Шаги:

❖Загрузка данных

- Проверка, что признак равен нулю
- Преобразуем buy_time к дате

❖Краткий разведочный анализ данных

- Проверка на пропуски
- Распределения (целевой переменной target, признака vas_id, диапазоны дат трейна и теста, id абонентов)
- Поиск дубликатов id

❖Сохраняем в pkl

❖Загружаем из pkl

❖Исследование повторяющихся id

Шаги:

- ❖ Объединение merge_asof nearest (модель LAMA, LAMA + stacking)
- ❖ Объединение merge_asof backward (модель LAMA)
- ❖ Объединение merge_asof forward (модель LAMA)
- ❖ Исследование дублирования id
- ❖ XGBoost (hold out, cross-validation, бустинг dart)
- ❖ Результаты
- ❖ Обучение модели на всём датасете и сохранение для прогноза
- ❖ Формирование индивидуальных предсказаний для абонентов

Опробованы модели:

- На основе фреймворка LightAutoML от Sber (linear_l2 – линейная с L2 регуляризацией (ridge), lgb – LightGBM с гипер-параметрами по умолчанию, lgb_tuned – LightGBM с гипер-параметрами оптимизированными Optuna, стэкинг: ridge + LightGBM + CatBoost состеканные с LightGBM (Optuna) + CatBoost)

- Опробованы варианты объединения датафреймов с помощью merge_asof (nearest, backward, forward)

- Опробован XGBoost с типами бустинга (gbtree - по умолчанию, dart - не склонный к переобучению)

- Выбрана модель XGBoost как показавшая наивысшее качество.

Принцип предложений для абонентов:

- Модель обучается на всей обучающей выборке.
- Модели подаются на вход каждый класс подключённой услуги.
- По каждому варианту услуги считается вероятность подключения.
- Выбирается услуга, имеющая наибольшую вероятность подключения

Вывод:

Алгоритм можно улучшить, установив порог вероятности для рекомендации услуги. Таким образом, если услуга с максимальной вероятностью для данного абонента тем не менее, низка (порог необходимо подбирать ориентируюсь на потребности бизнеса), услугу не стоит подключать. Это позволит избежать негативного эффекта снижения лояльности клиентов от «несработавшей» рекомендации.