コンペ２　Home Credit Default Risk　説明文

〇最終採点結果点数：0.76024

〇やったこと

1.　train データと test データの結合

　不均衡データであることに加え、一括で 2. をするために concat で最初に結合した。

2.　データの可視化と分析

　全てのデータを可視化した。また特徴量のピアソン相関とランダムフォレストを用いた重要な特徴量の選定を行った。

参考：https://www.kaggle.com/code/nozomuk/home-credit-complete-eda/notebook

3.　前処理と特徴量作成

　カテゴリカル変数の encording 手法については、one hot encording と label encoding を試したが、one hot encording の方が良い結果が出たため、そちらを選択。

　2. で得られた情報をもとに、重要な特徴量を活かせる特徴量と既存の特徴量から計算できる債務不履行に関係しそうな特徴量を追加。この部分は何度も試行錯誤した。ハイパーパラメータの選定の次に大事であった部分であると考える。

4.　機械学習モデルの作成

　optuna を利用したハイパーパラメータの最適化を行った。これが今回のコンペでは必須だったように感じる。

　不均衡データであるため、StratifiedKFold クラスによる交差検証を行った。

　様々なモデルを利用した結果、一番性能の良かった LightGBM のみを使用。ただしアンサンブルなどは行っていないため、より良い組み合わせはあったかもしれない。

今回のコンペでは、3. と 4. を何度も何度も繰り返した。

〇感想

　また、Chat GPT が素晴らしい。自分は初心者であるため、あまり自分でコードを書けないのだが、何をしたいのか？どのようなコードを書きたいのか？を明確にして、利用するとかなり制度の高い解答が返ってくる。それらから学ぶことはたくさんあった。今回のデータに対してどのような順番でアプローチしていくかに関しては、今までの GCI の授業やトップランカーのコードや解説を参考にした。

参考：https://www.kaggle.com/c/home-credit-default-risk/notebooks