日報（2019/11/19）

株式会社イノヴァストラクチャー

小林　悠

目的

営業費分析と予測の目的

- 毎年、予算の基準値や単価作成など多大な労力を要している
  → 人の手で長い時間をかけている

- 機械学習で予測し、参考値の算出、延いては自動化を図る

- 作りたいもの

扱ったデータについて

元データ

- 1,105,715件（加工前）
- AC（勘定項目） #ユニーク 197項目
- Dept（地区） #ユニーク 1200か所
- Date（年月）

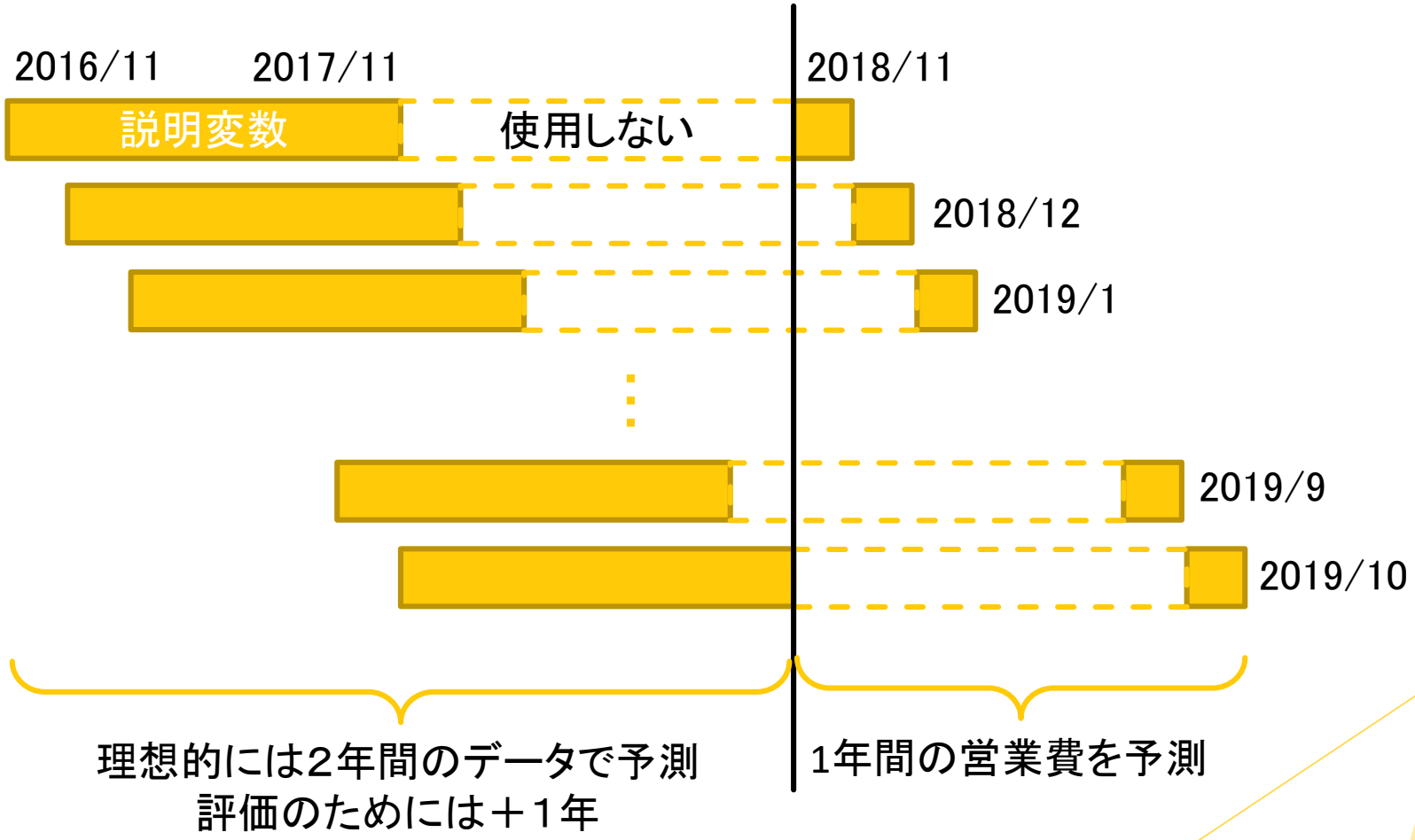
  期間：2015/12 – 2019/10
- Value（費用）

  負の値もある e.g. 収入、減給、積立金など

予測と評価の方針

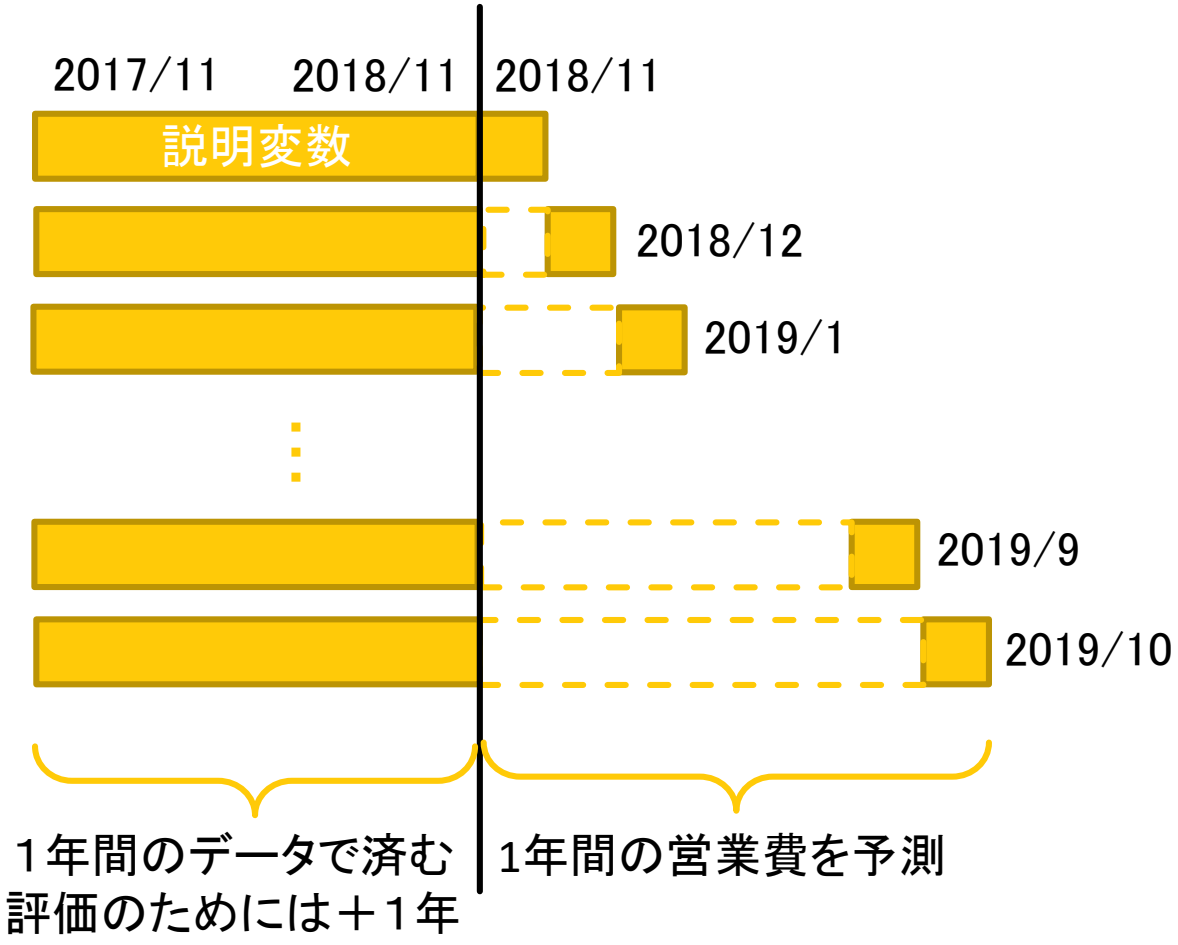
項目×地区×年月(1年後)の費用を予測

- 予測のイメージ



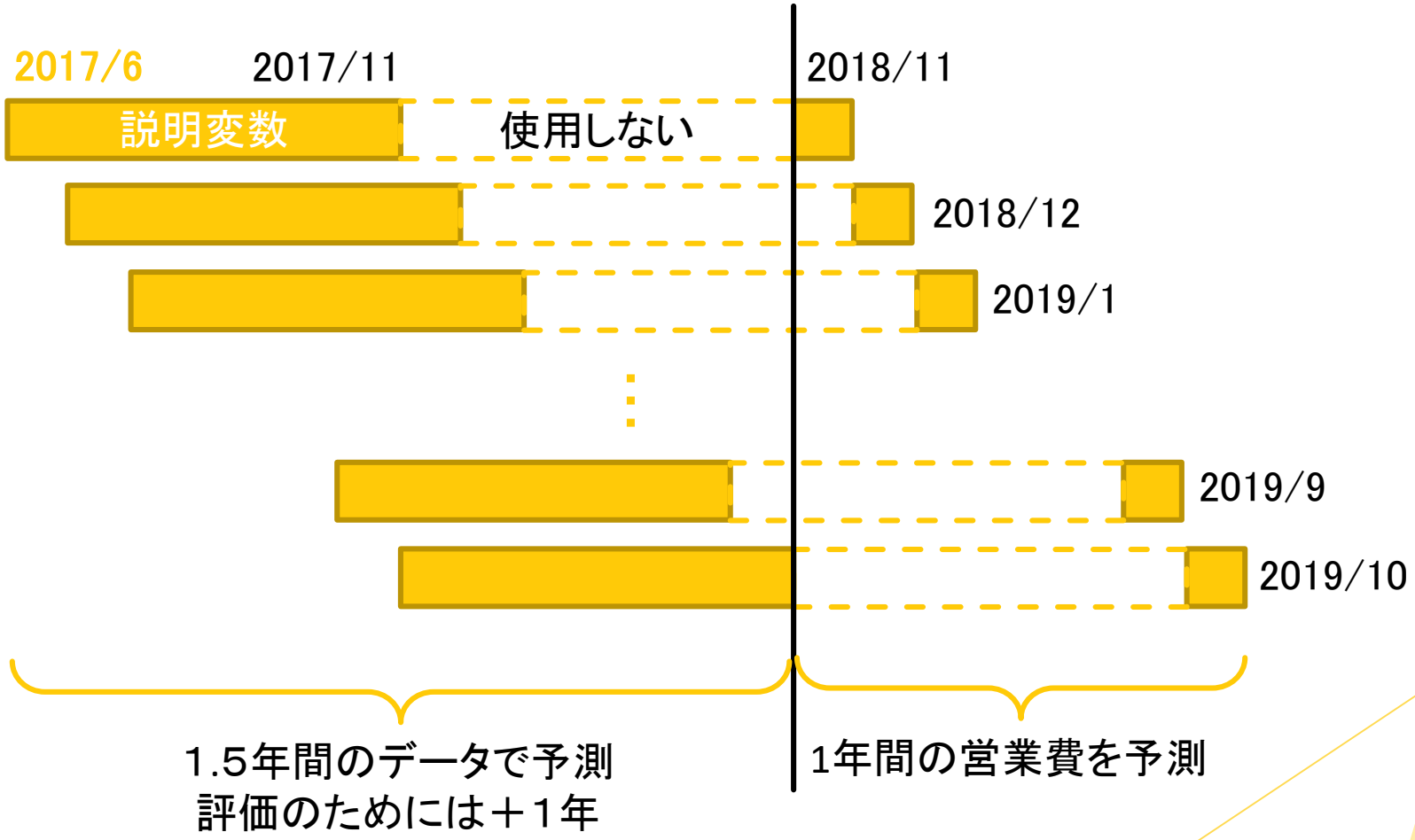
妥協案A １年間のデータから翌年をすべて予測

- 予測のイメージ

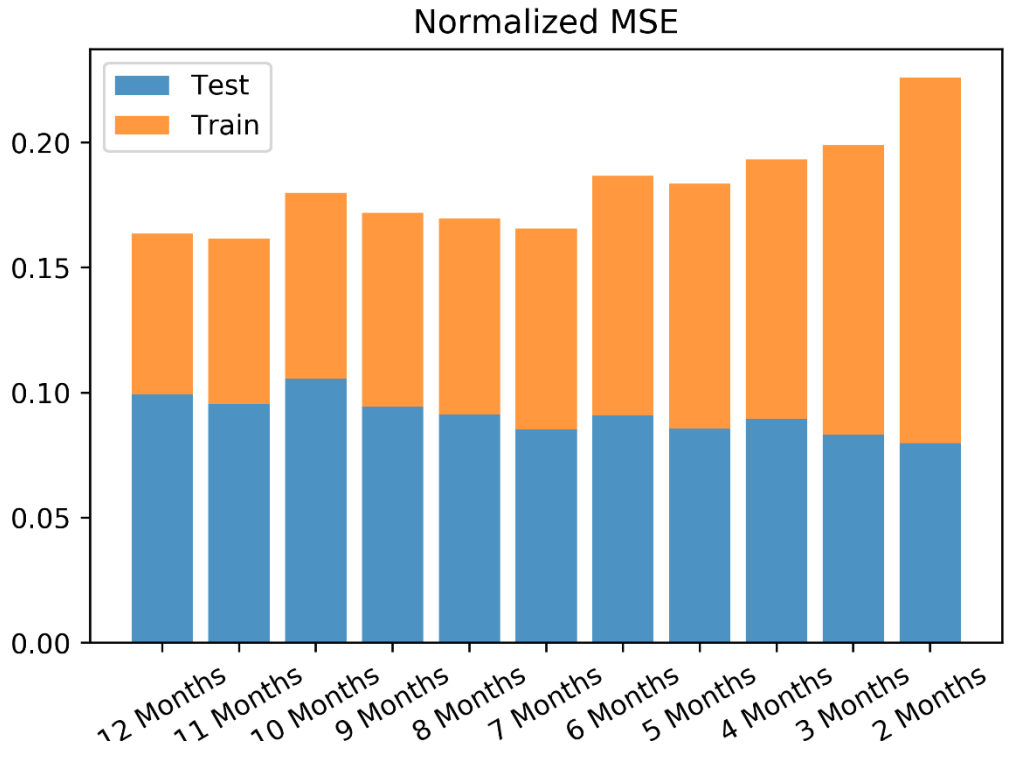


妥協案B 説明変数の期間 or 予測の遅延を短くする

- 予測のイメージ



1か月予測の結果

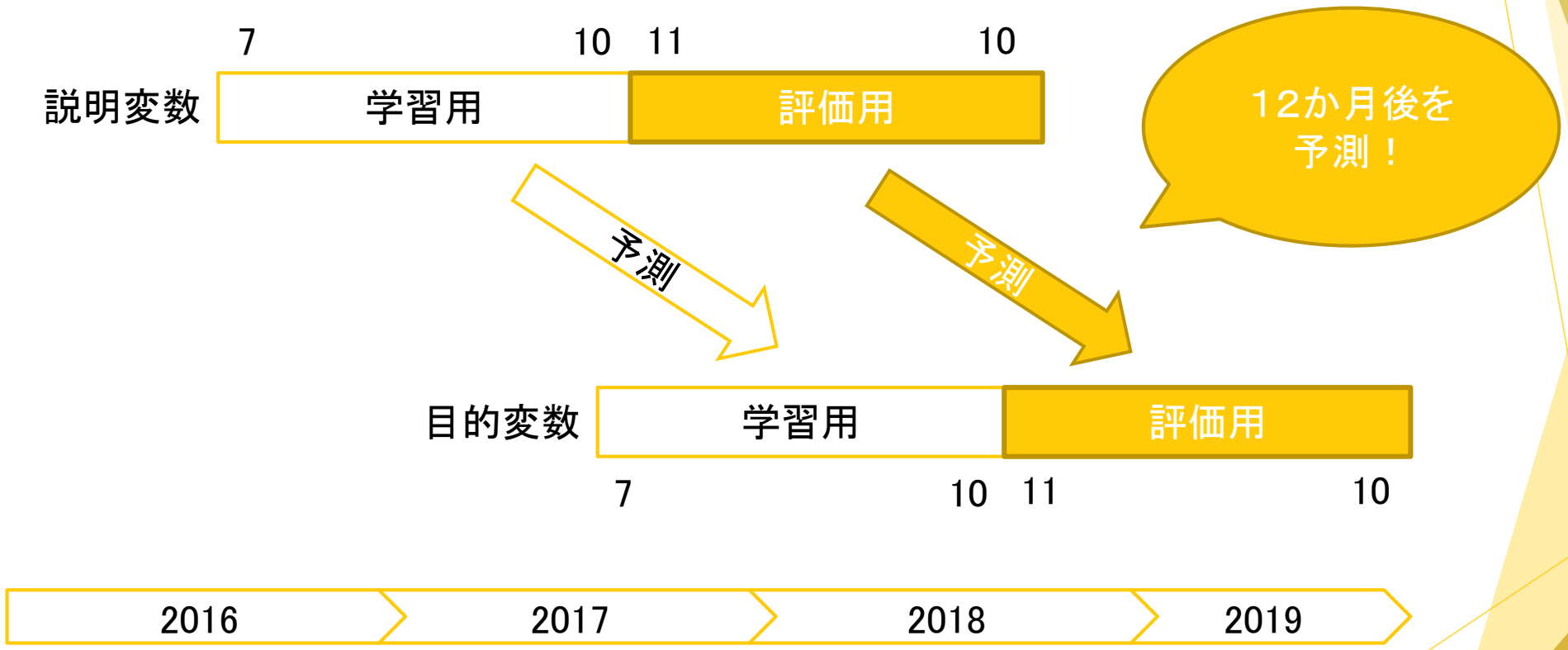


- 訓練用データ
  - 考慮する期間が長くなるほど
  - 誤差は小さくなっている
- 評価用データ
  - 考慮する期間を長くしても
  - 誤差が小さくならない
  - 半年間くらいが妥当か

説明変数の期間を短くする方針Bにする！

評価のしかた

- 学習とテスト（評価）



評価指標

- 平均二乗誤差（MSE：Mean Squared Error）

  （予測値 − 正解）^{2} の平均

- 平均絶対誤差（MSE：Mean Absolute Error）

  | 予測値 − 正解 |（絶対値）の平均

データ全体で
どのくらい誤差が出たか

小さいほど良い！

- 予測値（x軸）と正解（y軸）の散布図

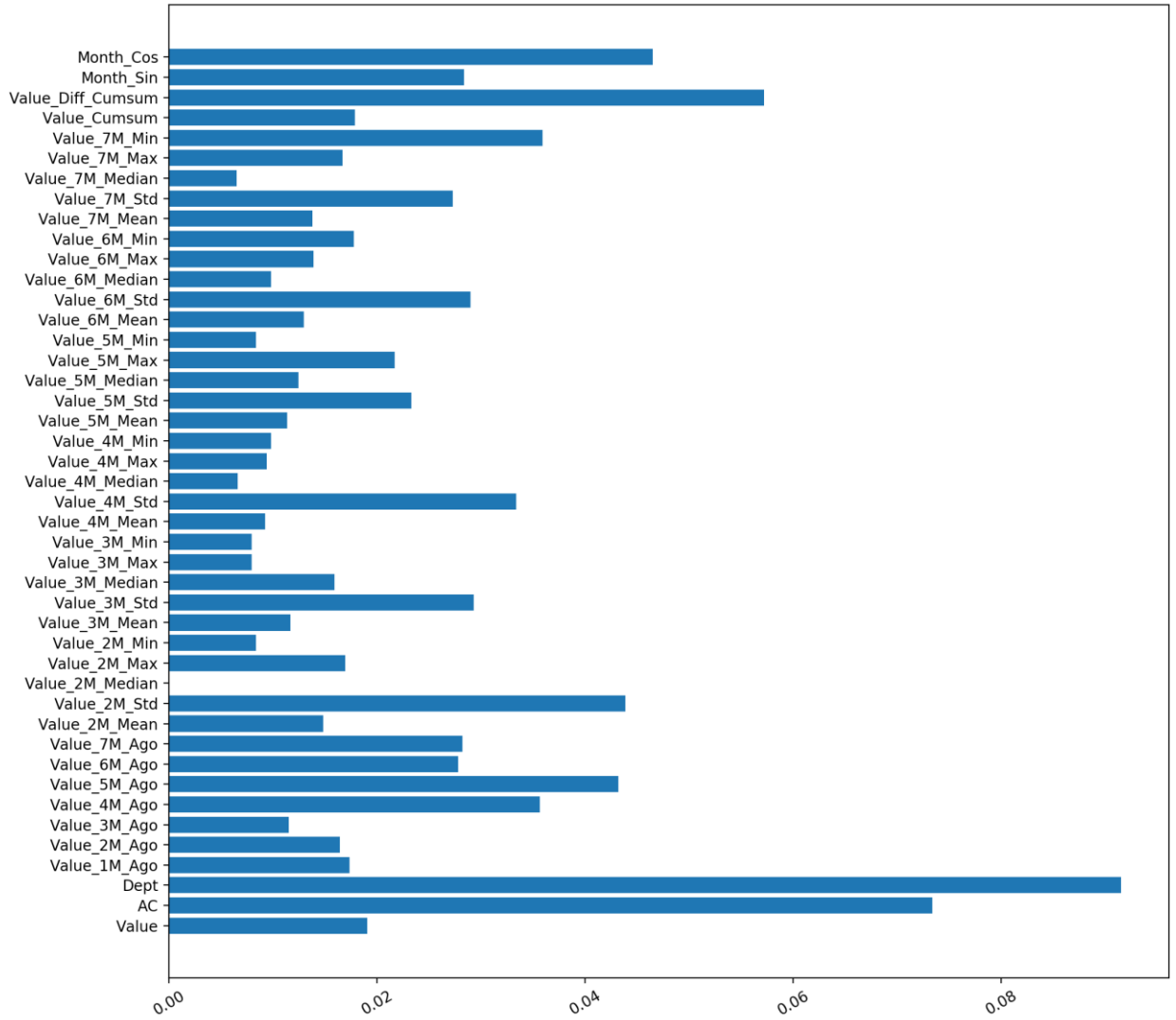
  理想的な予測では「予測値＝正解」となるので

  直線 y = x に近く散布しているか確認

特徴量の設計

特徴量として使った移動統計量など

- 過去の値
- 平均
- 標準偏差
- 中央値
- 最大値
- 最小値
- 累積和
- 差分の累積
- 月のSin、Cos



予測結果

誤差

- 平均二乗誤差（単位をそろえるために平方根をとっている）
  - ✓ Train

    678032.7862511722
  - ✓ Test

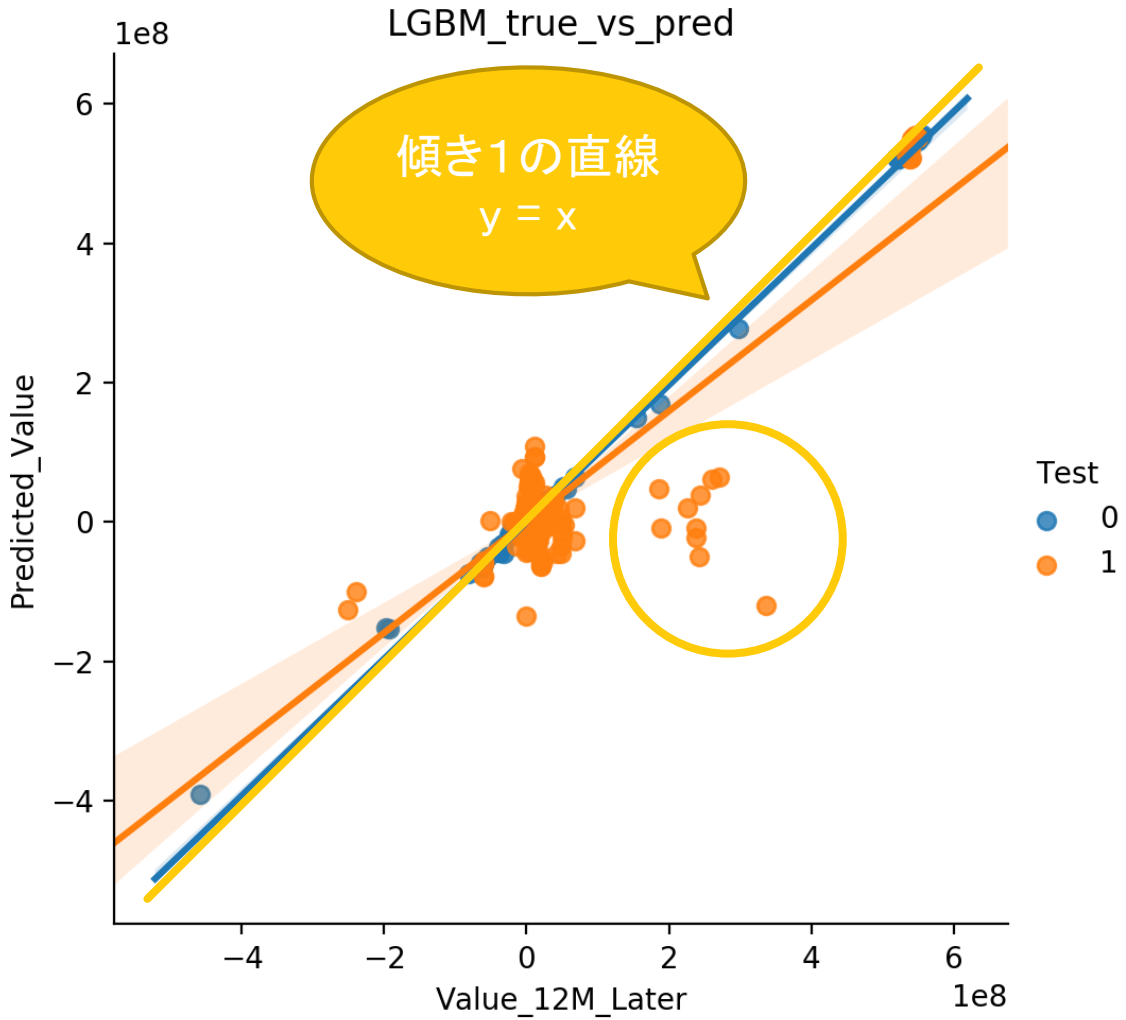
    2167669.4083287506

- 平均絶対誤差
  - ✓ Train

    260.7358382376772
  - ✓ Test

    352.3213493464561

予測の評価



- 訓練用データ
  - ほぼ1対1対応で予測できている
- 評価用データ
  - 回帰直線が下振れ

【原因】

- ✓ 2億付近で低く予測している
- ✓ 原点付近にも誤差あり