Решение задачи кредитного скоринга с помощью градиентного бустинга на данных карточных транзакций клиента

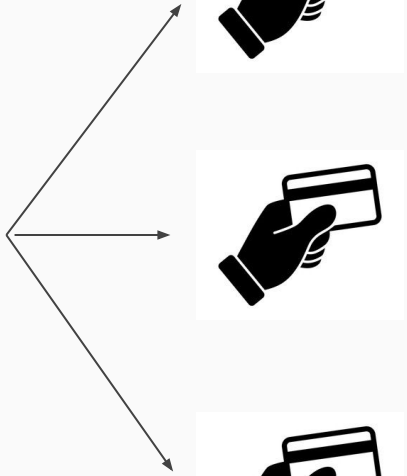
дипломная работа Соколова Александра

техническая часть

Содержание

1. Постановка задачи
2. Цели
3. ML
4. DL
5. Результаты
6. Что не успел реализовать

Постановка задачи



Продукты
Супермаркет
3000 р
21.04.2021
12:00 *(офлайн)*

Стрижка
Барбершоп
1500 р
21.04.2021
16:00 *(офлайн)*

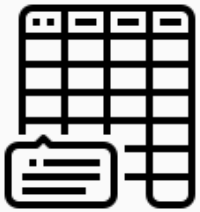
Билеты
Агенство
20000 р
21.04.2021
19:00 *(онлайн)*

Цели

1. войти в число медалистов соревнования - AUC ROC на привате не ниже 0.7616840
2. реализовать прототип оптимального по размеру и качеству ML-решения в виде web приложения на heroku
3. использовать для оценки моделей метрики f1, AUC PRC, а также матрицы ошибок и другие метрики пытаясь максимально приблизить задачу к реальной, так как метрика AUC ROC не совсем состоятельна в задачах по кредитному скорингу

Описание датасета

1 500 000 объектов



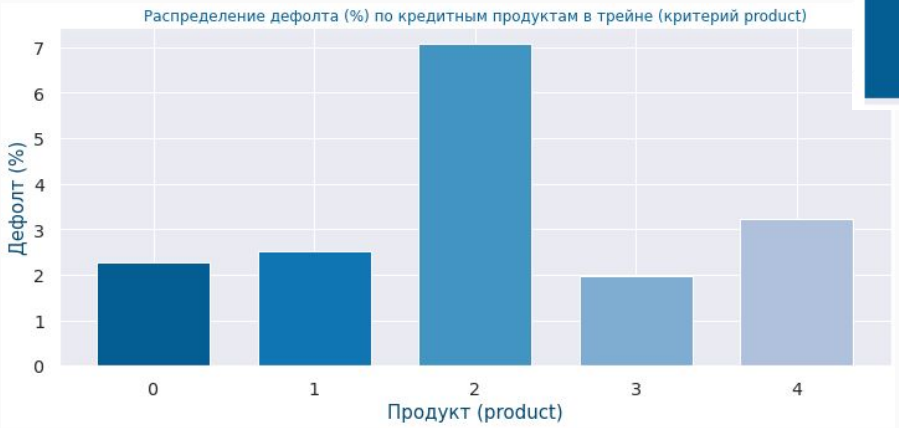
450 000 000 строк

6 ГБ в формате parquet

тестовая выборка смещена по времени

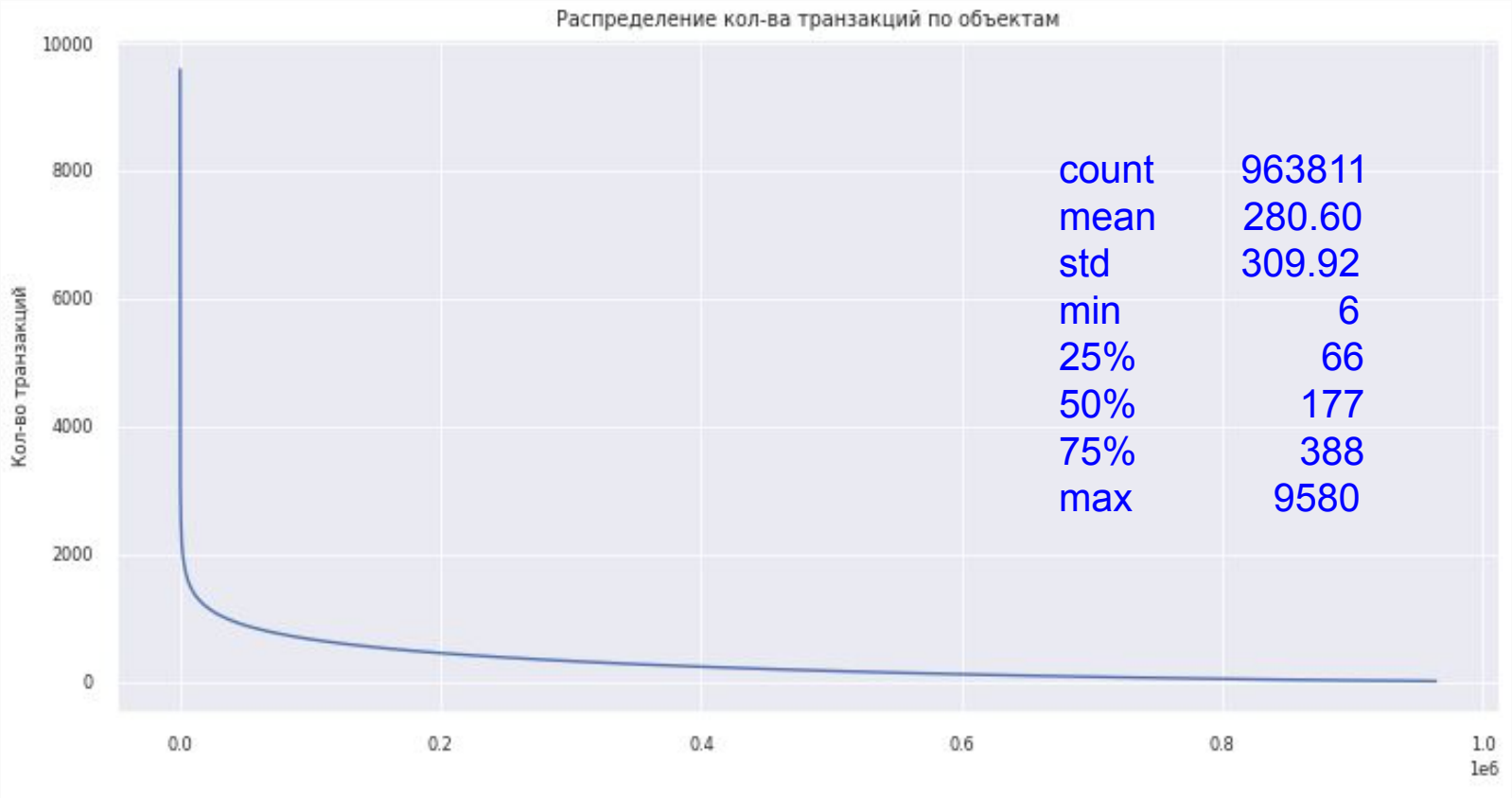
EDA по признаку ‘product’ (кредитный продукт)

Нулевой и первый кредитные продукты занимают около 80% всех кредитов



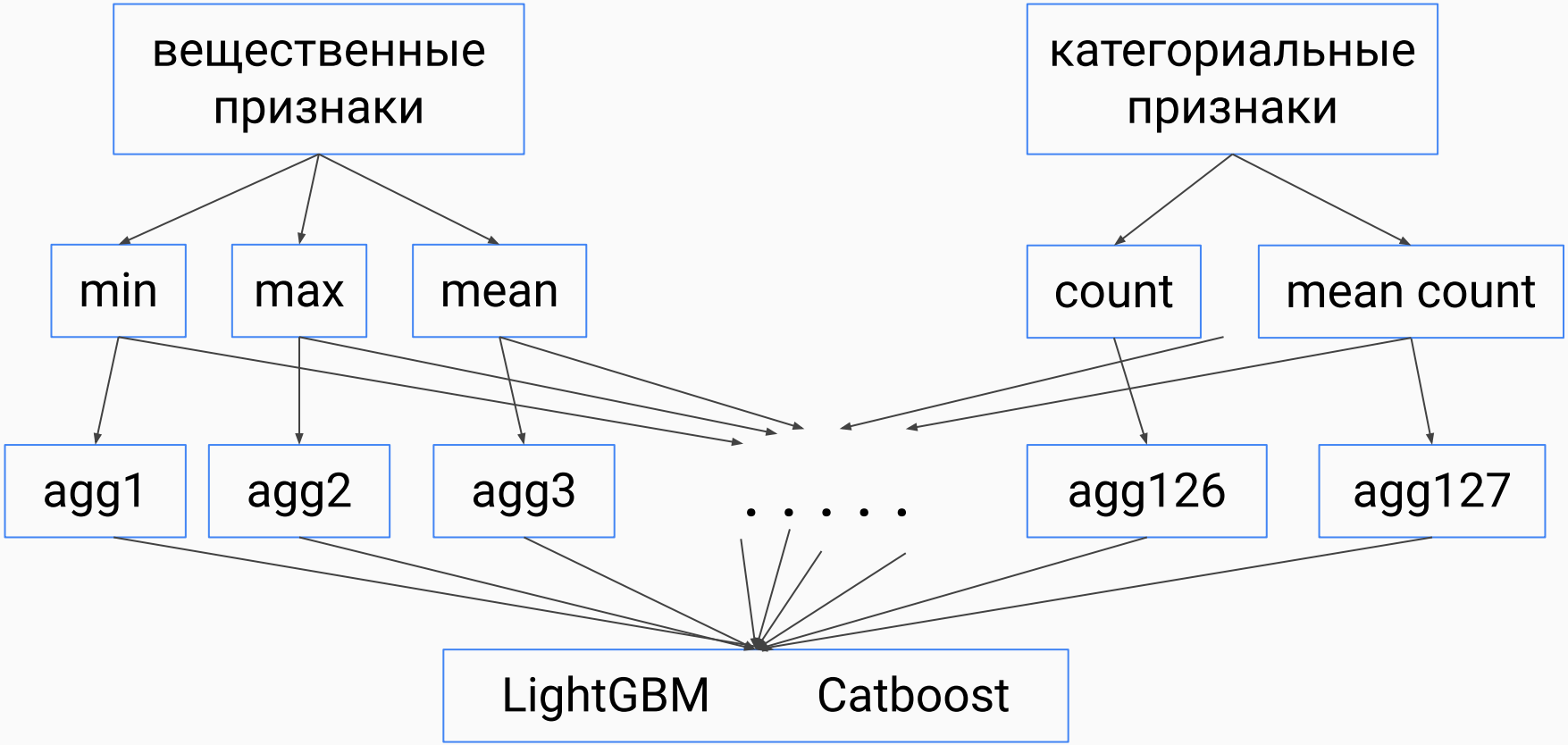
Процент дефолта по продукту 2 (7.1%) почти в три раза выше среднего значения по продуктам (2.7%)

EDA по количеству транзакций на один объект

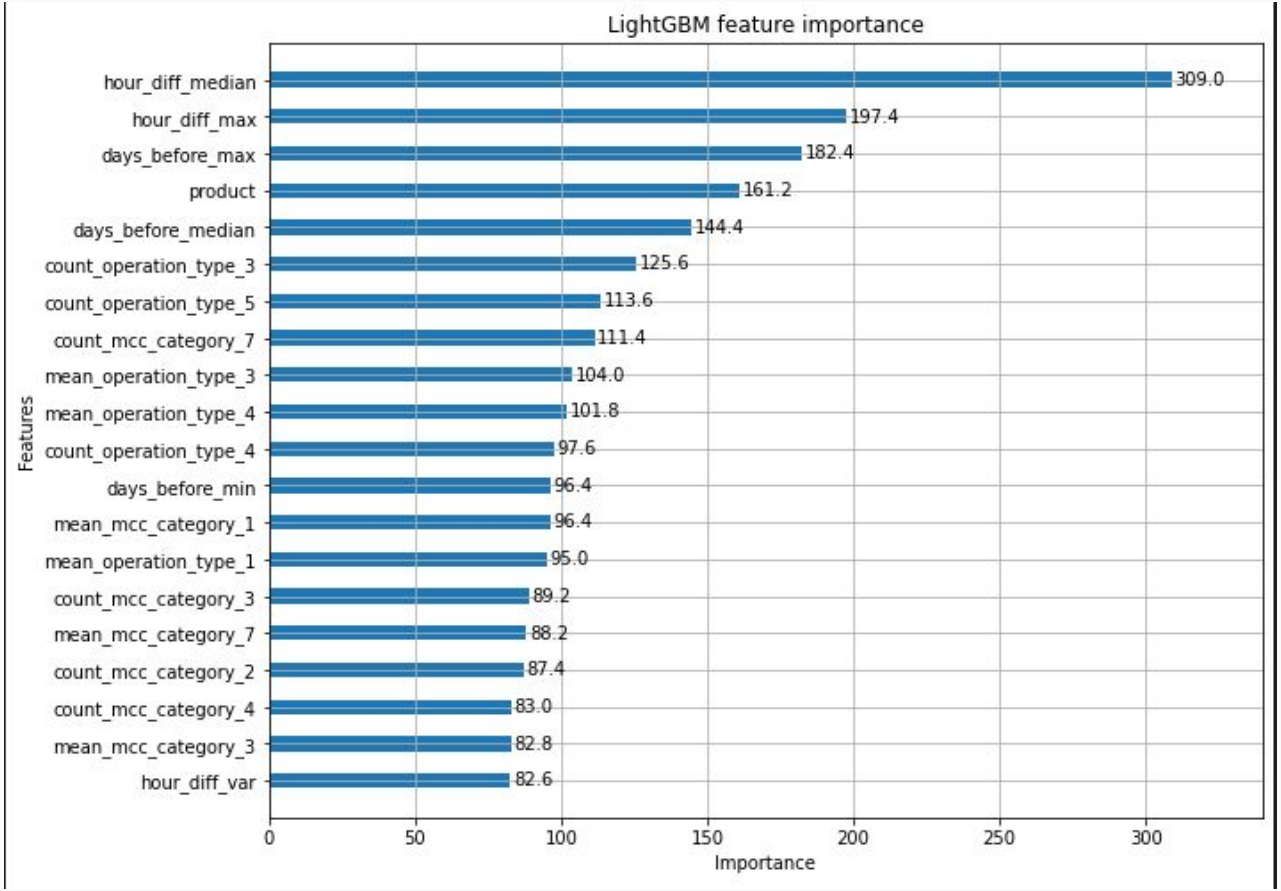


count	963811
mean	280.60
std	309.92
min	6
25%	66
50%	177
75%	388
max	9580

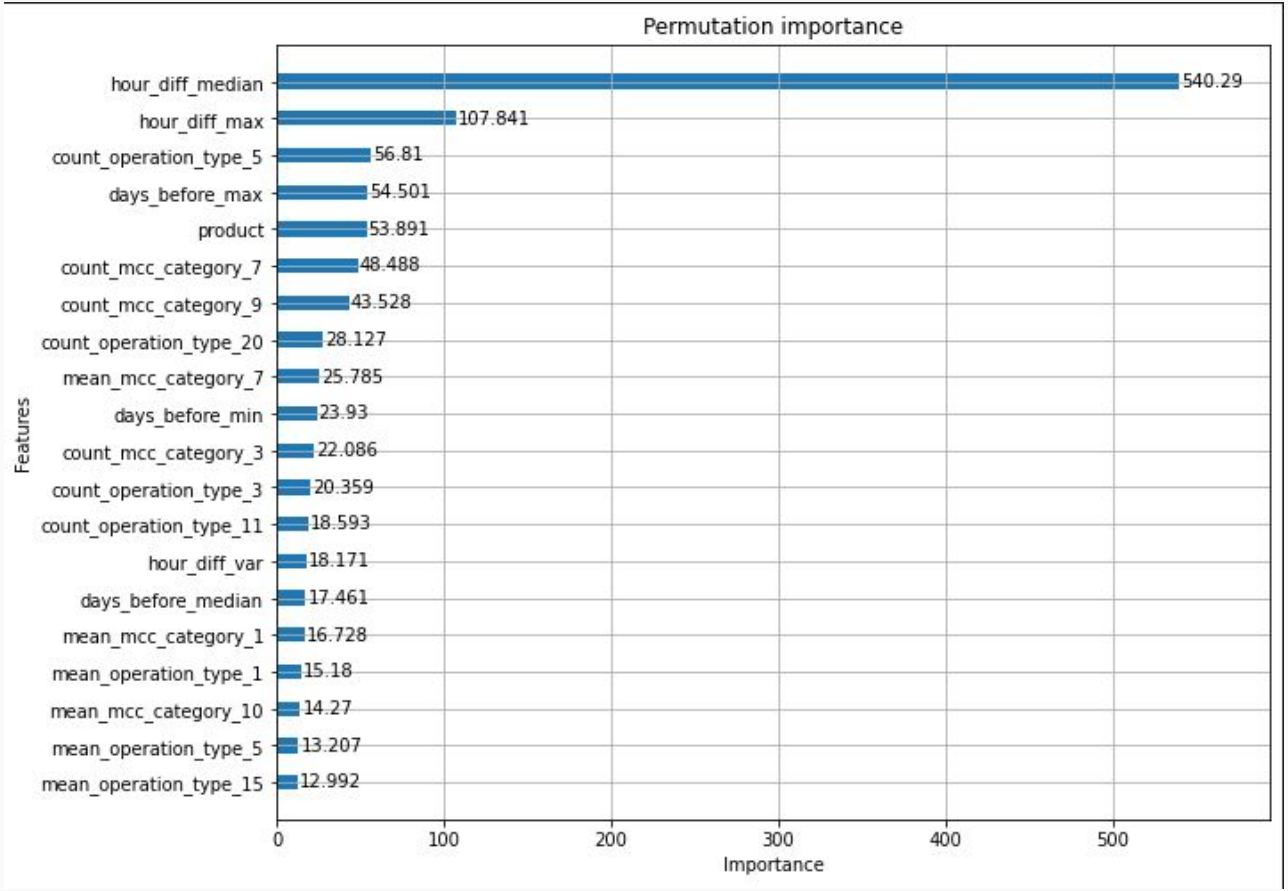
Препроцессинг для ML



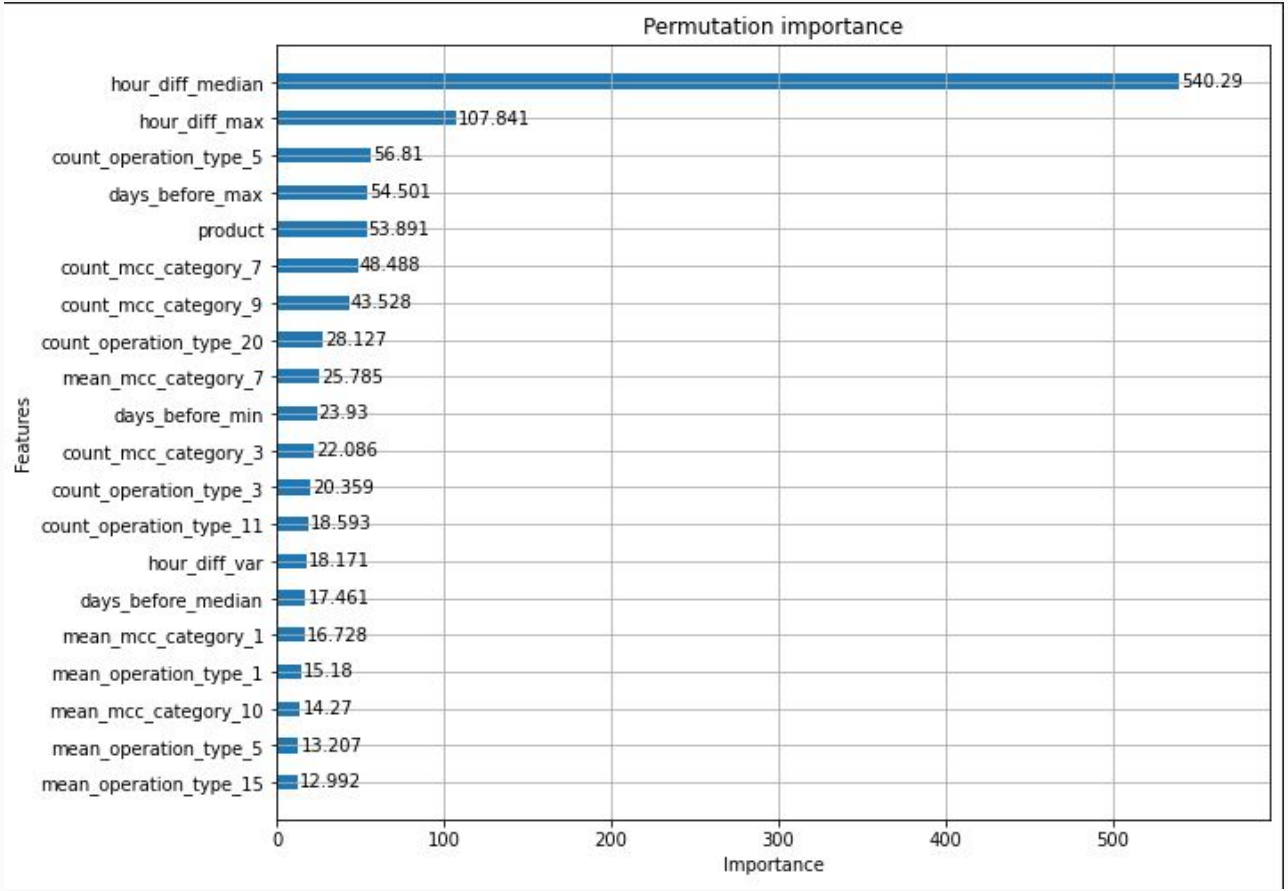
Отбор признаков (LightGBM feature importance)



Отбор признаков (Feature permutation importance)



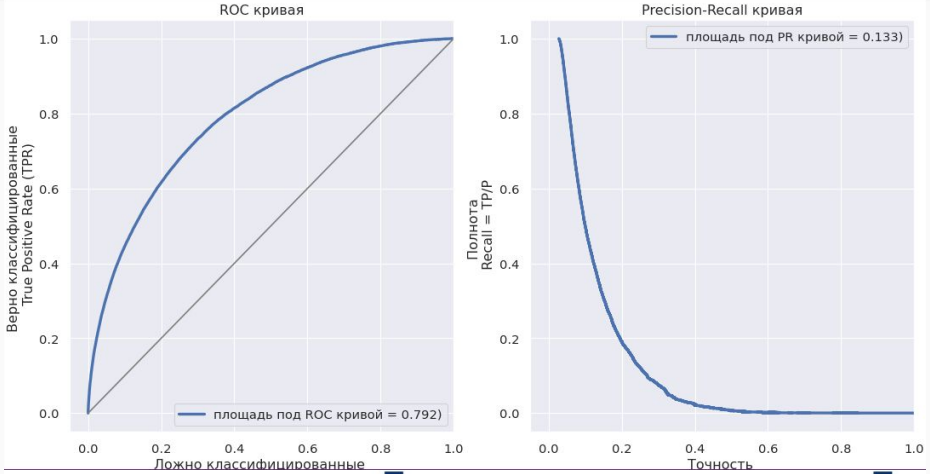
Отбор признаков (Target permutation importance)

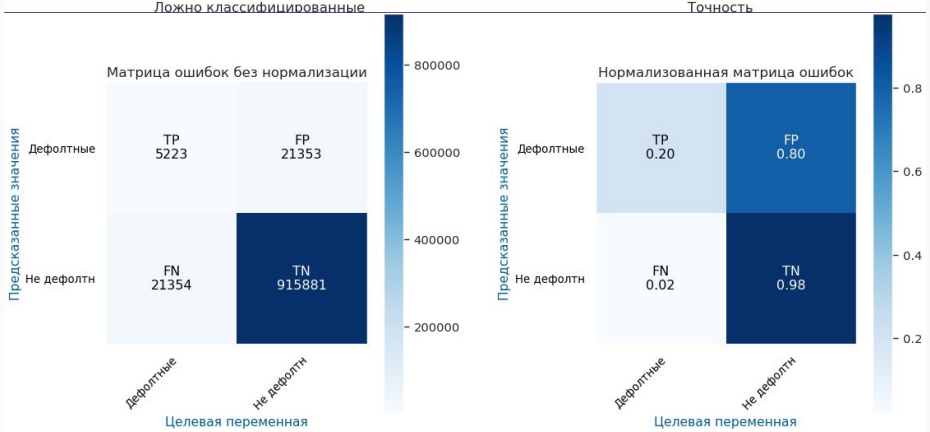


Промежуточные результаты по ML

	Метод отбора признаков	Кол-во признаков	Train	Val	Public Test
LightGBM	-	127	0.8	0.77	0.737
Catboost	-	127	0.77	0.764	0.732
LightGBM	feature permutation	59	0.796	0.769	0.735
LightGBM	target permutation	64	0.792	0.765	0.733

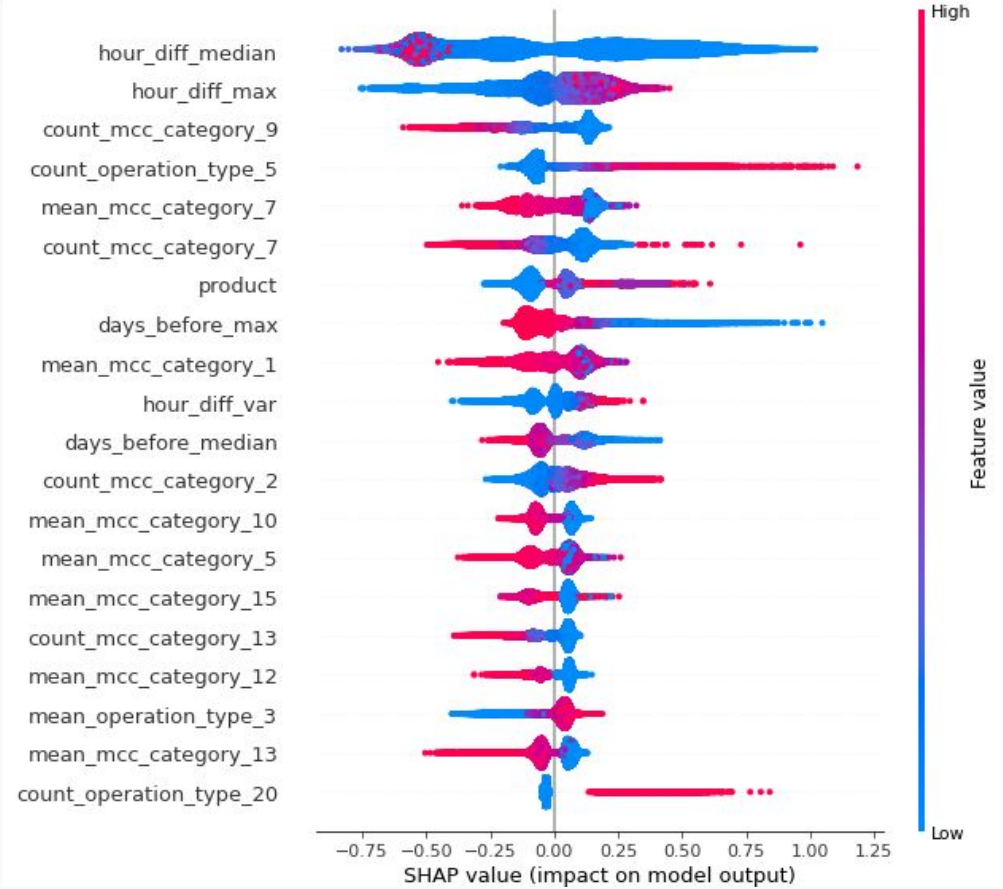
Анализ моделей с помощью расширенного списка метрик ML



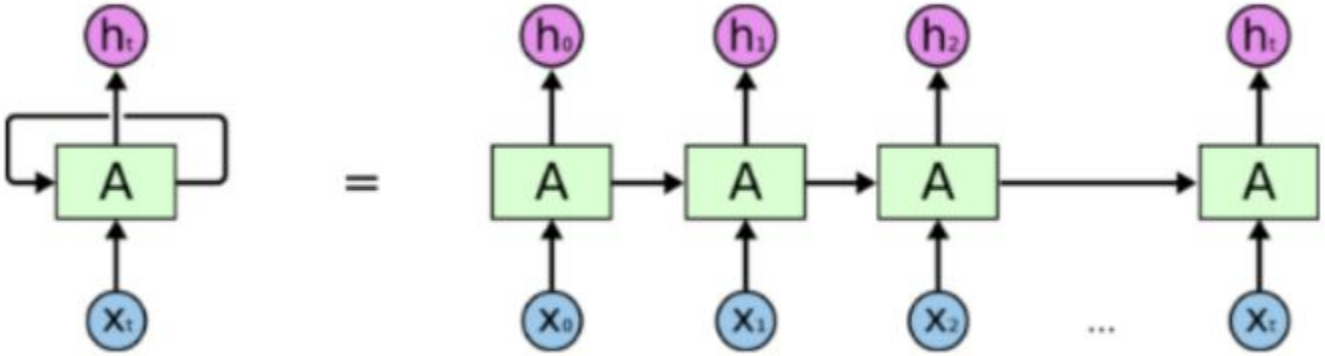


	model1	model2	model3	model4
Кол-во признаков	127	64	59	52
accuracy	0.9560	0.9558	0.9556	0.9556
balanced accuracy	0.5900	0.5885	0.5868	0.5864
precision	0.2026	0.1998	0.1965	0.1957
recall	0.2026	0.1998	0.1965	0.1957
Ошибка II рода	0.7973	0.8001	0.8034	0.8042
f1_score	0.2026	0.1998	0.1965	0.1957
roc_auc	0.7963	0.7920	0.7921	0.7875
roc_prc	0.1378	0.1352	0.1326	0.1323

Интерпретация прогнозов модели (shap)

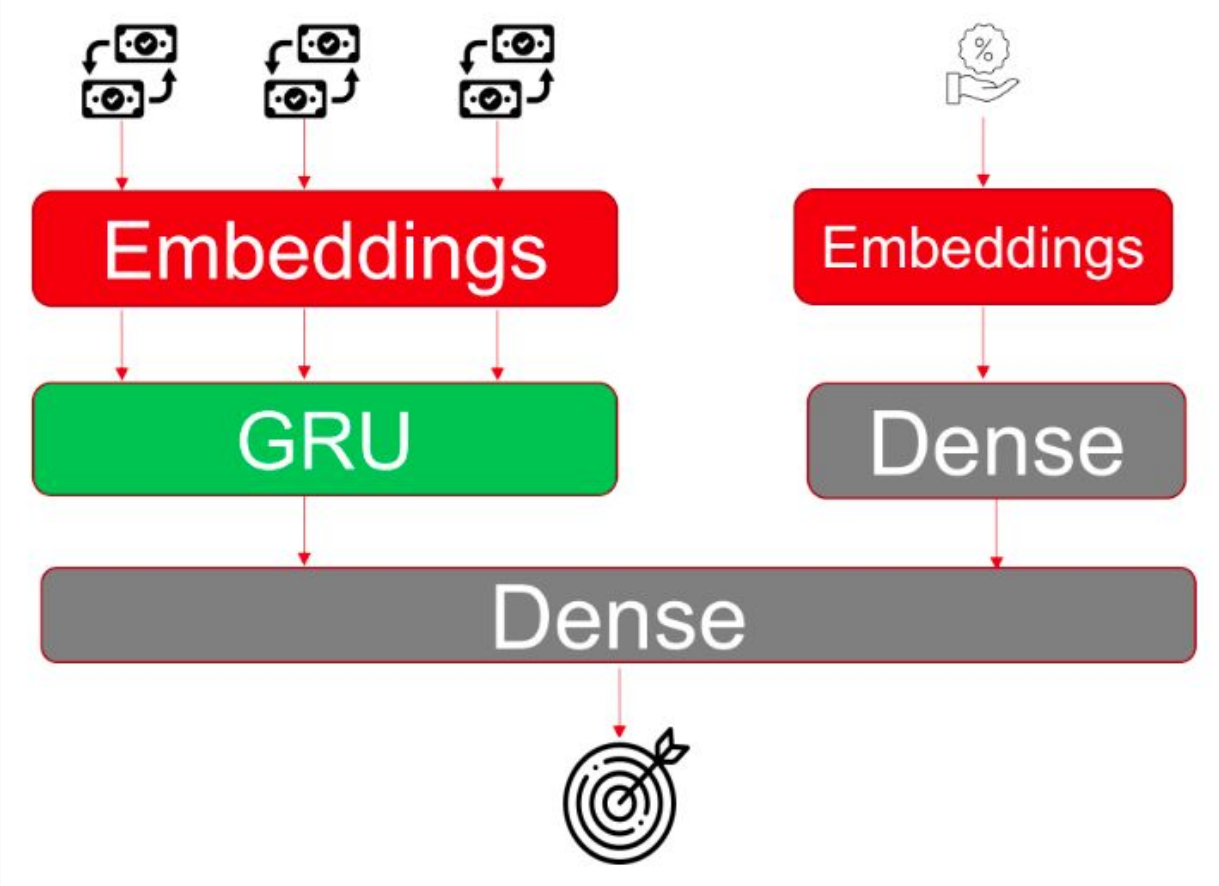


Рекуррентная нейронная сеть

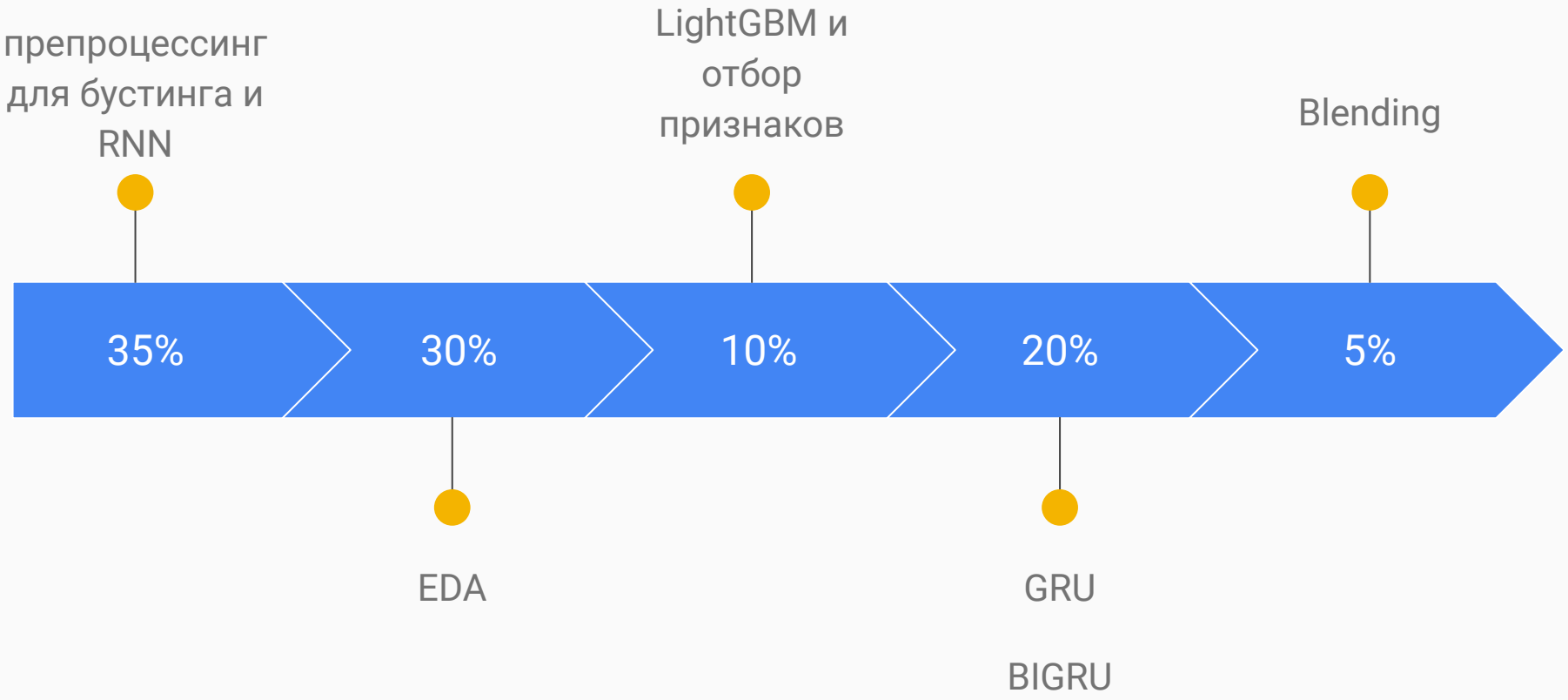


An unrolled recurrent neural network.

Архитектура рекуррентной нейронной сети



Тайминг


препроцессинг для бустинга и RNN
35%
EDA
30%
LightGBM и отбор признаков
10%
GRU
BIGRU
20%
Blending
5%

Результаты

- score = 0.7660770

  бронзовая медаль

- прототип

- сравнения моделей на расширенных метриках

Что не успел реализовать

Попробовать вместо блендинга использовать прогнозы нейросети как отдельные признаки и на них построить модели ML

Спасибо за внимание