



МЕГАФОН

КУРСОВОЙ ПРОЕКТ

ВЕРОЯТНОСТЬ ПОДКЛЮЧЕНИЯ УСЛУГИ

Алена Кухта | alenakukhta@yandex.ru
<https://github.com/AlenaKukhta/Megafon-course-project>

КРАТКОЕ СОДЕРЖАНИЕ

1. Задание
2. Подготовка и обработка данных
3. Выбор и сравнение моделей
4. Оценка результатов
5. Подход составления индивидуальных предложений для выбранных абонентов

ЗАДАНИЕ

ПОСТРОИТЬ АЛГОРИТМ, КОТОРЫЙ ДЛЯ КАЖДОЙ ПАРЫ ПОЛЬЗОВАТЕЛЬ-УСЛУГА
ОПРЕДЕЛИТ ВЕРОЯТНОСТЬ ПОДКЛЮЧЕНИЯ УСЛУГИ

ДАННЫЕ:

DATA_TRAIN.CSV

- id – идентификатор абонента: всего 831 653, в том числе 806 613 уникальных
- vas_id – подключаемая услуга: 8 различных
- buy_time – время покупки
- target – целевая переменная: 1 – подключение услуги, 0 - отказ

FEATURES.CSV.ZIP

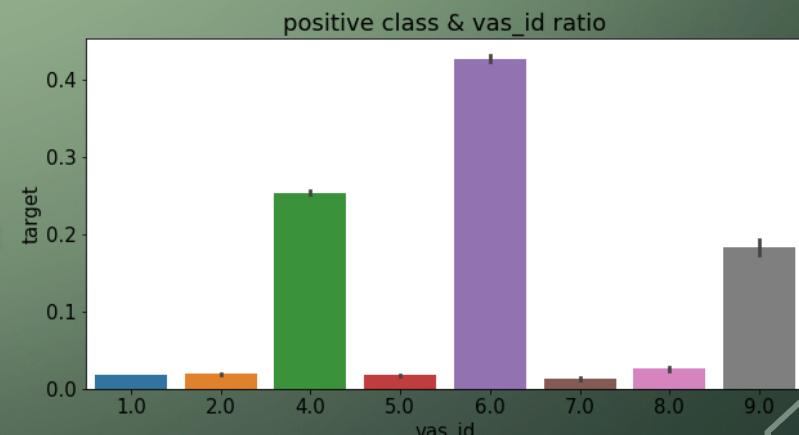
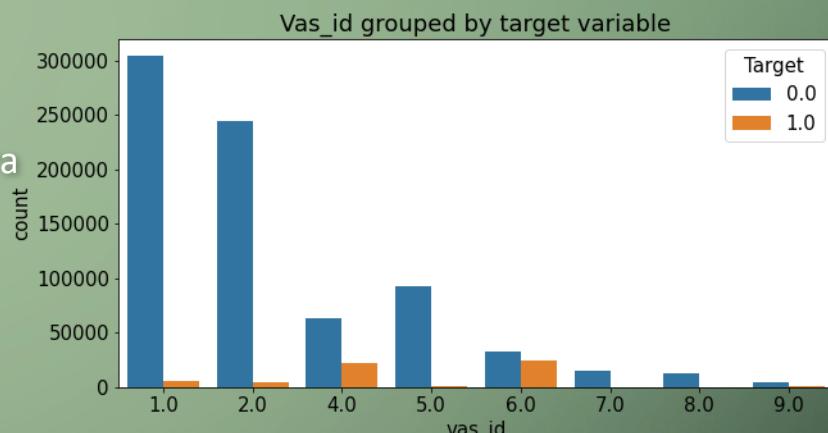
- id – идентификатор абонента
- 252 анонимизированных признака

DATA_TEST.CSV

- id – идентификатор абонента: всего 71 231, в том числе 70 152 уникальных, 67 013 не было в train
- vas_id – подключаемая услуга: 8 различных, как в train
- buy_time – время покупки

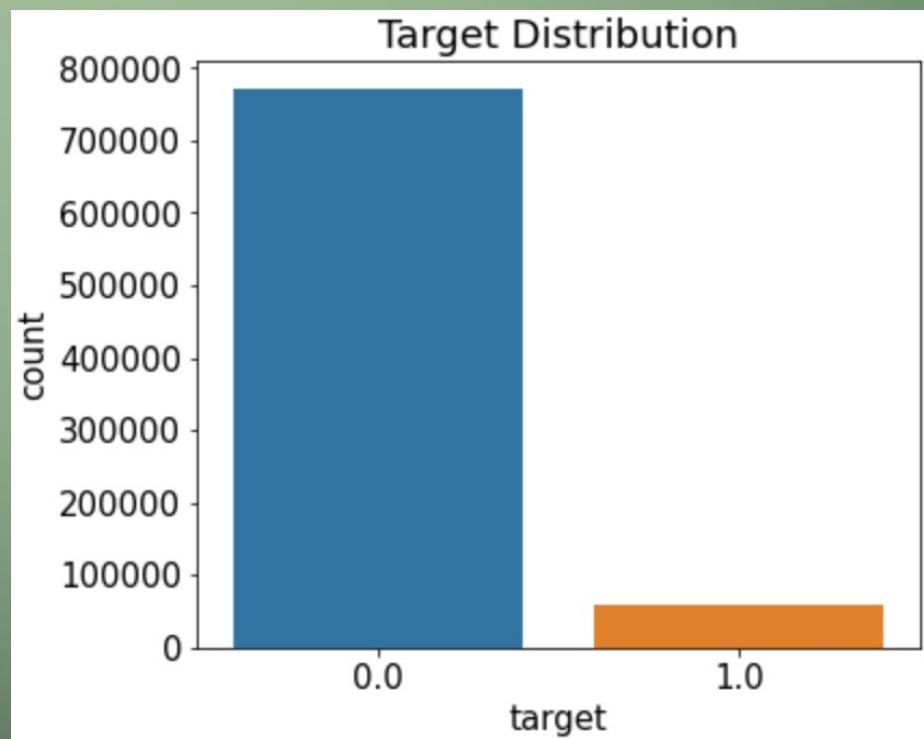
ПОДГОТОВКА И ОБРАБОТКА ДАННЫХ

- 1) Уменьшение features.csv до списка id, общих по train и test
- 2) Объединение data_train и features по id абонента
- 3) Создание новых признаков:
 - дата
 - доля подключения услуг по сравнению с отказами
 - количество предложений абоненту и промежутка времени между предложениями
 - информация о ранее предложенных услугах
 - самые высокие значения анонимизированных признаков
- 4) Количество анонимизированных признаков сокращено до 10 (метод PCA)



ЧУВСТВИТЕЛЬНЫЕ МОМЕНТЫ

- 1) Дисбаланс классов. Нужно корректировать веса
- 2) Разрыв времени предложения услуги и формирования профиля абонента
- 3) Значительный объем данных, снижающий скорость их обработки, увеличивающий сложность вычислений. Нужно оптимизировать ресурсы
- 4) Большая часть признаков анонимизированы и нормализованы. Их смысл не однозначен при наличии корреляции



МОДЕЛИ

ПОДБОР ПАРАМЕТРОВ ОСУЩЕСТВЛЯЛСЯ С ПОМОЩЬЮ GRIDSEARCHCV

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

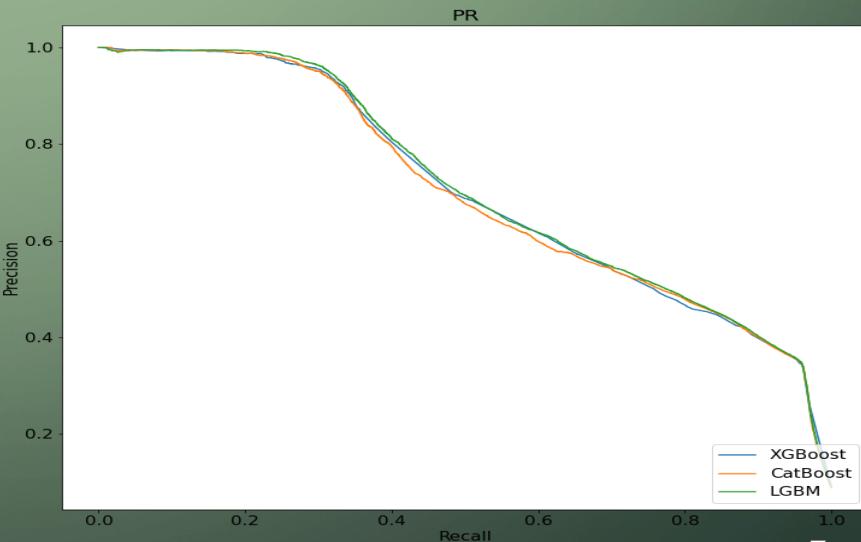
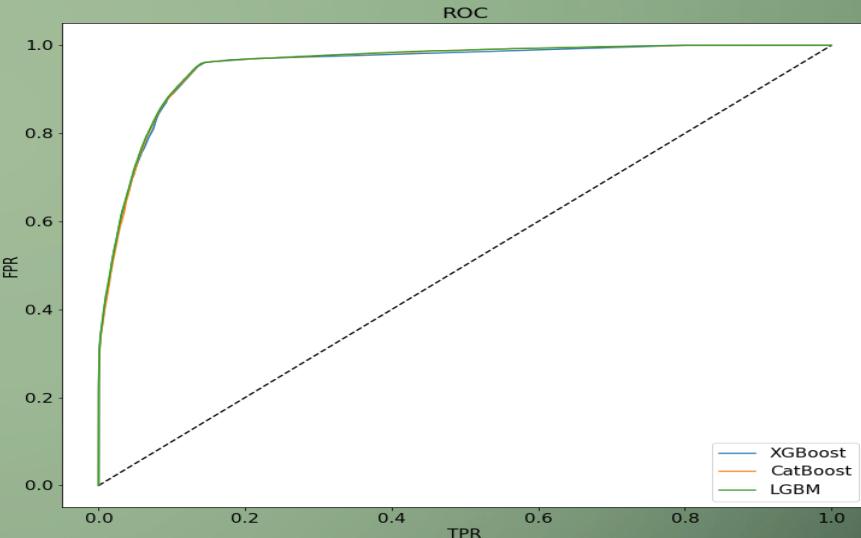
```
objective='binary',  
max_depth=13,  
n_estimators=100,  
num_leaves = 100,  
learning_rate=0.045,  
scale_pos_weight = 1.7935,  
reg_lambda = 0.2
```

СРАВНЕНИЕ МОДЕЛЕЙ

XGBOOST : AUC_PR = 0.706
AUC_ROC = 0.953

CATBOOST : AUC_PR = 0.703
AUC_ROC = 0.955

LGBM : AUC_PR = 0.711
AUC_ROC = 0.956

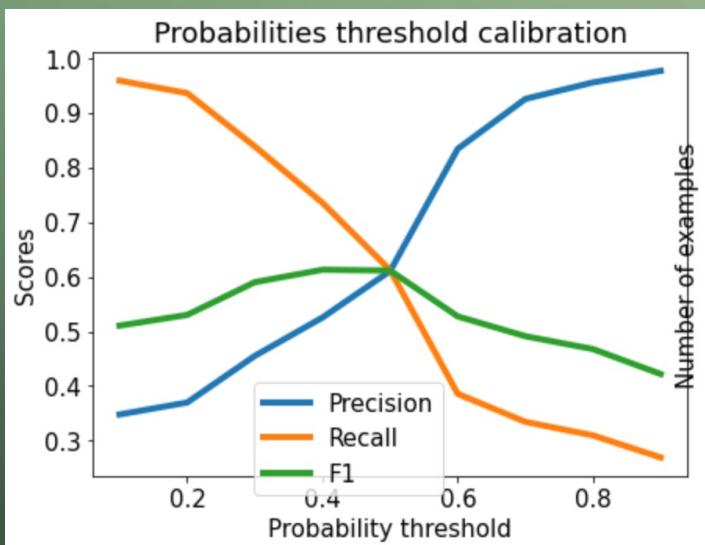


ВЫБОР МОДЕЛИ – LGBMCLASSIFIER

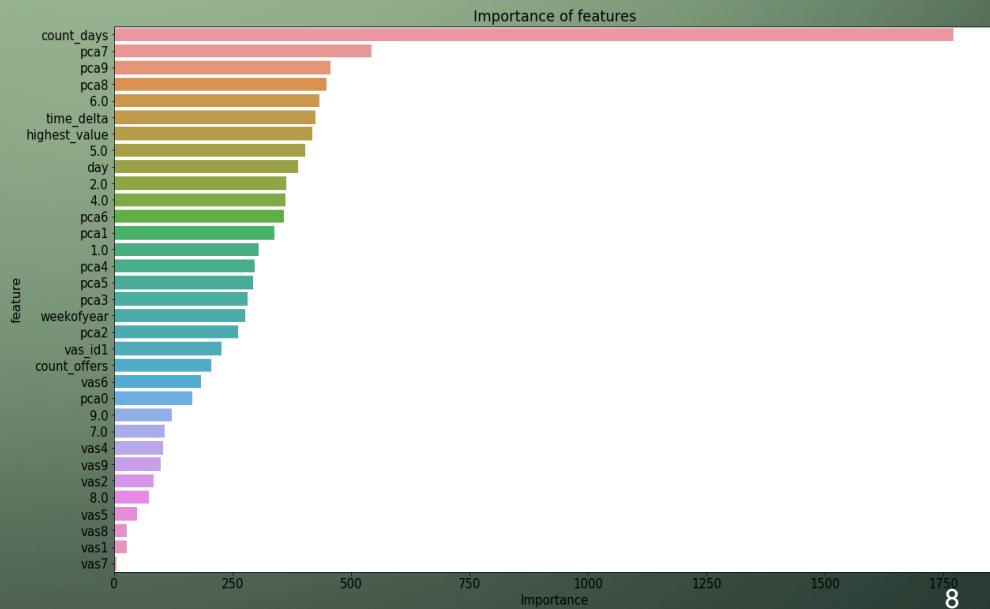
ПО СРАВНЕНИЮ С XGB И CV:

- ОПТИМАЛЬНЫЙ ПОРОГ – 0.49
- F1-МАКРО НА ВАЛИДАЦИОННОМ ДАТАСЕТЕ ПРИ ЭТОМ ПОРОГЕ – 0.79069

- ЛУЧШАЯ СКОРОСТЬ РАБОТЫ
- ЭКОНОМИЯ ВЫЧИСЛИТЕЛЬНЫХ РЕСУРСОВ
- МЕНЬШЕЕ ПОТРЕБЛЕНИЕ ПАМЯТИ



Алена Кухта | alenakukhta@yandex.ru



ИНДИВИДУАЛЬНЫЕ ПРЕДЛОЖЕНИЯ АБОНЕНТАМ

- Для оптимального соотношения precision-recall по позитивному классу, можно обращаться к абонентам со скор > 0.425
- Для бизнеса это может быть: снижение порога для захвата рынка и роста recall или повышение порога для экономии ресурсов и роста precision
- Клиенту, не заинтересовавшемуся услугой, нужно предложить другую
- Необходимо определить клиентов, на взаимодействие с которыми целесообразно вкладывать ресурсы (Uplift моделирование)



МУЛЬТИКЛАССОВАЯ КЛАССИФИКАЦИЯ С LGBMCLASSIFIER

- Позволяет подобрать услугу для выбранных пользователей

id	best_service	1	2	4	5	6	7	8	9
862975	8	0.104866	0.091815	0.066337	0.128631	0.007982	0.008014	0.494432	0.097922

- Позволяет определить абонентов для конкретной услуги

id	best_service	1	2	4	5	6	7	8	9
2521946	1	0.825278	0.014130	0.112268	0.020391	0.026789	0.000145	0.000573	0.000427
3686508	1	0.817292	0.051909	0.093723	0.017688	0.008805	0.003694	0.001798	0.005091
3242547	1	0.805830	0.053394	0.075873	0.041312	0.003206	0.009934	0.000945	0.009505

- Не добавляет уверенности в выборе абонента, в которого целесообразно вкладывать ресурсы

БЛАГОДАРЮ ЗА ВНИМАНИЕ!

РЕШЕНИЕ – [HTTPS://GITHUB.COM/ALENAKUKHTA/MEGAFON-COURSE-PROJECT](https://github.com/ALENAKUKHTA/MEGAFON-COURSE-PROJECT)

Алена Кухта | alenakukhta@yandex.ru