

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

Holiday_Package_Prediction

Trip  and Travel  Company need viable business model to expand customer base.



[Data Card](#) [Code \(17\)](#) [Discussion \(1\)](#) [Suggestions \(0\)](#)

About Dataset

Context

"Trips & Travel.Com" company wants to enable and establish a viable business model to expand the customer base. One of the ways to expand the customer base is to introduce a new offering of packages. Currently, there are 5 types of packages the company is offering - Basic, Standard, Deluxe, Super Deluxe, King. Looking at the data of the last year, we observed that 18% of the customers purchased the packages. However, the marketing cost was quite high because customers were contacted at random without looking at the available information. The company is now planning to launch a new product i.e. Wellness Tourism Package. Wellness Tourism is defined as Travel that allows the traveler to maintain, enhance or kick-start a healthy lifestyle, and support or increase one's sense of well-being. However, this time company wants to harness the available data of existing and potential customers to make the marketing expenditure more efficient.

Content

What's inside is more than just rows and columns. Make it easy for others to get started by describing how you acquired the data and what time period it represents, too.

Usability

10.00

License

[CC0: Public Domain](#)

Expected update frequency

Monthly

Tags

[Travel](#)

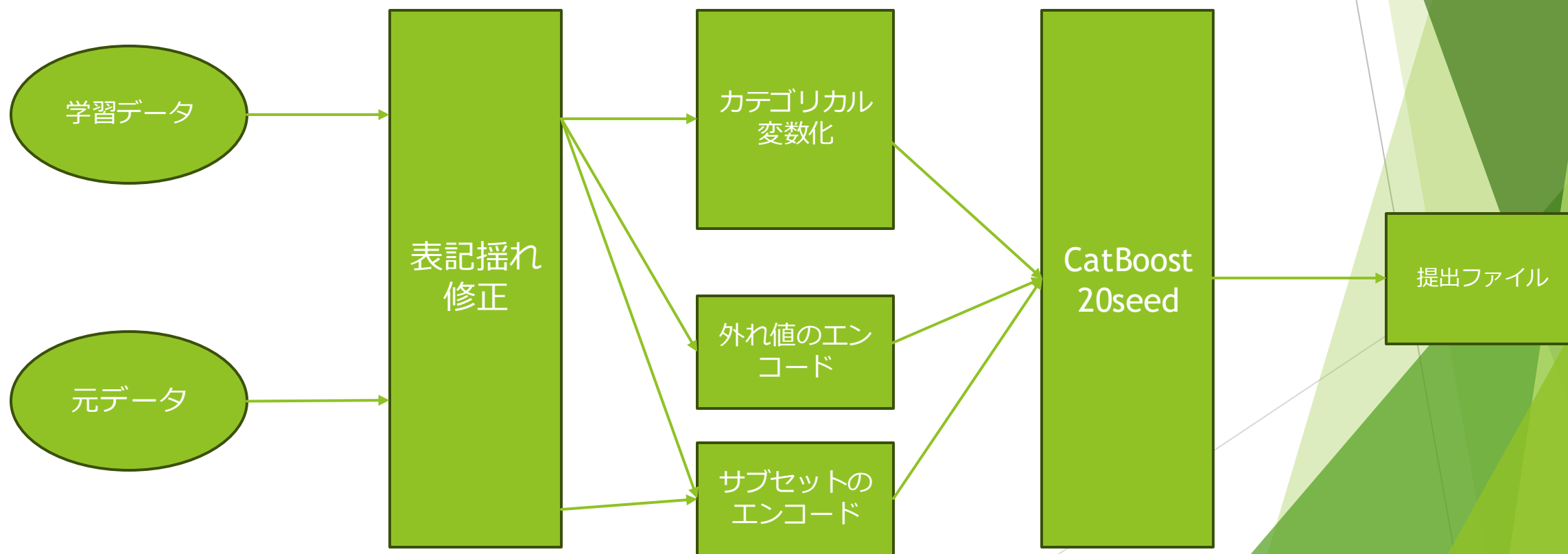
[Hotels and Accommodations](#)

[Holidays and Cultural Events](#)

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 요약

- ▶ 使用 모델은 CatBoost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でアベレー징



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) FDU A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:

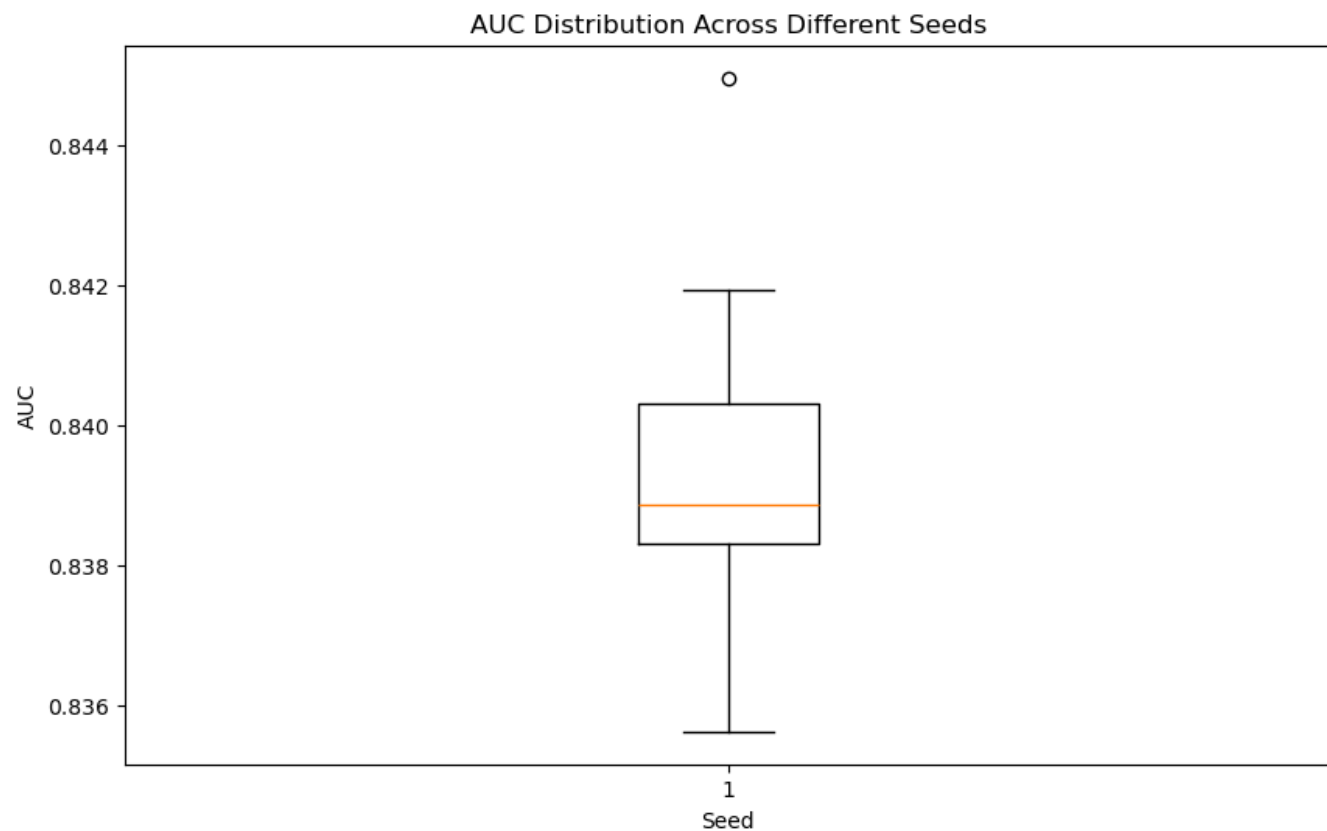
- CPU: 254
 - GPU in PairLogitPairwise and YetiRankPairwise modes: 32
 - GPU in all other modes: 128
- ▶ その他のハイパーパラメータはdepthのみ4に変更した

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ことを考えてたくさんの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と思う
- ▶ 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