GCI Competition2 解説文

MAEDADSK

最初に、配布された”HomeCredit_columns_descriptiontrain”で各変数の説明を確認
し、”pandas_profiling”を用いて、各変数の内容、欠損値や外れ値について確認しました。
次に、train データを LightGBM にかけ、”feature_importance”の”gain”から、各変数の重要度を確認、以下リンクを参考にし、新たな特徴量を作成し、その”feature_importance”を確認しました。

〇参照したコード

https://www.kaggle.com/code/jsaguiar/lightgbm-with-simple-features/script

https://www.kaggle.com/code/sangseoseo/oof-all-home-credit-default-risk

https://www.kaggle.com/code/jsaguiar/lightgbm-7th-place-solution/script

https://github.com/NoxMoon/home-credit-default-risk/blob/master/notebooks/lgb2.ipynb

最終的に 4 つのモデルを作成しました。各モデルで、k-fold - LightGBM を行い、out-of-fold prediction も同時に作成し、アンサンブル時に使用しました。
LightGBM のパラメータは、Optuna を使用し、改善されるモデルには適用しました。

〇各モデルの特徴

モデル１：雇用、年齢、収入、ローン額、家族構成等から、５つの特徴量を作成。

モデル２：”EXT_SOURCE_1,2,3”から新たな特徴量を作成。

モデル３：”EXT_SOURCE_1,2,3”の最小、最大、平均、中央、分散、
また、組織、教育、職業、年齢、性別ごとに、いくつかの特徴量の中央値、標準偏差、平均を新たな特徴量として作成。

モデル４：”AMT_INCOME_TOTAL”について、いくつかの特徴量で平均をとり、その平均値との差分の特徴量を作成。TARGET 値に偏りがある不均衡データのため、ダウンサンプリングにより均衡化した上で学習。

最後にモデル１～４をアンサンブルしました。CV 値が高い順に重み付けしています。
アンサンブルモデル（AUC 値）：0.766173

〇結果（AUC 値）

パブリックボード：0.76818

プライベートボード：0.76185