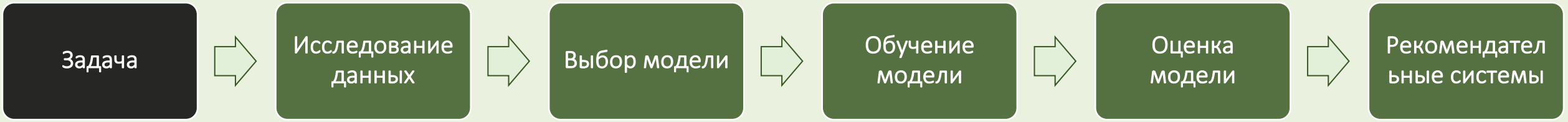
ПОИСК LOOK-ALIKE-АУДИТОРИЙ НА ОСНОВЕ АНАЛИЗА ДАННЫХ О КЛИЕНТАХ МАГАЗИНОВ «ПЕРЕКРЁСТОК»

X5 Tech

Drill Team

Ильназ Зиязиев	ilnazziaziev@gmail.com
Иван Богуш	shigurui2@gmail.com
Станислав Безвытный	stasbezv@gmail.com
Федор Бардин	fedya_bardin@mail.ru



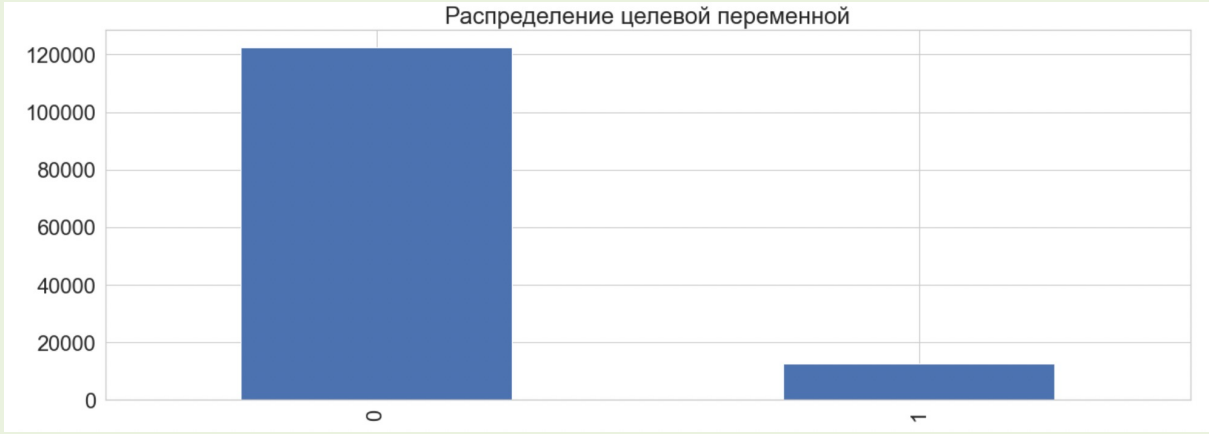
Построить модель на основе классификации данных, которая производит скоринг участников программы лояльности и предсказывает, кто из них максимально похож на людей, недавно присоединившихся к «Клубу полезных привычек».

Этапы выполнения:

- Провести EDA
- Отобрать модели и сравнить качество предсказывания
- Выбрать итоговую модель и добиться высокой точности классификации
- Оценить конечную модель
- Установить возможные способы взаимодействия с целевой аудиторией

Задача → **Исследование данных** → Выбор модели → Обучение модели → Оценка модели → Рекомендательные системы

- Всего наблюдений – 135 061
- Признаков – 149
- Целевая переменная является бинарной
- Была построена корреляционная матрица на сырых данных и на данных после предобработки
- Наблюдается дисбаланс классов



	is_in_club	rto_sum	rto_prod_sum	rto_prod_perc	checks_prod_perc
is_in_club	1.000000	0.166556	0.189555	0.102809	0.118443
rto_sum	0.166556	1.000000	0.876590	0.085134	0.336942
rto_prod_sum	0.189555	0.876590	1.000000	0.385240	0.516866
rto_prod_perc	0.102809	0.085134	0.385240	1.000000	0.639830
checks_prod_perc	0.118443	0.336942	0.516866	0.639830	1.000000

Задача → Исследование данных → **Выбор модели** → Обучение модели → Оценка модели → Рекомендательные системы

В ходе решения задачи выбора модели для классификации пользователей были рассмотрены три модели:

- RandomForestClassifier
- LogisticRegression
- LGBM Classifier

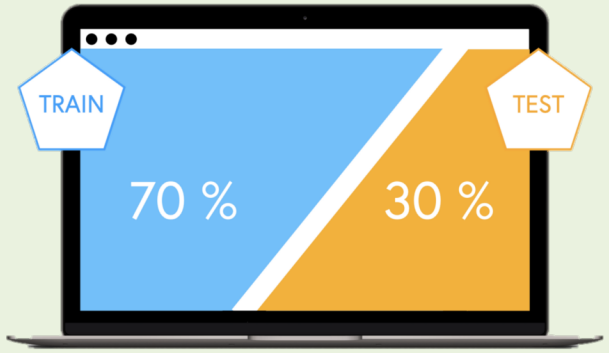
LightGBM

Преимущества LGBM, в сравнении с двумя другими:

- быстрая скорость обучения
- более низкое использование памяти
- совместимость с большими наборами данных

В случае использования RandomForest без подбора параметров, целевая метрика составила 0.32%,в случае LogisticRegression значение метрики оказалось еще ниже, 0.03%. В то время как LGBM Classifier продемонстрировал наилучшее значение метрики без подбора параметров – 4.61%. Таким образом был выбран алгоритм LGBM Classifier

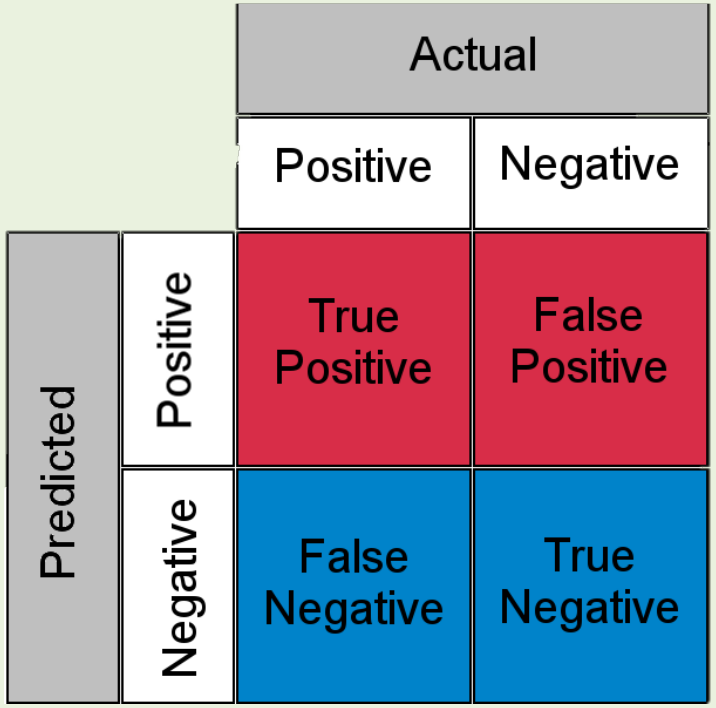
Задача → Исследование данных → Выбор модели → **Обучение модели** → Оценка модели → Рекомендательные системы



Обучение модели происходило на тренировочной части датасете, который был получен путём деления всего датасета в соотношении 70:30

Для подбора гиперпараметров был использован open-source framework - Optuna

OPTUNA



Test score:
f1: 26.02
precision: 17.40
recall: 51.60

Train score:
f1: 34.76
precision: 25.34
recall: 55.34

Для оценки модели мы использовали рекомендованные метрики F1 и precision.

Учитывая тот факт, что пользователи с меткой «0» могут быть, потенциально, новыми членами клуба, работу алгоритма оценить сложно.

Дополнительно мы решили оценить получившуюся по метрике recall

Задача → Исследование данных → Выбор модели → Обучение модели → Оценка модели → Рекомендательные системы

Варианты взаимодействия с целевой аудиторией
(предложение вступить в клуб):

- Через смс и письмо на почту, указанные при регистрации
- После оформление товаров через кассу самообслуживания при использовании карты лояльности
- Информирование о возможности вступить в клуб на чеках

Все предложенные варианты подразумевают то, что покупателям будет выгодно вступить в клуб. Возможно, стоит предложить в качестве подарка некоторое количество бонусных баллов

После анализа, предобработки и обучения модели мы смогли предсказать, кто из покупателей является потенциально возможным членом клуба здоровых привычек.

Также были предложены меры по взаимодействию с этими пользователями.

Решение состоит из:

- исполняемого кода в виде файла Jupyter Notebook
- файла .csv с предсказанием по пользователям
- презентации в формате .pdf
- readme файл