Signate Cup
1st place solution

自己紹介

名前：ひーーご/皆倉　諒

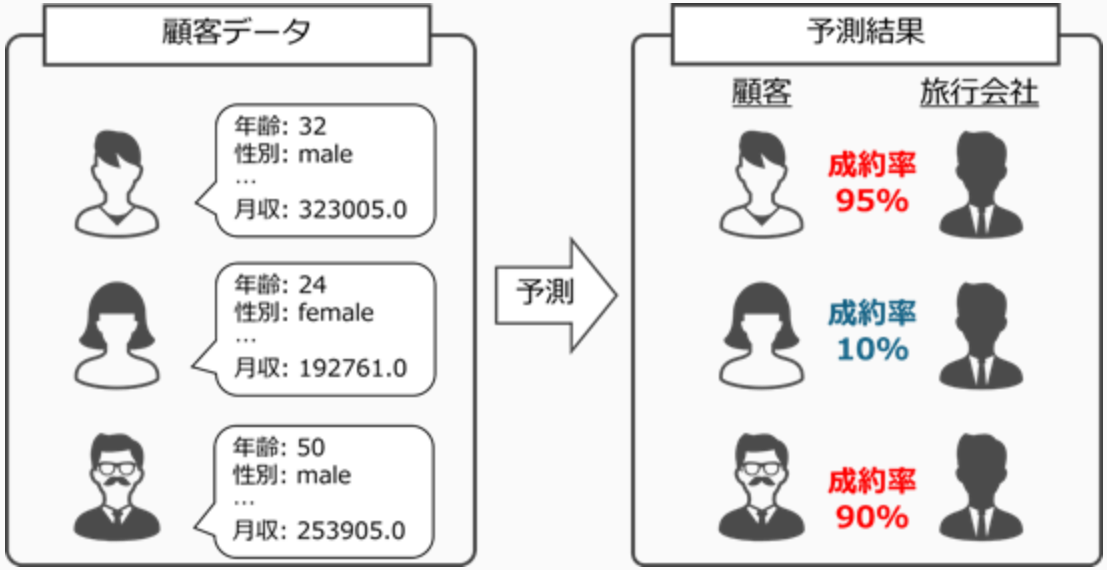
所属：肥後銀行　デジタルマーケティング部

経験：大学時代は確率論を専攻
ノーコードツールを使ってのデータ分析業務は2年
pythonを触り始めてから1年強

概要

URL:https://signate.jp/competitions/1376

課題：旅行パッケージの成約率の予測
期間：8月1日～9月1日
学習データ：3489データ×18カラム

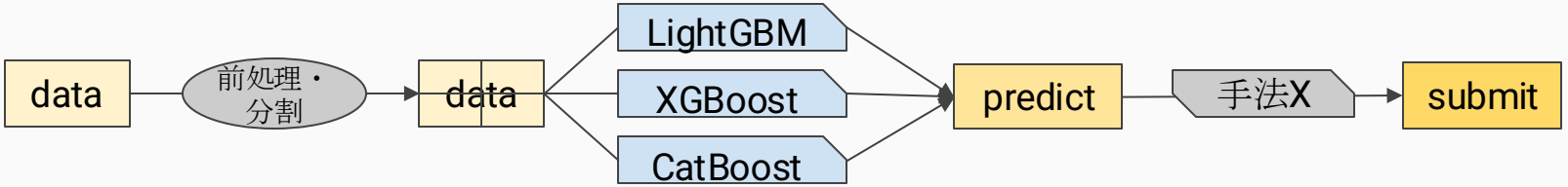


ベースモデル

◉データの下処理

・表記揺れの統一（ex.結婚フラグは未婚、離婚、独身の3種へ）
・null値の補完（ex.年齢については20代などは25へ）
・年齢は18～60にクリッピング(15歳と65歳が発生した)

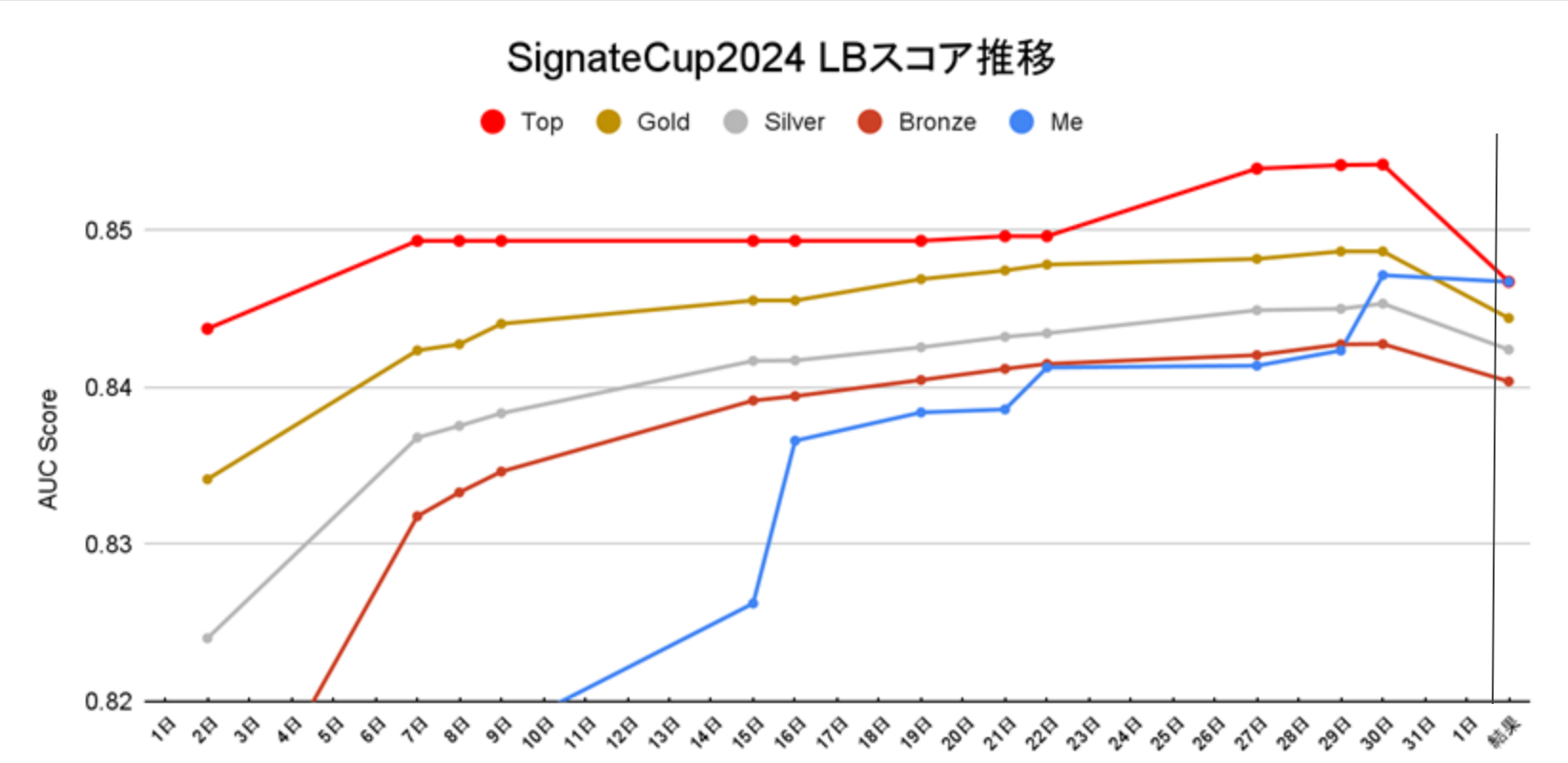
◉ベースモデルのイメージ



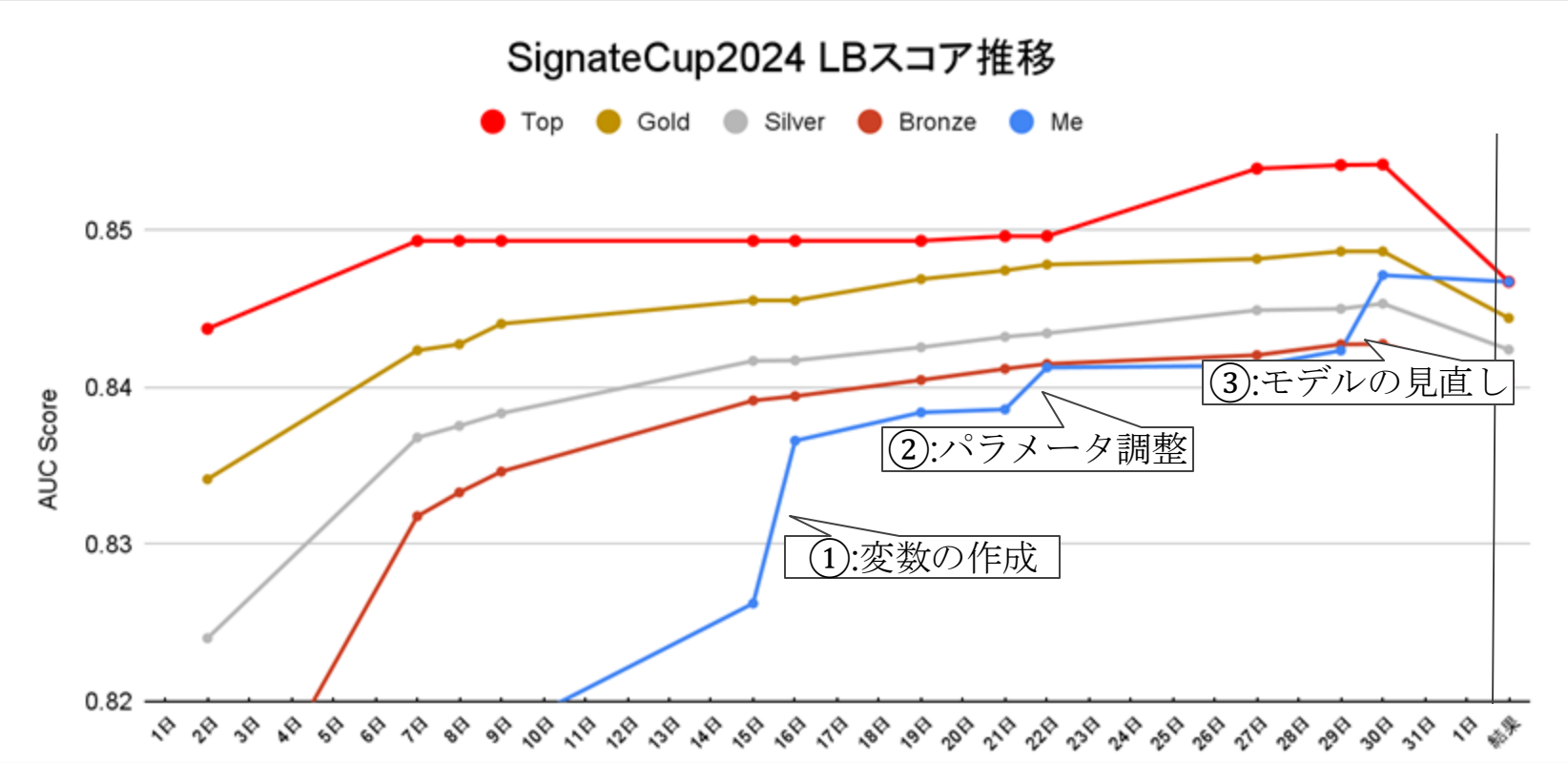
（手法Xの例）

・平均値をとる
・CV値が最大となるように配分を調整して線型結合する
・スタッキングモデルを調整する

スコアの推移


SignateCup2024 LBスコア推移
Top
Gold
Silver
Bronze
Me
AUC Score
0.85
0.84
0.83
0.82
1日
2日
3日
4日
5日
6日
7日
8日
9日
10日
11日
12日
13日
14日
15日
16日
17日
18日
19日
20日
21日
22日
23日
24日
25日
26日
27日
28日
29日
30日
31日
1日
結果

スコアの推移


SignateCup2024 LBスコア推移
Top
Gold
Silver
Bronze
Me
AUC Score
0.85
0.84
0.83
0.82
1日
2日
3日
4日
5日
6日
7日
8日
9日
10日
11日
12日
13日
14日
15日
16日
17日
18日
19日
20日
21日
22日
23日
24日
25日
26日
27日
28日
29日
30日
31日
1日
結果
①:変数の作成
②:パラメータ調整
③:モデルの見直し

スコア上昇の要因と考えられること

①における上昇の要因：新たな変数の作成
- 年齢と収入から相関の近似直線を作成
  そこからその人の収入が年齢の割に多いのかを考慮する変数を作成
- 給与、年齢、交渉時間など数値列をいくつかの区分に分けたもの

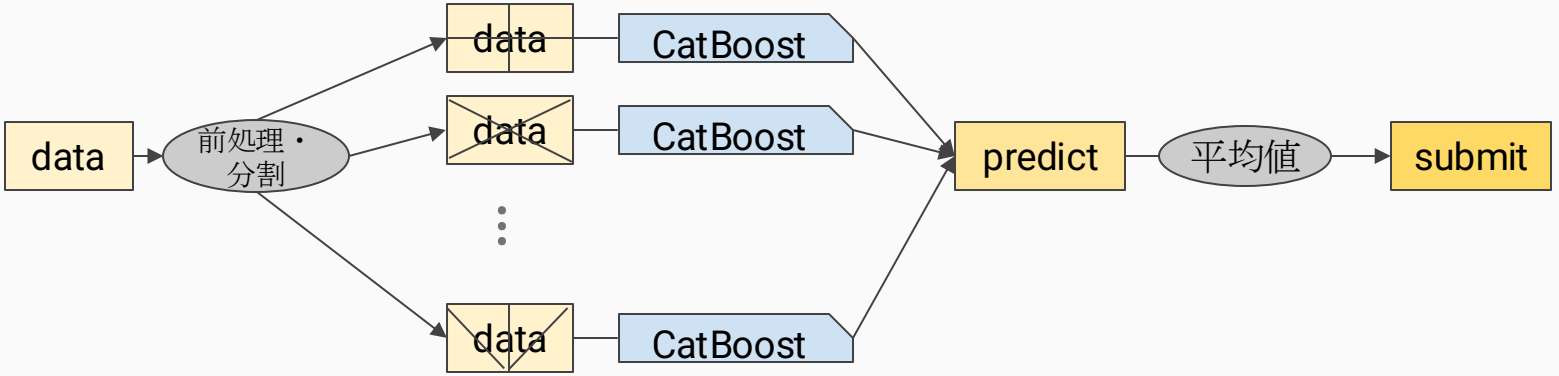
②における上昇の要因：パラメータの調節、欠損データの扱い
- 各モデルの決定木の深さを１に設定
- 各モデルの学習率を１に設定
- 欠損を含むデータを学習データから除外

スコア上昇の要因と考えられること

③における上昇の要因：モデル自体の構造の見直し

・3種のモデルを使っていたものをCatboostのモデルのみに変更、
データの分割方法を変えながらそれぞれの分割に対しCatboostモデルを作成
各モデルから予測された結果の平均値をとることで提出データとした

◉モデルのイメージ



スコア

	CV (LGBM)	CV (XGB)	CV (cat)	CV (predict)	LB	PB
モデル①-A	0.8386	0.8388	0.8402	0.8418	0.8372	0.8374
モデル①-B	0.8423	0.8411	0.8425	0.8440	0.8386	0.8382
モデル②-A	0.8428	0.8321	0.8457	0.8474	0.8413	0.8371
モデル②-B	0.8404	0.8301	0.8466	0.8475	0.8414	0.8367
	CV (cat-❶)	CV (cat-❷)	CV (cat-❸)	CV (predict)	LB	PB
モデル③-A （最終提出）	0.8476	0.8417	0.8422	0.8533	0.8471	0.8467
モデル③-B （最終提出）	0.8429	0.8478	0.8429	0.8562	0.8464	0.8450

あまりうまくいかなかったこと

◉新たな変数の作成

- 役職ランクと勧められたホテルのランクの関係性から作成した変数
- カテゴリ変数列を組み合わせた新たなカテゴリ列を作成

ex）`out_df['age*Des'] = out_df['AgeBlock']+out_df['Designation']`

◉モデルの構造の見直し

- CV,LBを参考に交差検証数の探索
- 損失関数をAUC以外に変えてみてモデルを作成
- 'boosting_type'を’gbdt’以外を使用
- ニューラルネットワーク型のモデルの構築

まとめ・感想

・過剰適合しないようにモデルを工夫したことが今回の結果に繋がったと思う
（大局的な傾向を捉えるようなモデルを作成することを意識しました）

・実際、CVを上げることに意識した加工や特化した変数はうまく作用しない傾向にあった

・運の要素も大きかったが、今回の結果をバネに今後とも成長していきたい

・第2回金融データチャレンジでは暫定30位→118位と悔しい思いをしたので、リベンジできて嬉しいです！！

ご清聴ありがとうございました