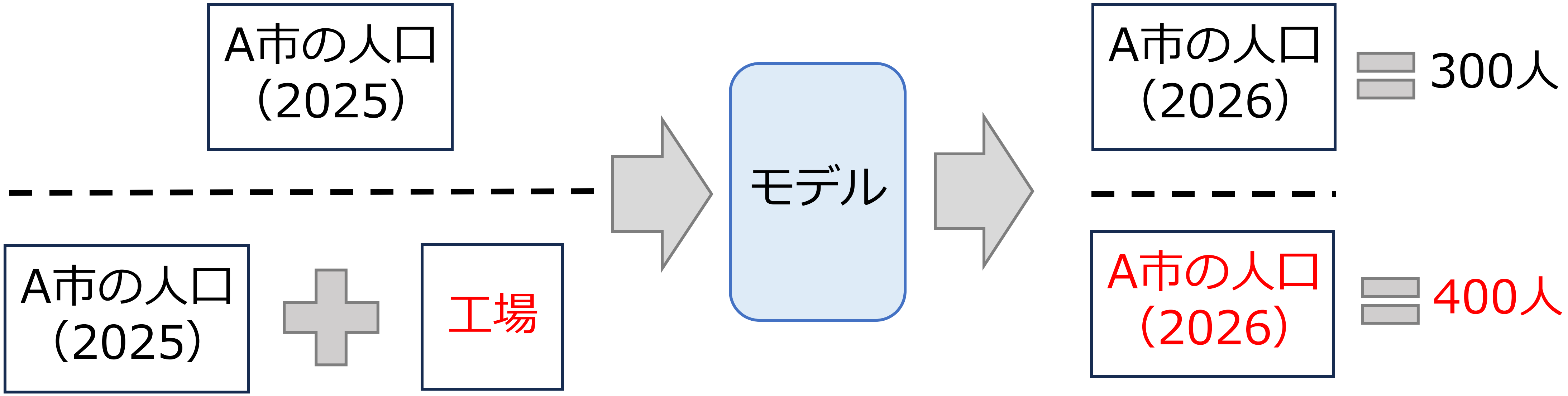
成果発表

2025/9/19

熊本大学大学院　自然科学教育部 M1
氏原　涼太

課題の趣旨

- 課題の趣旨
  - 人口とイベントの関係をモデル化
  - 将来のシナリオから人口を予測
    - 例）2026年A市に工場ができた場合の人口の変化を予測



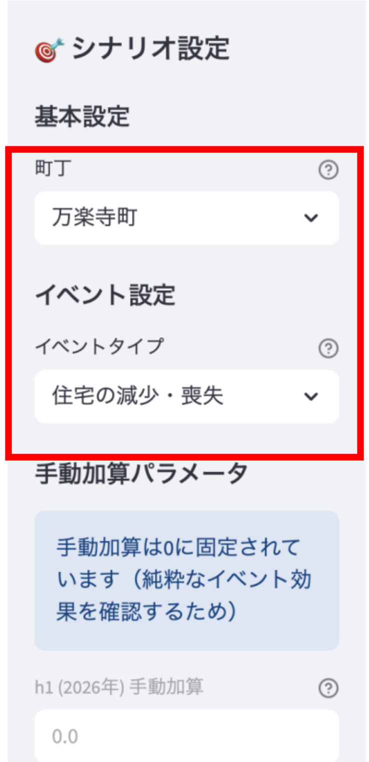
評価軸

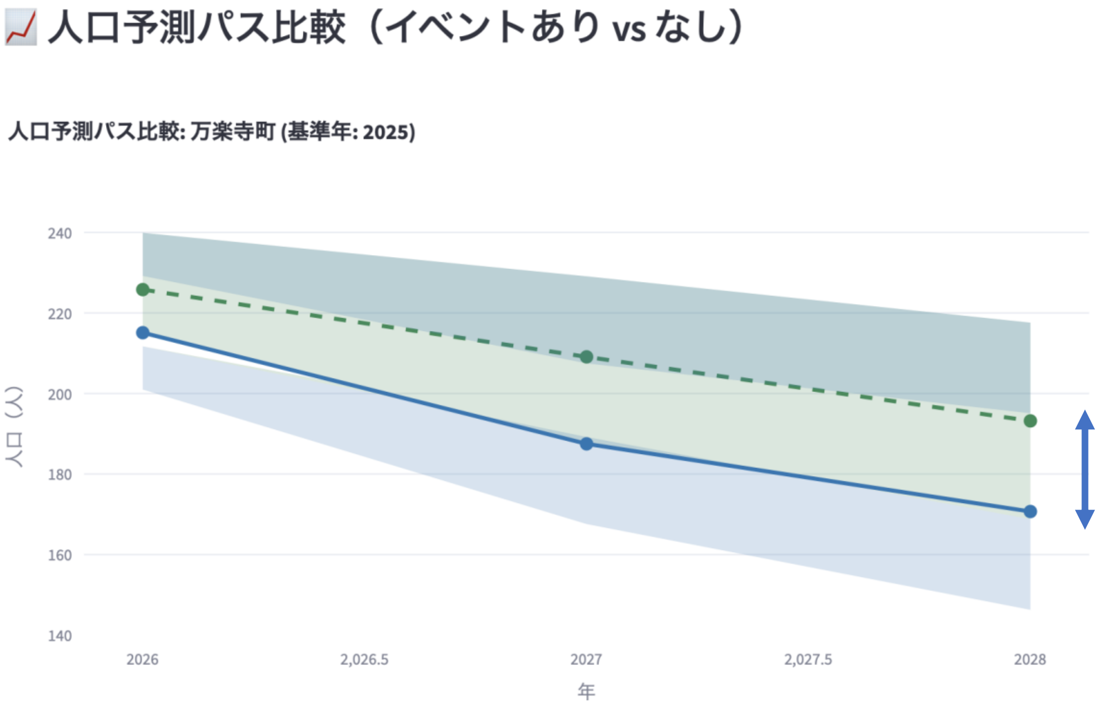
- 評価軸
  - **再現性**
    - データ，コード，手順が明確で第三者が再現可能
  - **説明可能性**
    - イベント寄与の内訳を説明可能
  - **精度の妥当性**
    - 単純ベースラインに対して，同等以上の安定性

成果物

- ダッシュボード（一部）

町丁と
イベント
を選択





22.5人の差

データ選定

- 熊本県熊本市
  - 町長単位で収集，モデル化
  - 1998年～2025年

- 選定理由
  - 地域に対する知識がある
    - 例1）A町にショッピングモールが建った
    - 例2）何年に地震があった外国人が増えた

  - 知識に基づくモデル設計が可能

データ前処理

- 20??年の政令市化に伴う表記揺れ
  - 例１）A町→A町（東区）
  - 例２）NULL→B町
  - 例３）C町～～丁目→C町

- 表記揺れの統一
  - A町（東区），A町（西区）はA町として扱う
  - B町は学習からは除外
  - C町~~丁目はC町として扱う

  - **最終的に669町丁　X　28年　の人口データ**

データ前処理

- 異常値を検出
  - 自然減少以外の人口の変動を検知
  - 方法を説明
- イベントデータ
  - 年次，町丁に基づく出来事
    - 約260件
    - 公的資料・公開情報を整理

イベント
マンション・アパート竣工
大型商業施設開業
学校・病院等新設
．．．

モデル設計

- モデル１：**二方向固定効果モデル**
  - **単純なイベントだけの効果を推定**
  - 予測値(i, t)＝町丁 i らしさ ＋ 年 t らしさ ＋ イベントの効果

$$\boxed{y_t^i = \alpha_i + \gamma_t + \beta}$$

y_t^i:予測値
α_i:地区 i らしさ
γ_t:年 t らしさ
β :イベント効果

- 最小二乗法でα_i， γ_t， βを推定
  - $\alpha_i = -10$：町丁 i は毎年-10人ずつ減っている
  - $\gamma_t = 20$：年 t は全町丁で平均20人増えている
  - $\boldsymbol{\beta} = \mathbf{30}$**：あるイベントが起きると30人増える**

モデル設計

- モデル２ ：**LightGBM**
  - **イベントを抜いた自然増減を捉え，将来を予測**
  - 勾配ブースティング決定木（GBDT）の改良版
    - 勾配降下法　＋　アンサンブル学習　＋　決定木

- 選定理由
  - 非線形・相互作用を拾える
  - 解釈しやすい
  - 学習・推論が高速

LightGBMの図

特徴量設計

- 特徴量
  - 人口（前年度，今年度）
  - イベント（マンション・アパート竣工など）
  - 時系列ラグ：イベントが何年後まで効果があるか
  - 外国人人口

分析設計

- 目的変数
  - 人口増減率

$$\boxed{r_t^i = \frac{\hat{y}_t^i - \hat{y}_{t-1}^i}{\hat{y}_{t-1}^i}}$$

$\hat{y}_t^i$ ： t 町の i 年の人口
r_t^i ： t 町の i 年の人口増減率

- 将来予測の算出方法

$$\boxed{y_t^i = \hat{y}_{t-1}^i(1 + r_t^i)}$$

y_t^i ： t 町の i年の人口予測値

  - 人口の大きさを考慮した予測が可能
    - 例）1000人から1100人　と　100人から200人は訳が違う

学習

- 学習方式
  - 時系列クロスバリデーション
    - 学習期間が一年ずつ伸びていく
    - (ステップ１）1999年-2017年→2018年を予測
    - (ステップ２）1999年-2018年→2019年を予測

- 損失関数
  - Huber損失
    - 外れ値にロバスト

$$L_\delta(y) = \begin{cases} \frac{1}{2}(y-\hat{y})^2 & (|y-\hat{y}| \leq \delta) \\ \delta\left(|y-\hat{y}| - \frac{\delta}{2}\right) & (|y-\hat{y}| > \delta) \end{cases}$$

評価方法

- 評価指標
  - 強みが異なる4つの指標で総合的に評価

評価指標	値範囲	目安	意味
MAE	0~無限	小さいほど良い	誤差の絶対値の平均
$RMSE$	0~無限	小さいほど良い	誤差の二乗の平均の平方根
$MAPE$	0~無限	小さいほど良い	実際の値に対して何%ずれているか
R^2	-無限~1	大きいほど良い	1：予測と実測値が同じ 0：平均値で予測したのと同じ 負：平均値予測より悪い

評価方法

- 比較対象（ベースライン）
  - ランダムウォーク（ドリフトなし）
    - 最低ライン比較対象

$$\boxed{N_t^i = \widehat{N}_{t-1}^i}$$

  - 移動平均
    - ファ

$$\boxed{N_t^i = \frac{1}{k}\sum_{j=1}^{k} \widehat{N}_{t-j}^i}$$

結果

- 精度
  - ベースラインに対して全ての指標で高い性能

モデル	$\boldsymbol{MAE}$ ↓	$\boldsymbol{RMSE}$ ↓	$\boldsymbol{MAPE}$ ↓	$\boldsymbol{R^2}$ ↑
ランダムウォーク	21.88	44.67	2.55	-1.01
移動平均	19.01	37.71	1.90	-0.30
LightGBM	**<u>2.62</u>**	**<u>11.08</u>**	**<u>0.70</u>**	**<u>0.87</u>**

結果

- 将来予測のデモを紹介
  - 図

課題

- **構造変化年の外れ**（住居表示変更・再開発・災害直後）への補正が未実装
- **将来イベントデータの運用設計**（更新頻度・責任部署・入力様式）
- **評価設計の強化**（W-MAPEの徹底、人口ビン別%誤差の可視化）

まとめ

- 課題
  - 人口とイベントの関係をモデル化
- 目的

補足資料

二方向効果モデル

- 最小二乗法
  - 残差二乗和（RSS）を最小にする係数を求める

$$RSS(\alpha,\gamma,\beta) = \sum_{i,t}\left(\hat{y}_{i,t} - \left(\alpha_i A_i + \gamma_t D_t + \beta E_{i,t}\right)\right)^2$$

  - 残差 $u = Y - X\theta$

$$X = \begin{bmatrix} A_1 & D_1 & E_{1,1} \\ A_1 & D_2 & E_{1,2} \\ \vdots & \vdots & \vdots \\ A_n & D_2 & E_{n,2} \end{bmatrix}$$

$$\theta = \begin{bmatrix} \alpha \\ \gamma \\ \beta \end{bmatrix}$$

$\hat{y}_{i,t}$:実測値
α_i:町丁 i らしさ
A_i:町丁の0/1
γ_t:年 t らしさ
D_t:年の0/1
β :イベント効果
$E_{i,t}$:イベントの0/1

二方向効果モデル

- 残差二乗和（RSS）を最小化＝微分が0

$$\boxed{RSS(\theta) = (Y - X\theta)'(Y - X\theta)}$$

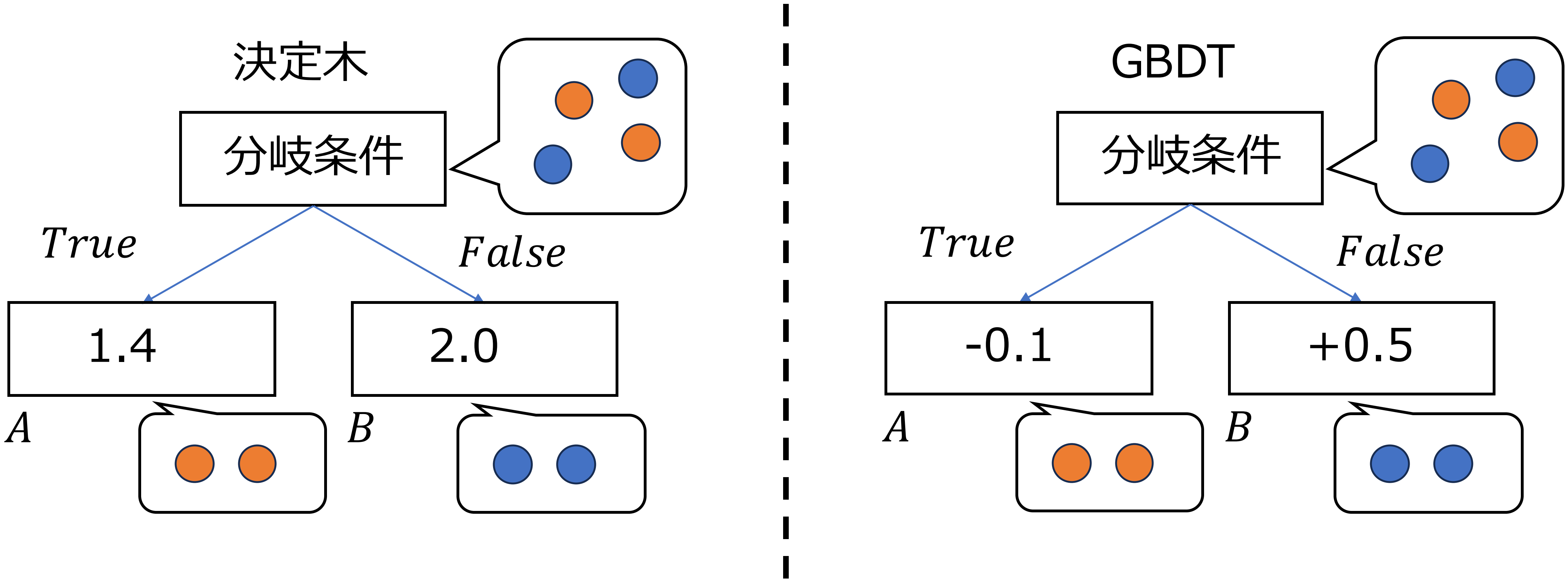
$$\boxed{\begin{aligned}\frac{dRSS(\theta)}{d\theta} &= (Y - X\theta)'(Y - X\theta) \\ &= -2X'Y + 2X'X\theta = 0\end{aligned}}$$

- $\theta = (X'X)^{-1}X'Y$　を解くと，イベント効果 β を推定可能

勾配ブースティング決定木（GBDT）

- 決定木：葉ノード=出力値
- GBDT：葉ノード=残差

実測値の平均=1.5

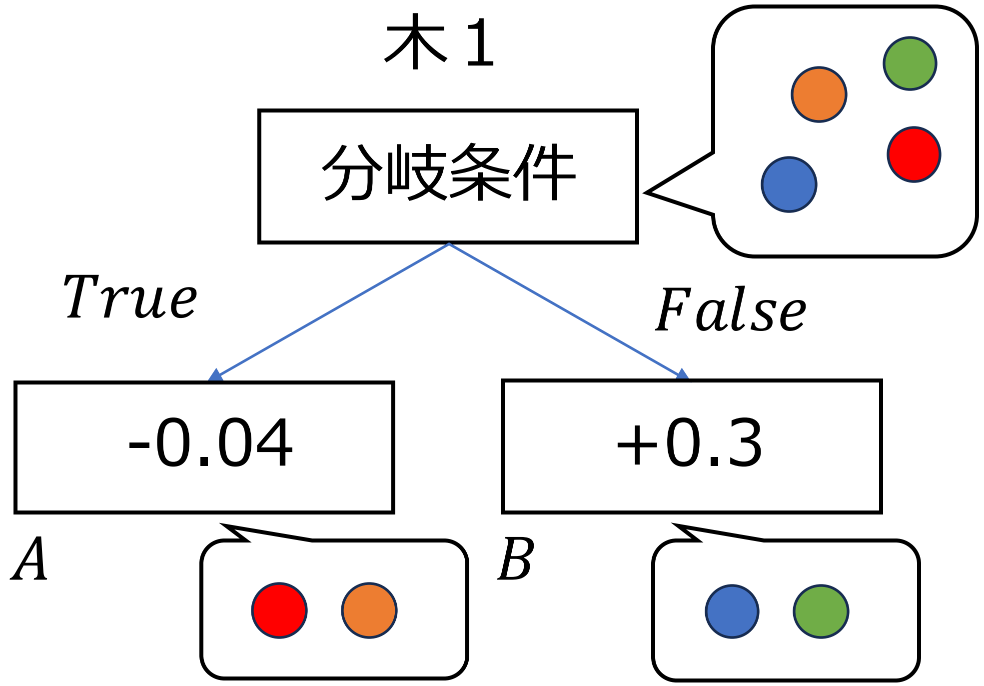


勾配ブースティング決定木（GBDT）

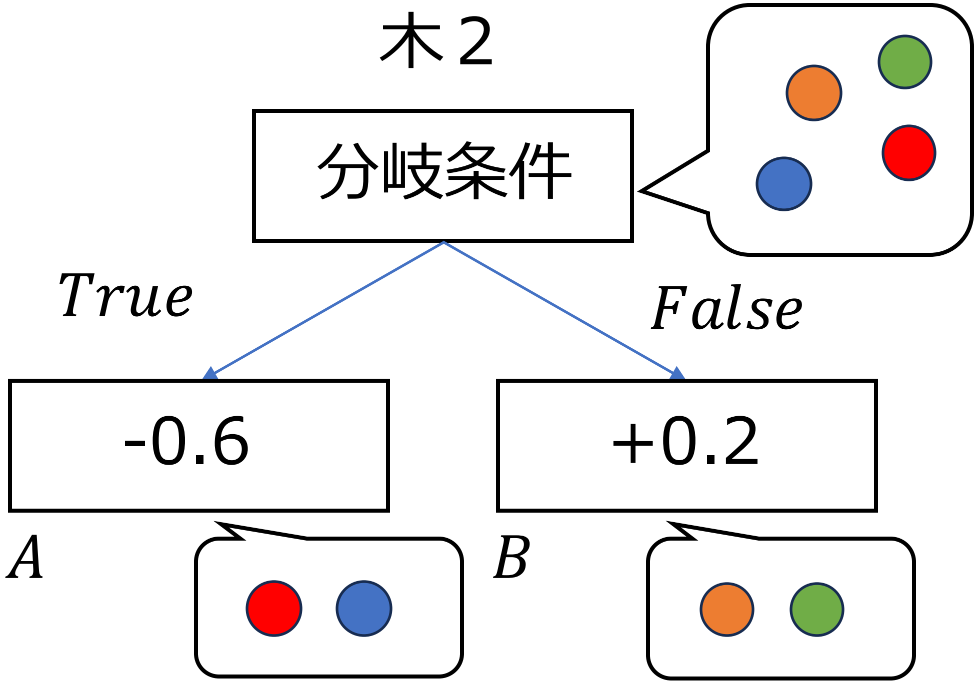
- GBDT：決定木＋アンサンブル学習＋勾配降下法
  - の実測値＝1.6　　　　の計算式=1.5-0.04-0.6=1.4

実測値の平均=1.5

木１



木２



勾配ブースティング決定木（GBDT）

- 分岐の決定法
  - 分岐： $\boxed{x_{i,j} < t_j}$
  - 特徴量分繰り返す
    1．閾値 t_j を変動
    2．損失 L を計算
  - Lが最も小さい（特徴量,閾値）のペア＝分岐

例)平均二乗誤差

$$L = \frac{1}{n}\sum_{i=1}^{n}\left(y_{i,j} - \bar{y}_j\right)^2$$

勾配ブースティング決定木（GBDT）

- 葉の出力
  - 出力： c=1.703
  - 全データの平均値 F_0 を計算
  - 分岐終了
    - 残差 r を計算
    - r に学習率 γ をかけて， F_0 に足す

$$r_k = \frac{1}{cout(Leef_k)} \sum_{i \in Leef_k} y_i - F_0$$

$$F_1(x) = F_0 + \gamma r_{k \in Leef_k}$$

- ２本目以降
  - 新しい決定木に従って出力

$$F_n(x) = F_{n-1} + \gamma r_{k \in Leef_k}$$

損失関数

- Huber損失
  - 閾値δが大きいほど外れ値に強い
  - 重い外れのあるデータにロバスト
    - 政令市化による町丁区域の変更などに対応

$$L_\delta(y) = \begin{cases} \frac{1}{2}(y-\hat{y})^2 & (|y-\hat{y}| \le \delta) \\ \delta\left(|y-\hat{y}| - \frac{\delta}{2}\right) & (|y-\hat{y}| > \delta) \end{cases}$$

y：予測値
$\hat{y}$：実測値
δ：閾値（0~1）

評価指標

- MAE（ Mean Absolute Error：平均絶対誤差）
  - 計算式

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}|$$

  - 範囲：0~無限 / 小さいほど良い
  - 強み：ハズレ値に引っ張られにくい
  - 弱み：大外しを強く罰しない

評価指標

- RMSE（Root Mean Squared Error：二乗平均平方根誤差）
  - 計算式

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}$$

  - 範囲：0~無限 / 小さいほど良い
  - 強み：大外しに敏感
  - 弱み：少しの外れ値で数字が大きくなりやすい

評価指標

- MAPE（Mean Absolute Percentage Error：平均絶対パーセント誤差）
  - 計算式 $$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$
  - 範囲：0%~無限% / 小さいほど良い
  - 強み：スケールに依存しない
  - 弱み： yが小さいと爆発

評価指標

- R^2 （決定係数）
  - 計算式

$$\boxed{R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2}}$$

  - 範囲：-無限~1 / 大きいほど良い
  - 強み：モデルが分散をどれだけ説明したかを表す
  - 弱み：集計の仕方で値がぶれる

結果

- アブレーション
  - 外国人〜〜

モデル	MAE ↓	RMSE ↓	MAPE ↓	R^2 ↑
フルモデル	21.88	44.67	2.55	-1.01
移動平均	19.01	37.71	1.90	-0.30
LightGBM	**<u>2.62</u>**	**<u>11.08</u>**	**<u>0.70</u>**	**<u>0.87</u>**