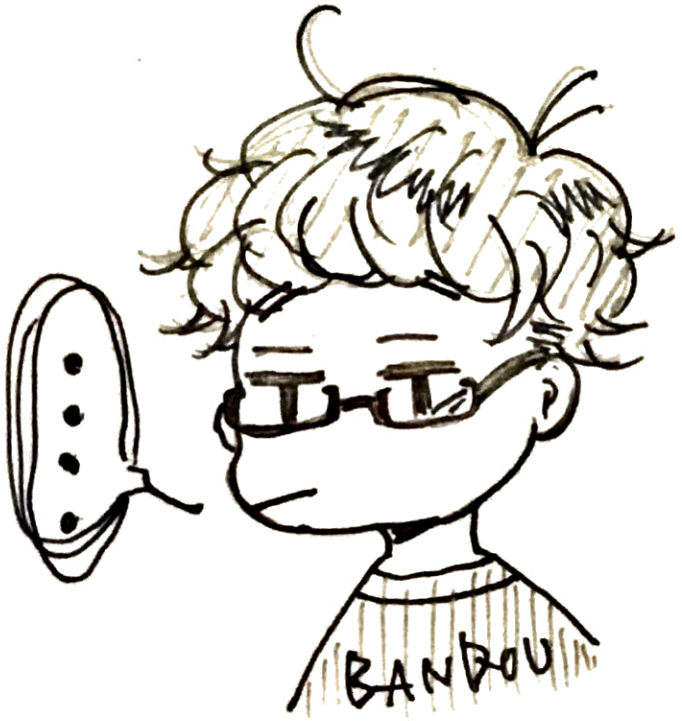
SIGNATE Student Cup 2021春【予測部門】

2nd place solution

でるぶ

twitter: @mathmandlb

自己紹介



坂東篤明 / でるぶ

京都府立医科大学5年

趣味は音楽鑑賞、ギター、写真、水泳、乗り鉄 etc.

好きな音楽のジャンルは Metal

他にはRock や Pops、EDMを聞きます。

目次

モデル

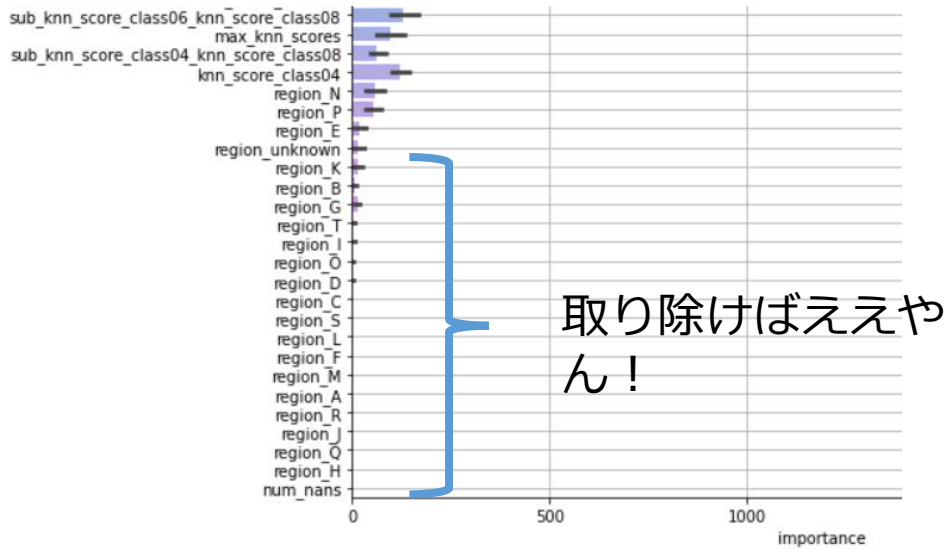
- Simple_lgb_model
- Pseudo_model
- Tabnet_model

以上3モデルを算術平均でアンサンブル

特徴量作成

lightGBMやXGBoostのような「GBDT」は、ノイズとなる特徴量を追加しても精度が落ちづらいと言われている。

重要度の少ない特徴量でも、取り除くと精度が落ちることがある。



CV: 0.66 -> CV: 0.54 😇

ある程度たくさん
特徴量を追加していこう

特徴量作成

- nagiss氏の特徴量
  - num_nas, tempo, one-hot region, CountEnc region, LabelEnc region, log tempo
  - agg_zscore_[ ]_grpby_region
  - standardscaled_[ ]
  - knn
- 自作特徴量
  - standardscaled_[ ]を2つずつ取り出し加減乗除した特徴量('2つ組みの特徴量'と名付ける)
  - standardscaled_[ ]を3つずつ取り出し和、積を計算した特徴量('3つ組みの特徴量'と名付ける)
  - standardscaledの代わりにRankGauss
- うまく行かなかったもの
  - standardscaled_[ ]特徴量に対して、行の統計量（"sum", "max", "min", "mean", "median", "mad", "var", "std", "skew", "kurt"）

特徴量作成

- ’2つ組み、3つ組みの特徴量’

GBDT系は加減乗除、とくに乗除の関係性を捉えるのが難しいため、四則演算の特徴量を追加すると良い。

A	B
1	4
2	5
6	2

A	B	A+B	A-B	A*B	A/B
1	4	5	-3	4	0.25
2	5	7	-3	10	0.4
6	2	8	4	12	3

特徴量作成

- Rankgauss

数値変数を順位に変換したあと、順位を保ったまま半ば無理矢理に正規分布となるように変換する手法。

他の正規化標準化といったスケーリングが分布の形状は変えずに縮尺を調整するだけに対し、これは分布の形状も変形。

NN系のモデルを作成する際の変換として、通常の標準化よりも良い性能を示すとのこと。

間隔尺度に変換されてしまうので、必要な情報量が削がれてしまう可能性がある。

特徴量選択

- ただ、2つ組みの特徴量を追加したあたりではCVは良くなるが、3つ組みも追加すると逆にCVは下がった。

- ある程度特徴量はどんどん追加してよいが、さすがにノイズとなる特徴量が多すぎると精度は下がり始める。

そこで、特徴量を選択する必要がある。

特徴量選択

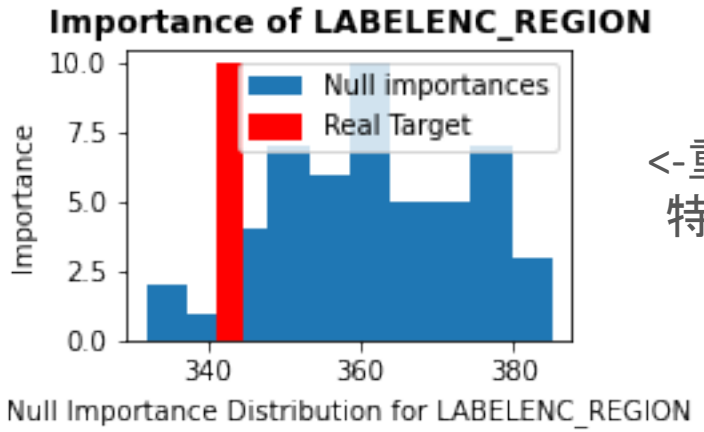
- null importanceを使用して特徴量を選択していった。

- umapなどの次元削減ではスコアがよくならなかった。

特徴量選択

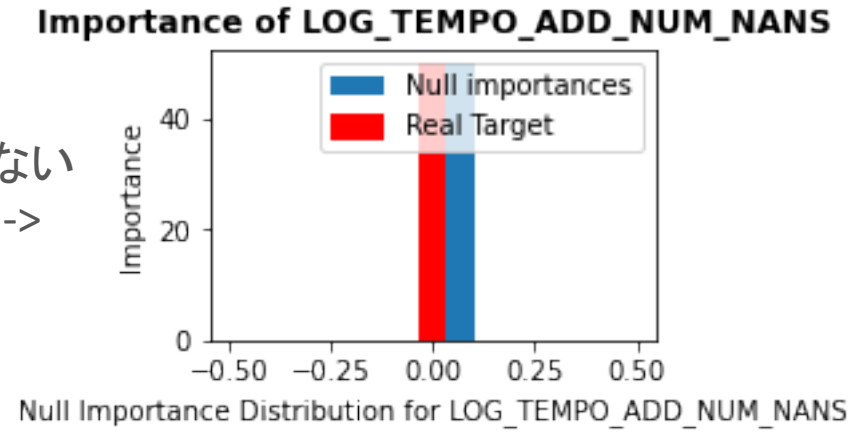
- null importance

目的変数をシャッフルして学習させた場合の重要度をnull importanceとして基準にし、目的変数をシャッフルしてない通常の重要度をactual importanceとして、この違いを重要度とする方法。それぞれの特徴量についてactual importanceがnull importanceの何パーセンタイル点にあるかで重要かどうかを判断。



<-重要な特徴量

重要でない特徴量 ->



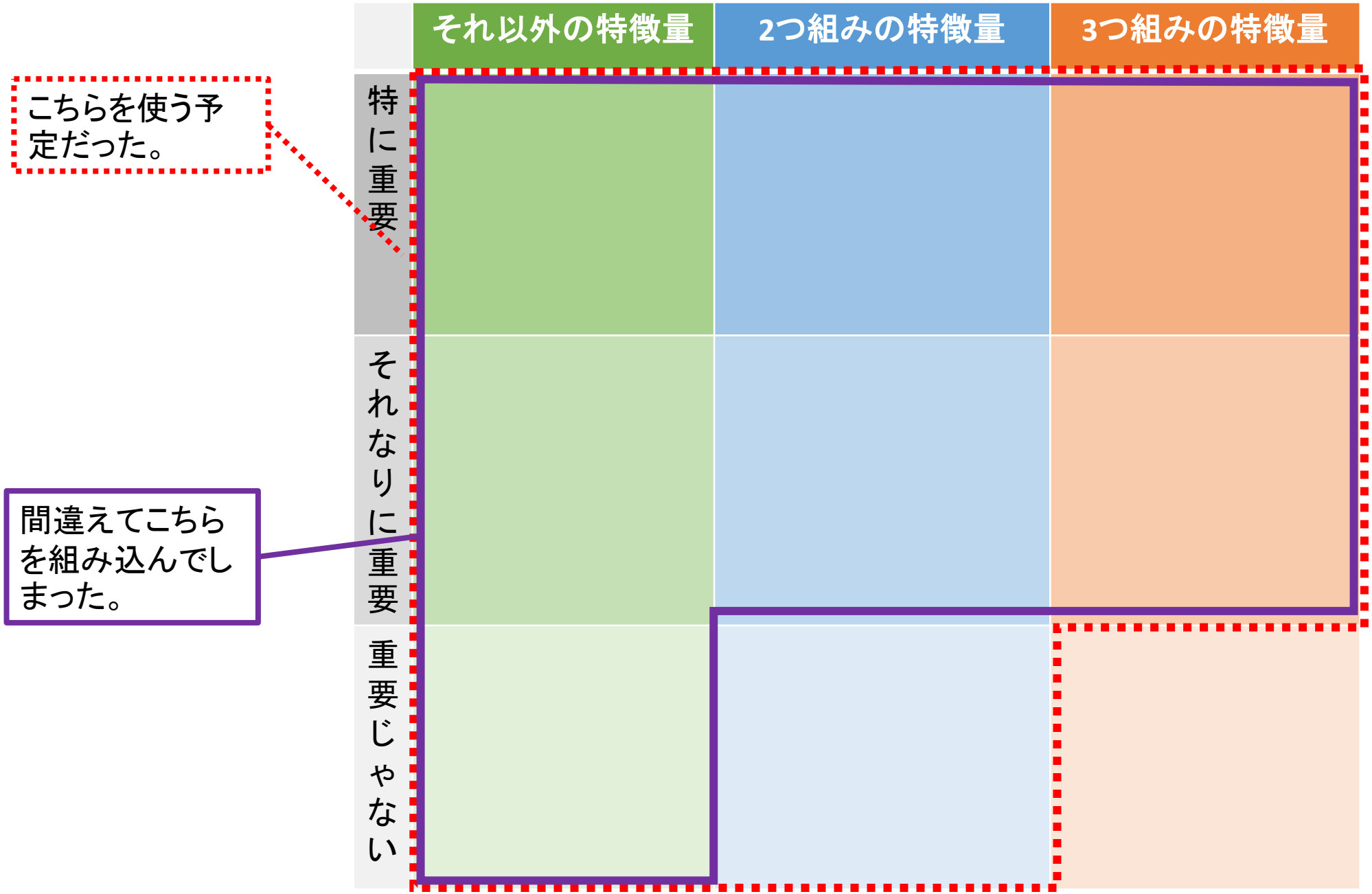
各モデルの紹介

Simple_lgb_model

- nigss氏のLGBM+KNNモデルがベース

- '2つ組みの特徴量', '3つ組みの特徴量'をすべて追加したのち、null importanceの重要度が低い特徴量を削除（手違いでこちらをアンサンブルに組み込んでしまった）

- '2つ組みの特徴量'をすべて追加, '3つ組みの特徴量'のうちnull importanceの重要度が高い特徴量のみ追加（よりよいCVが出ていたのでこちらを使う予定だった）


こちらを使う予定だった。
間違えてこちらを組み込んでしまった。
それ以外の特徴量
2つ組みの特徴量
3つ組みの特徴量
特に重要
それなりに重要
重要じゃない

Pseudo_model

- nigss氏のPseudo-Labelingモデルがベース

- 上記モデルでは0.95, 0.925, 0.9, 0.875, 0.85とどんどんpseudo labelを増やしているが、さすがに過学習してしまうのではないか？と思ったため、0.95のみにしてみた。

  (train : test = 4046 : 4046 -> 4503 : 3589と457だけ増加した。)

- Stratified K-foldのseed値を変えた。

Tabnet_model

- tabnet

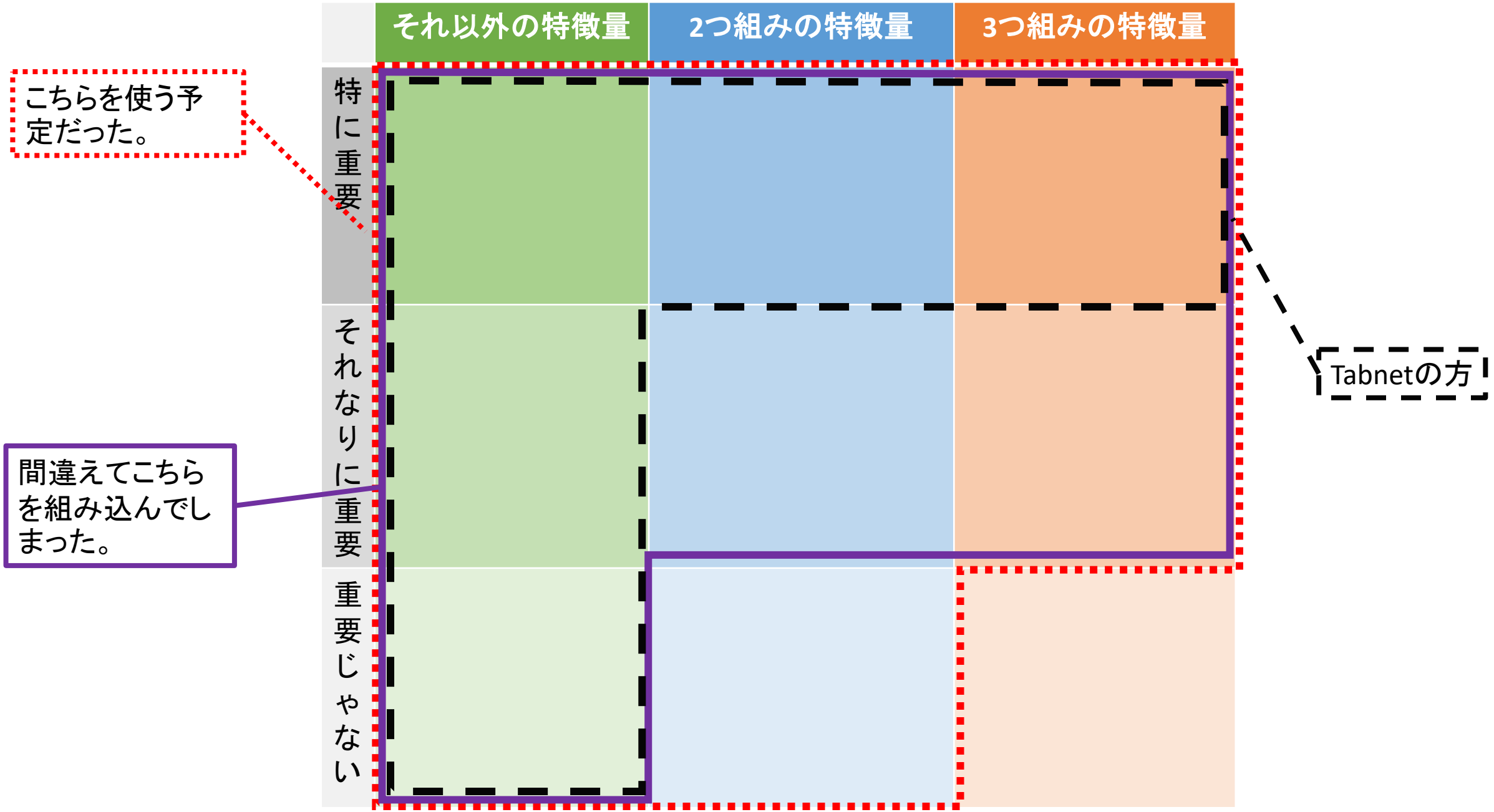
表データ向けのディープラーニングモデルでTree-basedとDNNのいいとこ取りをしたようなモデル。

Tree-based（決定木をベースにしたアルゴリズム）の解釈可能性を持ちつつ、 大きなデータセットに対してDNNのような高い性能を持つ。

今回はうまく実装できなかったが、pretrainを行うことでよりよい精度になりそうな気がしている。

Tabnet_model

- 欠損値補完 (他の変数からlightGBMで予測して埋めた)
- nagiss氏の特徴量
- standardscaledの代わりにRankGauss変換したもの
- '2つ組みの特徴量', '3つ組みの特徴量'のうち、null importanceの重要度が結構高いもののみ採用(特徴量増やしすぎても過学習するだけなので、lightGBMより少なめ)


それ以外の特徴量
2つ組みの特徴量
3つ組みの特徴量
特に重要
それなりに重要
重要じゃない
こちらを使う予定だった。
間違えてこちらを組み込んでしまった。
Tabnetの方

要点

- train:test = 1:1とtrainが少ないのでpseudo labelingは効くだろうとおもった。
- lightGBMなどの木モデルでは特徴量間の四則演算が効く事が多い(2つ組みや3つ組みのこと)
- lightGBMは特徴量をたくさん増やしてもうまいこと特徴量を選んでくれるのでどんどん作る。
- とはいえ3つ組みまで足すと増えすぎてCVが悪くなったのである程度は特徴量選択をしないとノイズになるだけ。

    -> 今回はumap等の次元削減よりもnull importanceがうまく行った

- アンサンブルの多様性のためにrandom seed averageもする。
- アンサンブルの多様性のためにNN系も追加。今回はtabnetをためした。
- Trust CV

参考

- 「kaggleで勝つデータ分析の技術」
- 「Rank Gaussという正規化手法」
  https://deoxy.hatenablog.com/entry/2020/12/03/235759
- 「Null Importanceを用いた特徴量選定」
  https://qiita.com/trapi/items/1d6ede5d492d1a9dc3c9
- 「TabNetとは一体何者なのか？」
  https://zenn.dev/sinchir0/articles/9228eccebfbf579bfdf4

ご清聴ありがとう
ございました