Описание задачи

Задача: *определение вероятности подключения услуги для каждой пары пользователь-услуга*

Исходные данные

data_train/data_test	
id	идентификатор абонента
vas_id	подключаемая услуга
buy_time	время покупки
target	целевая переменная (1/0)

train – данные за 4 месяца
test – данные за последующий месяц

features	
id	идентификатор абонента
features_list	признаки

Последовательность решения задачи

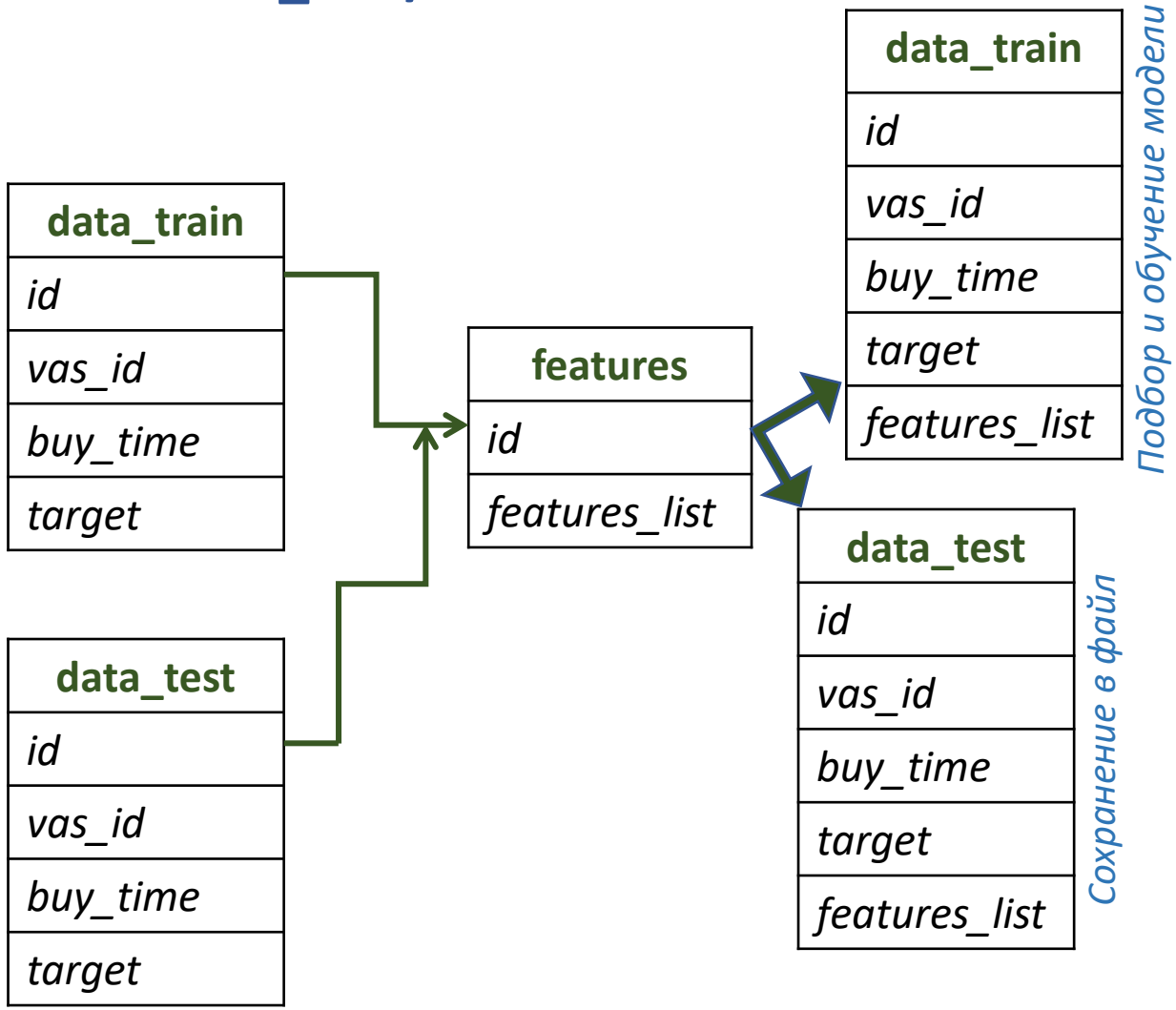
- *Загрузка и объединение данных*
- *Обзор признаков*
- *Предобработка признаков*
- *Подбор оптимальной модели*
- *Обучение оптимальной модели*
- *Расчёт метрик и анализ полученных результатов*
- *Сохранение модели*
- *Запуск и тестирование модели*

Результаты

- ✓ *Модель, сохраненная в файл*
- ✓ *Скрипт для запуска модели и сохранения результатов её работы*

Объединение данных. Работа с признаками

Совмещение датасетов data_test/train и features



Обзор и предобработка признаков

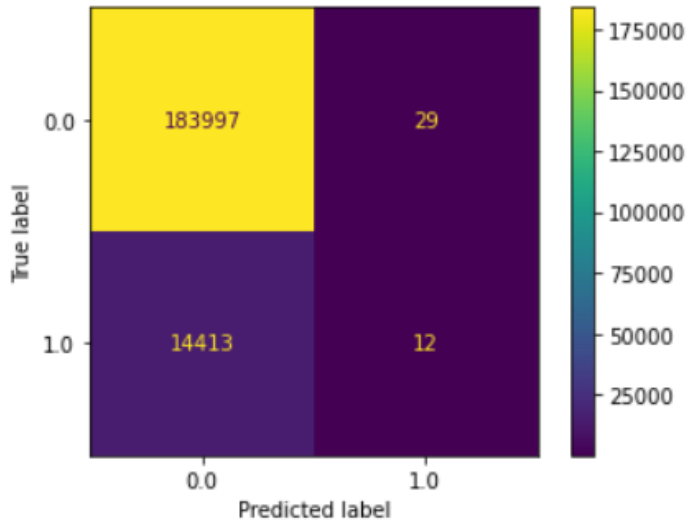
Вид признака	Количество	Вид предобработки
Константные	5	Не принимают участие в обучении модели
Бинарные	0	Не требуется
Категориальные	1	**OneHotEncoder**
Вещественные	249	**StandardScaler**
Итого: 255 признаков		

Подбор оптимальной модели

Логистическая регрессия

LogisticRegression

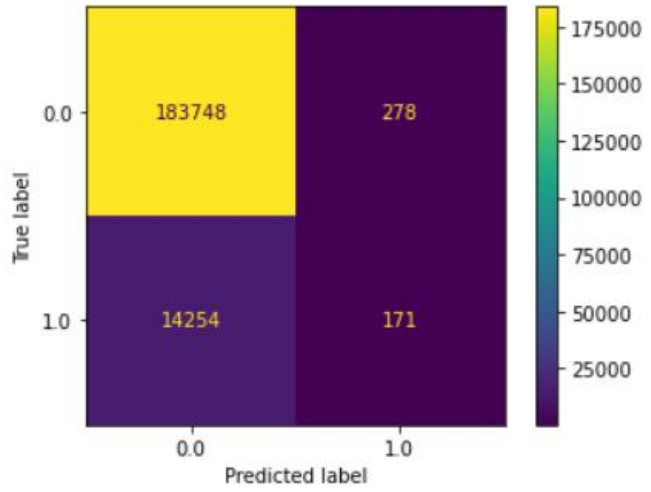
	precision	recall	f1-score	support
0.0	0.93	1.00	0.96	184026
1.0	0.29	0.00	0.00	14425
accuracy			0.93	198451
macro avg	0.61	0.50	0.48	198451
weighted avg	0.88	0.93	0.89	198451



SGD классификатор

SGDClassifier

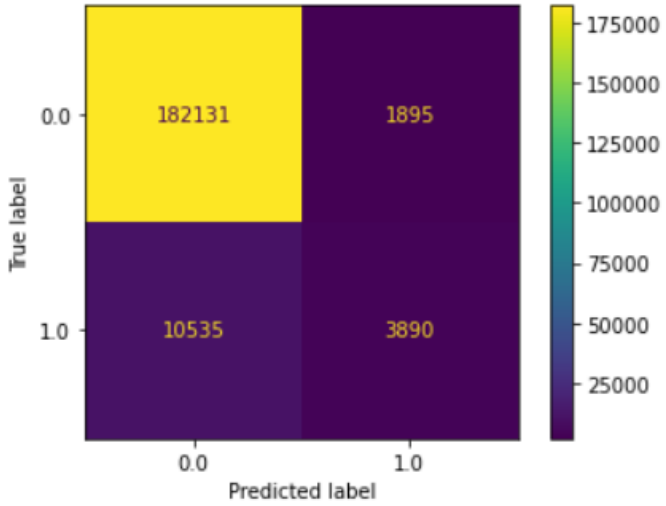
	precision	recall	f1-score	support
0.0	0.93	1.00	0.96	184026
1.0	0.38	0.01	0.02	14425
accuracy			0.93	198451
macro avg	0.65	0.51	0.49	198451
weighted avg	0.89	0.93	0.89	198451



Градиентный бустинг

GradientBoostingClassifier

	precision	recall	f1-score	support
0.0	0.95	0.99	0.97	184026
1.0	0.67	0.27	0.38	14425
accuracy			0.94	198451
macro avg	0.81	0.63	0.68	198451
weighted avg	0.93	0.94	0.92	198451



Вывод: Наибольшая метрика **f1-score** (*macro невзвешенная*) наблюдается при использовании модели градиентного бустинга. Поэтому в качестве оптимальной модели используем **GradientBoostingClassifier**

Обучение оптимальной модели

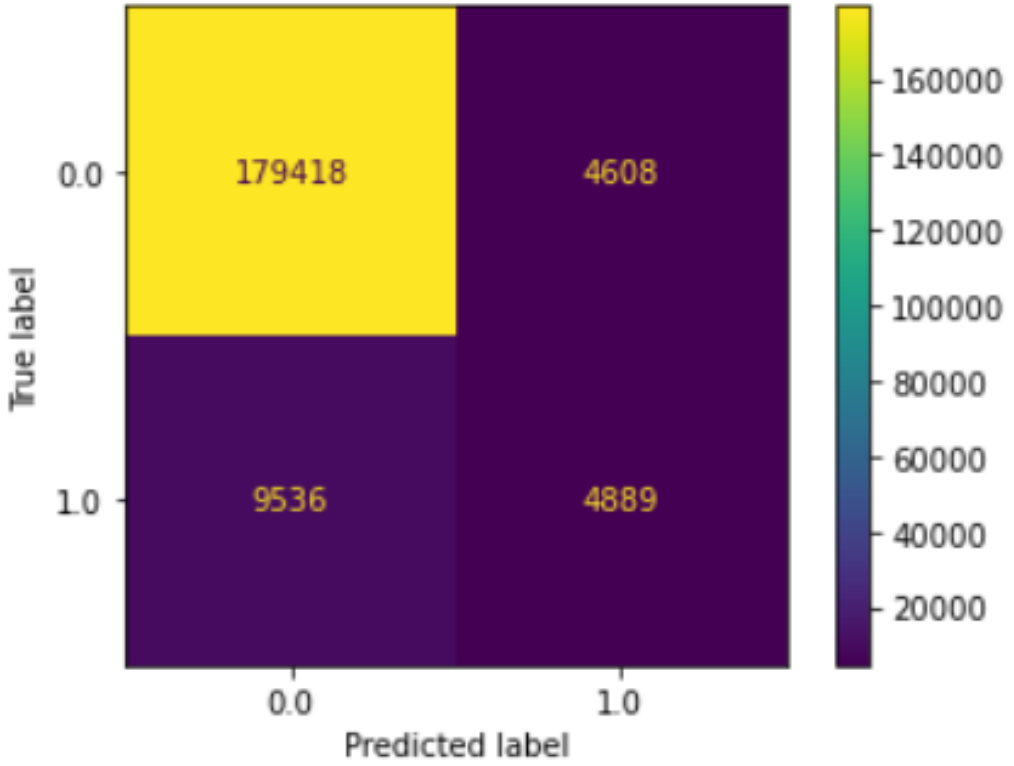
Параметры модели

GradientBoostingClassifier	
learning_rate	0,3
n_estimators	300
min_samples_split	2
max_depth	8

Расчёт метрик

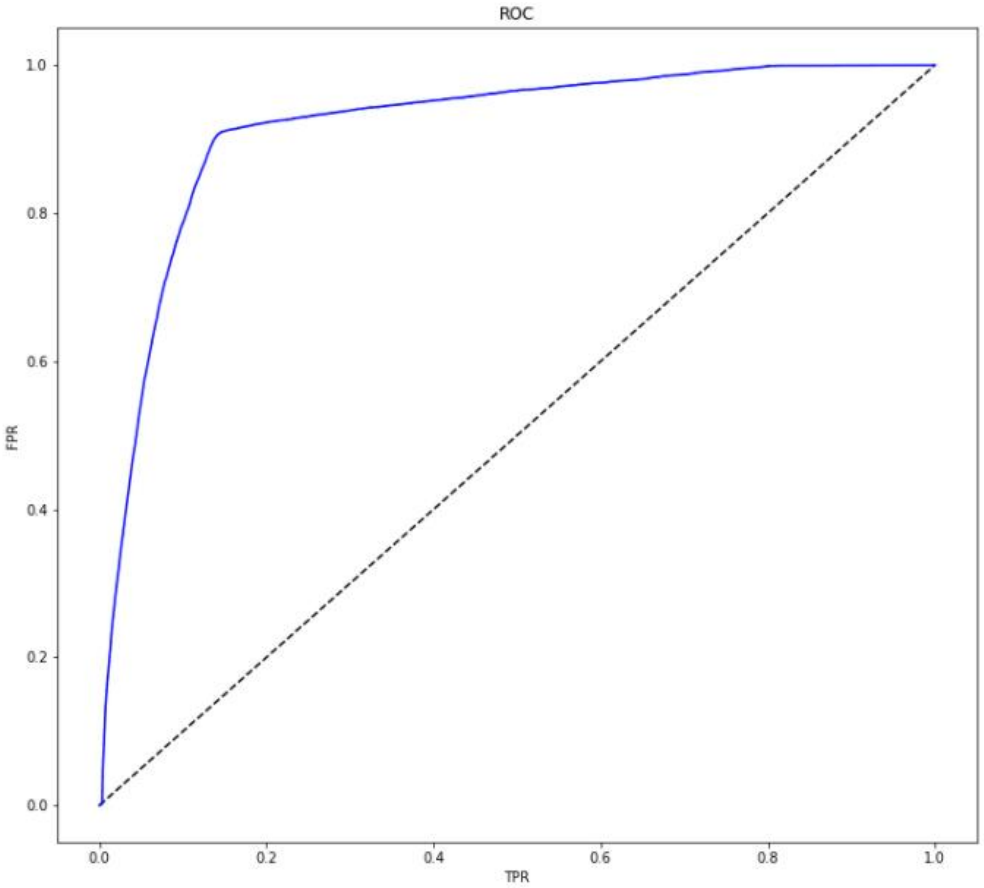
	precision	recall	f1-score	support
0.0	0.95	0.97	0.96	184026
1.0	0.51	0.34	0.41	14425
accuracy			0.93	198451
macro avg	0.73	0.66	0.69	198451
weighted avg	0.92	0.93	0.92	198451

Матрица ошибок



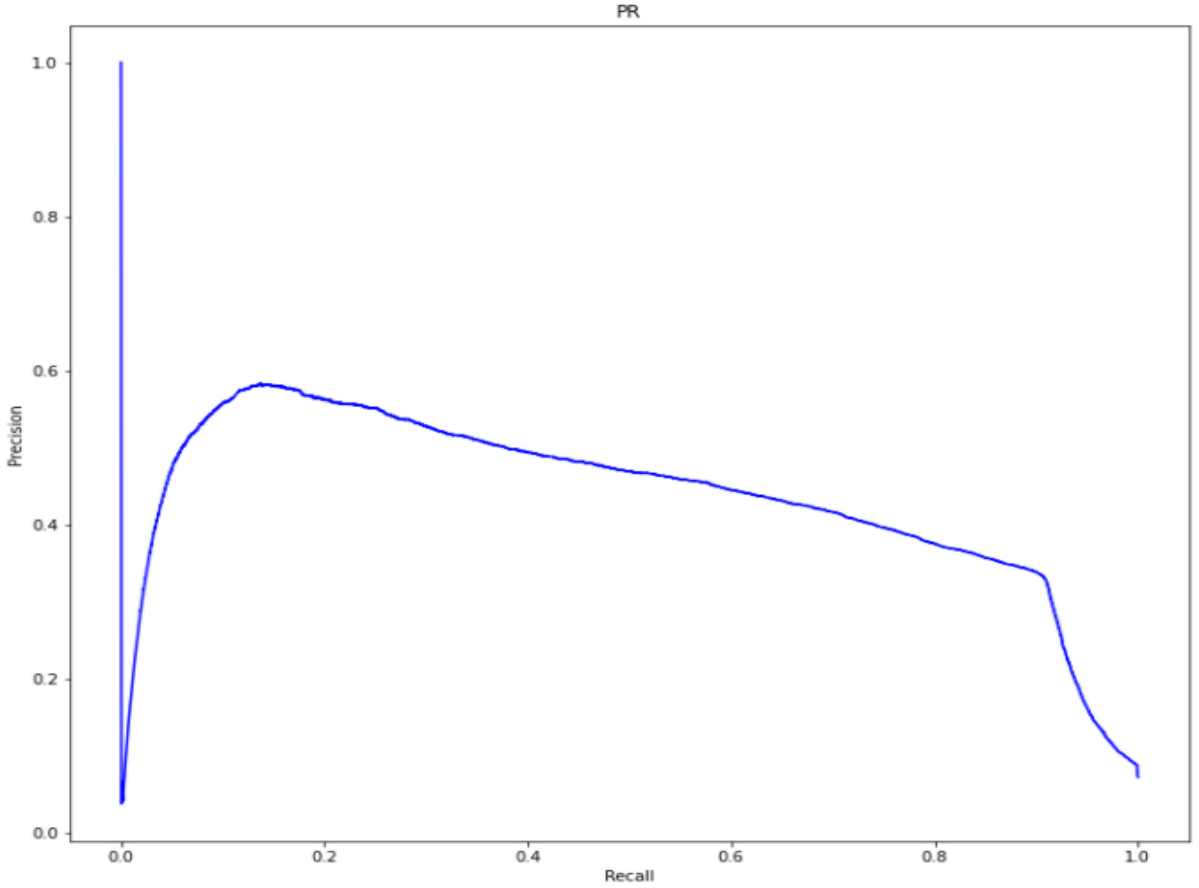
Результаты

ROC-кривая



AUC_ROC=0,917

PR-кривая



AUC_PR=0,436

Модель сохранена в формате dill (model.dill). Запуск осуществляется с помощью скрипта run_model.py, который нужно запускать через виртуальное окружение. Результатом работы модели является файл с вероятностями для каждого клиента (answers_test.csv)