Курсовой проект от «МегаФон»

Расчет вероятности подключения услуги

Абакумова Яна | abakumovayana@yandex.ru

https://github.com/YanaAbakumova/Projects/tree/main/Megafon_project

Содержание:

Абакумова Яна | abakumovayana@yandex.ru

Задача:

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

Данные:

1.DATA_TRAIN.CSV

- id - идентификатор абонента (831 653 итого , из них 806613 уникальных);
- buy_time - время покупки (с 09.07.2018 по 31.12.2018);
- vas_id - подключаемая услуга (8 различных);
- target - целевая переменная, где 1 означает подключение услуги, 0 - отказ;

2. FEATURES.CSV

- id - идентификатор абонента (4 512 528 итого);
- buy_time - время покупки с 09.07.2018 по 21.01.2019;
- 252 анонимизированных признака

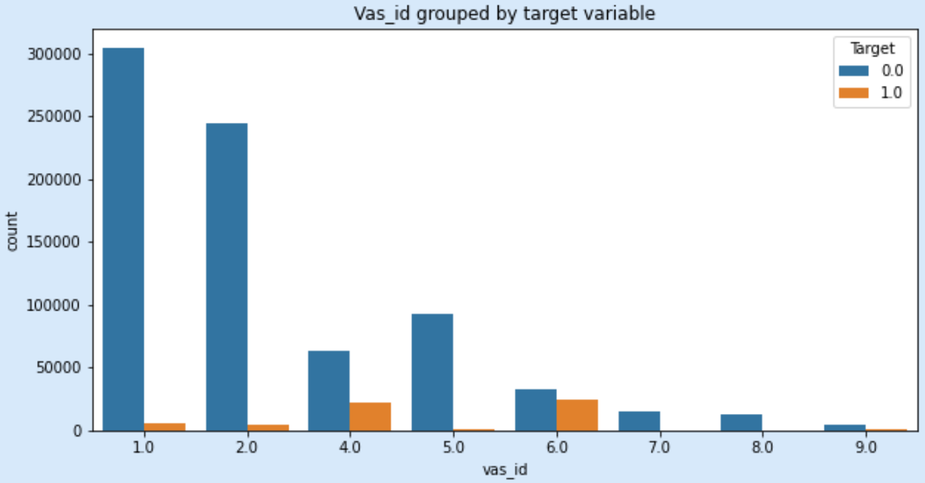
3.DATA_TRAIN.CSV

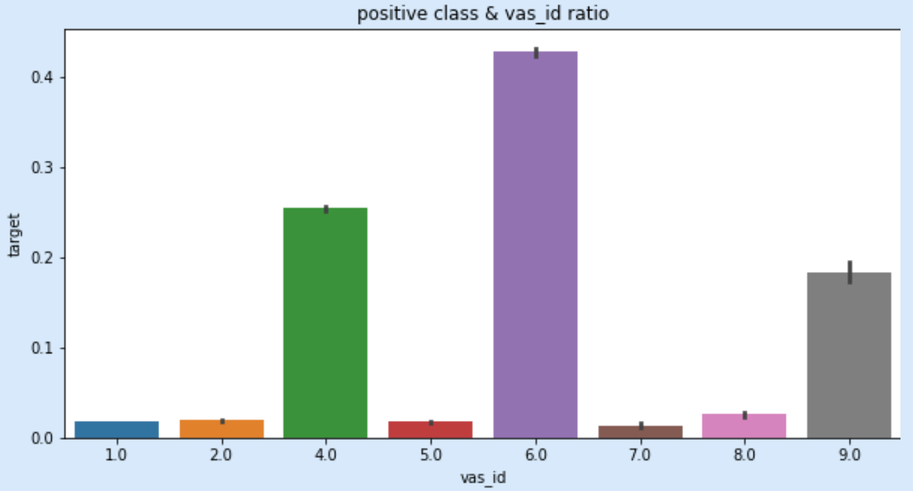
- id - идентификатор абонента (71231 итого, из них 70152 уникальных, 67013 не было в train);
- buy_time - время покупки (с 07.01.2019 по 21.01.2019);
- vas_id - подключаемая услуга (8, как и в train);

Абакумова Яна | abakumovayana@yandex.ru

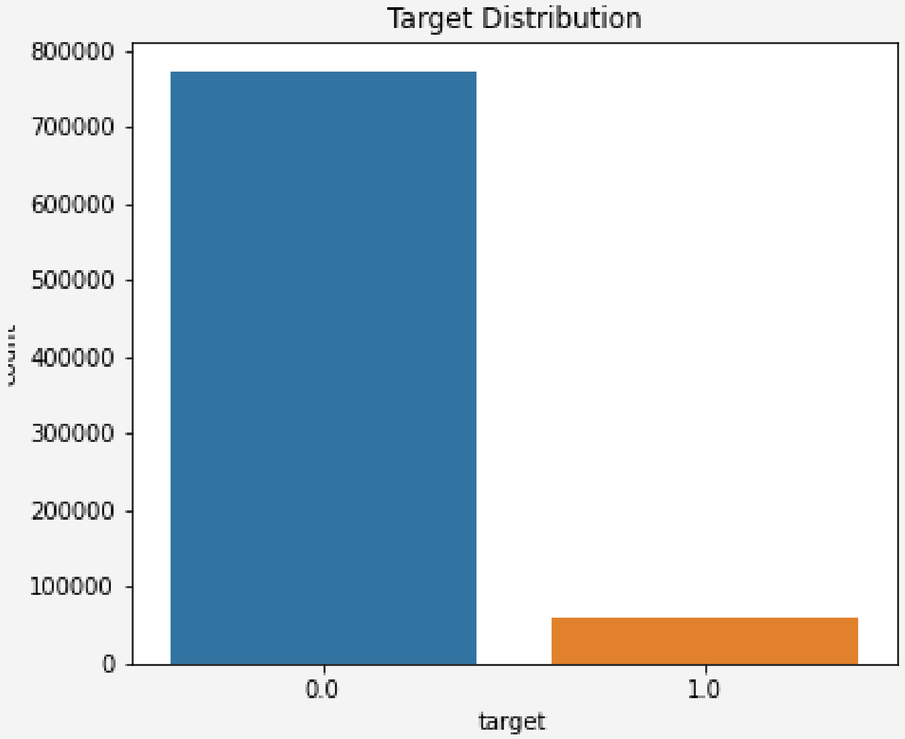
Подготовка и обработка данных:

1. Уменьшение features.csv до списка id только из train + test
2. Объединение data_train + features по id пользователей
3. Создание новых признаков:
   - дата: счёт дней, день месяца, неделя года
   - доля подключений услуг по сравнению с отказами
   - подсчет кол-ва предложений абоненту и промежутка времени между предложениями
   - информация о том, какие услуги уже предлагались ранее
   - необычные (самые высокие) значения анонимизированных признаков
4. Количество анонимизированных признаков уменьшено до 10 методом PCA





Абакумова Яна | abakumovayana@yandex.ru



Сложности:

1. Значительный дисбаланс между классами. Важно скорректировать веса.
2. Время предложения услуги не совпадает с временем формирования профиля пользователя. Тем не менее, временной промежуток всего полгода.
3. Большой объем данных снижает скорость их обработки, увеличивает сложность вычислений, важно оптимизировать ресурсы
4. Большинство признаков анонимизированы и нормализованы, их смысл не очевиден, но видна высокая корреляция между собой.

Абакумова Яна | abakumovayana@yandex.ru

Модели

Подбор параметров осуществлялся с помощью GridSearchCV

dmlc XGBoost

n_estimators=425, max_depth=6, learning_rate=0.005,
reg_lambda=0.8, reg_alpha=0.8, scale_pos_weight=3, random_state=13,
eval_metric='logloss', importance_type='weight'

CatBoost

silent=True, iterations=160, learning_rate=0.03, depth=7,
l2_leaf_reg=4, auto_class_weights='Balanced', eval_metric='F1',
early_stopping_rounds=50, random_state=42

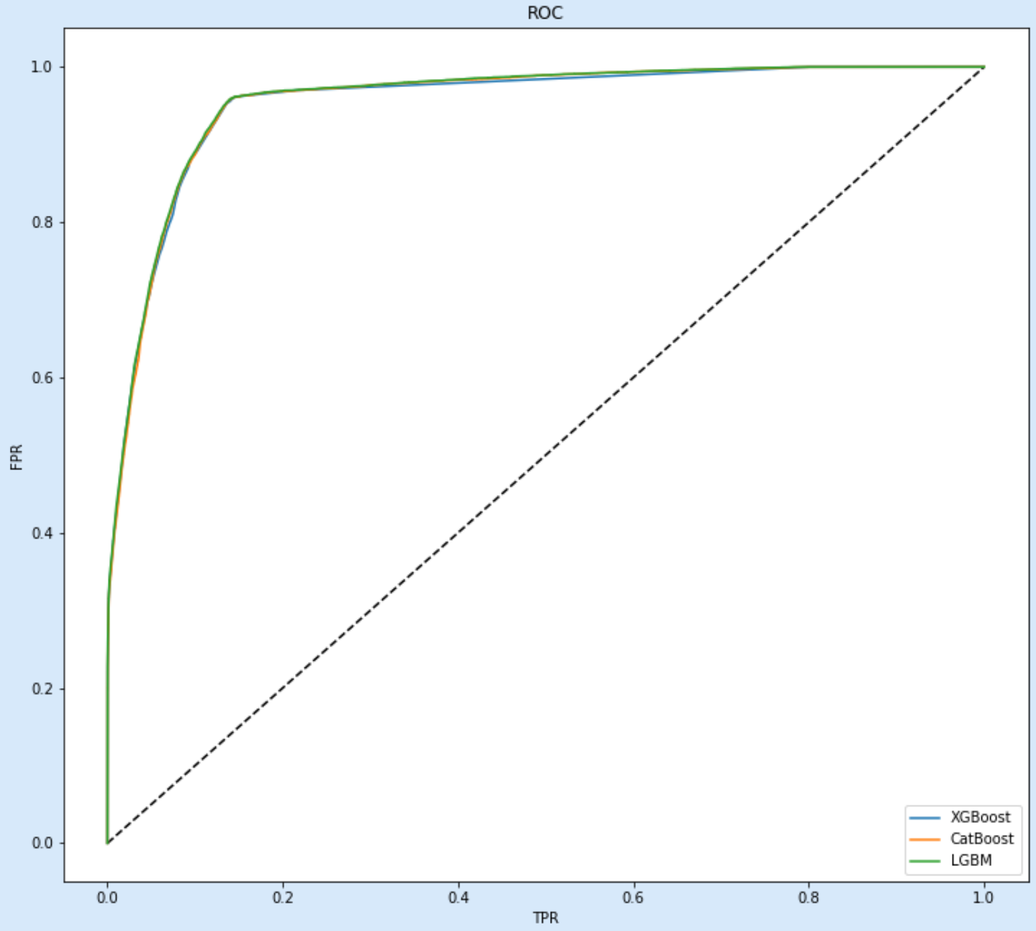
LightGBM

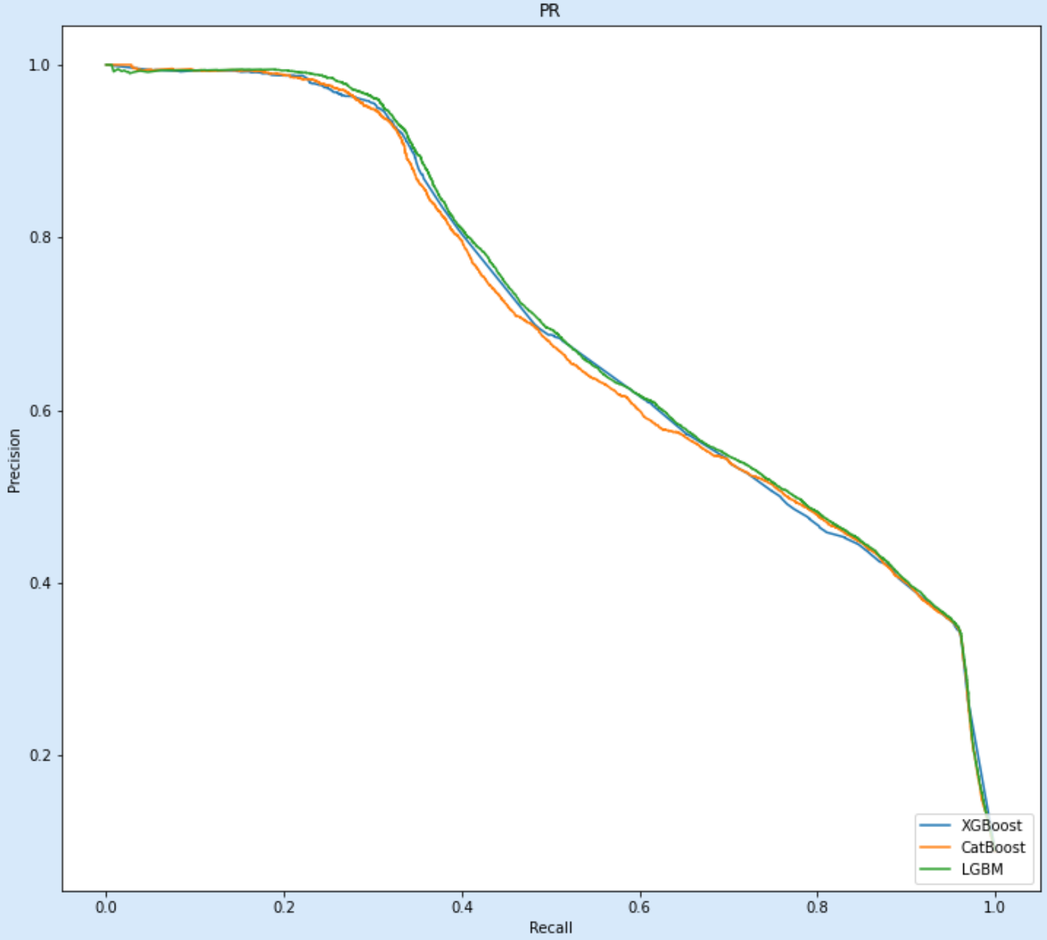
objective='binary', max_depth=13, n_estimators=100,
num_leaves = 100, learning_rate=0.045,
scale_pos_weight = 1.7935, reg_lambda = 0.2

Абакумова Яна | abakumovayana@yandex.ru

Результаты:

XGBoost : AUC_PR = 0.706, AUC_ROC = 0.953,
CatBoost : AUC_PR = 0.702, AUC_ROC = 0.955
LGBM : AUC_PR = 0.711, AUC_ROC = 0.956

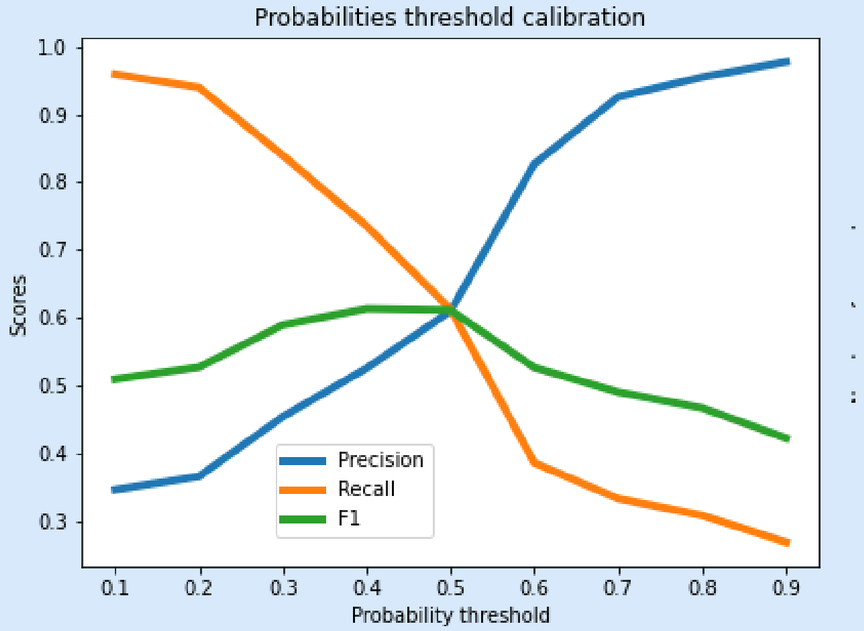


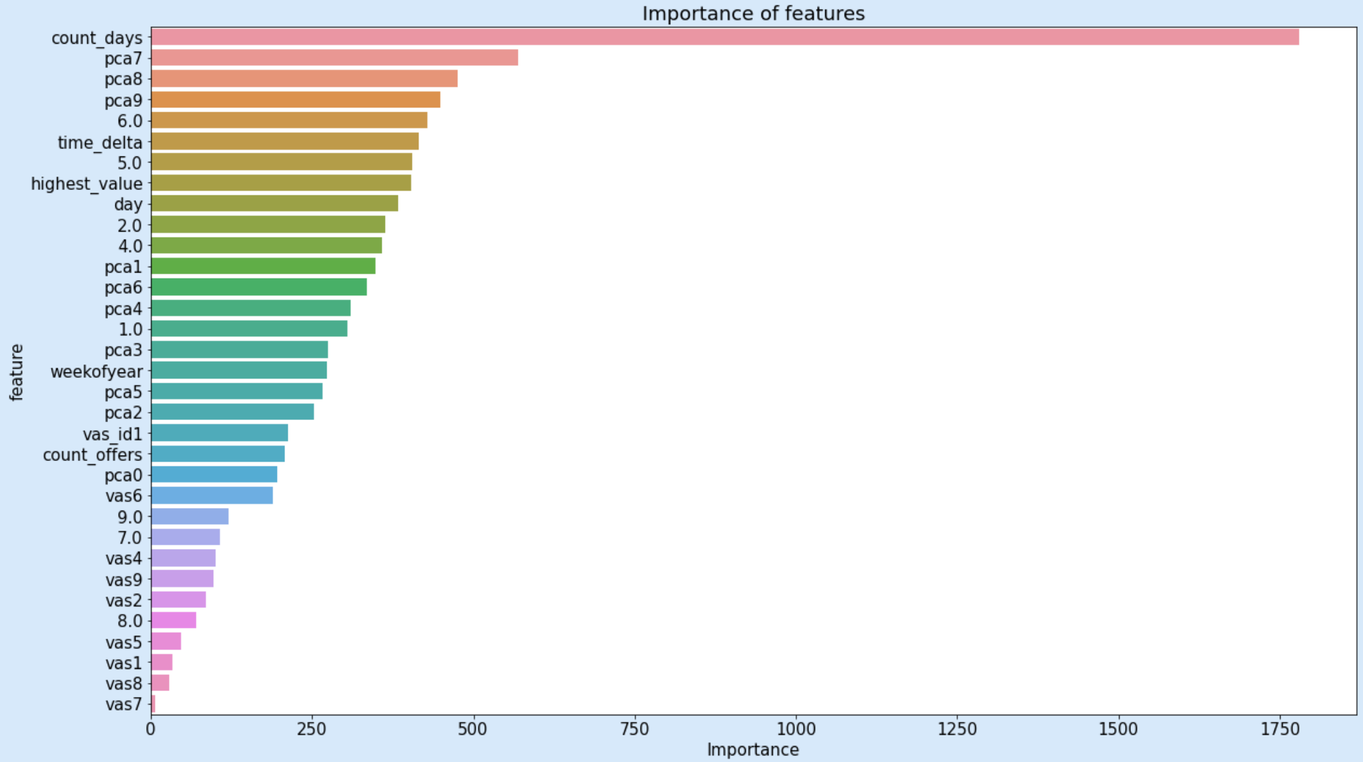


Абакумова Яна | abakumovayana@yandex.ru

Выбор: LGBMClassifier

- Оптимальный порог: 0.49
- F1-макро на валидационном датасете при этом пороге: 0,7911





По сравнению с XGB и CB:

- Лучшая скорость работы (04.39 с)
- Экономия вычислительных ресурсов
- Меньшее потребление памяти

Абакумова Яна | abakumovayana@yandex.ru

Индивидуальные предложения абонентам

Каким абонентам стоит предлагать услуги?

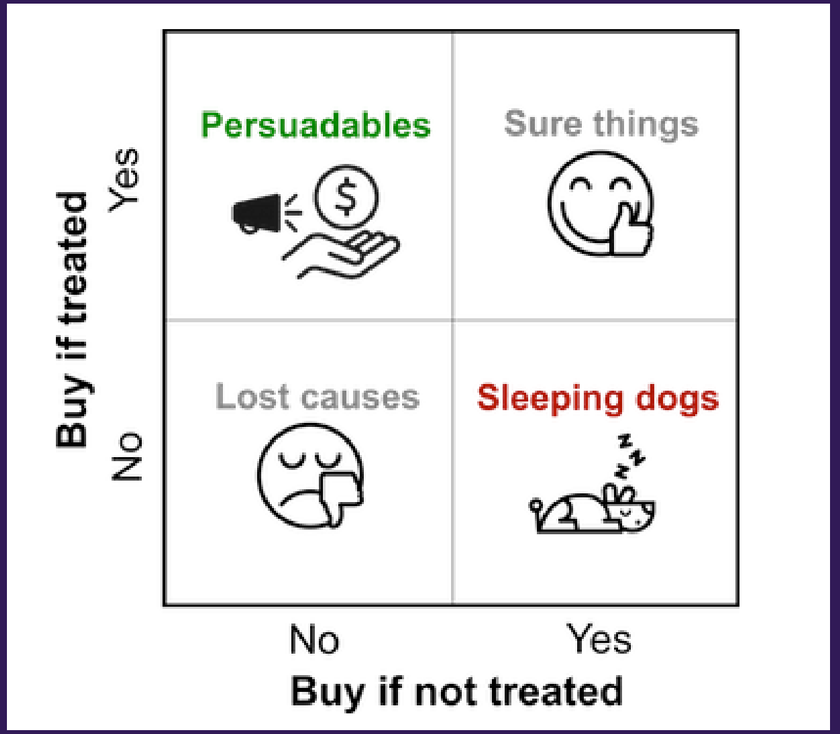
- С точки зрения ML, чтобы добиться оптимального соотношения precision-recall по позитивному классу, можно звонить клиентам, набравшим скор > 0.423
- С точки зрения бизнеса могут быть разные варианты:

- снижение порога для захвата рынка и роста recall
- повышение порога для экономии ресурсов и роста precision.

- Возможно, если клиента не заинтересовала эта услуга, ему интересна другая, нужно предложить её.

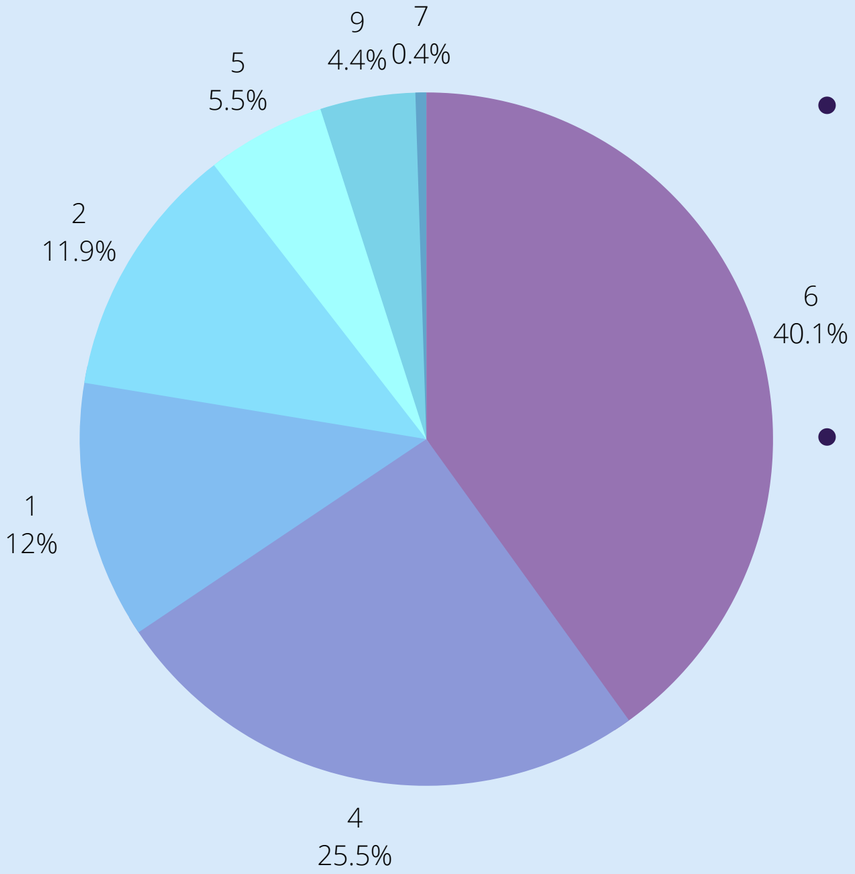
- Возможно, клиенту не интересна ни одна из услуг или клиент не настроен общаться.

Важно определить, на взаимодействие с какими абонентами имеет смысл тратить ресурсы (решение задачи Uplift)



Абакумова Яна | abakumovayana@yandex.ru

Мультиклассовая классификация с LGBMClassifier



- Позволяет подобрать услугу для выбранных пользователей

id	best_service	1	2	4	5	6	7	8	9
862975	8	0.058014	0.055012	0.021933	0.025325	0.003256	0.003395	0.799158	0.033909

- Определить, каким абонентам лучше предложить конкретную услугу

id	best_service	1	2	4	5	6	7	8	9
3814346	1	0.829745	0.025227	0.099069	0.034631	0.003106	0.000659	0.003532	0.004030
1286899	1	0.828157	0.024882	0.112372	0.018246	0.004297	0.006880	0.000737	0.004429
2025852	1	0.818740	0.047081	0.094031	0.016927	0.002638	0.012438	0.000656	0.007488

- Но не позволяет с уверенностью сказать, что на клиента имеет смысл тратить ресурсы

Абакумова Яна | abakumovayana@yandex.ru

Спасибо за внимание!

Решение:
https://github.com/YanaAbakumova/Projects/tree/main/Megafon_project

Абакумова Яна | abakumovayana@yandex.ru