CUP IT 2022
Data Science
Финал

Устранение ошибок в данных по двум гипотезам повысило основные метрики. В итоге мы выделили 13102 человека, наиболее похожих на участников клуба

Гипотеза 1. Обработка шума в *данных* повысит качество работы модели

Гипотеза 2. Обработка шума в *признаках* повысит качество работы модели

Baseline

LightGBM		
Precision	Recall	F1
0.888	0.783	0.826

LightGBM		
Precision	Recall	F1
0.892	0.850	0.868
+0.45%	+8.55%	+5.01%

- тратят меньше среднего
- сумма чеков может умеренно варьироваться
- не претендуют на кешбек, но чувствительны к акциям и проявляют интерес к полезному питанию

- тратят больше среднего
- сумма чеков может значительно варьироваться
- выполняют условия предоставления кешбека и в целом заинтересованы в полезном питании

Постепенное знакомство части аудитории с бонусами клуба позволит в перспективе увеличить конверсию и не потерять потенциальных участников

Предсказанные клиенты

Акция: 6293

Приглашение: 6809

Процесс разработки включал в себя следующие этапы:

CSV → Анализ данных и их особенностей → Извлечение новых признаков → Сплит

Аномалии → Замена некорректных значений → Трейн сэт

Корреляция фич → Отбор признаков по важности → Тест сэт

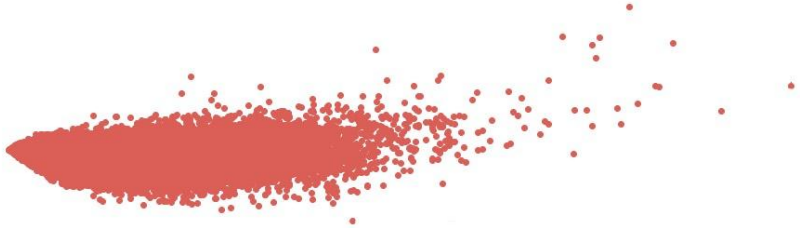
Обучение → Оценка качества и тюнинг → Разметка и вывод данных → CSV

Output: Список клиентов Перекрёстка с оценкой схожести с участниками Клуба Полезных Привычек и разметкой механизмами дальнейшего взаимодействия

Использовался тестовый набор | Использовался тренировочный набор | Новый этап

Гипотеза 1. Обработка шума в данных повысит качество работы модели

Было сделано: выделение аномалий



One Class SVM строит разделяющую границу так, чтобы по одну сторону находилась большая часть данных, а по другую - оставшиеся выбросы

Local Outlier Factor вычисляет локальное отклонение плотности каждой точки данных по отношению к ее соседям

Сохранены особенности

Выше метрики

Доработано: замена некорректных данных (отрицательных значений)

10% строк имеют некорректные значения

11.5% из них - участники клуба

14.3 тыс. всего

После обработки не грозят снижением качества предсказаний модели

Невозможно интерпретировать и извлечь смысл, поскольку ошибки носят случайный характер

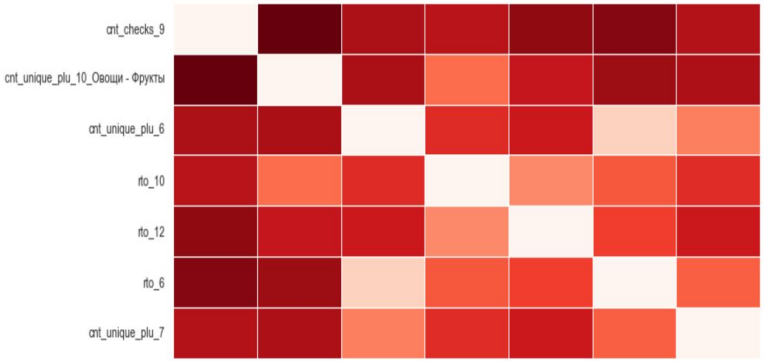
Небольшое количество относительно общего объема данных, но значительное кол-во строк

Simple Imputer заполняет выбранные некорректные значения на основе наиболее часто встречающихся

Удаление аномалий и замена некорректных данных позволят модели не переобучаться на специфических наблюдениях и в целом показывать лучшие результаты на тестовых данных

Гипотеза 2. Обработка шума в признаках повысит качество работы модели

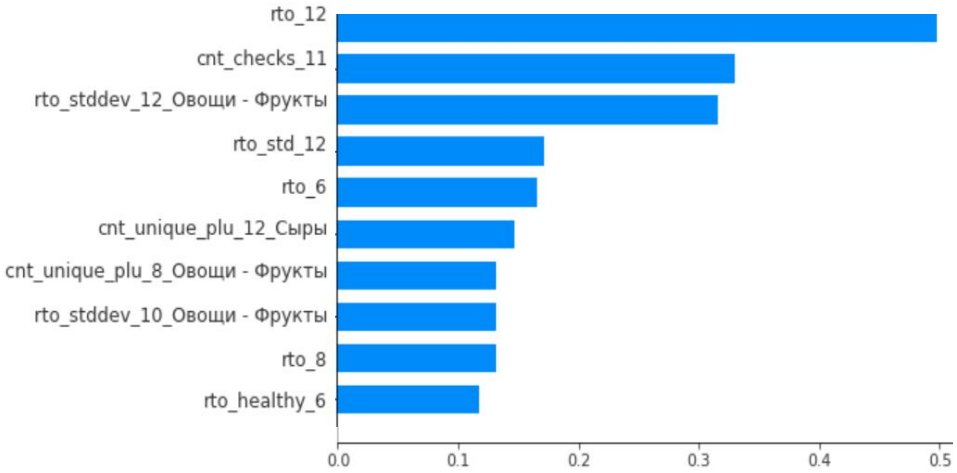
Было сделано: уменьшение размерности за счет устранения высокой корреляции



Добавили две новые колонки, рассчитанные из старых признаков

Проанализировали корреляцию между признаками и оставили наиболее важные

Доработано: отбор признаков по важности



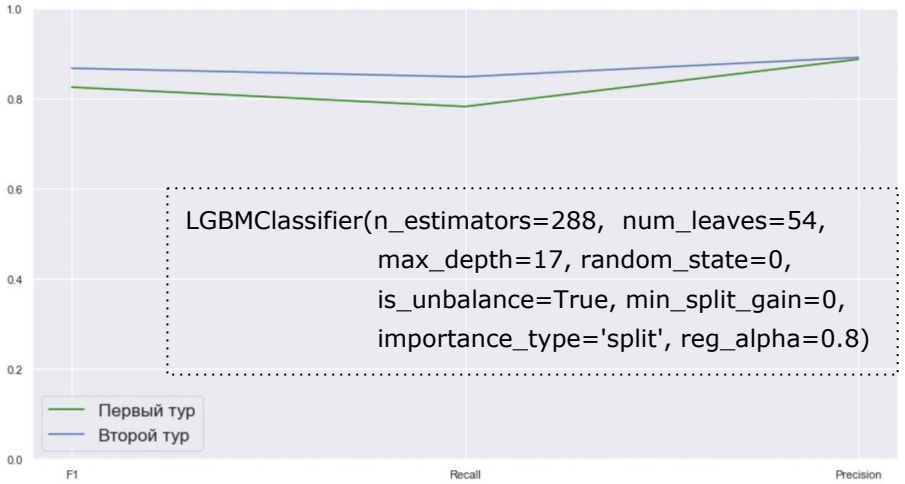
Отбор топ-80% признаков с помощью критерия Фишера оказал наибольшее влияние на метрики качества модели

Обработка признаков снизит риск переобучения и принятия решения моделью на основе шума

Мы выбрали LightGBM - высокоэффективную и гибкую модель и увеличили метрики за счет обучения на дополнительных данных и проверки двух гипотез

LightGBM Classifier - это leaf-wise grow реализация градиентного бустинга - ансамблевого алгоритма на основе решающих деревьев, последовательно уменьшающих ошибку модели

- Обработали шум в данных
- Понизили размерность признаков
- Сделали кросс-валидацию
- Использовали регуляризацию



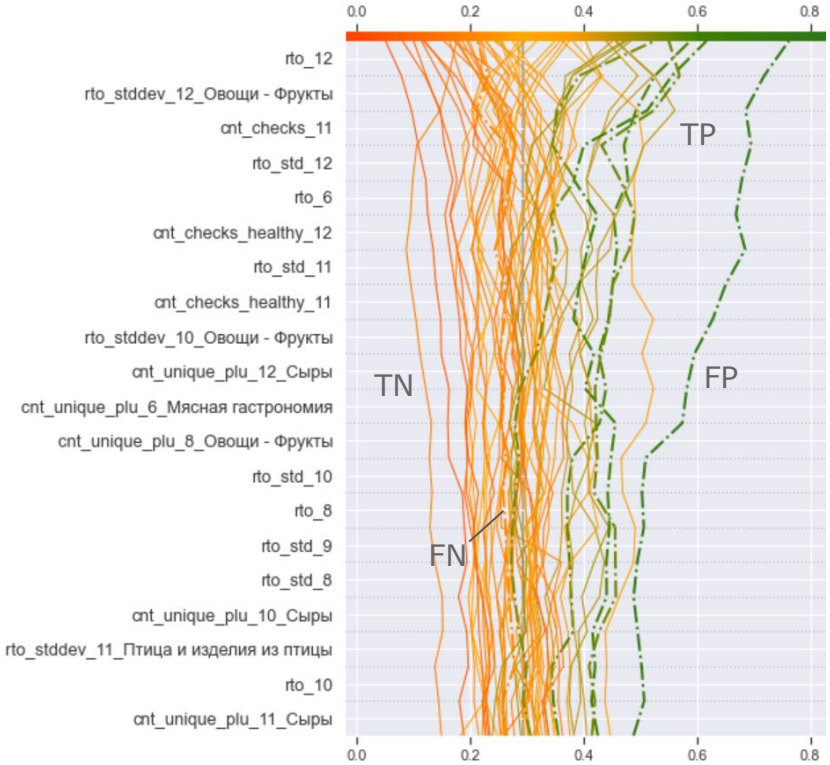
До

Precision	Recall	F1
0.888	0.783	0.826

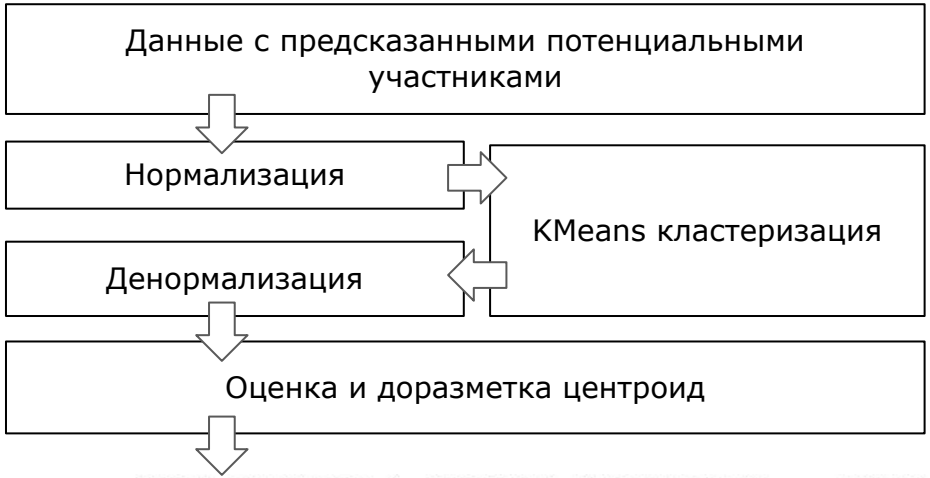
После

Precision	Recall	F1	Кол-во позитивных лейблов при обучении	Время обучения
0.892	0.850	0.868	5164 / 40519-test_labels.shape[0]	20 сек
+0.45%	+8.55%	+5.01%		

Чтобы выбрать оптимальный способ взаимодействия с потенциальными участниками клуба, мы разделили положительные лейблы на 5 кластеров



^ Влияние наиболее важных фич на процесс классификации



	rto_6	cnt_checks_6	rto_std_6	rto_healthy_6	method
0	5315.293270	7.310176	489.536881	0.368737	приглашение
1	2825.056174	6.204293	254.354934	0.167270	акция
2	8856.115390	7.063830	767.089150	0.332650	приглашение
3	22980.375587	17.535932	1193.448352	0.313797	приглашение
4	10935.371790	16.202064	566.958594	0.250087	акция

^ Центроиды с механизмом взаимодействия

One-Zero

Мичурин Артём

РЭУ им. Г.В. Плеханова
Бизнес-информатика

Data Engineer @МТС

8(916)3176642
amichurin0@gmail.ru

Исаева Диана

РЭУ им. Г.В. Плеханова
МОиАИС

Supply Chain Cup 2021
HQ 15%

8(964)0497904
dii.grase@yandex.ru

REU DS Club

Попова Нина

РЭУ им. Г.В. Плеханова
Бизнес-информатика

Supply Chain Cup 2021
HQ 15%

8(989)2666821
popovaninam@yandex.ru

Агишев Владимир

РЭУ им. Г.В. Плеханова
ПМИ

Финалист Cup IT 2022
Data Science

8(905)3960344
agishev1961@gmail.com