作戦

今回のコンペは特徴量，データの数が非常に多いため計算コストを重視して機械学習を行った．そのため特徴量作成の前にモデルで LightGBM を利用することを決め，LightGBM に適した特徴量作成を行うことにした．Home Credit Default Risk に関するドメイン知識がないため Kaggle やその他のインターネット上のサイトを参考にして，計算コストを削減しつつ，精度はなるべく下げないことを考えてコードを書いた．以下で行っとことを述べる．

EXT_SOURCE

EXT_SOURCE の 3 つのカラムは TARGET との相関が高いが，詳細は不明なデータであるため様々な四則演算や統計量を追加してスコアが上がったものを採用した．最終的にはそれぞれの差の絶対値，2 つの平均，3 つの平均を特徴量として残した．EXT_SOURCE 以外のカラムとの交差項を追加することも考えたが計算コストを考えて断念した．

エンコーディング

NAME_EDUCATION_TYPE ， NAME_INCOME_TYPE ， NAME_TYPE_SUIT ， NAME_FAMILY_STATUS，NAME_HOUSING_TYP の 5 つのデータに対して目的変数(TARGET が 1)の割合が高い順に手作業で数値を割り当てていった．その他のユニークな値が多いカテゴリカルデータは One-hot Encoding を行った．

特徴量追加

AMT_CREDIT(クレジット額)を 5 つにグループ分けして，所属するグループの平均で AMT_ANNUITY(ローン年金額)を割った値を新たな特徴量として追加した．その他 AMT_INCOME_TOTAL や DAYS_BIRTH などのカラムや，カテゴリカルデータに関しても同様の処理を行ったが効果が無いと判断して不採用とした．

モデル

モデルは LightGBM を用いた．今回の TARGET には偏りがあり，一般的な K-fold 法を行うと各 Fold で偏りが生じてしまう．そこで各 fold の TARGET の割合が均等になるように分割することができる Stratified K-fold を利用することで公平な評価を行った．特徴量作成だけでは AUC スコアが頭打ちになったため，スタッキングを行うことでスコア向上を狙った．スタッキングは学習用データを複数モデルで予測した結果を，第 2 段階のモデルの特徴量として与えることでスコアの向上を図る手法である．しかし，今回は LightGBM を前提とした特徴量作成を行ったため，その他のモデルの AUC スコアは低く，スタッキングを行ってもスコアの向上にはつながらなかったため不採用とした．