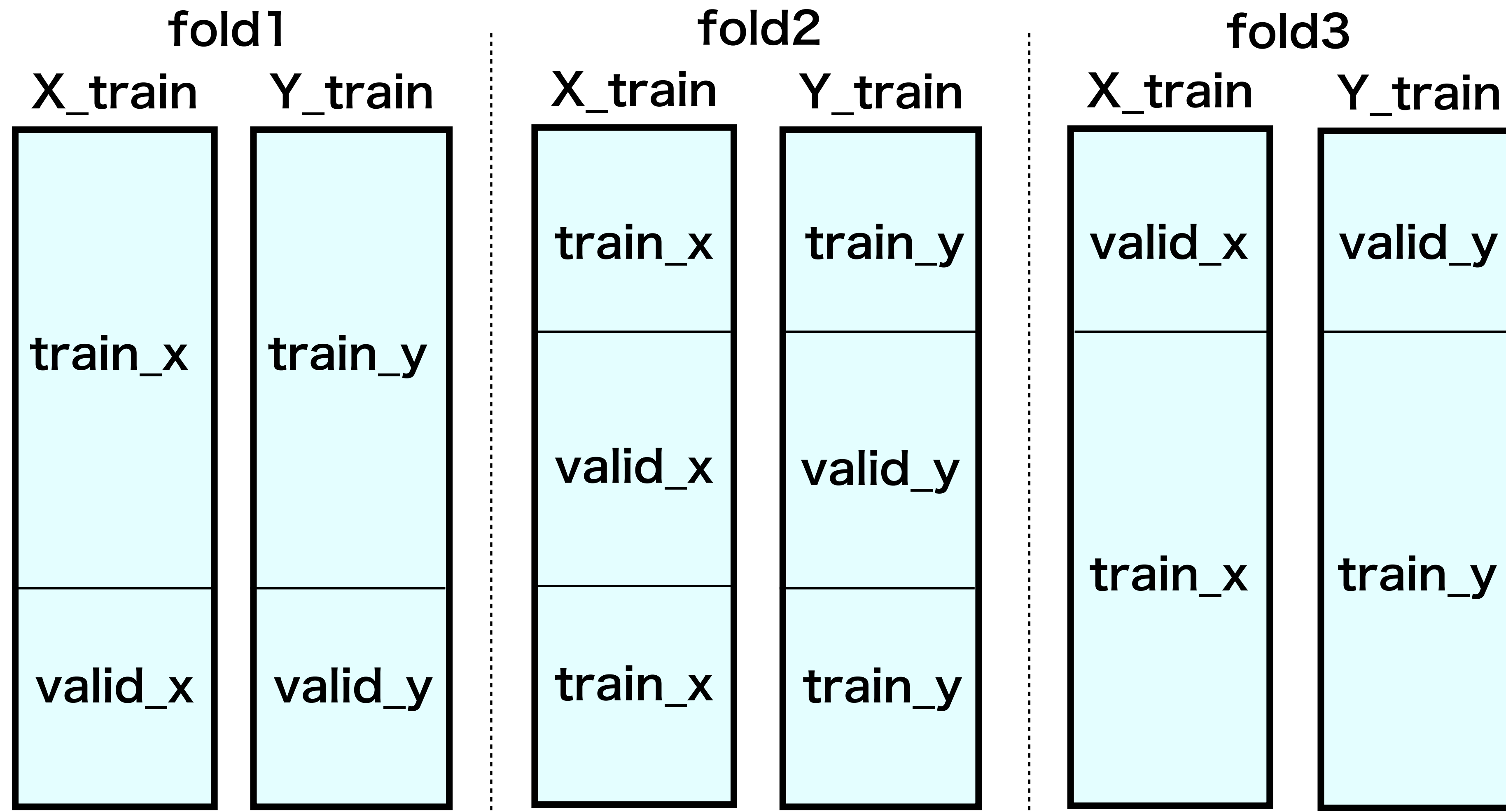


# ホールドアウト法



ホールドアウト法は、  
trainもvalidも固定され無駄がある

# k分割交差検証 (k=3)





# C Santander Value Prediction Challenge で金メダルを獲得しました（解説）

- コンペ期間：
- 期間：6月中旬～8月下旬

# データサイズ

- テーブルデータのコンペです。
- トレーニングデータ：4,459行×4,993列
- テストデータ：49,342行×4,992列
- 小さいのでとてもやりやすい。

# 評価指標

(Root Mean Squared Logarithmic Error)

$$\epsilon = \sqrt{\frac{1}{n} \sum_{i=1}^n (\log(p_i + 1) - \log(a_i + 1))^2}$$

$\epsilon$  : スコア

$n_i$  : データ数

$p_i$  : 予測値

$a_i$  : 実際の値

$\log(x)$  : ログは自然対数

# データの特徴 1

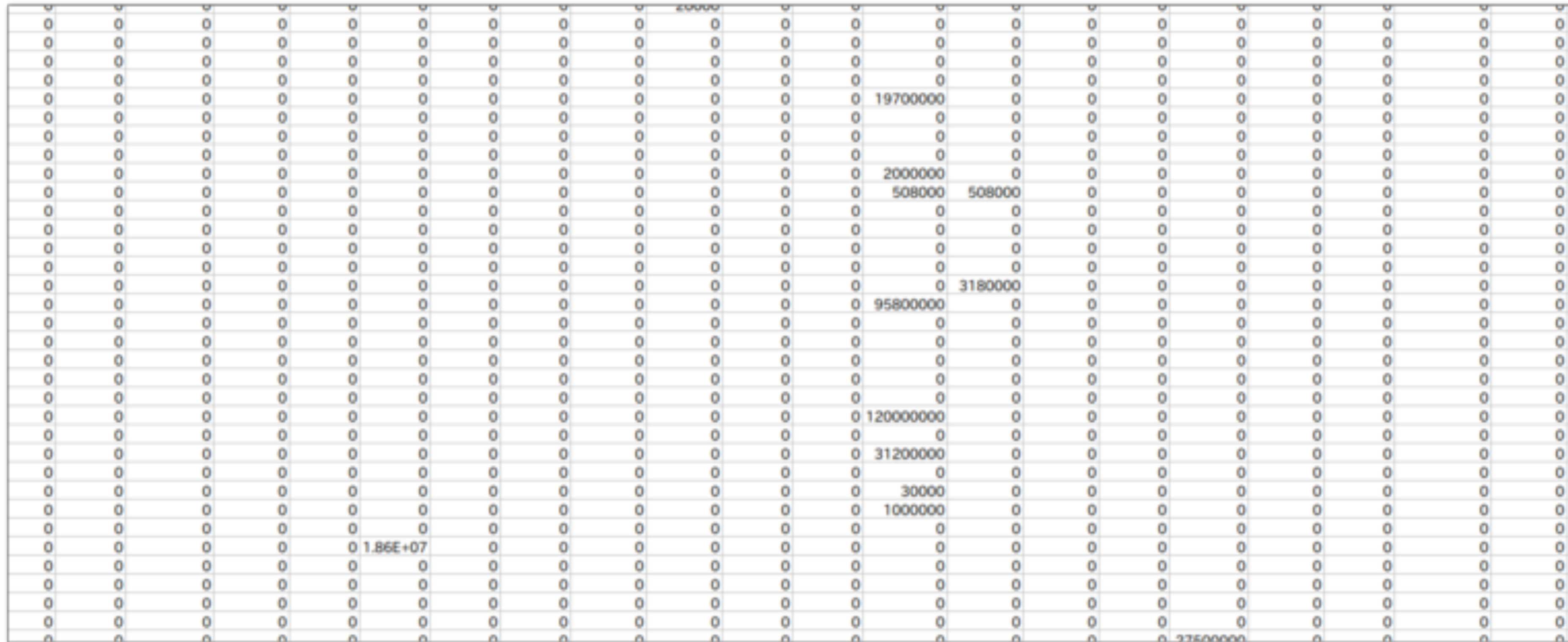
- ・ 匿名化されている

ID	target	48df886f9	0deb4b6a8	34b15f335	a8cb14b00	2f0771a37	30347e683	d08d1fbe3	6ee66e115
000d6aaf2	38000000.0	0.0	0	0.0	0	0	0	0	0



# データの特徴 2

- ゼロが多い



The image shows a large data table with many rows and columns. The table is mostly filled with zeros, indicating a high degree of sparsity. Some non-zero values are visible, such as 19700000, 2000000, 508000, 508000, 3180000, 95800000, 120000000, 31200000, 30000, 1000000, 1.86E+07, and 37500000. The table is organized into a grid with alternating light and dark gray background colors for the cells.



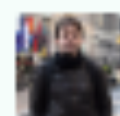
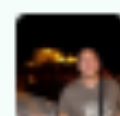
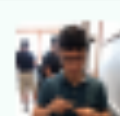


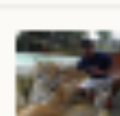
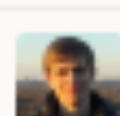
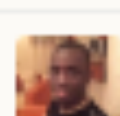
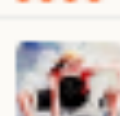
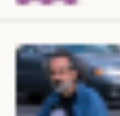

- trainの約96.8%がゼロ
- testの約98.6%がゼロ

# コンペのリーダーボードの流れ と解法

# 当初の1ヶ月弱くらいの間

- ・ 全員1.3付近
- ・ 何をやってもほとんど効かないのでお手上げ状態

# 7月中旬 リークが疑われるようになる

#	△1w	Team Name	Kernel	Team Members	Score ?	Entries	Last
1	▲2108	Simo Korkolainen			0.84	4	1d
2	—	Overfitting_maniac			1.29	32	15h
3	—	No Magic No Gain			1.30	45	7h
4	▼3	Manel			1.36	28	9h
5	▲8	Marcelo Tamashiro			1.37	49	1h
6	▼1	Alex Caveny			1.37	40	3d
7	new	Morgan Gough			1.37	17	21h
8	▲48	Roger Fed-Error			1.37	24	9h
9	▲596	Mykhailo Matviiv			1.37	16	2h
10	▲2	Get the Magic or Die Tryin'			1.37	43	2h
11	▲272	SCP-055			1.37	11	15h
12	new	Vicens Gaitan			1.37	11	2d
13	▲24	Prachant Kikani			1.37	20	14h



# 7月下旬 リークの情報が共有される

- ・ 赤の列がtarget

ID	target	f190486d6	58e2e02e6	eeb9cd3aa	9fd594eec	6eef030c1	15ace8c9f	fb0f5dbfe
7862786dc	3513333.3	0	1477600	1586889	75000	3147200	466461.5	1600000.0
c95732596	160000.0	310000	0	1477600	1586889	75000	3147200.0	466461.5
16a02e67a	2352551.7	3513333	310000	0	1477600	1586889	75000.0	3147200.0
ad960f947	280000.0	160000	3513333	310000	0	1477600	1586888.9	75000.0
8adafbb52	5450500.0	2352552	160000	3513333	310000	0	1477600.0	1586888.9
fd0c7cfc2	1359000.0	280000	2352552	160000	3513333	310000	0.0	1477600.0
a36b78ff7	60000.0	5450500	280000	2352552	160000	3513333	310000.0	0.0
e42aae1b8	12000000.0	1359000	5450500	280000	2352552	160000	3513333.3	310000.0
0b132f2c6	500000.0	60000	1359000	5450500	280000	2352552	160000.0	3513333.3
448efbb28	1878571.4	12000000	60000	1359000	5450500	280000	2352551.7	160000.0
ca98b17ca	814800.0	500000	12000000	60000	1359000	5450500	280000.0	2352551.7
2e57ec99f	307000.0	1878571	500000	12000000	60000	1359000	5450500.0	280000.0

# 7月下旬 リークの情報が共有される

ID	target	f190486d6	58e2e02e6	eeb9cd3aa	9fd594eec	6eef030c1	15ace8c9f	fb0f5dbfe
7862786dc	<u>3513333.3</u>	0	1477600	1586889	75000	3147200	466461.5	1600000.0
c95732596	<u>160000.0</u>	310000	0	1477600	1586889	75000	3147200.0	466461.5
16a02e67a	<u>2352551.7</u>	<u>3513333</u>	310000	0	1477600	1586889	75000.0	3147200.0
ad960f947	<u>280000.0</u>	<u>160000</u>	3513333	310000	0	1477600	1586888.9	75000.0
8adafb52	<u>5450500.0</u>	<u>2352552</u>	160000	3513333	310000	0	1477600.0	1586888.9
fd0c7cfc2	<u>1359000.0</u>	<u>280000</u>	2352552	並び替えると2つ下の 行にtargetが!!				1477600.0
a36b78ff7	<u>60000.0</u>	<u>5450500</u>	280000					0.0
e42aae1b8	12000000.0	<u>1359000</u>	5450500					310000.0
0b132f2c6	500000.0	<u>60000</u>	1359000					3513333.3
448efbb28	1878571.4	12000000	60000					160000.0
ca98b17ca	814800.0	500000	12000000	60000	1359000	5450500	280000.0	2352551.7
2e57ec99f	307000.0	1878571	500000	12000000	60000	1359000	5450500.0	280000.0
f190486d6	58e2e02e6	eeb9cd3aa	9fd594eec	6eef030c1	15ace8c9f	fb0f5dbfe		

# 7月下旬 からコンペ終了まで

- どのようにデータを並び替えて、
- リーク（データないにtargetがあること）を正確にたくさんみつけるか
- ということを、やる勝負に。

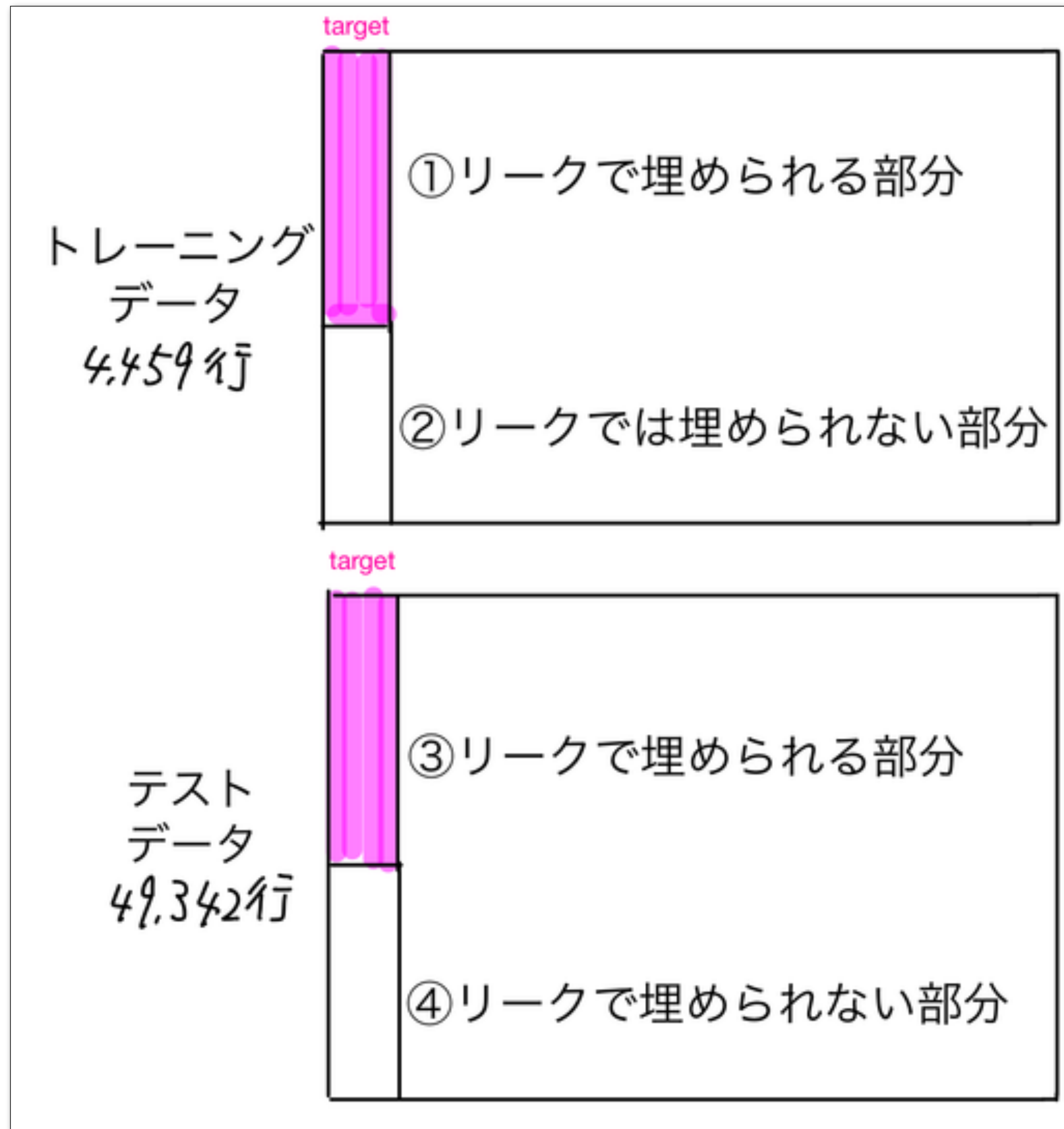


# リークの探し方

- 緑の線のように2つずらして検索。一致したら下のピンクが上のtarget







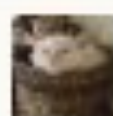



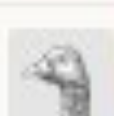
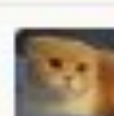








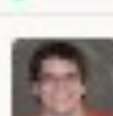
ID	target	f190486d6	58e2e02e6	eeb9cd3aa	9fd594eec	6eef030c1	15ace8c9f	fb0f5dbfe
7862786dc	3513333.3	0	1477600	1586889	75000	3147200	466461.5	1600000.0
c95732596	160000.0	310000	0	1477600	1586889	75000	3147200.0	466461.5
16a02e67a	2352551.7	3513333	310000	0	1477600	1586889	75000.0	3147200.0
ad960f947	280000.0	160000	3513333	310000	0	1477600	1586888.9	75000.0
8adafbb52	5450500.0	2352552	160000	3513333	310000	0	1477600.0	1586888.9
fd0c7cfc2	1359000.0	280000	2352552	160000	3513333	310000	0.0	1477600.0
a36b78ff7	60000.0	5450500	280000	2352552	160000	3513333	310000.0	0.0
e42aae1b8	12000000.0	1359000	5450500	280000	2352552	160000	3513333.3	310000.0
0b132f2c6	500000.0	60000	1359000	5450500	280000	2352552	160000.0	3513333.3
448efbb28	1878571.4	12000000	60000	1359000	5450500	280000	2352551.7	160000.0
ca98b17ca	814800.0	500000	12000000	60000	1359000	5450500	280000.0	2352551.7
2e57ec99f	307000.0	1878571	500000	12000000	60000	1359000	5450500.0	280000.0

# 私の最終の結果



- ③7865行をリークで埋めて、
- ④は、一番スコアが高いカーネルを1.1倍して提出
- (①と③の平均を比較すると、③のほうがやや大きかったなので、リーダーボードを頼りに探索)

# 最終結果




1	▲ 8	Linear Regression Only	 	0.51	180	1h
2	▲ 53	adilism		0.52	26	3m
3	▲ 36	anatoly		0.52	43	1h
4	▲ 17	chenhan zhang		0.52	27	12h
5	▲ 107	Vladimir Larin [ods.ai]		0.52	92	2d
6	▲ 1	Paradox		0.52	136	2h
7	▲ 36	verycourage		0.52	33	3h
8	▲ 7	currypurin		0.52	45	4m
9	▼ 4	Raymond Zeng		0.52	119	5m
10	▲ 19	Zuper	  	0.52	63	9h
11	▲ 21	yurodiviy		0.52	52	5m
12	▲ 35	kamitsu	    	0.52	82	34m
13	▲ 38	Academics looking for real jobs	 	0.52	145	7m



# まとめ

- ・ 色々なコンペがあります
- ・ 初心者でも金メダルをとれることもあります

# E.1 称号

- Grand master 金 5、内 1 つはソロ 
- Master 金 1、銀 2 
- Expert 銅 2 
- Contributor プロフィールの完成等
- Novice Kaggleに登録

※<https://www.kaggle.com/progression> より

# E.2 メダルの獲得条件

	0~99 チーム	100~249 チーム	250~999 チーム	1000 チーム以上
ブロンズ	上位40%	上位40%	上位100人	上位10%
シルバー	上位20%	上位20%	上位50人	上位5%
ゴールド	上位10%	上位10人	上位10人と 0.2%*1	上位10人と 0.2%*1

※<https://www.kaggle.com/progression> より

# E.2 Kaggle用語集

- EDA(Exploratory Data Analysis)
- Kaggle api
- Kaggle learn
- kaggler-ja
- Kaggler-ja Wiki
- Kaggle Rankings
- regonn&curry.fm

# F データ分析の勉強方法（解説）

# F データ分析の勉強方法（解説）

- ・ 実践から入るのが 1 番だと思います
- ・ 実践した後であれば、難しい本を読んでも、理解できることが多いです
- ・ 自分がやりたいように学ぶのが 1 番だと思います

D HomeCreditコンペ 銀メダルを獲得  
得するするするために行ったこと（解説）

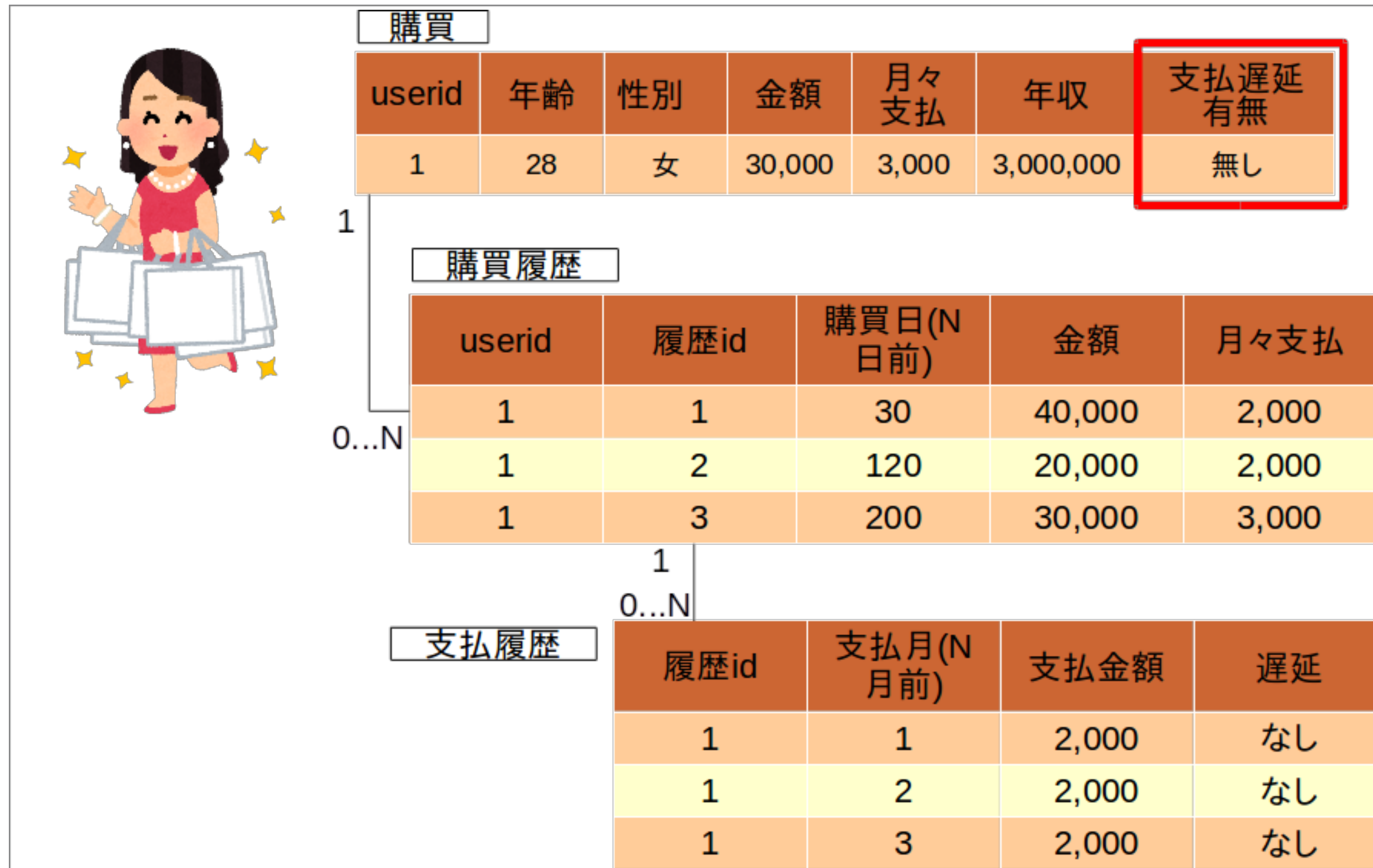
# D HomeCreditコンペ 概要

- ユーザーが契約したローンを遅延なく返せるか予測
- 評価指標はAUC (参考 : [http://www.randpy.tokyo/entry/roc\\_auc](http://www.randpy.tokyo/entry/roc_auc))

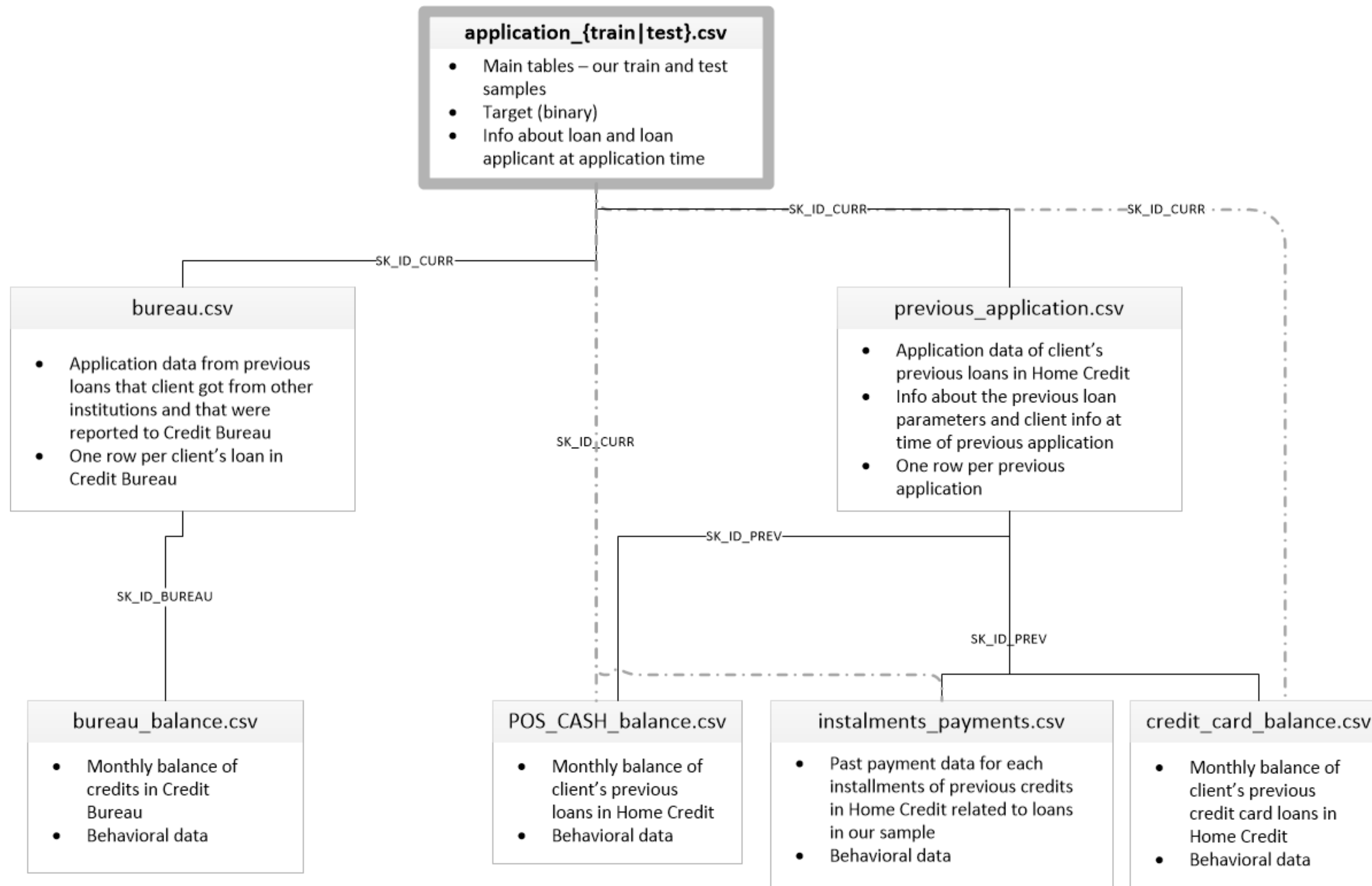


# 概要

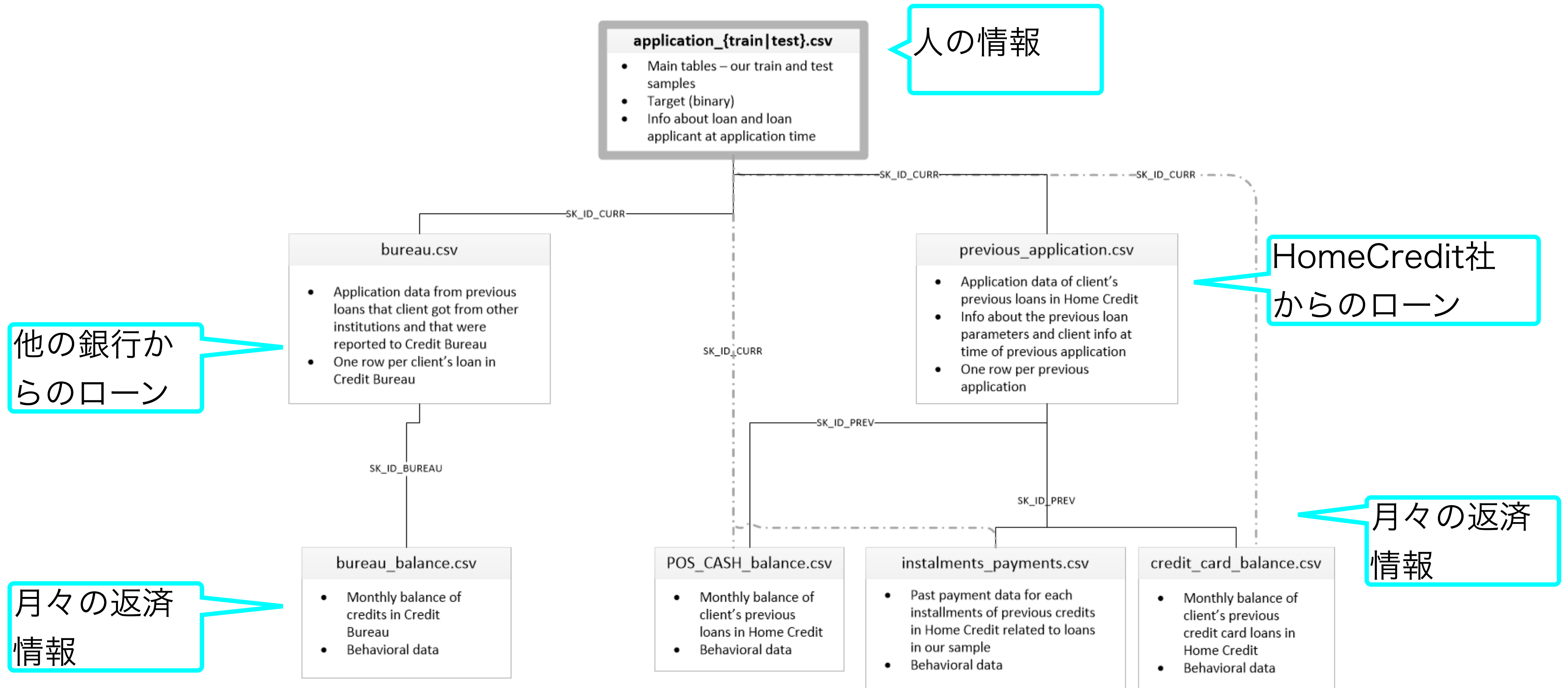
- ・ 赤字を当てる
- ・ 複数テーブル (csvファイル) がある



# Data



# Data



# くるびーさんの手法

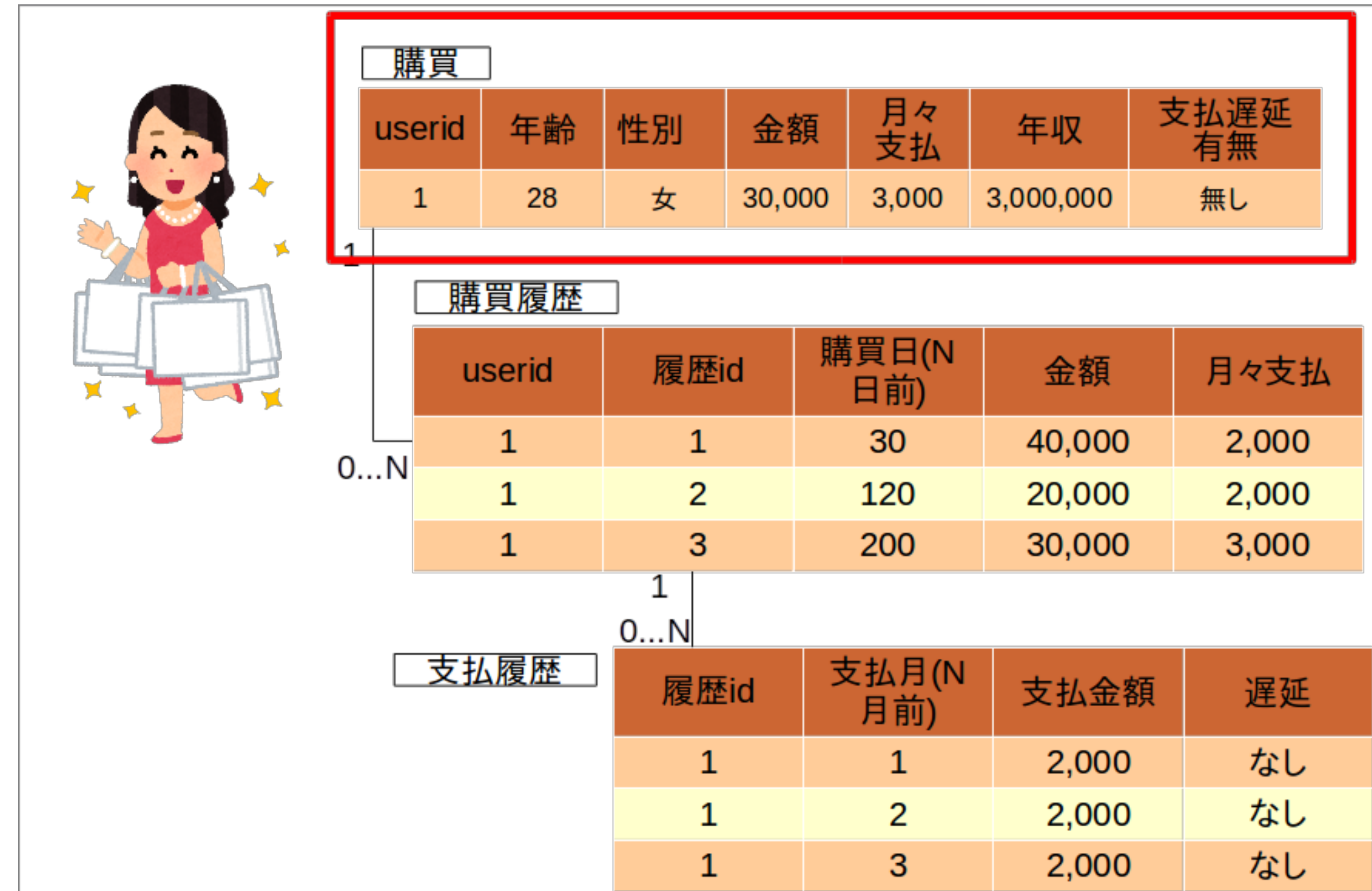
- LightGBMに特徴を徐々に足して行って、精度を確認していく
- 最後に特徴選択やパラメータチューニングを行う

# くるびーさんの手法

## 4.1 重要な3特徴だけで予測

1.信用スコアのような特徴がこのコンペでは提供されており、重要な特徴であった

4.2 ファイル全部使って、LightGBMにチャレンジ





# くるびーさんの手法

## 4.3.1 全ファイルを機械的に集計

- ・1人で複数の購買履歴がある
- ・購買履歴の平均や最大などを特徴にする
- ・Kernelが参考なる

購買履歴									
userid	履歴id	購買日(N日前)	金額	月々支払					
1	1	30	40,000	2,000					
1	2	120	20,000	2,000					
1	3	200	30,000	3,000					

↓


userid	件数	購買日平均	購買日最大	購買日最小	購買日合計	金額平均	金額最大	金額最小	金額合計
1	3	117	200	30	350	30,000	40,000	20,000	90,000

# くるぴーさんの手法

## 4.3.2 複数の項目同士で特徴を作る

- ・子供の数と年収から、子供1人あたりの年収など、特徴間の関係を捉えた特徴を作る

購買		
userid	子供の数	年収
1	2	5,000,000
2	1	5,000,000
3	4	5,000,000



userid	子供1人あたり年収
1	2,500,000
2	5,000,000
3	1,250,000

# くるびーさんの手法

## 4.3.3 複数の項目同士で特徴を作る (カテゴリ変数)

- ・ターゲットエンコーディング

購買		
userid	持ち家	支払遅延 有無
1	あり	1
2	あり	0
3	なし	0
4	なし	0
5	なし	1

↓

userid	持ち家 (encoding)
1	0.5
2	0.5
3	0.33
4	0.33
5	0.33



# くるびーさんの手法

## 4.3.3 複数の項目同士で特徴を作る (カテゴリ変数)

- ・ドメイン知識による数値変換

購買	
userid	学歴
1	大卒
2	大学院卒
3	高卒

↓

userid	就学年数
1	22
2	24
3	18

# くるびーさんの手法

## 4.3.4 時系列アプローチ

- ・直近〇日という特徴を作成
  - ・直近〇日の金額
  - ・直近〇日の件数

購買履歴

userid	履歴id	購買日(N 日前)	金額	月々支払
10	1	15	100,000	5,000
10	2	20	100,000	5,000
10	3	30	80,000	4,000
10	4	40	70,000	3,500
10	5	70	50,000	5,000
10	6	80	60,000	3,000
10	7	100	40,000	4,000
10	8	200	30,000	3,000

userid	過去30 金額	過去30 件数	過去90 金額	過去90 件数	過去365 金額	過去365 件数
10	280,000	3	460,000	6	530,000	8

# くるびーさんの手法

## 4.3.4 特徴選択

- ・特徴を作ると、簡単に数千の特徴になってしまう
- ・LightGBMやXGboostは多くの特徴があっても、通常大きな問題はないが、有害な特徴は取り除くとスコアが改善する

# くるびーさんの手法

## 4.3.5 パラメータチューニング

- Light GBMのパラメータは多数存在
- パラメータチューニングによりスコアが改善する
  - hyperopt
  - 今ならoptunaが候補か

# くるびーさんの手法

## 4.3.5 アンサンブル

- ・Light GBMのboostingを変え、またseedを変えてアンサンブルしている
- ・相関の低いアルゴリズムをアンサンブルすると、スコアが改善するため、一般に、ニューラルネットと勾配ブースティング（LightGBM等）とのアンサンブルで大きくスコアが改善する。

# HomeCreditコンペに挑戦してみる（ハンズオン）

- ノートブック
  - <https://www.kaggle.com/currypurin/simple-lightgbm>