SIGNATE Cup 2024

学生部門1位(全体7位)解法

目次

- 自己紹介
- コンペ概要
- 解法
- まとめ

自己紹介

ユーザー名:matu997

名前:松本 大樹

所属:電気通信大学1類　学部2年

趣味:サイクリング、3Dモデリング

コンペ概要

- 顧客データを元に旅行パッケージの成約率を予測

学習データ	3489 列 × 18 行
テストデータ	3556 列 × 17 行

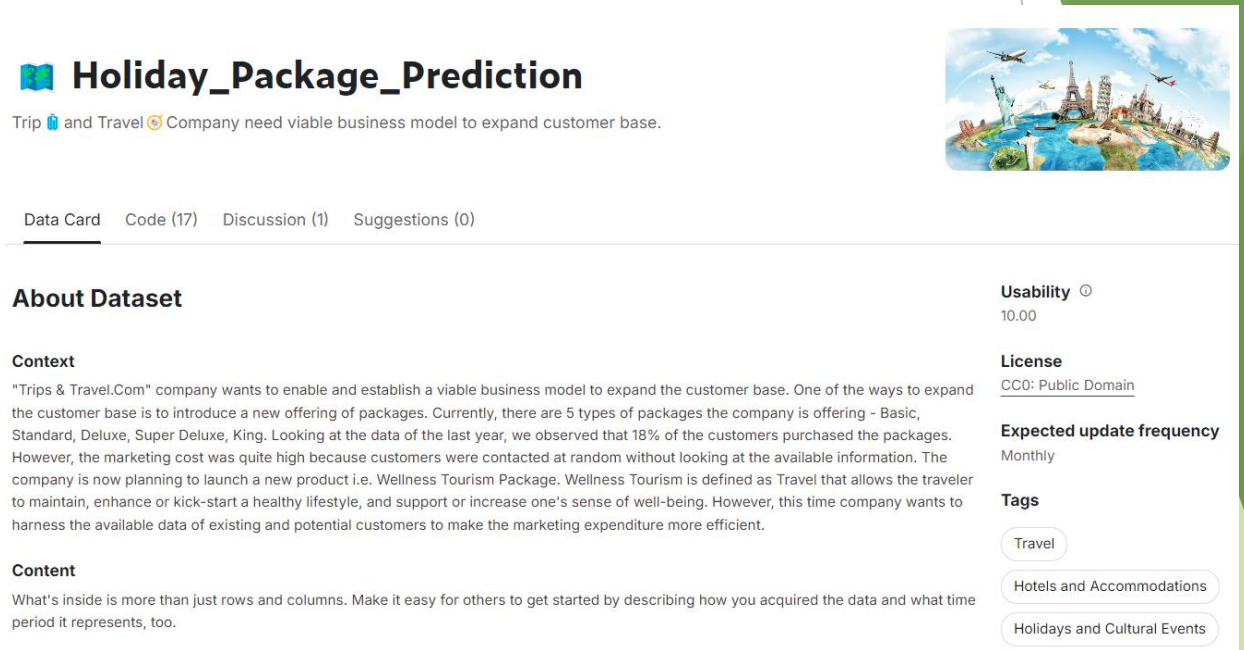
- 評価指標はAUC

今回のデータの特徴として合成されたデータセットである

kaggleデータセット

4888 列 × 20 行

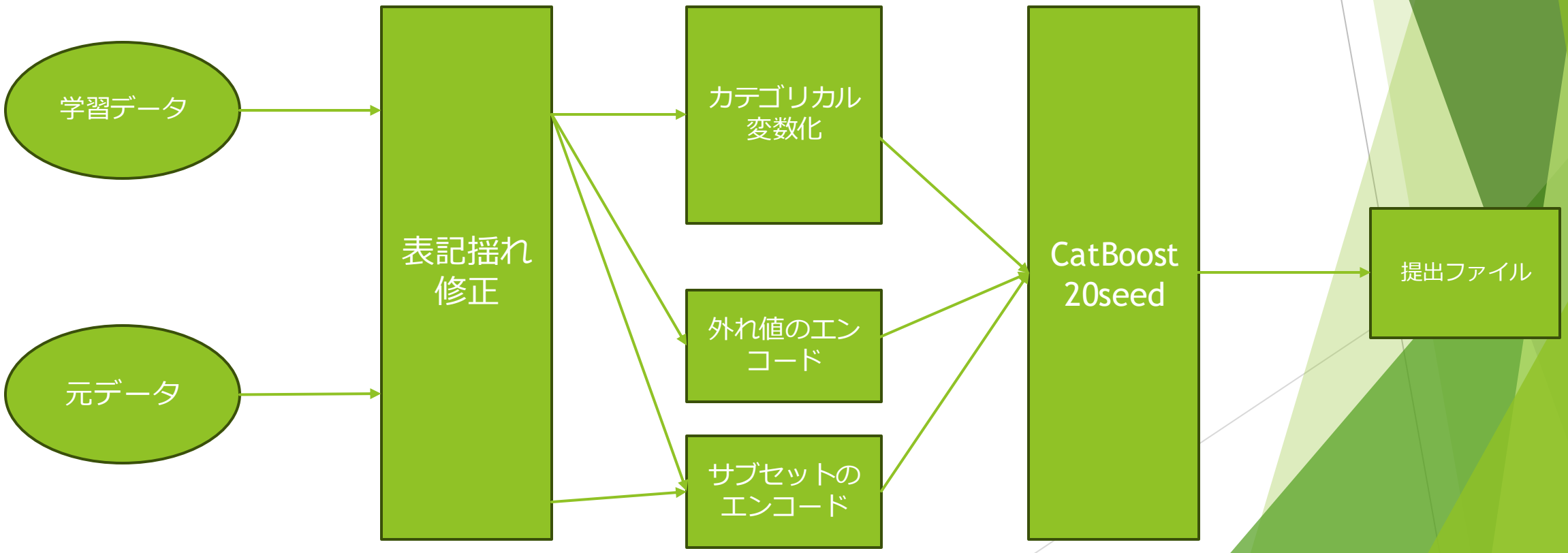
→与えられたデータの customer_infoを3つに分離すると行数が揃う



customer_info		MaritalStatus	OwnCar	NumberOfChildrenVisiting
未婚 車未所持 子供なし	→	Unmarried	0	0
離婚済み、車所持、子供無し		Divorced	1	0

解法の要約

- 使用モデルはCatBootのみ
- 合成データであることを着目して2種類の特徴量作成
- シェイクの影響を少しでも抑えるため20seedでアベーレージング



特徴量作成

- 合成データであることを生かした特徴量
- 主に2種類、150個程を過去の類似コンペを参考に作成

(参考) FDUA第2回金融データ活用チャレンジ 1位解法
https://qiita.com/negoto_coder/items/2793d772825f94319cb3
kaggle Playground Series - Season 4, Episode 1　2nd place solution
https://www.kaggle.com/competitions/playground-series-s4e1/discussion/472496

1つ目

- 成約率が平均より大きく外れている内容をエンコーディング

  例(年齢=22歳)

Age	Prodtakenの平均値
22	0.5348837209302325
その他	0.1375507835171213

学習時にモデルが取りこぼさないことを意図して追加

2つ目

- サブセットを作成してその値が元データにあるかどうかをエンコーディング

  例　('Age', 'CityTier', 'MaritalStatus’)

age	CityTier	MaritalStatus			ProdTaken平均
56	1	Divorced	→	1	0.1213
50	2	Unmarried		0	0.2058

全部のサブセットを特徴量として追加すると数が膨大(2^17=13万個以上)になるため平均から離れているものを特徴量として追加

モデリング

- 使用したモデルはCatBoost

理由　合成データは数字自体の意味が薄くなる

→カテゴリー変数として扱えるCatBoostが強い

CatBoost

- 全ての変数をint型に変換した上でcat_features = np.where(X.dtypes != np.float)[0]でカテゴリー変数として扱う
- CatBoostはGPUを用いて学習するとハイパーパラメータが変化するので揃える

例) max_bin

Default value

The default value depends on the processing unit type and other parameters:

- CPU: 254
- GPU in PairLogitPairwise and YetiRankPairwise modes: 32
- GPU in all other modes: 128

- その他のハイパーパラメータはdepthのみ4に変更した

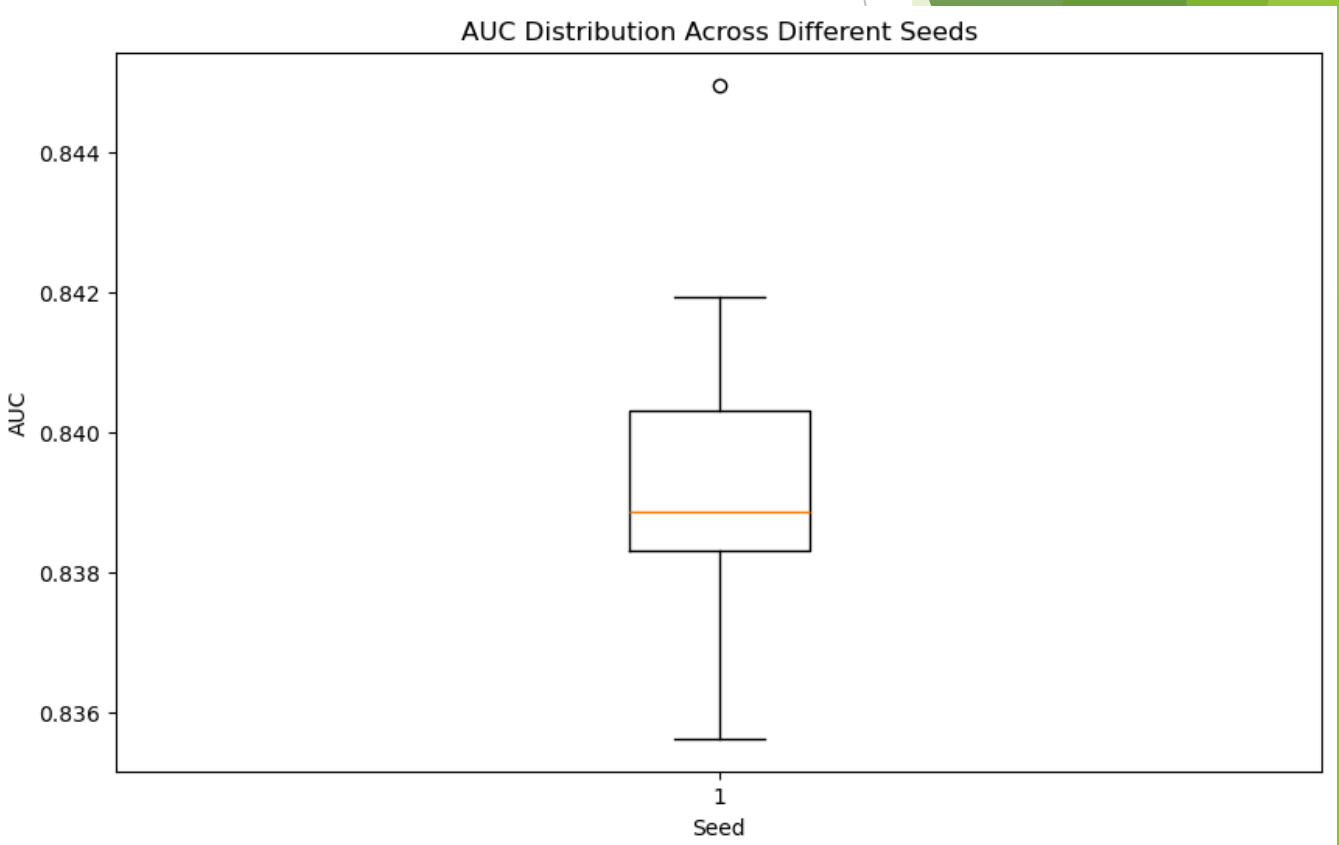
参考：CatBoost on GPU のひみつ
https://www.slideshare.net/slideshow/20230923lt-secret-of-catboost-on-gpu-tawara-261335048/261335048

CV

- 不均衡データ→ StratifiedKFold(5 folds)を使用
- CVとLBが一致しないどころかCVのseedを変更するだけでスコアが大きく変化する

→同様のことが最終結果でも起こると考えてたくさんのseedで平均を取ることに

→最終的には20seed×5 folds= 100個のファイルの平均値を提出



まとめ

- 合成データに着目した特徴量、LBにオーバーフィットしないように注意したことが大きかったように思う
- 今回の結果に満足せずにもっと色々なコンペに参加して腕を磨いて行きたい