GCI Summer 2023 2nd Competition: Home Credit Default Risk

stkz_e_ai

2023 年 7 月 6 日

1 Feature Engineering

今回のコンペでは EXT_SOURCE の重要度が非常に高く、これらの特徴量を利用して新たな特徴量を作ることが重要であると考えた。

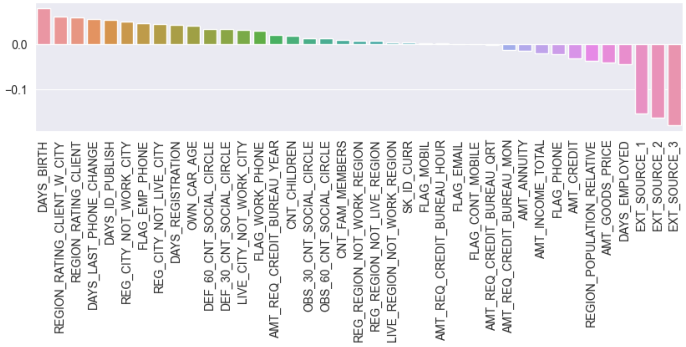

図 1　数値特徴量と TARGET の相関係数

特に図 2 でわかるように EXT_SOURCE_3 は TARGET との相関が高いためこれを有効に利用する方法を考えた。

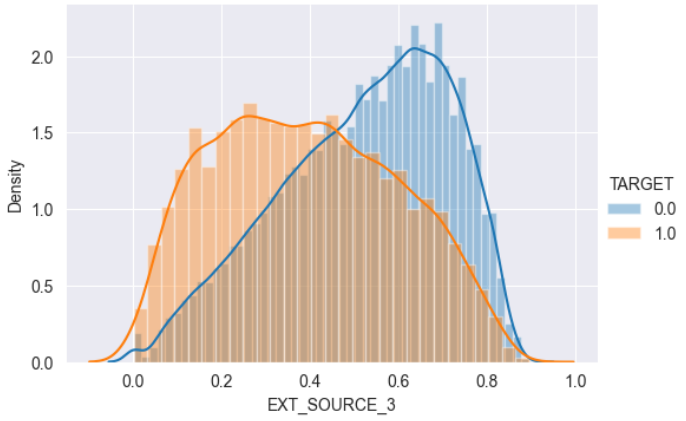

図 2　EXT_SOURCE_3 の分布

今回のデータには沢山のカテゴリカル特徴量があったので、それらに対し EXT_SOURCE_3 を用いて TargetEncoding を行ったが、これはあまり効果がなかった。結局、EXT_SOURCE の平均や差分を取った特徴量のみを採用した。

他には、Kaggle の Discussion にあった”1st place solution”を参考に四則演算特徴量を採用した。また同じく Discussion 上に、今回のカテゴリカル特徴量には LabelEncoding よりも One-Hot Encoding が適していることが示されていたので、これを採用した。

また興味深いことに DAYS_EMPLOYED の分布を見てみると、負の値が多い中 365243 という数値が多数存在したのでこれを NaN に置き換えた。最終的な特徴量は 157 個になった。

2 Modeling

今回は LightGBM を用いた。これは多くの人が採用していることに加え NaN を処理しなくても使用することができるからである。本来は XGBoost も用いてアンサンブルを行いたかったが時間の都合でることが出来なかった。

2.1 Hyper Parameter

今回も KFold と Optuna を用いて最適化した。データの偏りが激しいため StratifiedKFold を持ちた。また、いくつかの乱数シードを試していい感じのを選んだ。

3 Result

表 1　採用モデルのスコア

AUC	0.76447
Public Score	0.76814
Private Score	0.76105

PublicScore から PrivateScore が大きく下がったので、過学習してしまったと考えられる。今回は時間の都合上アンサンブルや特徴量選択などを行うことが出来なかったのが心残りである。

参考文献

[1] https://qiita.com/sh-tatsuno/items/8eca6fbf1de5e66794f0

[2] https://www.kaggle.com/c/home-credit-default-risk/discussion/64821

[3] https://speakerdeck.com/hoxomaxwell/home-credit-default-risk-2nd-place-solutions

[4] https://www.kaggle.com/competitions/home-credit-default-risk/discussion/66010

[5] https://www.kaggle.com/code/aantonova/797-lgbm-and-bayesian-optimization/notebook

[6] https://github.com/paveltr/home_credit_default_risk/blob/master/home_credit_solution.ipynb

[7] https://github.com/Hirochon/GCI2020-Summer/blob/master/Competition2/3rd_place_solution.ipynb