Модель определения вероятности подключения услуги

С.Панкратов, 2022

Задача

Построить алгоритм, который для каждой пары пользователь-услуга определит вероятность подключения услуги

Данные

Данные train и test разбиты по периодам – на train доступно 4 месяцев, а на test отложен последующий месяц.

Отдельным набором данных является нормализованный анонимизированный набор признаков, характеризующий профиль потребления абонента.

Метрика

F-мера, с усреднением “macro”

Оптимизация памяти, отбор признаков, устранение дисбаланса классов

Применение модуля Dask для работы с большой (20Gb) таблицей признаков (>250), изменение типов числовых данных	потребление памяти снизилось в 3-4 раза
Создание новых признаков, их **отбор** с использованием алгоритма BorutaShap (shadow-features, shap-values, stat-tests)	из почти 260 признаков выбрано 10 важных без потери в качестве предикта
Решение проблемы дисбаланса классов без under/over-sampling или SMOTE, с помощью расчета весов классов (CatBoost - “SqrtBalanced”), что позволяет модели не переобучаться и не терять важную информацию	

Модель CatBoost, ее преимущества и гиперпараметры

Преимущества:

1. Хороший bias / variance tradeoff;
2. Возможность обучения на GPU;
3. Возможность извлечения сложных нелинейных связей между признаками и таргетом;
4. Позволяет получать хорошие предсказания с параметрами по умолчанию;
5. Обработка пропусков автоматически;
6. Низкая чувствительность к выбросам и др.

Главные гиперпараметры подобраны аналитически (без автоподбора, например, optuna):

1. Iterations (количество деревьев в ансамбле) - 500;
2. learning rate (поправка к gradient residual) - 0.34
3. depth (глубина каждого дерева) - 6;
4. grow_policy (способ построения дерева) - SymmetricTree (одинаковый критерий сплита на каждом уровне дерева - таблица)

Результаты обучения, кросс-валидация

1. Кросс-валидация на 5 фолдах;
2. Низкое отклонение от среднего значения метрики;

```
1-fold score = 0.7681
2-fold score = 0.7668
3-fold score = 0.7680
4-fold score = 0.7672
5-fold score = 0.7681
Точность перекретсной оценки: 0.7677 +/- 0.0005
CPU times: total: 15min 2s
Wall time: 1min 20s
```

1. Приемлемые значения метрик;
2. Отсутствие переобучения;

Train

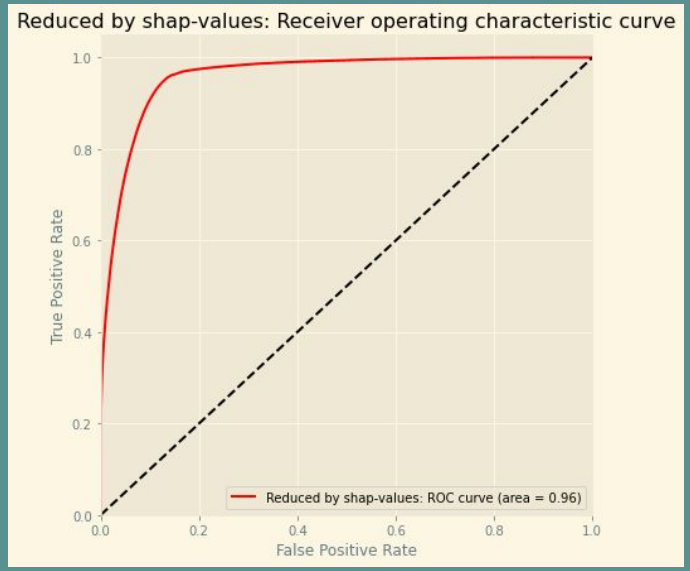
	precision	recall	f1-score	support
0.0	0.9816	0.9337	0.9570	617174
1.0	0.4771	0.7756	0.5908	48149
accuracy			0.9222	665323
macro avg	0.7293	0.8546	0.7739	665323
weighted avg	0.9451	0.9222	0.9305	665323

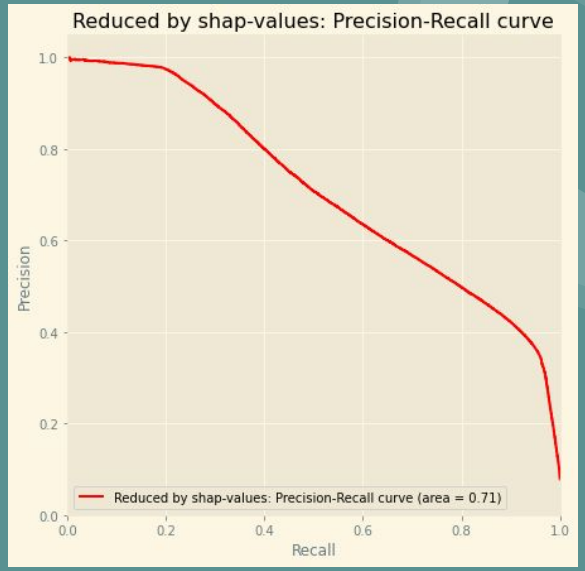
Test

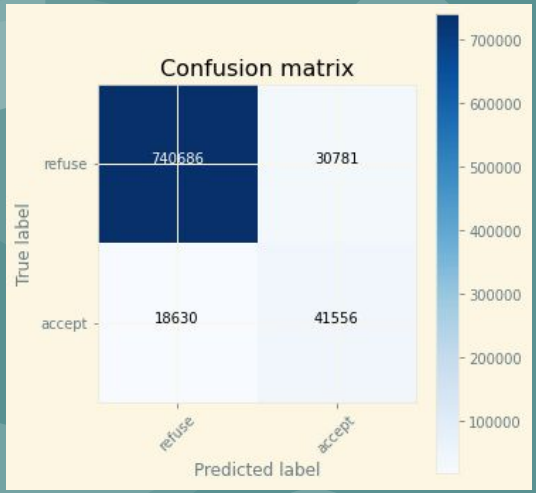
	precision	recall	f1-score	support
0.0	0.9808	0.9327	0.9562	154293
1.0	0.4704	0.7658	0.5828	12037
accuracy			0.9207	166330
macro avg	0.7256	0.8493	0.7695	166330
weighted avg	0.9439	0.9207	0.9291	166330

Best thres: 0.5593782580216716, f1-score: 0.5867, fmacro: 0.7736

Финальные метрики и калибровка порога вероятности







Обучение на полном наборе данных и калибровка вероятности повысило качество метрики

```
Best thres: 0.6113188028649806, f1-score: 0.6272, fmacro: 0.7974
```

Построение пайплайна Luigi

Две задачи:

1. Загрузка тестового набора данных и его обработка для передачи в модель;
2. Создание таблицы предсказаний (id, vas_id, buy_time, target) путем применения предобученной модели CatBoostClassifier.

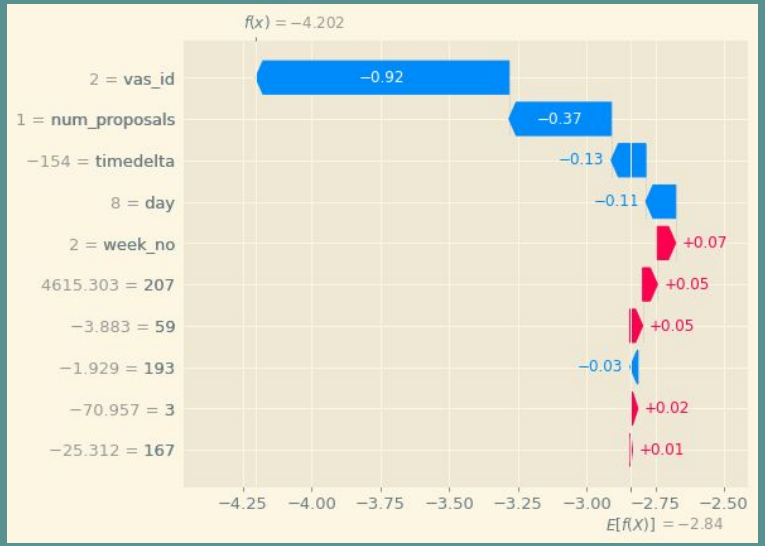
	Name	Details	Priority	Time	Actions
DONE	Transformer	path_data_train=data_test.csv, path_features=features.pkl	0	03.08.2022, 00:34:43	
DONE	Forecaster	path_data_train=data_test.csv, path_features=features.pkl, model_name=cat_red, threshold=0.6113188028649806	0	03.08.2022, 00:34:43	

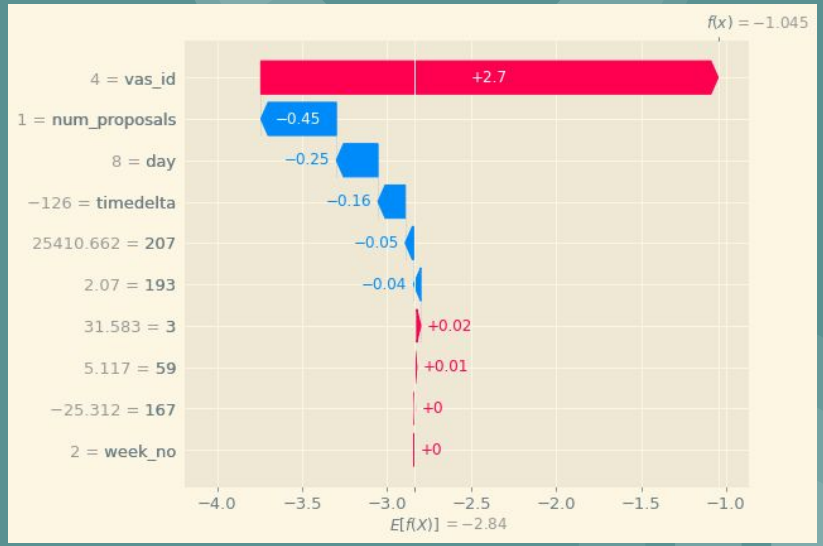
Интерпретация работы модели для предложения услуги пользователю

Ввиду отсутствия бизнес-контекста в предложенной задаче, принято решение о направлении предложения пользователю на основе метки класса (не вероятности, т.к. положительных меток немного) с учетом shap-values

Дополнительно:

1. проверять факт подключения услуги перед предложением;
2. не предлагать услугу пользователю, отказавшемуся несколько раз ранее;
3. В зависимости от величины профита и затрат устанавливать уровень порога вероятности (precision/recall).





* vas_id сильно влияет на предсказание, в декабре была некая акция, что может давать смещение в дальнейшем: можно обучить несколько моделей на каждой услуге либо взять выборку за год.

Как применить?

Запустить скрипт get_predictions.py в терминале (требуются файлы features.pkl, prop_dict.pkl, cat_red.cbm, data_test.csv):

$ python get_predictions.py

Либо с указанием параметров:
$ python get_predictions.py -p data_test.csv -f features.csv -m cat_red -t 0.6113188028649806

Данное сообщение = успех

```
Scheduled 2 tasks of which:
* 2 ran successfully:
    - 1 Forecaster(path_data_train=data_test.csv, path_features=features.pkl, model_name=cat_red, threshold=0.6113188028649806)
    - 1 Transformer(path_data_train=data_test.csv, path_features=features.pkl)

This progress looks :) because there were no failed tasks or missing dependencies

===== Luigi Execution Summary =====
```