SLW

統計的学習による意思決定システム

-- 知識発見モデルへの実践 --

徐 良為 (Xu Liangwei)
(liangweixu2205@gmail.com)

2023-11-18

自己紹介

- 名前：　徐 良為（じょ　りょうい）
- 出身：　中国上海
- 現在
  - NTTデータ数理システム　非常勤技術顧問
  - 個人事業主（屋号＝Statistical Learning Workshop）
    - **機械学習の新しい技術、仕組への研究・実験**
    - **Youtubeに動画作成、統計・機械学習の普及・最新の研究動向など**

- 学歴
  - 学部：中国上海交通大学コンピュータサイエンス
  - 修士＆博士：東京大学大学院情報工学研究科
- 職歴
  - NTTデータ数理システム　データマイニング部 部長
    - 統計・機械学習・データ解析の受託・システム開発

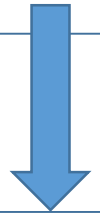
機械学習の問題点

■モデル構築に、膨大な数のサンプルが必要

- 人間の子供に、数個の「机」を見せるだけで、「机」と認識　⇦　「知識」の遺伝

■モデルの不可読性（ブラックボックス）

- 予測パフォーマンスの獲得に見合うモデルの複雑度合？



妨げ

■自らの知識発見、発明などに必要な論理的な推論

■知識の「遺伝」

■人間とのコラボレーション

■意思決定の道具としても限界

提案： ルールの線形結合モデル

■知識ルール： 条件（説明変数）➡ 結論（目的変数）

BMI(肥満度）が３０以上となると、高血圧になる確率は 0.7
塩分摂取量は８ｇ 以上の場合は、平均収縮期血圧は 150

■ルール

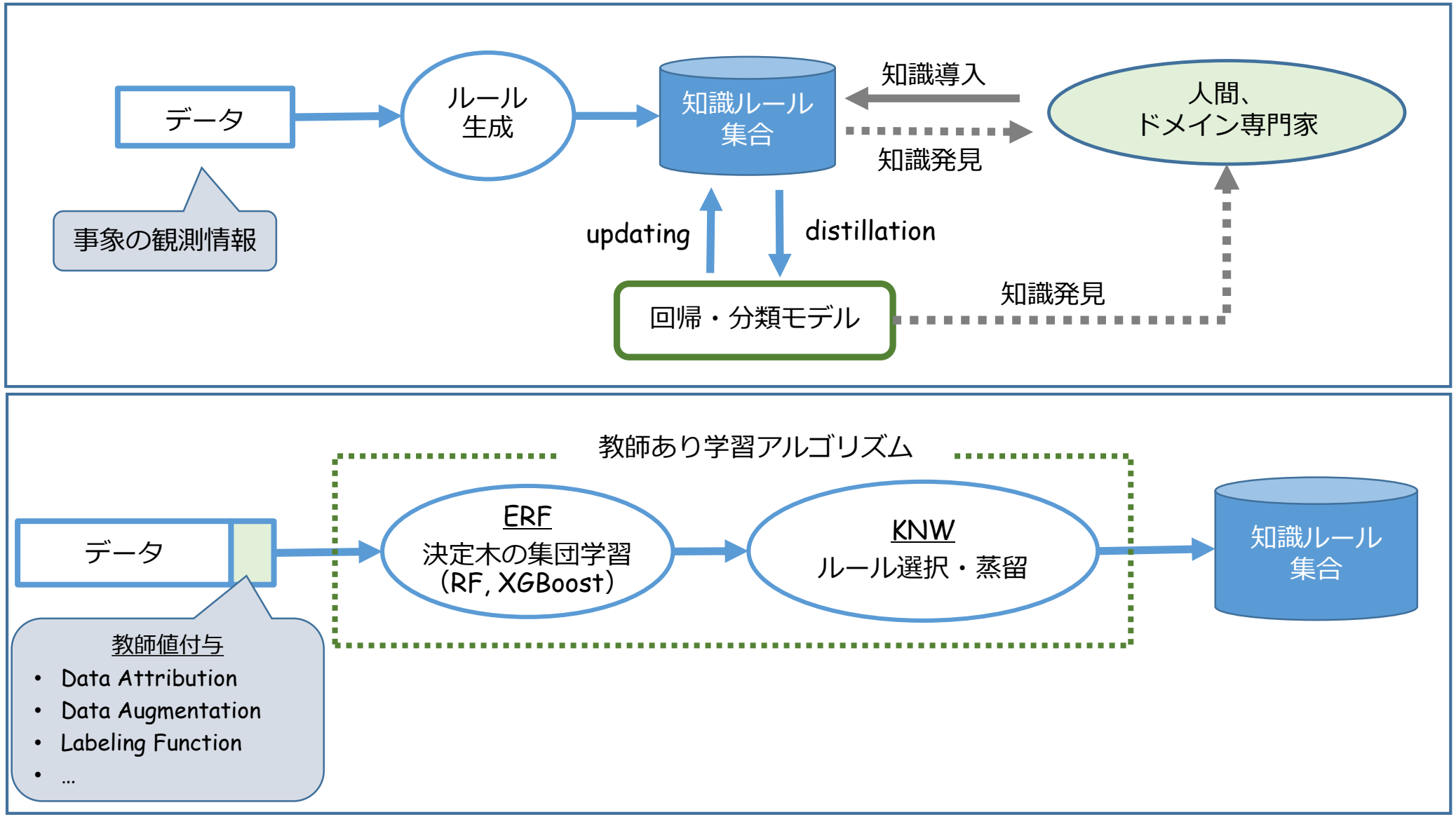
$$r(x) = 1 \quad if\ cond(x)$$
$$= 0 \quad otherwise$$

x は説明変数を表し、 $cond(x)$はxに関する条件式を表す

■モデル(ルールの線形結合)

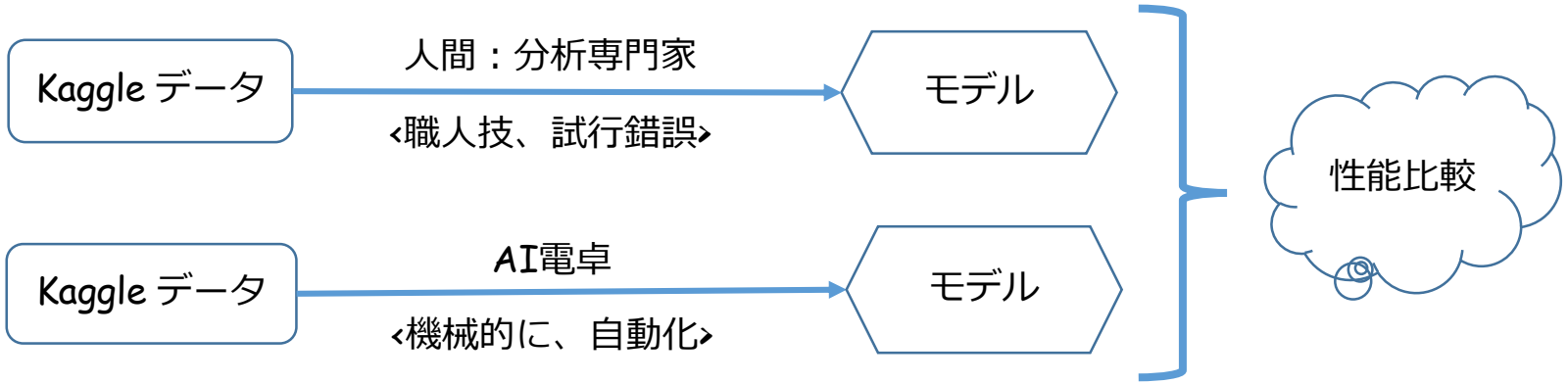
$$f(x) = \beta_0 + \sum_{i=1}^{m} \beta_i r_i(x)$$

AI電卓： データ➡モデル



AI電卓の精度検証： Kaggleデータ分析コンテスト

- Kaggleとは
  - Google社が運営、企業や、研究者がデータを投稿し、世界中の統計・機械学習・サイエンティストがデータに最適な予測モデルを競い合う
  - 参加者
    - 世界最大規模（？）、約１０万人参加
    - 情報科学者、統計学者、経済学者、数学者など
  - 様々な業種から様々なテーマ
    - 金融、流通、生物・医療・バイオ、製造、メディア・広告、自然言語処理、有償テーマも多数
  - 日本でも、Kaggleの攻略法、体験談などに関する書籍多数
- 目的



検証1： House Price

■不動産に関する属性から、不動産価格を予測するモデル

■学習用データ（正解を含むデータ）

- ➢属性数（列数）：81
- ➢行数：1460件

■検証用データ（正解未知）

- ➢属性数（列数）：80
- ➢行数：1459件

■モデル評価方法

学習データから構築したモデルを用いて、検証データに対して不動産価格を予測し、予測結果の精度

データ特徴

Viewer - train.csv

train.csv

	snPorch	ScreenPorch	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	SalePrice
1	0	0	0	NA	NA	NA	0	2	2008	WD	Normal	208500
2	0	0	0	NA	NA	NA	0	5	2007	WD	Normal	181500
3	0	0	0	NA	NA	NA	0	9	2008	WD	Normal	223500
4	0	0	0	NA	NA	NA	0					140000
5	0	0	0	NA	NA	NA	0	1				250000
6	320	0	0	NA	MnPrv	Shed	700	1				143000
7	0	0	0	NA	NA	NA	0	8	2007	WD	Normal	307000
8	0	0	0	NA	NA	Shed	350	11	2009	WD	Normal	200000
9	0	0	0	NA	NA	NA	0	4	2008	WD	Abnorml	129900
10	0	0	0	NA	NA	NA	0	1	2008	WD	Normal	118000
11	0	0	0	NA	NA	NA	0	2	2008	WD	Normal	129500

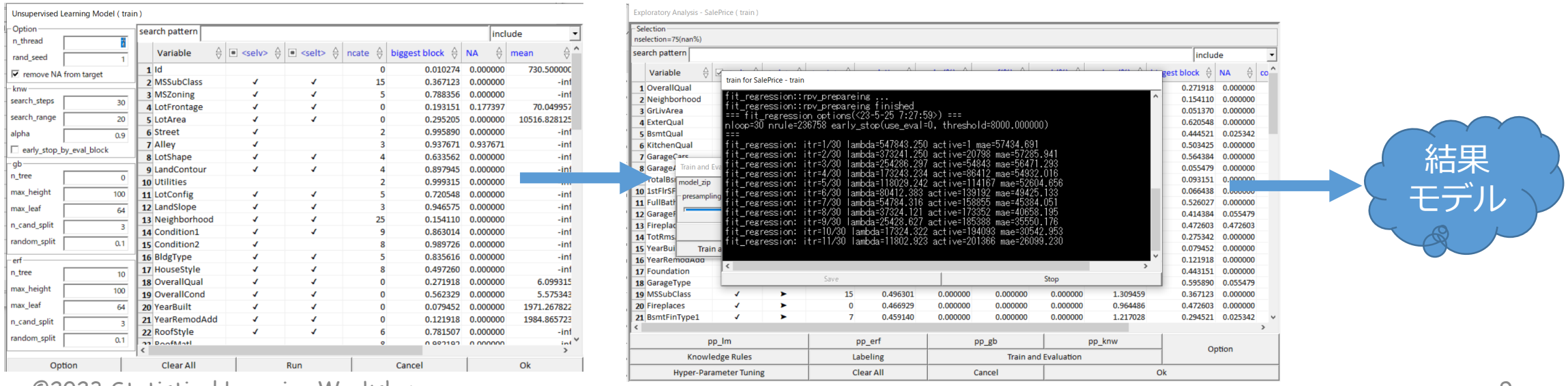
予測対象：販売価格

nrow=1,460, ncol=81

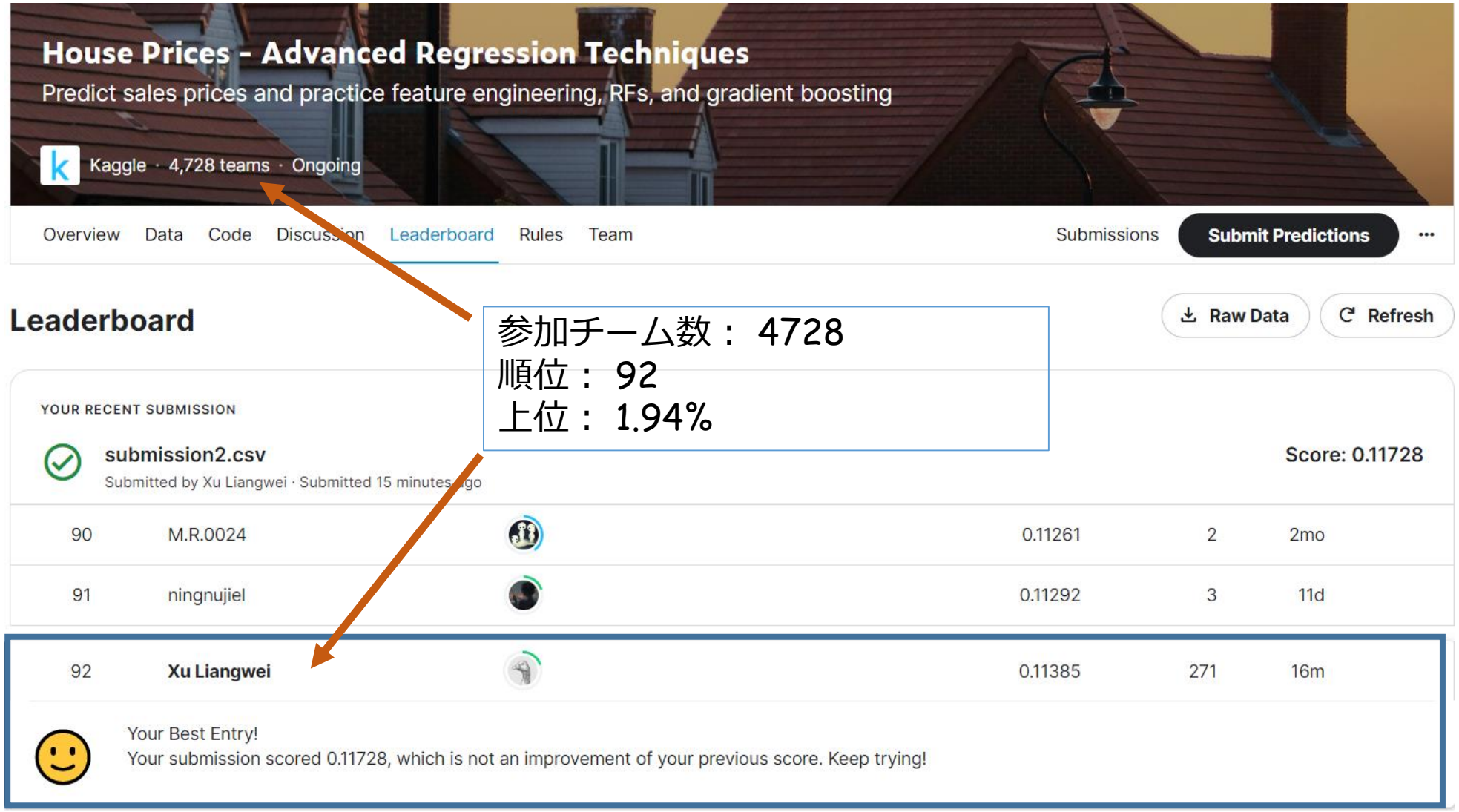
- 学習用、検証用両方のデータに、欠損（NA）が多数存在
- 予測対象（不動産価格）の分布にかなりの格差がある
- 不動産価格に関連性が薄い属性が多く見受けられる

AI電卓でのモデル構築様子

- 下記のような前処理を一切行わず
  - 欠損の補填
  - 予測対象（不動産価格）の分布への事前調整
  - 特徴量抽出
- 学習用＆検証用データをそのまま導入、学習＆予測



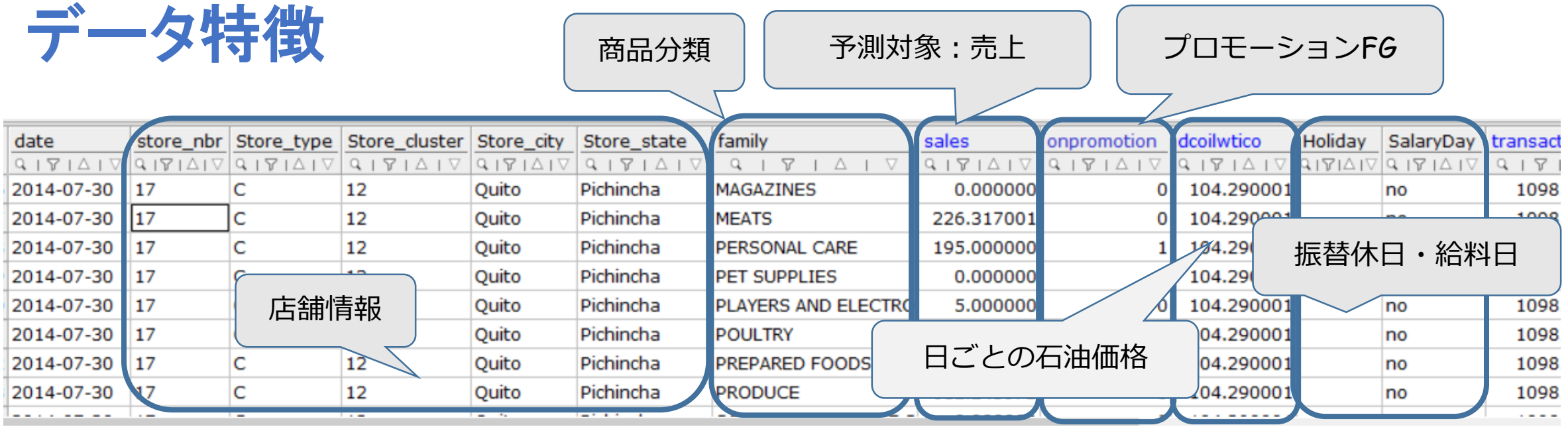
Kaggleからの成績評価(2023-5-25)



検証2： Store Sales ＝ 時系列データ分析

- 食料品チェン店（grocery store）の日々の売上実績から、将来の売上を予測する
- 店舗数（store_nbr）＝54、商品分類（family）数=33
- 店舗毎、商品分類毎の売上に関する時系列数＝54*33=1782
- 学習用データ（正解を含むデータ）
  - 売上履歴（日数）：1680日
  - 行数：294万件、属性数（列数）：１７
- 検証用データ（正解未知）
  - 予測対象日数：16日
  - 行数：28512件（＝店舗数*商品分類数*予測対象日数）
- モデル評価方法
  学習データから構築したモデルを用いて、検証データに対して売上を予測し、予測精度

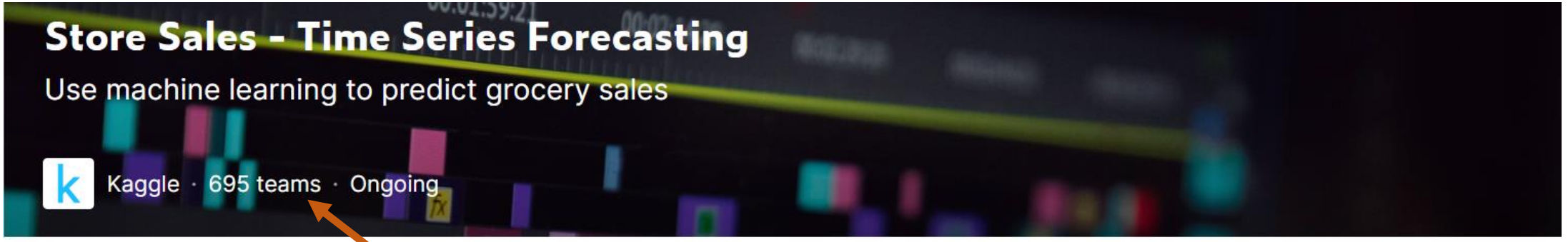
データ特徴



date	store_nbr	Store_type	Store_cluster	Store_city	Store_state	family	sales	onpromotion	dcoilwtico	Holiday	SalaryDay	transact
2014-07-30	17	C	12	Quito	Pichincha	MAGAZINES	0.000000	0	104.290001		no	1098
2014-07-30	17	C	12	Quito	Pichincha	MEATS	226.317001	0	104.290001			
2014-07-30	17	C	12	Quito	Pichincha	PERSONAL CARE	195.000000	1	104.29			
2014-07-30	17			Quito	Pichincha	PET SUPPLIES	0.000000		104.29			
2014-07-30	17			Quito	Pichincha	PLAYERS AND ELECTRC	5.000000	0	104.290001		no	1098
2014-07-30	17			Quito	Pichincha	POULTRY			04.290001		no	1098
2014-07-30	17	C	12	Quito	Pichincha	PREPARED FOODS			04.290001		no	1098
2014-07-30	17	C	12	Quito	Pichincha	PRODUCE			04.290001		no	1098

- 時系列データ（店舗、商品ごと、日々売上）以外の「外部要因」も含まれる
- 予測対象（売上金額）の分布にかなりの格差（桁違い）がある
- データの行数は比較的に多い（３００万件前後）

Kaggleからの成績評価(2023-9-26)



Overview Data Code Discussion Leaderboard Rules Team Submissions **Submit Predictions** ···

参加チーム数： 695
順位： 41
上位： 5.89%

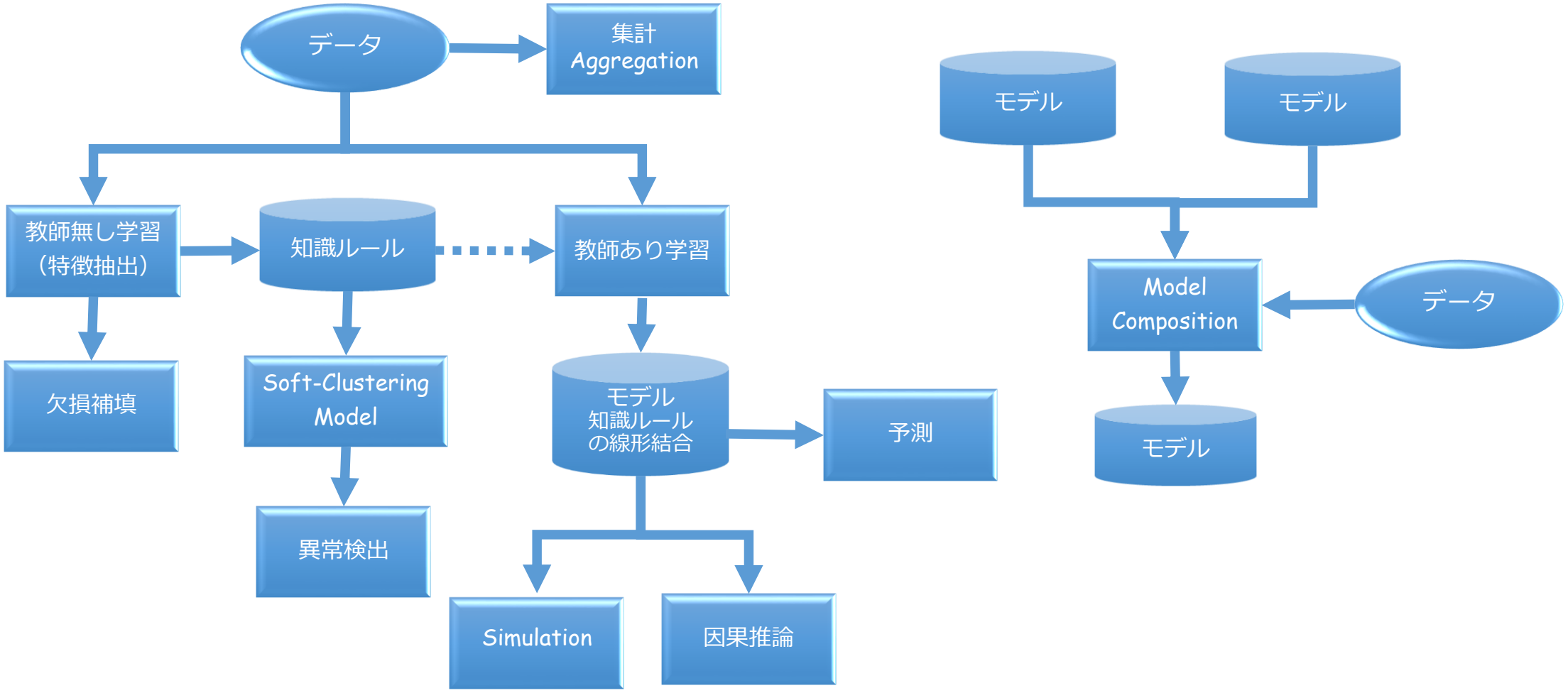
39	Will Gilchrist	0.39261	2	12d
40	Mario Refoyo López	0.39495	4	1mo
41	**Xu Liangwei**	0.39778	110	5m

Your Best Entry!
Your most recent submission scored 0.39778, which is an improvement of your previous score of 0.40014. Great job!

Tweet this

42	David Gilbertson	0.39842	20	1mo

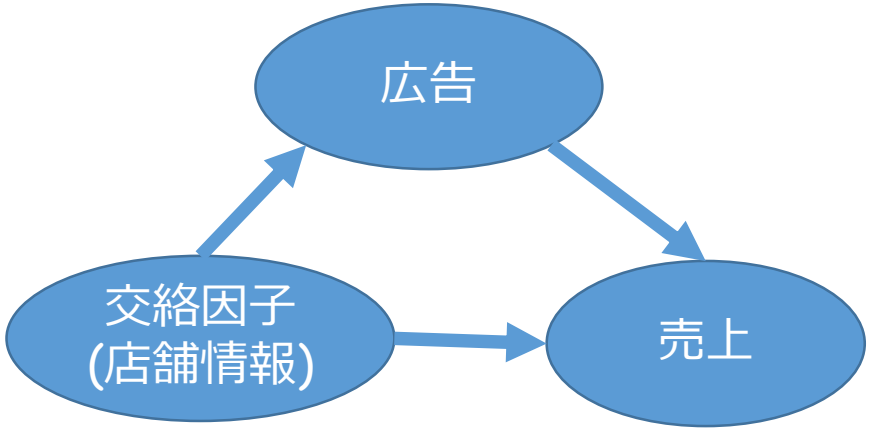
AI電卓モデル関連機能



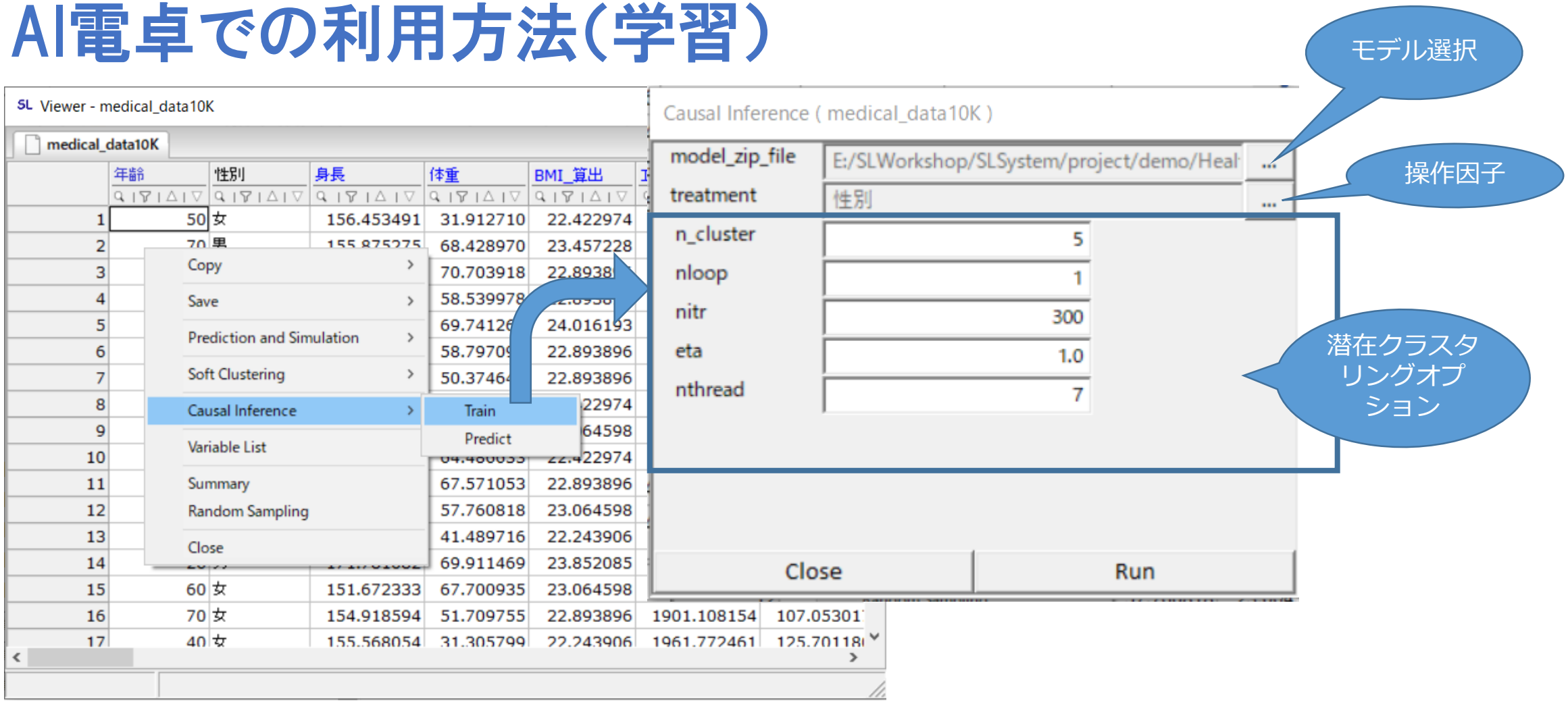
因果推論：店舗の選択（売上が広告から受ける影響）

店舗	競合店	人口構成	立地条件	…	売上	広告	売上.広告有	売上.広告無
s1	…	…	…	…	198	有	198	?
s2	…	…	…	…	55	無	?	55
s3	…	…	…	…	77	無	?	77
s4	…	…	…	…	160	有	160	?
s5	…	…	…	…	143	無	?	143
…	…	…	…	…	…	…	…	…

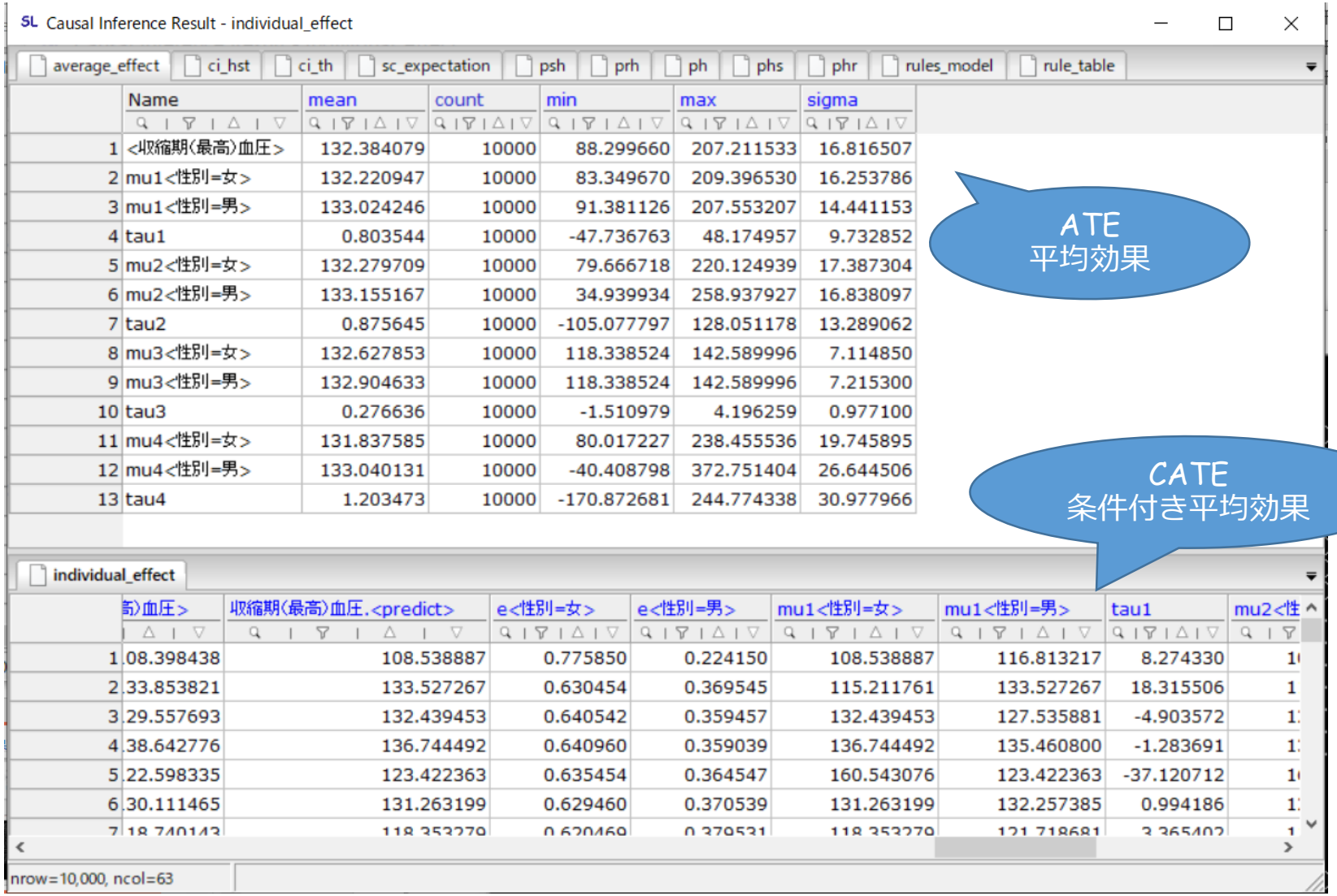
他の因子＝店舗情報
を考慮する



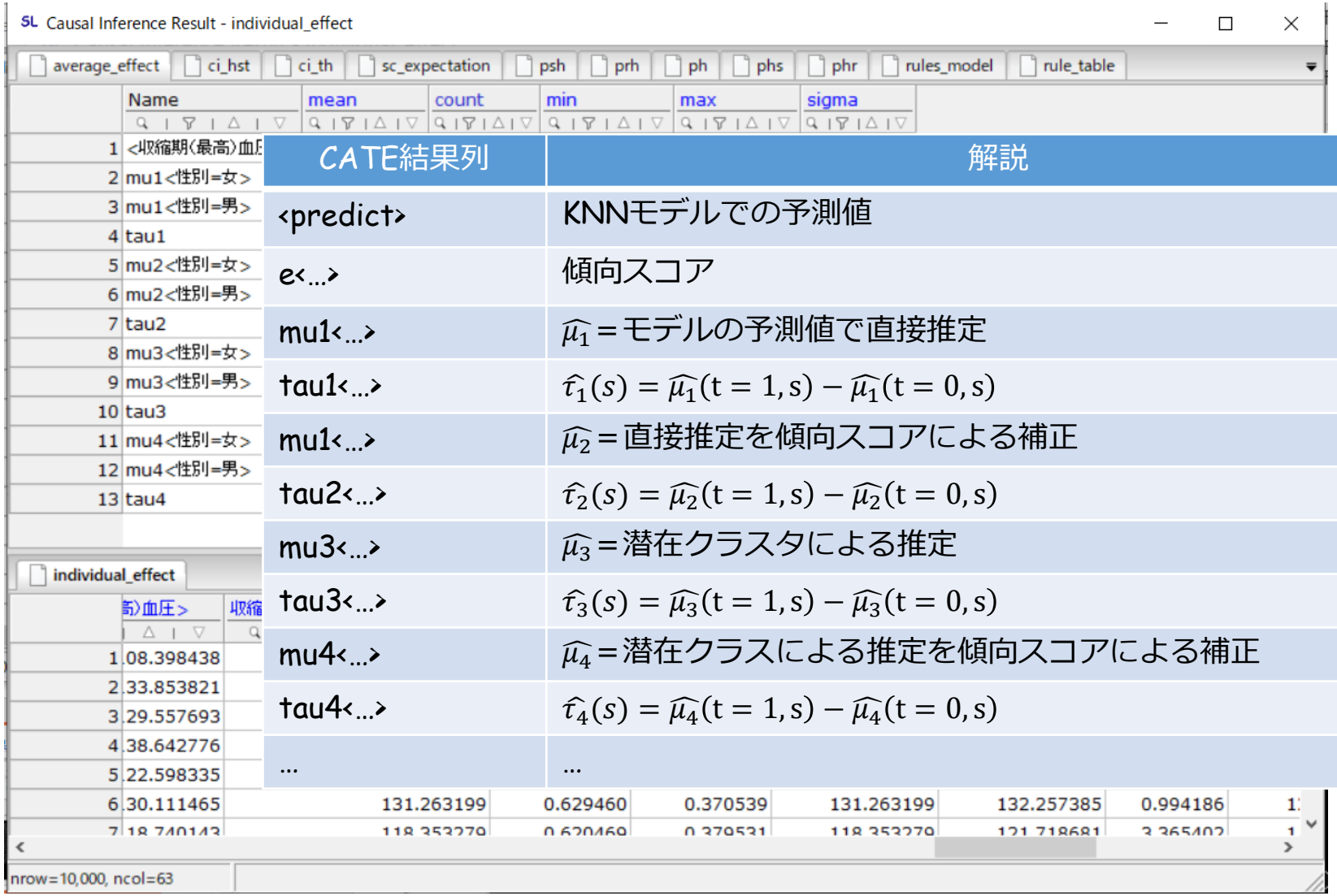
AI電卓での利用方法(学習)



AI電卓での利用方法(結果)



AI電卓での利用方法(結果)



CATE結果列	解説
<predict>	KNNモデルでの予測値
e<...>	傾向スコア
mu1<...>	$\widehat{\mu_1}$＝モデルの予測値で直接推定
tau1<...>	$\widehat{\tau_1}(s) = \widehat{\mu_1}(\mathrm{t}=1,\mathrm{s}) - \widehat{\mu_1}(\mathrm{t}=0,\mathrm{s})$
mu1<...>	$\widehat{\mu_2}$＝直接推定を傾向スコアによる補正
tau2<...>	$\widehat{\tau_2}(s) = \widehat{\mu_2}(\mathrm{t}=1,\mathrm{s}) - \widehat{\mu_2}(\mathrm{t}=0,\mathrm{s})$
mu3<...>	$\widehat{\mu_3}$＝潜在クラスタによる推定
tau3<...>	$\widehat{\tau_3}(s) = \widehat{\mu_3}(\mathrm{t}=1,\mathrm{s}) - \widehat{\mu_3}(\mathrm{t}=0,\mathrm{s})$
mu4<...>	$\widehat{\mu_4}$＝潜在クラスによる推定を傾向スコアによる補正
tau4<...>	$\widehat{\tau_4}(s) = \widehat{\mu_4}(\mathrm{t}=1,\mathrm{s}) - \widehat{\mu_4}(\mathrm{t}=0,\mathrm{s})$
...	...

参考動画(検索=「AI電卓」)

Youtube チャンネル

https://www.youtube.com/@ai_dentaku

動画（時間順）

1. 背景紹介
2. 背景紹介(コア技術)
3. データ可視化（見る）
4. 分析１：説明変数重要度計算、知識ルールを入力、編集、ラベリング
5. 分析２：モデル構築、評価、知識ルール抽出
6. 潜在クラスタ、モデルの予測根拠、知識発見
7. 因果推論(入門編)
8. 因果推論(AI電卓編)
9. Kaggleデータ分析：不動産価格予測
10. Conformal Prediction による予測信頼区間の推定

時系列予測に必要なラグ情報

- 機械学習のモデルは関数 $y = f(x)$ を求める過程である
- 時系列の場合は y は時間 t に依存することになり、通常 y_t と表現される
- 時間に依存することを利用、モデル（ f ）を構築するときに使える情報は、 x に加え、次の二種類の情報も利用可能
  - 時間軸（日付）から算出可能な情報（T）、例えば、週、月、旬、季節など
  - 歴史を表す（lag=ラグ）情報： 前日の売上（y_{t-1}）, 2 日前の売上（y_{t-2}）,…, k日前の売上（y_{t-k}）
    kは、ユーザ指定 1 より大きい自然数となります。
- 上記まとめると、y_t を予測するモデル（ f ）としては、次のようになる

$$y_t = f(y_{t-1}, y_{t-2}, \cdots, y_{t-k}, \mathrm{T}, x)$$