Abebe0210

# コンペ2説明

## 方針

Kaggle上位のnotebookやdiscussionを見ると、特徴量同士を四則演算して新たに作成した特徴量が有効であると分かった。
しかし、どのように四則演算を行った特徴量が有効かを解釈することが出来なかったため、機械的に四則演算した大量の特徴量をラッパー法により特徴量選択を行う方針とした。
四則演算に用いる特徴量は量的データを単位別に分類したもので、計算法ごとに組み合わせを変更した。

## 特徴量選択

### 特徴量重要度(Null Importance)

大量に生成した特徴量をそのままラッパー法にかけると計算コストが非常に大きくなるため、特徴量重要度を参考に厳選する。
通常の特徴量重要度ではノイズも評価されてしまう可能性もあるようなので、Shapに加えてそれぞれのNull Importanceも算出した。
Null Importanceとはターゲット変数をランダムにシャッフルして特徴量重要度を算出するという手順を数十回繰り返し、その平均値を求めたものだ。
予測において特定の特徴量が本当に重要であれば、その特徴量のNull Importanceよりも通常の特徴量重要度の方が高くなると言える。

### ラッパー法(Forward Selection)

前項の特徴量重要度に基づき絞り込んだ特徴量をラッパー法(Forward Selection)によりさらに選択する。
Forward Selectionとは最初に1つの特徴量を選び、スコアが最も向上する特徴量を順次追加していく方法である。
約500個の特徴量から32個まで絞り込みを行った。
しかし想定以上に計算コストが高く、特徴量の増加とともにモデルのKfoldの値も大きくしなければ選択が進まなかったため実際的な手法ではないと感じた。

## 機械学習モデルとハイパーパラメータ

機械学習モデルはLightGBMを使用した。
欠損値処理まで行うことができず、決定木モデルを使用することとなり、XGBoost、Catboostと比較して最もスコアが高くなったLightGBMを採用した。
アーリーストッピングを行うことがポイントだったと感じており、特別な特徴量作成やチューニングを行わずとも精度0.7を突破したと記憶している。
ハイパーパラメータチューニングにはOptunaを利用し、探索範囲は『Kaggleで勝つデータ分析』を参考にした。

# アンサンブル

同じく受講生のpetakumiさんのモデルとアンサンブルを行った。
アンサンブルの手法はそれぞれのモデルの予測値の平均を取るバギングを採用した。
私自身の1つのモデルとpetakumiさんの1つのモデルを、それぞれ異なる2つのseed値のKfoldで予測した計4つの予測値から、重みを調節して算出した平均値を最終予測とした。
Public Score 0.76917、Private Score 0.76158という結果となった。

# 反省点・改善法

- Forward Selectionを実際的な手法で行えなかった点
  →欠損値処理を行い、線形モデルでForward Selectionを実行するのが実際的か
- Valid Score, Public Score, Private Scoreに乖離が生じた点
  →複数の評価指標で評価、複数seedでのKfoldのScoreで評価、アンサンブル学習など
- アンサンブルの仮説・検証を十分に行えなかった点
  - 前提としてアンサンブルの技術的ハードルが高く、時間も不足していた
  - 残り提出回数に限りがあり、最良のものを提出できなかった
  - アンサンブルモデルの評価をPublic Score以外で行えなかった
  - スタッキングを試すことが出来なかった

## 参考

直接的なコードの引用はないが、以下のサイト・書籍を参考にした。


- https://www.kaggle.com/competitions/home-credit-default-risk/discussion/64821
- https://kurupical.hatenablog.com/entry/2018/09/10/221420#f-ef8d3589
- kaggleで勝つデータ分析の技術(書籍)