コンペ 2 Home credit 解説

Tanaka Shoma

0.結果

GCI のスコア：0.76085

1.コード作成の流れ

① ベースライン作成 → ② 特徴量エンジニアリング → ③ モデルチューニング
の流れでベースラインのスコアと比較しながら特徴量エンジニアリングとモデルチューニングを行いました。このあたりの手順は
「著 諸橋政幸, Kaggle で磨く機械学習の実践力, 2022」[1]を参考にさせていただきました。ベースラインは欠損値や外れ値に対して何も処理をせずパラメータも固定してモデルに学習させました。モデルは LightGBM を使用しました。様々なモデルを試した結果 LightGBM が良いスコアが出やすく計算速度も速いように感じました。特徴量については「２．効果があった特徴量」で記します。モデルチューニングには optuna を使用しました。ベイズ最適化や各種パラメータについての理解は浅いため正直行き当たりばったり感はありましたがチューニングを行う前と後ではスコアにかなり改善が見られました。

2.効果があった特徴量

各種 EXT_SOURCE や AMT_INCOME_TOTAL などベースライン作成をした際に重要度が高かった特徴量を用いて新たな特徴量を作成していきました。むやみに四則演算するのではなくある程度意味があるだろうと予測しながら特徴量を追加していきました。
【例】 総所得金額を世帯人数で割った値 (大きい方が貸し倒れしにくいと予想できる)
特に効果があったのは近傍法を用いて 4 つの特徴量の距離が近いもの 500 個の TARGET の平均を取るというものです。こちらは Home Credit Default Risk で 1 位の方のアイデアを参考にさせていただきました。[2]しかしこの特徴量を追加して重要度を見たところかなり重要度が大きくなっていることがわかり、過学習の恐れがあるのではないかと思いこの特徴を含むモデル、含まないモデルをそれぞれ作成しアンサンブルしました。アンサンブルにより public score が少し上昇しました。

3.やらかした話（オマケ）

実は今回私が最終的に提出したものは私の意図とは少し違うものとなってしまいました。本来は以下のようなコードを実行し(test_pred_1 は近傍法による特徴量を含んだモデルの予測結果、test_pred_2 は含まないモデルの予測結果)2 つのモデルの予測結果を足して 2 で割るはずでした。

```
#2つのモデルのアンサンブリング
test_pred = test_pred_2
test_pred['pred'] = test_pred_1['pred'] * 0.5 + test_pred_2['pred'] * 0.5
```

しかし 0.5, 0.5 の割合を 0.6, 0.4 など変更したらどうだろう、など同じセルで値を変えて試していると意図せぬうちに出力結果が変わってしまっていたのです。(前回の実行結果の影響を受けていると思われますが結局原因は分かっていません。) この出力結果をろくに確認もせずに提出したところ今までの public score 結果の中でベストスコアが得られ、そのままコンペが終了しました。このことに気がついたのはコンペが終了し入賞候補の連絡を受け、コードを確認していたときです。正直めちゃくちゃ焦りました。「このまま実行すれば同じ結果が得られる」と思っていたのにそうならなかったのです。「このまま再現性を示せないままだと失格になるのでは。。。」と絶望的な気持ちになりましたが何とか過去の記憶を頼りに自分がどんなふうに値を変えて何回セルを実行したのか見つけ出すことができました。結局提出コードは以下のような汚くて恥ずかしいコードになりました。

```
#2つのモデルのアンサンブリング
test_pred = test_pred_2
test_pred['pred'] = test_pred_1['pred'] * 0.5 + test_pred_2['pred'] * 0.5
test_pred['pred'] = test_pred_1['pred'] * 0.5 + test_pred_2['pred'] * 0.5
test_pred['pred'] = test_pred_1['pred'] * 0.6 + test_pred_2['pred'] * 0.4
test_pred['pred'] = test_pred_1['pred'] * 0.6 + test_pred_2['pred'] * 0.4
test_pred
```

超絶初歩的なミスですが
結論、再現性の確保は命。(最終提出は特に。)
二度と同じ過ちを繰り返さないよう教訓にしたいと思います。

参考文献

[1] 著 諸橋政幸, Kaggle で磨く機械学習の実践力,2022
https://www.amazon.co.jp/Kaggle%E3%81%A7%E7%A3%A8%E3%81%8F-%E6%A9%9F%E6%A2%B0%E5%AD%A6%E7%BF%92%E3%81%AE%E5%AE%9F%E8%B7%B5%E5%8A%9B-%E5%AE%9F%E5%8B%99x%E3%82%B3%E3%83%B3%E3%83%9A%E3%81%8C%E9%8D%9B%E3%81%88%E3%81%9F%E3%83%97%E3%83%AD%E3%81%AE%E6%89%8B%E9%A0%86-%E8%AB%B8%E6%A9%8B-%E6%94%BF%E5%B9%B8/dp/4865943269
[2] https://www.kaggle.com/c/home-credit-default-risk/discussion/64821