コード解説文 YKanenori98

CV スコア：0.763366 public LeaderBoard(LB)のスコア：0.76777 private LB のスコア：0.7611

第 1 回のコンペではチュートリアルを提出しただけなのでコンペに参加したのは実質今回が初めてになります。といっても時間が足りず多くのことには取り組めませんでした。

金融業界のドメイン知識は全く持っていないのでとりあえず Kaggle のコードやディスカッションを調べました。すると過去の GCI 受講者で上位入賞された方がコードを公開されていました。こちらに従って csv ファイルを作成し提出したところ public LB で 0.76768 という高いスコアが得られたのでこちらをベースラインとしました。

https://github.com/Hirochon/GCI2020-Summer/blob/master/Competition2/3rd_place_solution.ipynb

こちらのコードでは lightgbm を採用していましたが、今回 optuna を使ったハイパラ探索を勉強したかったことと他のモデルを試してみたかったので、以下のサイトなどを参考にして xgboost のハイパラを optuna で探索しました。

https://datatechlog.com/how_to_use_xgboost_and_optuna_for_parameter_optimization/#toc26

得られたハイパラで予測をしてみたところ public LB でのスコアが上がったので今回はこちらを提出しました。

他に optuna でハイパラ探索するときに StratifiedKFold を使うコードを lightgbm と xgboost の両方で試しましたが時間がやたらかかるうえにスコアは下がり、いいことがありませんでした。

あとは Kaggle で提出されていたコードを参考にして時間に関する特徴量を作ったり余分な特徴量を落としたりして試しました。が、試した中では上のものが結局一番 public LB でのスコアが

https://www.kaggle.com/code/jsaguiar/lightgbm-7th-place-solution

https://www.kaggle.com/code/hikmetsezen/base-model-with-0-804-auc-on-home-credit

良かったので採用しませんでした。ただこれらについては optuna でハイパラ探索するときスコアが落ちた StratifiedKFold の方を使ったので、もしかしたら hold-out 法で探索していたらこっちの方がスコアは上がったのかもしれません。あとは CV スコアと public LB のスコアが相関とれず悩みました。CV スコアの方を信じてもよかったのかもしれません。