HomeCreditコンペ説明文 - Yarto271

TL;DR（解法のキモ）

- すべての特徴量を一つずつ丁寧に観察し，観察結果に基づいた欠損値補完や変数変換，新たな特徴量の作成を行いました。
- EXT_SOURCE_1，EXT_SOURCE_2，EXT_SOURCE_3の欠損値を他の特徴量から予測して補完しました。
- 不均衡データであったためundersampling + baggingを適用しました。

特徴量エンジニアリング

- 特徴量の数が多かったですが，一つひとつの特徴量をすべて確認しました。
    - 主に，分布・欠損値の量・目的変数との関係を観察しました。
- 欠損値の補完方法は，観察結果をもとに決めました。
    - 特に，EXT_SOURCE_1，EXT_SOURCE_2，EXT_SOURCE_3は欠損が目立ち，分布が目的変数と強く関係を示していたので，予測によって補完することにしました。欠損しているか否かと目的変数の間に若干関係が見られたので，欠損しているか否かも新たな特徴量として加えました。
    - 最終的な特徴量の重要度では，EXT_SOURCE_1，EXT_SOURCE_2，EXT_SOURCE_3がトップ3を占めていたので，きちんと予測によって補完するという判断は正しかったように思います。
- カテゴリ変数はすべて単純なラベルエンコーディングを施しました。
- 対数や平方根を取るなどの変数変換を行い，特徴量の分布がなるべく正規分布に近づくようにしました。
- Kaggleでの1位解法(https://www.kaggle.com/c/home-credit-default-risk/discussion/64821)も参考にしながら，いくつか新しい特徴量を作成しました。
    - その中で，最終的な特徴量の重要度において比較的上位に食い込んだものとしては，(AMT_CREDIT / AMT_ANNUITY)や(AMT_GOOD_PRICE - AMT_CREDIT)がありました。

- うっかりSK_ID_CURRを消去し忘れていたことに締切前日に気づき，消去したところスコアがぐんと伸びました(Public LB 0.76284→0.76785)。肝を冷やしました。

モデル作成

- EXT_SOURCE_1，EXT_SOURCE_2，EXT_SOURCE_3の欠損値の予測には単純なXGBRegressorを用いました。
- 全体の予測には単純なXGBClassifierを用いました。
- ハイパーパラメータのチューニングは結果的に行えませんでした。多少試みたものの，スコアがむしろ悪くなってしまったので採用しませんでした。もう少し時間をかけてチューニングをすれば良いハイパーパラメータが手に入ったかもしれません。
- 今回のデータは目的変数が0に大きく偏っている不均衡データであったため，undersampling + baggingを適用しました。
    - 正例すべてと，正例と同じ数だけランダムに抽出した負例を合わせることで，正例と負例の数が等しい部分的なデータセットが構成できます。これを用いてモデルを学習します(undersampling)。しかしこれだと負例の大部分を捨てることになり損なので，改めて負例を抽出し直して学習するということを繰り返して複数のモデルを作成し，それらの推論結果を平均します(bagging)。これがundersampling + baggingです。
    - この手法は，現在Kaggleで開催中のICR - Identifying Age-Related Conditionsコンペに投稿されたDiscussion(https://www.kaggle.com/competitions/icr-identify-age-related-conditions/discussion/412507)を読んで知り，取り入れることにしました。このコンペはHomeCreditコンペの状況と類似しているため参考になりました。
    - モデル数は増やせば増やすほどスコアが上昇しましたが，それだけ学習に長い時間がかかるようになるので，300個くらいで止めました。学習時間に糸目をつけなければもう少しスコアの伸びしろがあったかもしれません。

その他

- 自分は第一回コンペで全く良い成績を収められなかったので，そのときの反省と教訓を生かして今回のコンペに取り組みました。
- Google Colab上でプログラムを書くのをやめて，使い慣れているVSCode上で書くことにしました。
- EDAを行うNotebookとモデルを作成するNotebookを分けました。
- EDAの結果や，参考になる資料の調査結果，スコアの推移などをまとめるのにNotionを活用しました。