

2nd Place Solution

飯田産業
土地の販売価格の推定

2019.10.11

三船 哲史

目次

コンペ概要

解法の紹介

- ✓ 解法の概要
- ✓ 特徴量
- ✓ モデル作成
- ✓ その他

目次

コンペ概要

解法の紹介

- ✓ 解法の概要
- ✓ 特徴量
- ✓ モデル作成
- ✓ その他

コンペの概要とデータ

- ✓ 土地（**+建物**）の販売価格を予測
- ✓ 学習用データ
 - 現場：2780件
 - 号棟：6461件（土地売り：360件）
- ✓ 評価用データ
 - 現場：1855件
 - 号棟：4273件（土地売り：216件）



評価指標

- ✓ MAPE（平均絶対誤差率）
- ✓ 目的変数が小さいほど誤差に厳しくなる

プラン	建物面積	契約金額	予測値	誤差
土地売り	0	10,000,000	11,000,000	10 %
3LDK	150 m ²	20,000,000	21,000,000	5 %

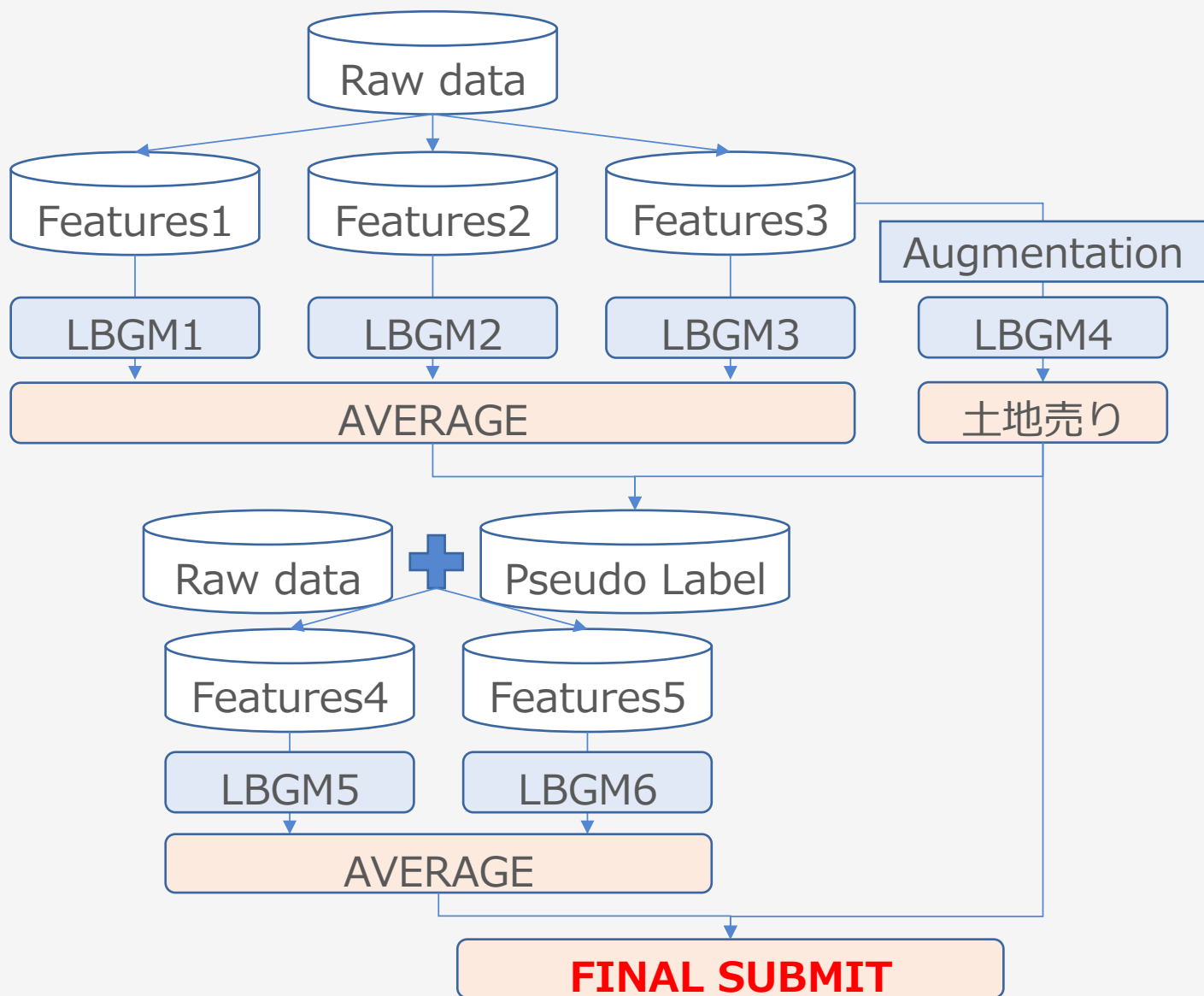
目次

コンペ概要

解法の紹介

- ✓ 解法の概要
- ✓ 特徴量
- ✓ モデル作成
- ✓ その他

解法の概要



- ✓ 特徴量セットを複数作成
- ✓ モデルはLightGBMのみ
- ✓ 土地売りデータのみ別モデルで予測値を置換
- ✓ **Pseudo Label**を用いて、評価用データも学習時に使用

目次

コンペ概要

解法の紹介

- ✓ 解法の概要
- ✓ 特徴量
- ✓ モデル作成
- ✓ その他

特徴量（前処理）

✓ 誤入力や表記ぶれを直す

- 埼玉県埼玉県さいたま市 → 埼玉県さいたま市
- ひばりが丘、ひばりヶ丘 → ひばりヶ丘に統一

✓ 特徴量の分割

プラン		階数	部屋数	納戸
土地売り	→	0	0	0
2F/3LDK		2	3	0

特徴量（ドメイン）

- ✓ 南側に道が接しているか
- ✓ 駅からの距離
 - 徒歩は80m/分、車は400m/分
 - 車で10分 → 徒歩で50分に変換して合わせる
- ✓ **路線価の欠損値をLightGBMで予測**
 - 路線価が0の現場が存在（倍率方式の地域？）

特徴量（一般的な手法）

✓ Aggregation（カテゴリで集約）

- Ratio : 土地面積100, 平均値50 → $\text{Ratio} = 100/50$
- Diff : 土地面積100, 平均値50 → $\text{Diff} = 100 - 50$

✓ 数値列の四則演算

- 基準価格 - 路線価, 土地面積 + 建物面積...

目次

コンペ概要

解法の紹介

- ✓ 解法の概要
- ✓ 特徴量
- ✓ モデル作成
- ✓ その他

モデル作成

✓ LightGBMのみ使用

- CatBoostを試すもスコアがいまいち
- NNも試せるようになりたい...

✓ パラメータ調整

- 手動チューニング

モデル作成（学習）

✓ Group 5-Fold CV

- 現場でGroup化
- TrainとValidに同じ現場が存在しないように

✓ 目的変数を $[0,1]$ で正規化し、Xentropyを使用

✓ 特徴量の評価

- CVの平均値とfeature_importance眺めてました
- 最終的にはfeature_importanceの上位N個を使用

目次

コンペ概要

解法の紹介

- ✓ 解法の概要
- ✓ 特徴量
- ✓ モデル作成
- ✓ その他

Pseudo Labeling

- ✓ データ量が少なかったため使用
- ✓ 全評価用データを学習用データに追加
- ✓ Pseudo Labelで学習した結果を再度Labelとして使用
- ✓ Pseudo Labelなし → 最終順位6位相当
 - ✓ スコア 8.39625 → 8.22340

Data Augmentation

- ✓ 少数だが、誤差が大きい土地売りへの対策
 - 土地売りの物件の面積と価格をN倍して水増し
 - 建物の価格を $1\text{m}^2:80000$ 円とみなす

建物面積	契約金額
200	40,000,000
100	16,000,000



建物面積	契約金額
0	24,000,000
0	8,000,000

-16,000,000

-8,000,000

Ensemble

- ✓ 特徴量セットをいくつか作成
- ✓ 各モデルはそれぞれSeedAverage(x5)
 - ✓ LightGBM (gbdt, Xentropy)
 - ✓ LightGBM (gbdt, RMSE)
 - ✓ LightGBM (dart, RMSE)

うまくいかなかったこと

- ✓ 住居表示に対応する緯度・経度を用いてkNN
- ✓ catboost
- ✓ 都市でGroupFold
- ✓ PCA等での次元削減
- ✓ 現場単位の合計価格を予測
- ✓ その他いっぱい

終わってから思いついたこと

- ✓ 評価指標に対して予測値を最適化できていない
 - MAPEでは得られた予測値より小さくした方が良い
 - 予測値 $\times 0.99$ すると手元のCVでスコア 0.08 向上

予測値	真の値	誤差
10,000,000	11,000,000	9.09
10,000,000	9,000,000	11.11

MAPE : 10.1

$\times 0.99$
→

予測値	真の値	誤差
9,900,000	11,000,000	10.0
9,900,000	9,000,000	10.0

MAPE : 10.0

まとめ

【特徴量作成の流れ】

- ① データクレンジング
- ② ドメインベース
- ③ 一般的な手法

【特に効いた取り組み】

- ① **路線価**の予測
- ② 部分的なデータの**水増し**
- ③ **Pseudo Labeling**

ありがとうございました