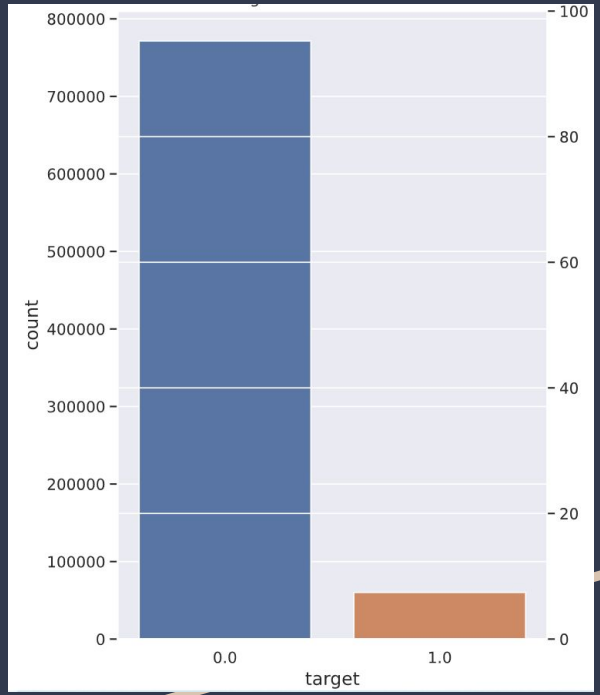
МОДЕЛЬ ОПРЕДЕЛЕНИЯ ВЕРОЯТНОСТИ ПОДКЛЮЧЕНИЯ УСЛУГ

Богданова Виктория

ДАННЫЕ МЕГАФОН

ДАННЫЕ



data_train.csv - размеченная выборка с данными об отклике абонентов на предложение подключения одной из услуг за 4 месяца.

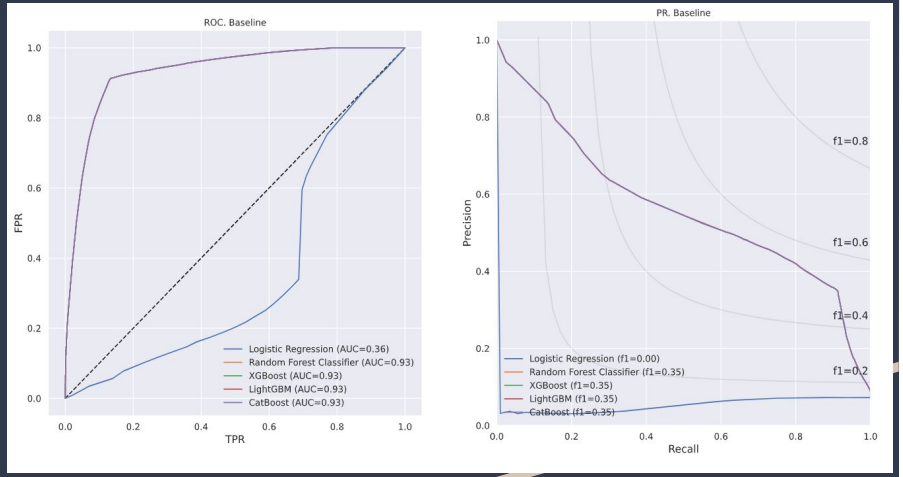
data_test.csv - тестовый набор данных за последующий месяц

features.csv - анонимизированный набор признаков, характеризующий абонента

ЗАДАЧА

Определить вероятность подключения услуги (скоринг осуществляется **функцией f1**, невзвешенным образом)

ЭТАП 1.
BASELINE MODELS

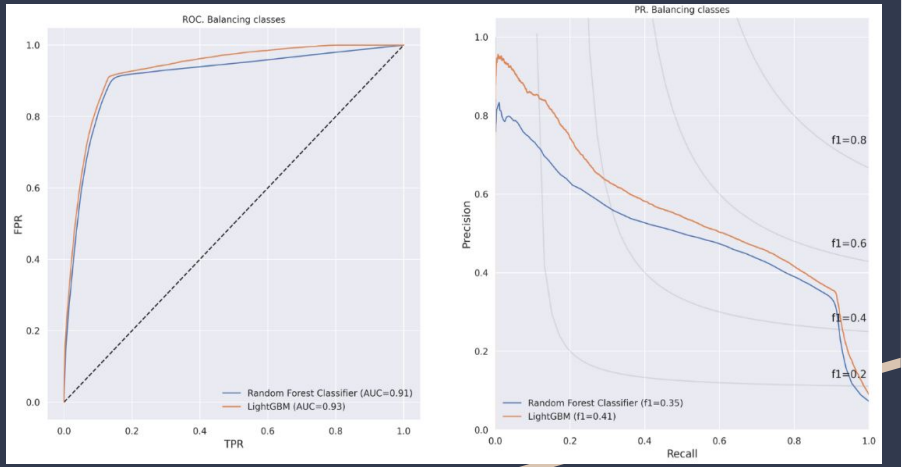


Построение базовых моделей **(без подбора параметров)** LogisticRegression, RandomForestClassifier, XGBClassifier, LGBMClassifier, CatBoostClassifier на исходном наборе данных.

В исходных данных целевая переменная обладает **дисбалансом классов** (92,76 / 7,24 %).

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

ЭТАП 2.
BALANCING CLASSES

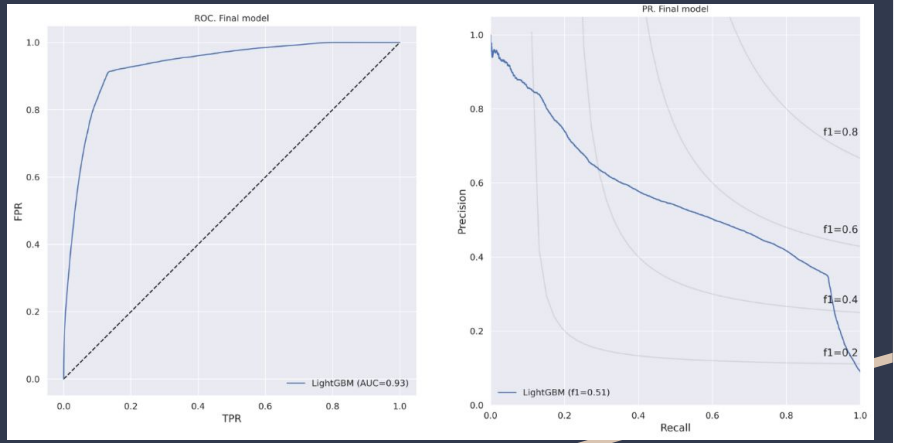


Построение моделей RandomForestClassifier и LGBMClassifier на наборе данных с дополнительными признаками из features.csv и **выравнивание балансов** класса методом **over sampling**.

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 ≈22,5 Gb до ≈2,7 Gb.

Использование дополнительных признаков и выравнивание баланса классов привело к **улучшению качества метрики** с 0,35 до 0,41 для LGBMClassifier.

ЭТАП 3.
RANDOMIZED SEARCH



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

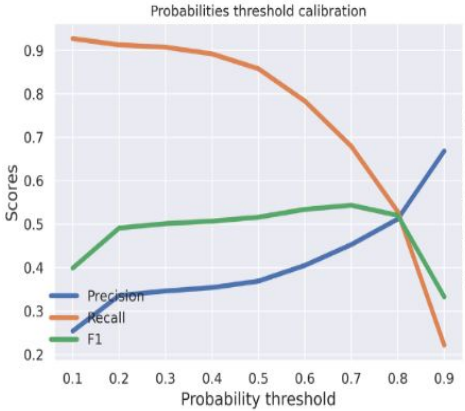
Найденные параметры модели LGBMClassifier позволило **улучшить качество метрики** до 0,51

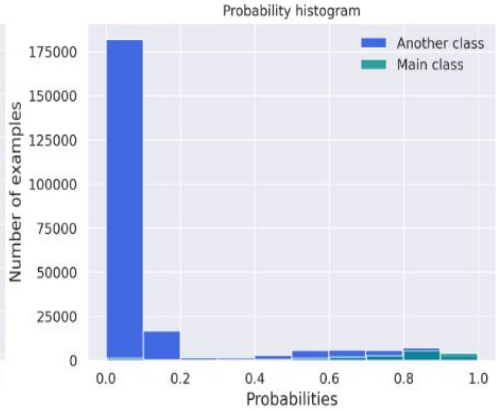
ЭТАП 4.
PROBABILITY CALIBRATION

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

ИТОГ

Построение моделей RandomForestClassifier и LGBMClassifier на наборе данных с дополнительными признаками из features.csv и **выравнивание баланса** класса.





В результате проделанной работы полученная модель имеет метрику **f1 = 0,557** при пороге вероятности **0,7**.