X5Group

Cup IT 2022

ПОИСК LOOK-ALIKE АУДИТОРИИ

X5Group

Для построения модели на основе классификации данных, которая производит скоринг участников программы лояльности мы:

- Провели обработку данных;
- ✓ Обучили несколько различных моделей и выбрали лучшую;
- ✓ Настроили гиперпараметры;
- ✓ Обучили повторно лучшую модель с выбранными;
- ✓ гиперпараметрами
- Визуализировали метрики качества
- ✓ Посмотрели на то, какие признаки модели были наиболее
- ✓ важны для классификации

Итоговая модель - CatBoost Classifier

Основные параметры:

boosting_type	Ordered
depth	3
l2_leaf_reg	4.758
learning_rate	0.056
max_ctr_complexity	6
n_estimators	306

- Предобработка данных
- Градиентный бустинг - CatBoost Classifier
- Балансировка классов + Undersampling
- Настройка гиперпараметров
- Подбор threshold

Результаты

F-score validation 0.71
F-score test 0.64

Почему CatBoost?

1. Библиотека позволяет получить отличные результаты с параметрами по умолчанию.
2. Обеспечивает повышенную точность за счет уменьшения переобучения.
3. Умеет “под капотом” обрабатывать пропущенные значения.

При подготовке данных мы столкнулись со сложностями и пробовали применять различные подходы к их решению.

X5Group

Проблема

- ✗ Классы покупателей сильно не сбалансированы. Клиенты, не вступившие в клуб составляют ~ 90% выборки.
- ✗ Большое количество пропущенных значений в столбцах. Для некоторых признаков >90%.
- ✗ Отрицательные значения в столбцах со стандартным отклонением суммы. Строки, в которых месячный товарооборот ниже суммы товарооборотов в категориях.

Решения

- ✓ Undersampling
  Oversampling
  Использование весов при обучении
- ✓ Обучение с CatBoost, который умеет обрабатывать NaN
  Заполнение пропущенных значений нулями
- ✓ Зануление отрицательных стандартных отклонений.
  Прибавление к месячному товарообороту разницы с суммой

Для решения проблемы несбалансированности значительно лучшего результата удалось добиться при помощи использования весов при обучении, не прибегая к Oversampling и Undersampling. Также заполнение пропусков нулями, зануление отрицательных значений и добавление новых признаков улучшило качество модели.

Мы провели анализ различных методов машинного обучения. Решая задачу поиска look-alike аудитории, наилучших результатов удалось добиться с CatBoost Classifier.

X5Group

	Встроенная обработка Nan	Балансировка классов	Поддержка регуляризации	Поддержка большого числа признаков	F1 -score
k-NN	—	—	—	—	0.05
Logistic Regression	—	+	+	+	0.07
Random Forest	—	+	+	+	0.14
XGBoost	+	+	+	+	0.55
CatBoost Classifier	+	+	+	+	0.71

CatBoost имеет наивысший f1 score среди всех моделей, протестированных нами. Он удовлетворяет всем необходимым для нас требованиям к модели.

X5Group

Для улучшения качества модели мы провели подбор гиперпараметров на кросс-валидации с использованием класса StratifiedKFold из sklearn и библиотеки Hyperopt.

F-score на обучающей и на валидационной выборках задаются следующим образом:

1. Для каждого фолда считается лучшее F1-score на обучении и валидации;
2. Берется среднее от этих значений соответственно

Цель: Добиться высокого F1-score и на трейне и на валидации

Для этого мы вводим функционал, зависящий от F1-score на обучении и на валидации. Подбираются гиперпараметры, минимизирующие функционал.

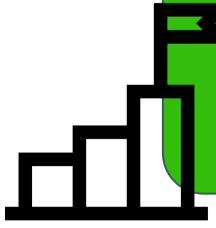
$$F(f1_{val}, f1_{train}) = -f1_{val} e^{-|f1_{val} - f1_{train}|}$$

Таким образом, мы получаем:

- **F1-score на трейне и на валидации, стремящиеся к 1**
- **Разница между F1-score уменьшается**

Параметры CatBoost Classifier, подобранные после оптимизации:

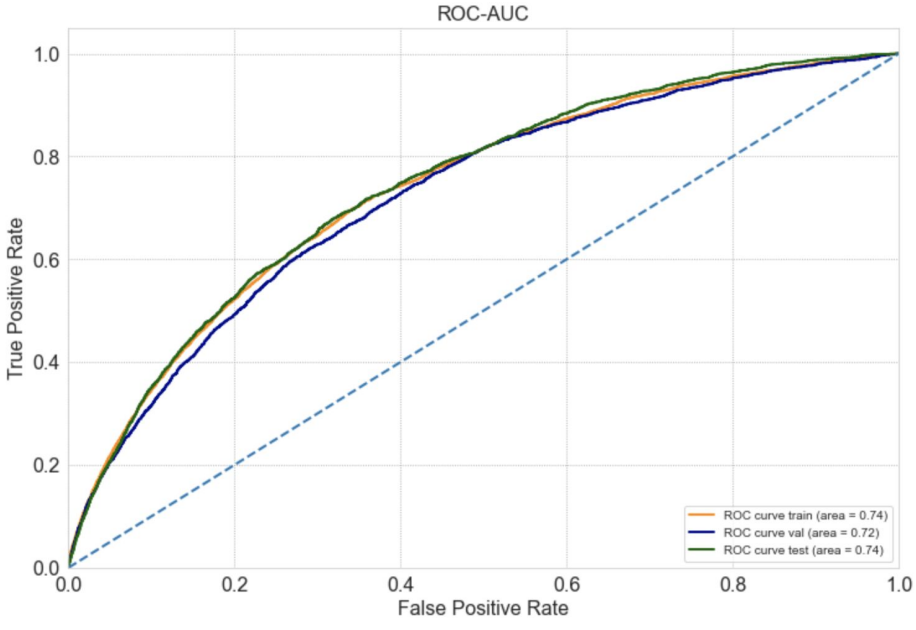
boosting_type	Ordered
depth	3
l2_leaf_reg	4.758
learning_rate	0.056
max_ctr_complexity	6
n_estimators	306



Результаты:

F1-score val: 0.71

F1-score test: 0.64



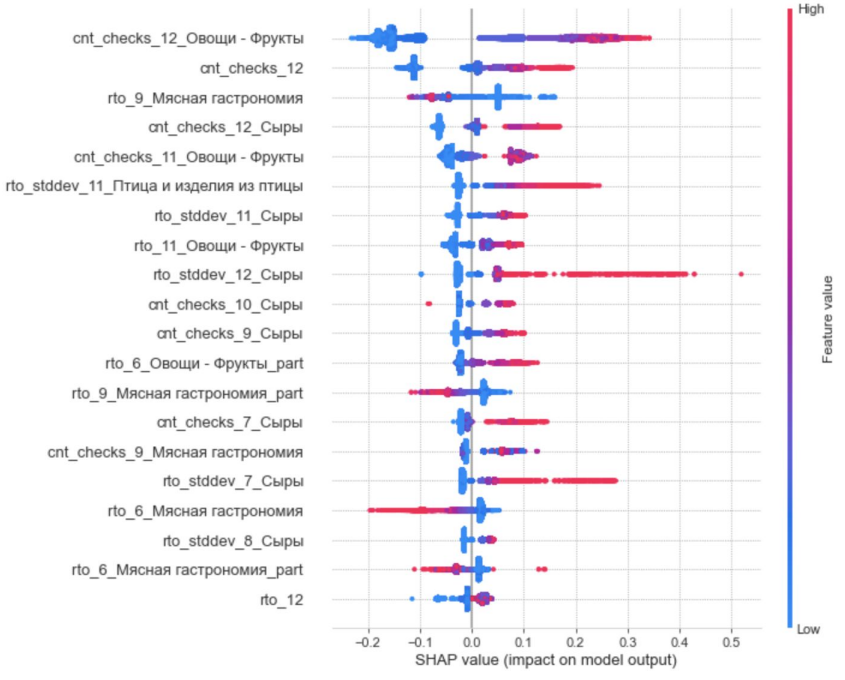
ROC-AUC на train, val и test практически не отличаются, таким образом модель имеет достаточно хорошую обучающую способность, а не подстраивается под train.

Threshold

Для улучшения качества модели мы подбирали порог отсечения классов (threshold). Для сравнения в таблице ниже приведены значения F1-score на валидации и на тесте для стандартного значения порога (0.5) и соответствующие лучшие значения после подбора.

Threshold	F-score val	F-score test
0.5	0.656	0.60
0.36	0.706	0.644
0.35	0.704	0.643
0.37	0.706	0.643

Feature importance



На данном графике показана важность фичей для принятия решения модели о распределении в класс 0 или 1.

Мельница

Серебрякова Софья

ВМК МГУ

Магистратура, 1 курс

Абдуллаева Ума

РГУНГ им. Губкина

Бакалавриат, 3 курс

Мкртчян Георгий

Сколтех

Магистратура, 1 курс

Иллюк Александр

Сколтех

Магистратура, 1 курс