第二回コンペ　説明文

今回のコンペの課題について、取り組んだ作業について説明します。

やったこと

①データの可視化、確認

今回の課題は特徴量自体と、関連性のありそうなカラムが多かったことから、まずはデータの可視化と欠損値の確認を行いました。グラフ等で数値がどのようになっているのかを見る中で、分析に使わないカラムを決めました。具体的には、ほとんどの人が所持し連絡が取れていた携帯電話に関するカラム（FLAG_MOBIL, FLAG_CONT_MOBILE）と、他のカラムと被っている要素が多かった本籍と住んでいる地域の違いについてのカラム（REG_REGION_NOT_LIVE_REGION, LIVE_REGION_NOT_WORK_REGION）を使わないカラムとして設定しました。

②特徴量エンジニアリング

公開されている Home Credit の GitHub を参考にしながら、四則演算に新しい特徴量を作成することを目指しました。

・EXT_SOURCE については、それぞれが外部データベースの正規化スコアであるということで、3 種類の平均、2 種類ずつの平均と差の絶対値を新たな特徴量としました。

・ローンの具体的な金額に関わるカラムである、AMT_CREDIT、AMT_ANNUITY、AMT_GOODS_PRICE について、それぞれの商や差を特徴量としました。

・DAYS_BIRTH を 365 で割り申請時の年齢（年単位）を算出し、OWN_CAR_AGE（自動車の年齢）を引いて、自動車を所持している年数という特徴量を作成しました。

・申込前の信用情報機関への紹介件数に関するカラム（AMT_REQ_CREDIT_BUREAU_〜）について、合計を求めました。また、その数値に 0 があるかないかが重要だと考え、それを確認するため、それぞれの積も新たな特徴量としました。

・NAME_TYPE_SUITE について、同行者がいるかいないかのデータに置き換えました。

・残りのカテゴリカル変数に対し、One Hot Encoding をしました。

③機械学習モデルの作成

StratifiedKFold を使った交差検証を行いました。Fold 等の数値は実際のスコアを見ながら調整しました。