Kaggle研修 成果報告

2023/2/27

Hitachi, Ltd. Data Scientist: Fuki Yamamoto

Table of Contents

1. Overview of Competition

- ✓ コンペ名: PetFinder.my - Pawpularity Contest
- ✓ データセット: 各ペットの画像データ、画像から抽出した特徴量のテーブルデータ
- ✓ タスク: ペット里親発見プラットフォームにおける、
  ペットごとのWebページへのアクセス統計から計算された”Pawpularity”を推定

id	subject focus	eyes	face	near	action	accessory	group
0007de18844b0dbbb5e1f607da0606e0	0	1	1	1	0	0	1
0009c66b9439883ba2750fb825e1d7db	0	1	1	0	0	0	0
0013fd999caf9a3efe1352ca1b0d937e	0	1	1	1	0	0	0
0018df346ac9c1d8413cfcc888ca8246	0	1	1	1	0	0	0
001dc955e10590d3ca4673f034feeef2	0	0	0	1	0	0	1

テーブルデータ

EDA

EDA

画像データ

2. Solution and Results

Solutions:

✓ 画像データのみでタスクを解く

✓ 目的変数をbinningして各binに対して層化KFoldCVを適用

✓ 0~1にスケールして二値分類問題として解く

✓ pre-trained ResNet18をfinetunigする

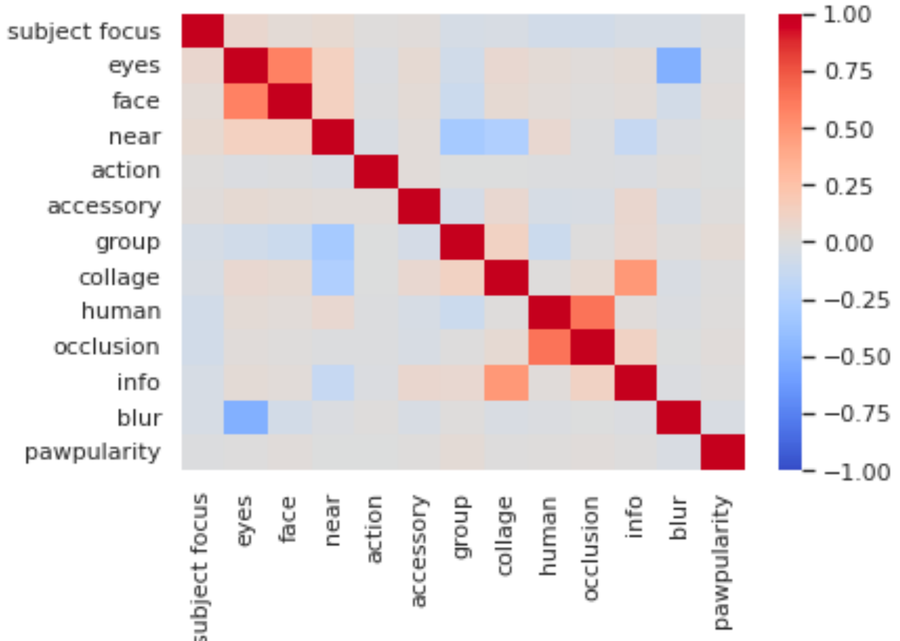
Results:

Public 19.22849 (2290 / 3537)

Private 18.90042 (2308 / 3537)

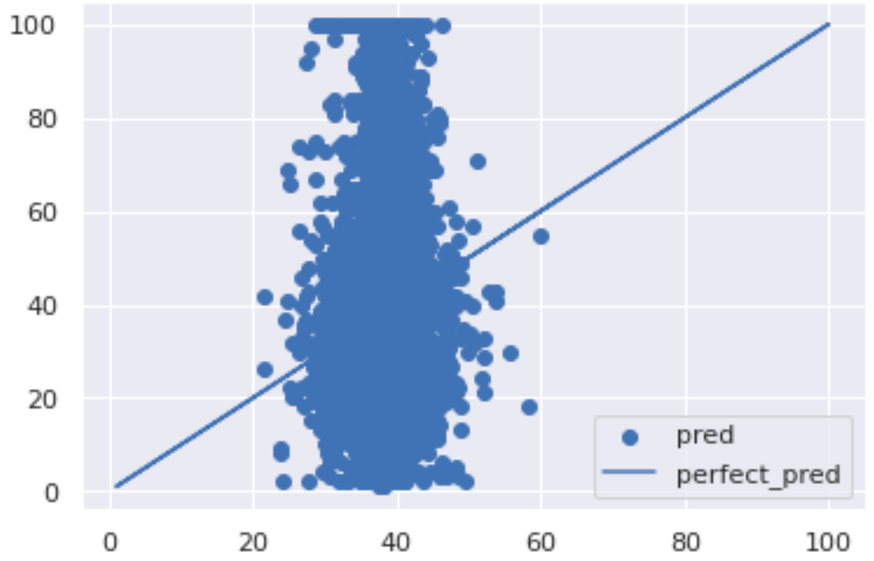
3-1. Only Tabular Data

- ✓ テーブルデータのみを使用（モデルはLightGBMを使用）
- ✓ 平均値で予測した場合 RMSE = 20.59



相関係数のヒートマップ

Pawpularityと相関を持つ変数はない



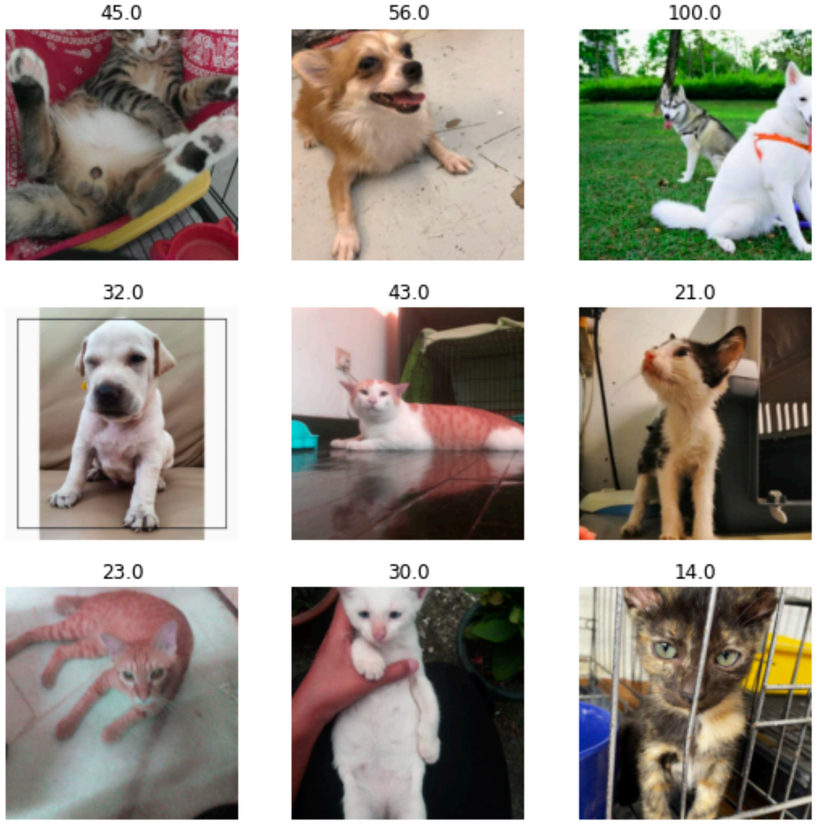
LightGBMによる予測 RMSE = 20.66

ほとんど一定の予測を行っている

数値も平均値予測とほとんど変わらない

3-2. Only Image Data

- ✓ 画像データのみを使用（ライブラリはfastaiを使用）
- ✓ 予測モデルはResNet18を選択 (速さと精度のバランスが良いため)

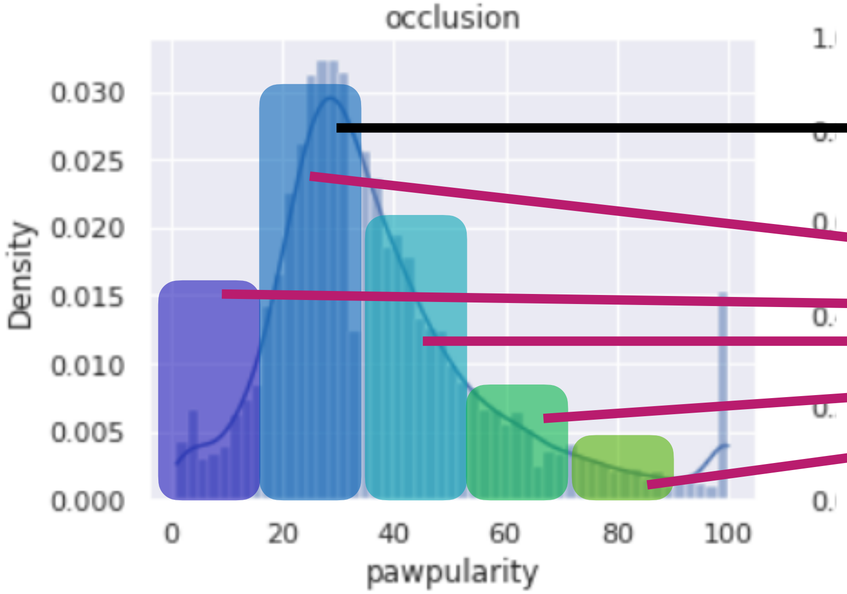


- ✓画像とPawpularityの組み合わせ
- ✓動物の種類に大きく影響を受けているように見える

3-3. Stratified Kfold

✓　目的変数”Pawpularity”をビニング（スタージェスの公式でbinの数を決定）

✓　各ビンに対して偏らないようにfoldを生成



右側に裾が重い分布

✓普通のKFoldCV

→ 偏ったデータを抽出する可能性あり

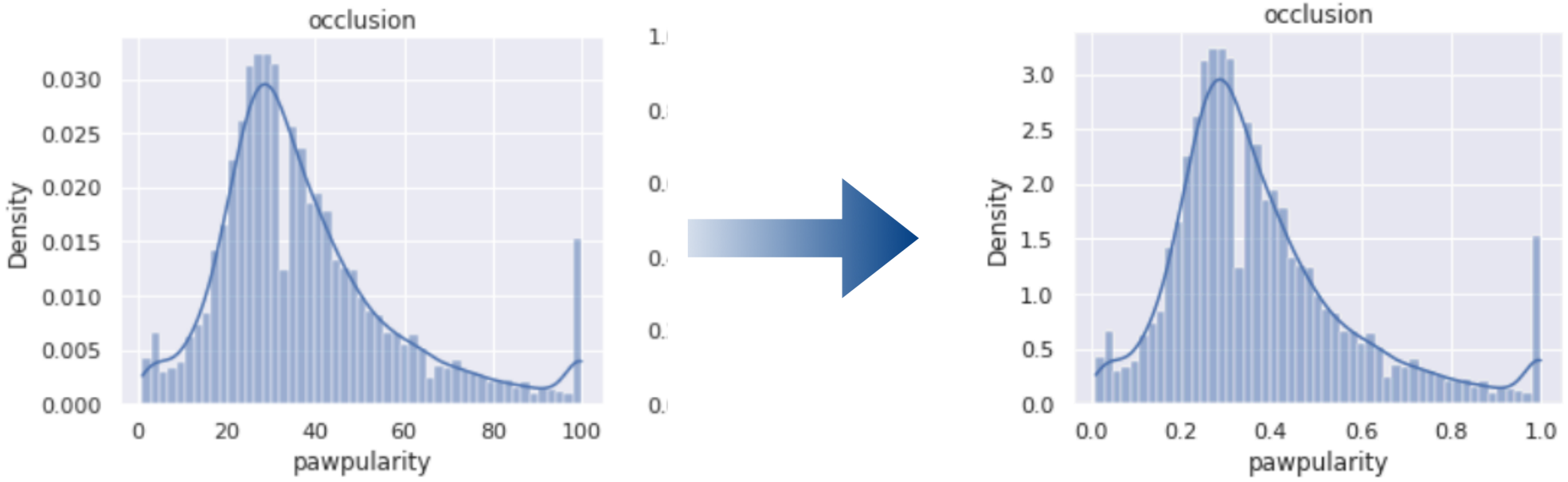
✓Stratified KFoldCV with Binning

→ 目的変数に対しては偏りなく抽出することを保証

3-3. Convert to Classiffication Task

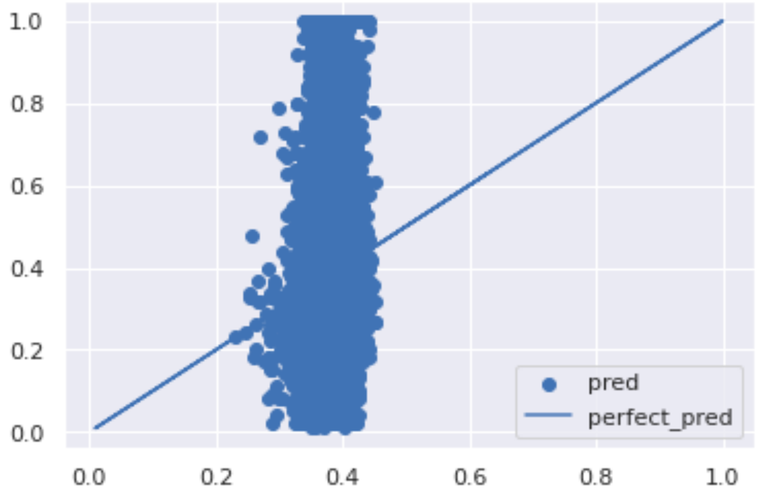
- ✓ 目的変数は0~100に整数値で分布している
- ✓ 100で割ることで0~1の定義域に変換することで二値分類として解けるようにする

  → 出力が0~1(0~100)になることを保証できる

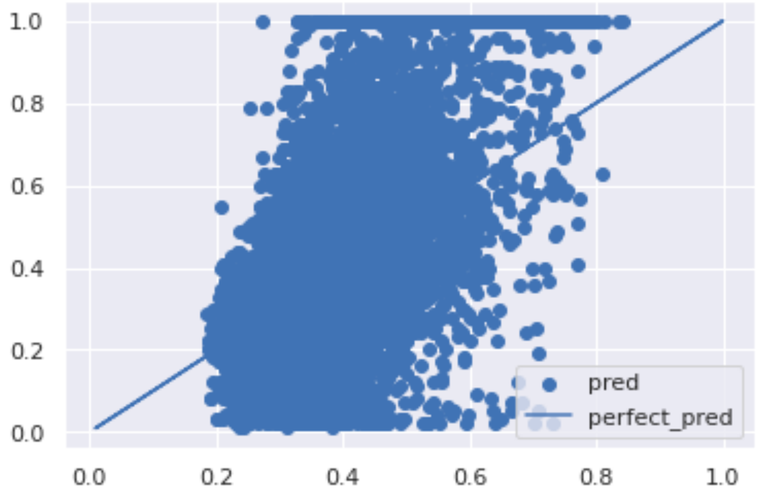


3-3. Pre-trained Model (Image task)

- ✓ timmに格納されているPretrainedモデルを使用
- ✓ Pretrainedモデルをfine tuningしたモデルが最も高精度を示した

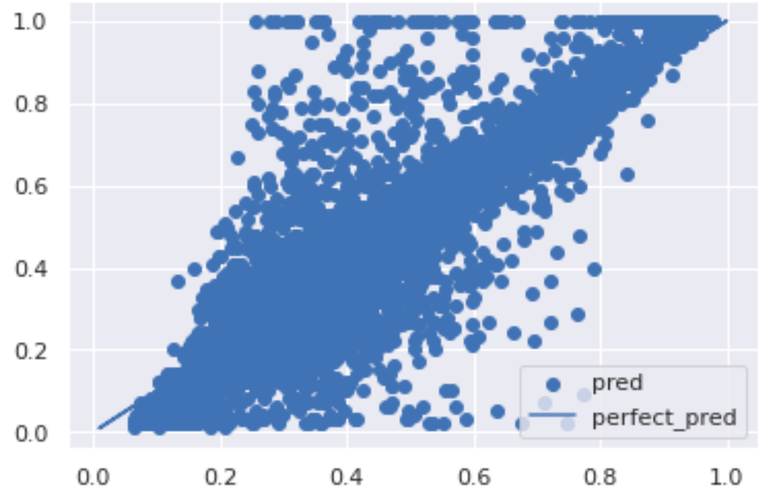


Pretrained modelなし



Pretrained model

+ Headerのみ学習



Pretrained model

+ Fine Tuning

4. Reflection

✓ 新人全体研修などと時期が重複してしまい時間が確保できなかった

✓ 画像タスクは初めてであったため, 基本を調べるのに時間が取られた

今後

✓ プライベートの時間を使用してコンペに取り組んでいきたい

✓ 自然言語など、他のタスクにも挑戦していきたい