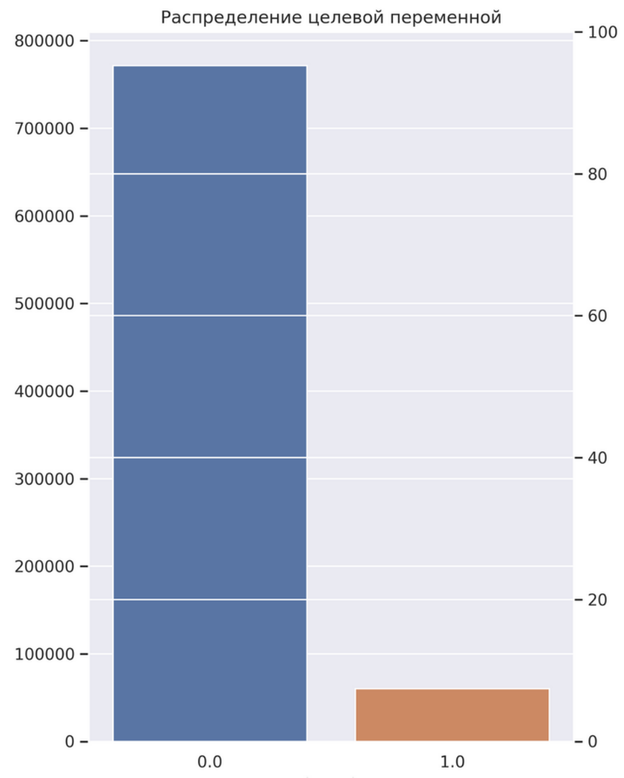
Модель определения вероятности подключения услуги

Дмитрий Рубцов

Данные

data_train.csv размеченная 4-х месячная выборка с данными об отклике абонентов на предложение подключения одной из услуг
data_test.csv тестовый набор набор данных
features.csv.zip нормализованный анонимизированный набор признаков, характеризующий профиль потребления абонента



Задача

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

Метрика

Скоринг будет осуществляться функцией f1, невзвешенным образом.

Этап 1. Baseline models

Этап 2. Balancing classes

Этап 3. Randomized Search

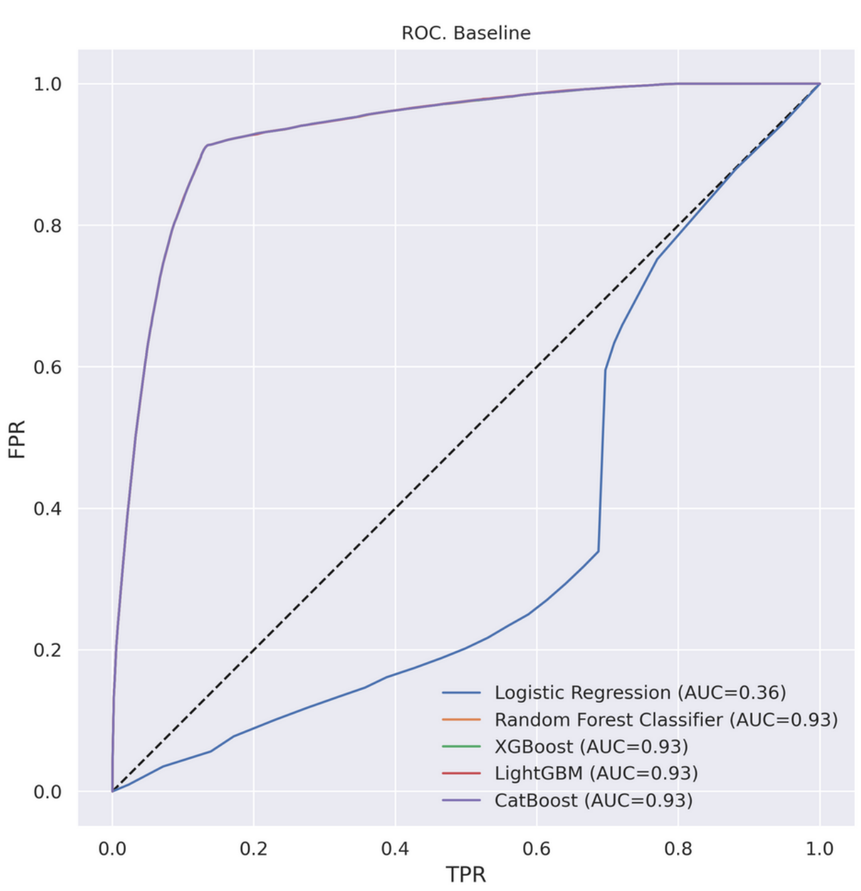
Этап 4. Probability calibration

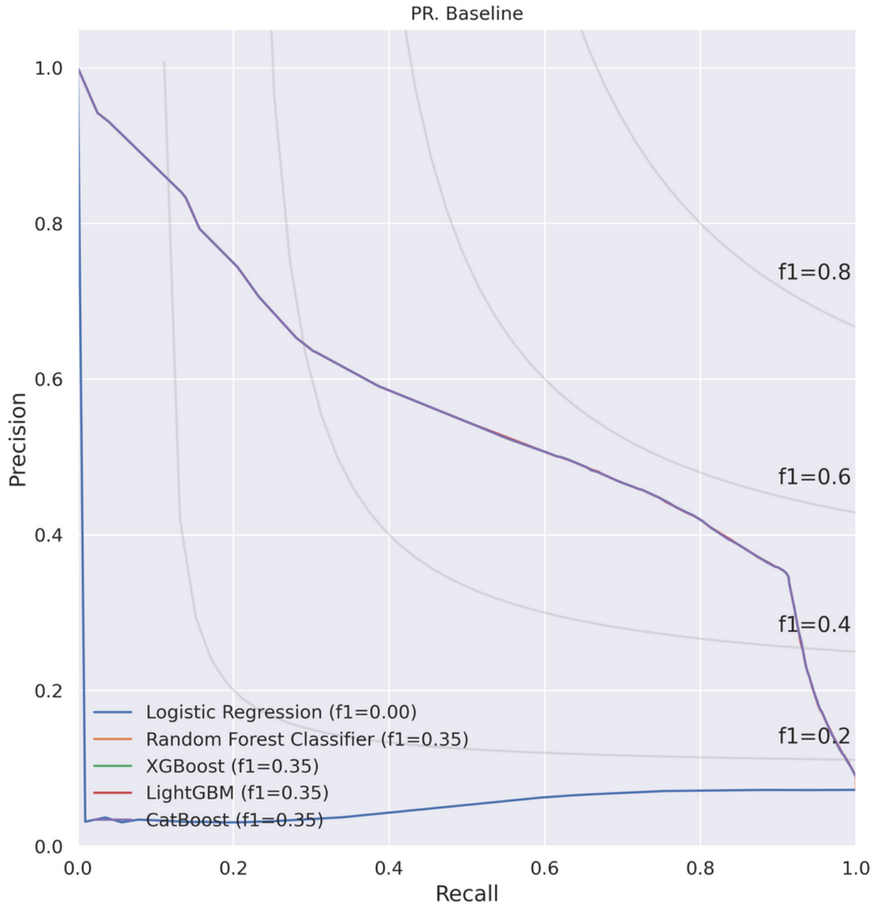
Этап 1. Baseline models

Построение базовых (коробочных) моделей Logistic Regression, Random Forest Classifier, XGBoost, LightGBM и CatBoost на исходном наборе данных с признаками vas_id, buy_time

В исходных данных целевая переменная обладает дисбалансаом классов (92.76% / 7.23%).

Дисбаланс классов в меньшей степени влияет на точность алгоритмов, основанных на деревьях решений (и их обобщениях – случайном лесе и градиентном бустинге), так как в этих алгоритмах дисбаланс целевой переменной влияет на меры неоднородности листьев, которые пропорциональны для всех классов.



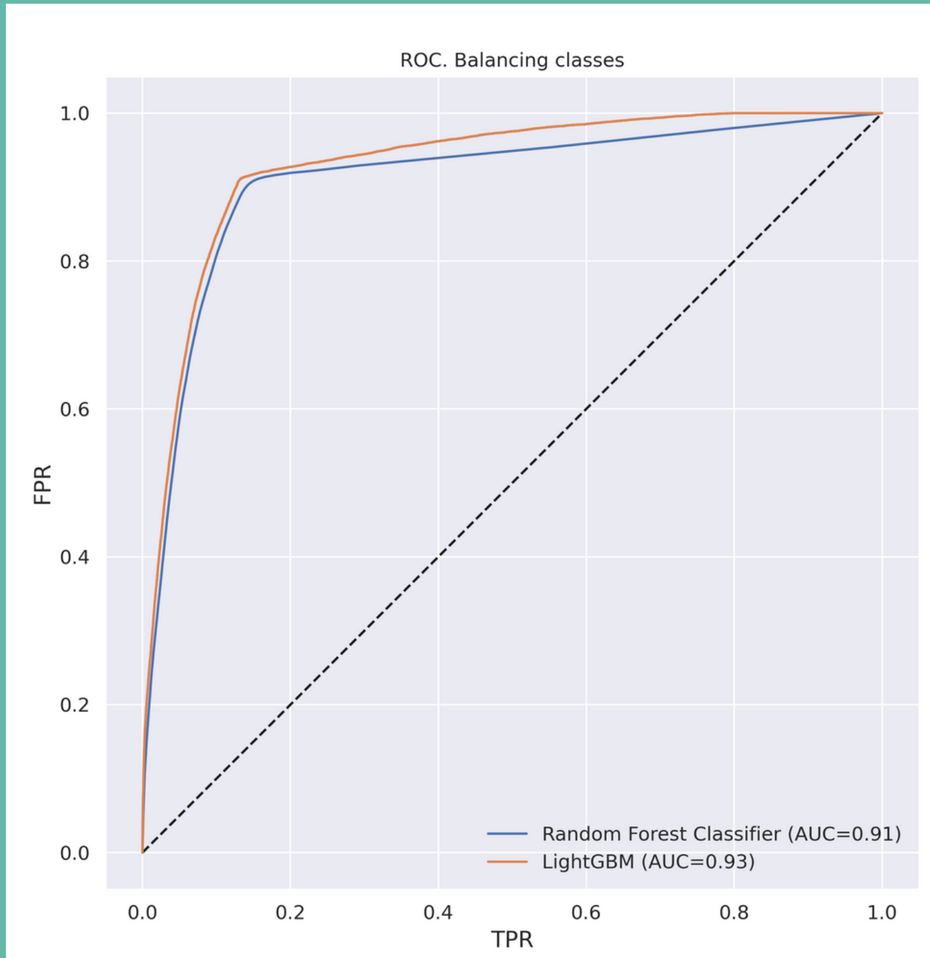


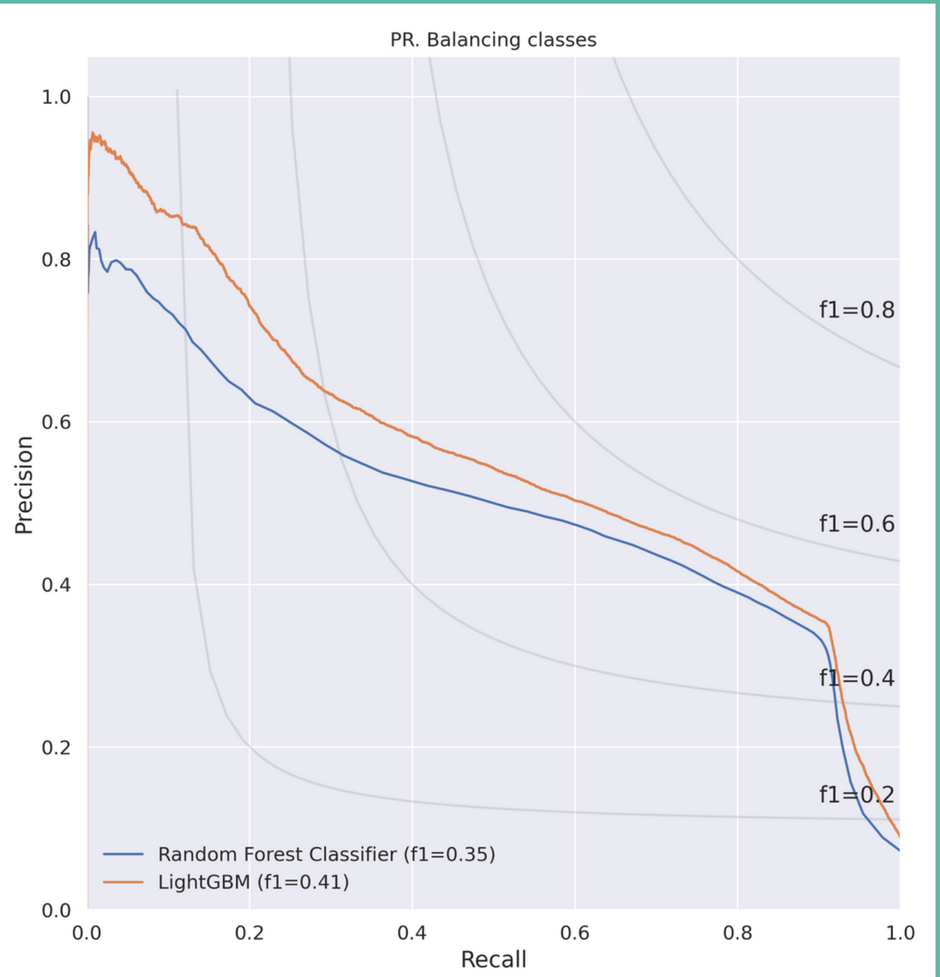
Этап 2. Balancing classes

Построение моделей Random Forest Classifier и LightGBM на наборе данных с дополнительными признаками из features.csv и выравнивание баланса классов методом over sampling

Оптимизация файла с дополнительными признаками с помощью использования фреймворка Dask и подбора оптимальных типов данных для каждого признака привела к уменьшению размера набора данных с ~22.5Gb до ~2.8Gb.

Использование дополнительных признаков и выравнивание баланса классов методом over sampling привело к улучшению качества метрики с 0.35 до 0.41 для LightGBM.

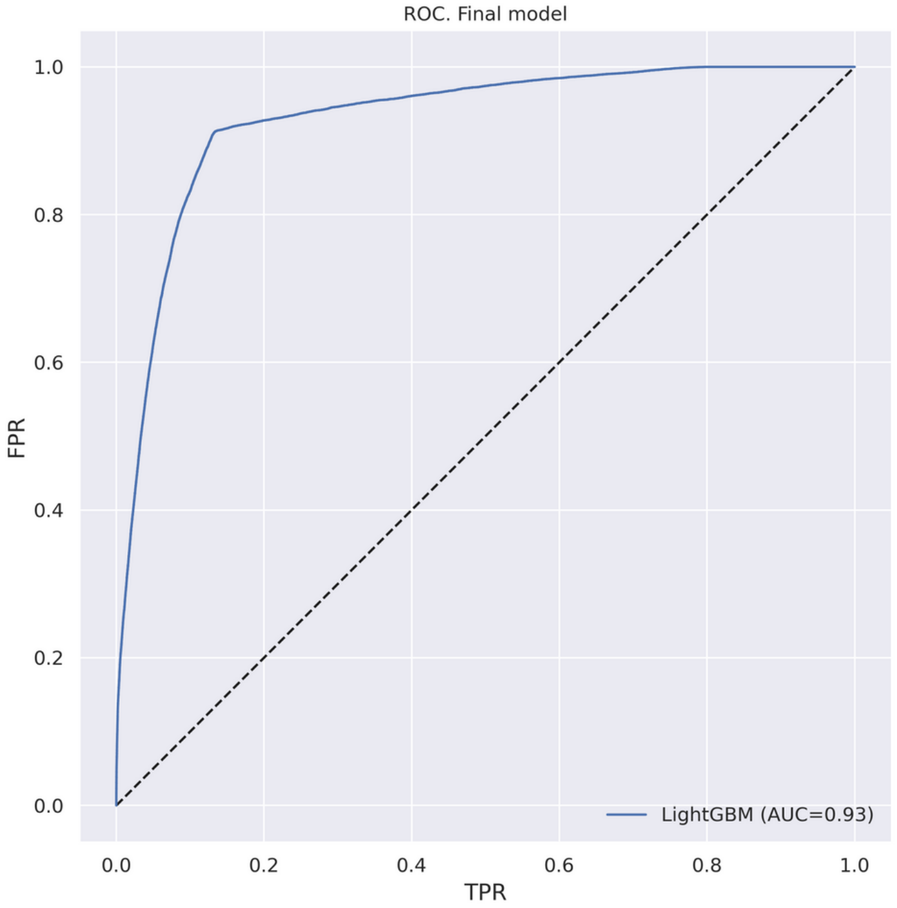


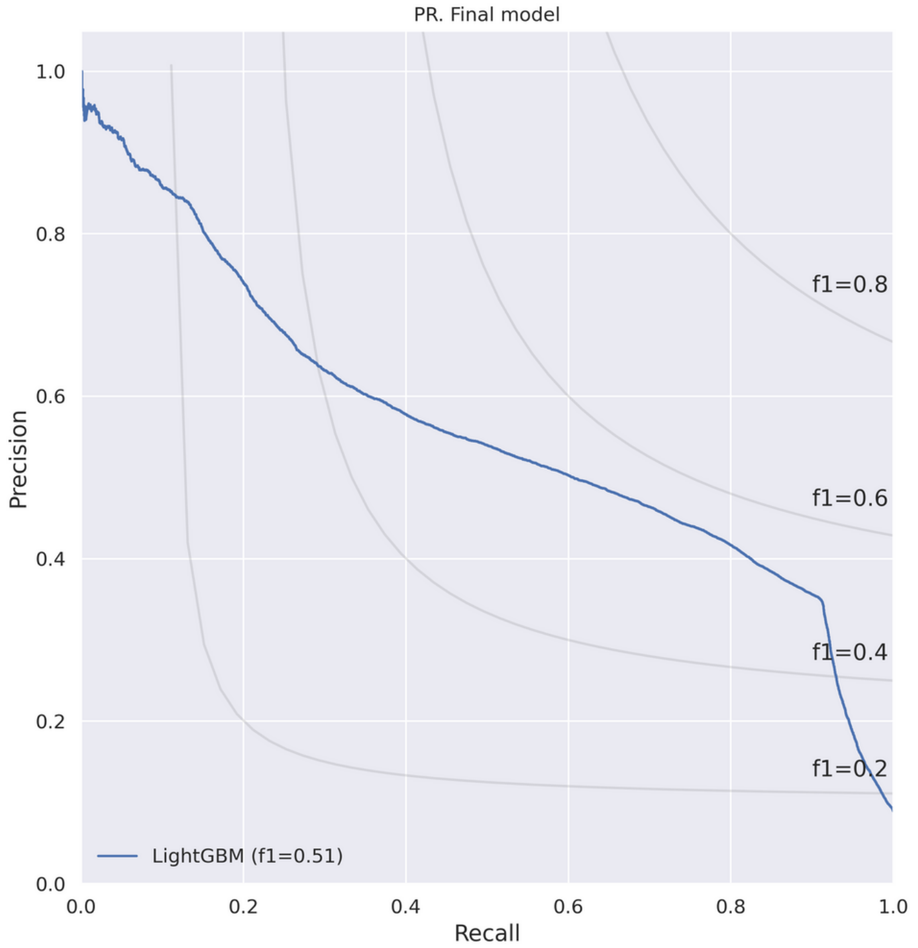


Этап 3. Randomized Search

Поиск оптимальных параметров модели определения вероятности подключения услуги путем случайного перебора параметров из заданного диапазона

Найденные параметры модели LightGBM позволили улучшить качество метрики до 0.51.

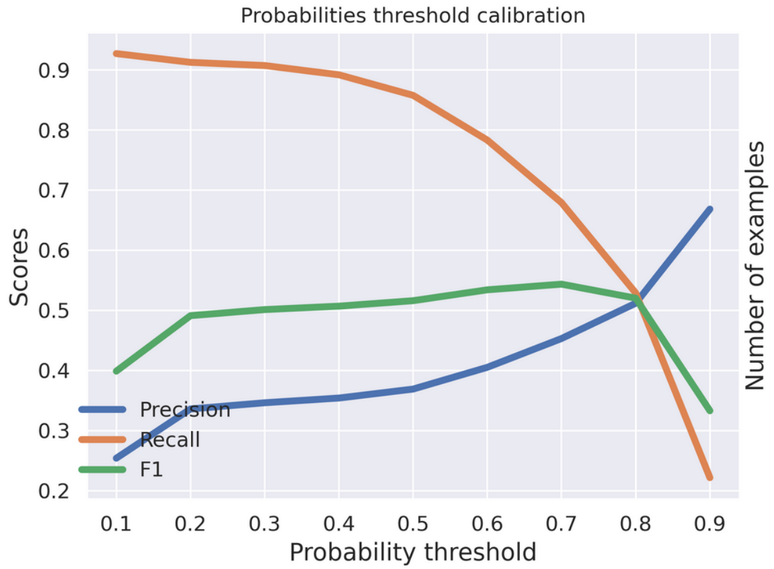


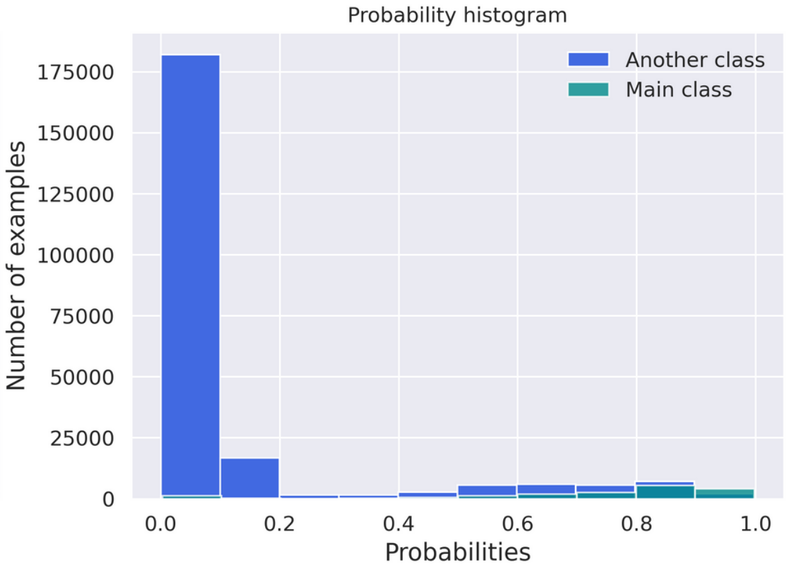


Этап 4. Probability calibration

Построение моделей Random Forest Classifier и LightGBM на наборе данных с дополнительными признаками из features.csv и выравнивание баланса класса методом over sampling

f1	precision	recall	probability
0.557	0.453	0.723	0.7
0.544	0.409	0.812	0.6
0.538	0.517	0.562	0.8
0.511	0.357	0.899	0.5
0.506	0.351	0.909	0.4
0.503	0.347	0.912	0.3
0.501	0.345	0.913	0.2
0.433	0.283	0.923	0.1
0.309	0.749	0.195	0.9





В результате полученная модель имеет метрику f1 =0.557 при пороге вероятности 0.7.