Home Credit Default Risk 解説

Private LB: 0.76043
Public LB: 0.76749

0. 全体を通して

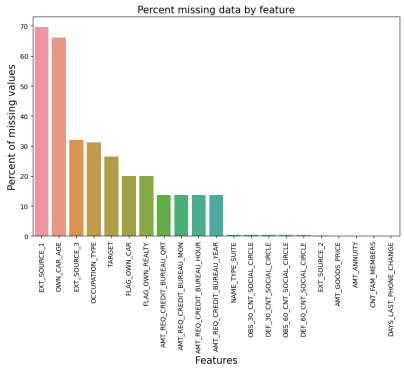
Jikky1618 です。今まで機械学習・コンペの経験は全くなければ、Python 経験もなかったため、今回のコンペは Python や機械学習の記事、今までの GCI の教材を読み漁り、勉強しながらの挑戦でした。今回の特徴量は 50 種類と、前回の Titanic と比べてかなり増えているため、特徴量を全て扱うことはかなり大変だと思いましたが、Description が用意されていた為、日本語に翻訳してなんとなく全体像を理解できました。

特徴量エンジニアリングをするにあたり、kaggle のコンペの notebook を沢山参考にしました。データは完全に同じではありませんが、詳細な EDA やドメイン知識を活用した特徴量の作成方法がたくさん見つかり、非常に役立ちました。

モデルの構築に関しては、自分の知識不足で分からない部分が多く、資料を参考にしながらモデルを構築しました。パラメータチューニングは時間的に余裕がなかったため、まだ改善の余地があると感じています。

1. 欠損値の扱い

全 50 種類の特徴量の中で欠損値が存在しているのは 21 種類。それぞれの欠損率のグラフを以下に示します。

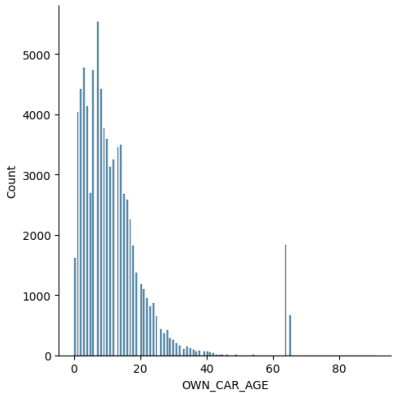


欠損率が 70% 近くある EXT_SOURCE_1 や OWN_CAR_AGE の欠損値の処理を無理やり行って、全ての欠損値を補完するのではなく、欠損値を含んでも動作する決定木系のモデルを使用することで対応を行います。

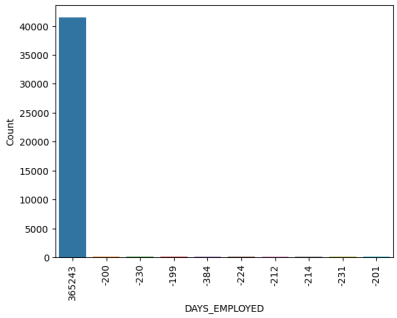
2. 特徴量エンジニアリング

かなり EDA に時間をかけた後、特徴量エンジニアリングに着手しました。
まず初めに外れ値や欠損値の代わりに入れられた値を全て欠損値として扱います。

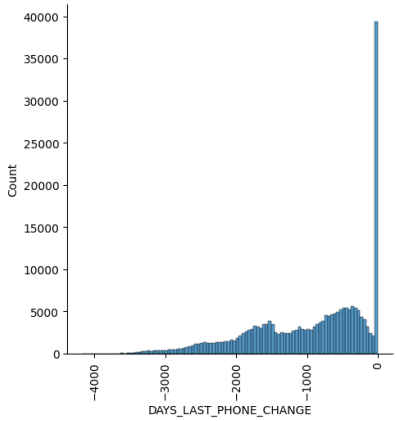
- OWN_CAR_AGE の 60 以上の値を全て外れ値として、欠損値扱いにする。



- DAYS_EMPLOYED の 365243 という値は欠損値の代わりに入れられた値だと考え、欠損値扱いにする。



- DAYS_LAST_PHONE_CHANGE の 0 という値も欠損値の代わりに入れられた値だと考え、欠損値扱いにする。



次に kaggle にあるコンペの notebook を参考にしながら、特徴量を作成します。本来は 100 種類くらい追加していましたが、なかなかスコアが伸びなかったので厳選して、最終的には 10 種類程度に落ち着きました。

- EXT_SOURCES をベースに四則演算特徴量の作成

EXT_SOURCES 同士の積と、EXT_SOURCES の平均値、分散を作成しました。他にも最小値、最大値、中央値を試しましたが、スコアが伸びなかったので不採用です。

- ドメイン知識に基づく特徴量の生成

完全に知識がなかったので、自力で考えて作成するのはかなり大変です。なので kaggle の notebook を参考に、大きく分けてローンのクレジット金額に関する特徴量、日数に関する特徴量、クライアントの収入に関する特徴量を作成しました。他にも days_birth から年齢を計算して 6 分割する特徴量も生成しましたが、あまり効果はありませんでした。

最後にカテゴリ変数を全て One Hot Encoding で数値変数に変換しました。

3. モデルの構築

モデルの選択に関してですが、決定木系のモデルの中で一番スコアが伸び、参考資料が豊富にあったものが LightGBM だったので、そちらを採用しました。また、tutorial.ipynb にもある通り、目的変数が不均衡データの為、StratifiedKFold クラスによる層化 k 分割交差検証を行いました。分割数は 5~10 を試して, 最もスコアが伸びた 8 分割を採用しました。

参考資料

- https://www.kaggle.com/code/nozomuk/home-credit-complete-eda/notebook
- https://www.kaggle.com/code/jsaguiar/lightgbm-7th-place-solution
- https://www.kaggle.com/code/jsaguiar/lightgbm-with-simple-features
- https://github.com/Hirochon/GCI2020-Summer/blob/master/Competition2/3rd_place_solution.ipynb