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

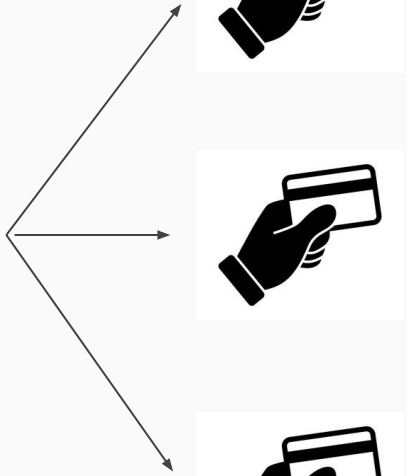
дипломная работа Соколова Александра

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)*

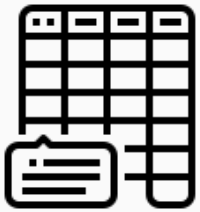
Билеты
Агенство
20000 р
21.04.2021
19:00 *(онлайн)*

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

Описание датасета

app_id	amnt	currency	operation_kind	card_type	operation_type	operation_type_group	ecommerce_flag	payment_system	income_flag	mcc	country	city	mcc_category	day_of_week	hour	days_before	weekofyear	hour_diff	transaction_number
0	0.465425	1	4	98	4	2	3	7	3	2	1	37	2	4	19	351	34	-1	1
0	0.000000	1	2	98	7	1	3	7	3	2	1	49	2	4	20	351	34	0	2
0	0.521152	1	2	98	3	1	3	7	3	2	1	37	2	4	20	351	34	0	3
0	0.356078	1	1	5	2	1	3	7	3	10	1	49	7	2	0	348	34	52	4
0	0.000000	1	2	98	7	1	3	7	3	2	1	49	2	4	16	337	53	280	5

app_id - идентификатор заявки на кредит

19 признаков в том числе 15 категориальных, 3 числовых и 1 индексный (transaction_number)

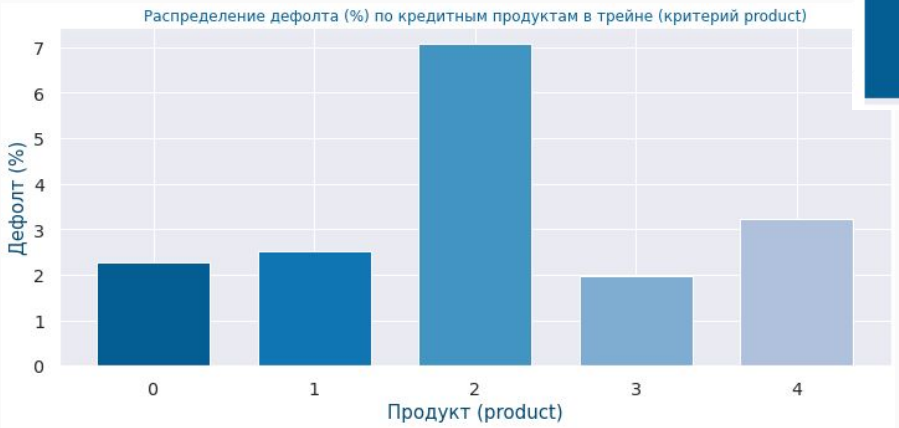
Один признак “product” в дата-фрейме target

Описание датасета (описание признаков)

amnt - нормированная сумма транзакции. 0.0 - соответствует пропускам
currency - идентификатор валюты транзакции
operation_kind - идентификатор типа транзакции
card_type - уникальный идентификатор типа карты
operation_type - идентификатор типа операции по пластиковой карте
operation_type_group - идентификатор группы карточных операций, например, деб. или кред. карта
ecommerce_flag - признак электронной коммерции
payment_system - идентификатор типа платежной системы
income_flag - признак списания/внесения денежных средств на карту
mcc - уникальный идентификатор типа торговой точки
country - идентификатор страны транзакции
city - идентификатор города транзакции
mcc_category - идентификатор категории магазина транзакции
day_of_week - день недели, когда транзакция была совершена
hour - час, когда транзакция была совершена
days_before - количество дней до даты выдачи кредита
weekofyear - номер недели в году, когда транзакция была совершена
hour_diff - количество часов с момента прошлой транзакции для данного клиента
transaction_number - Порядковый номер транзакции клиента

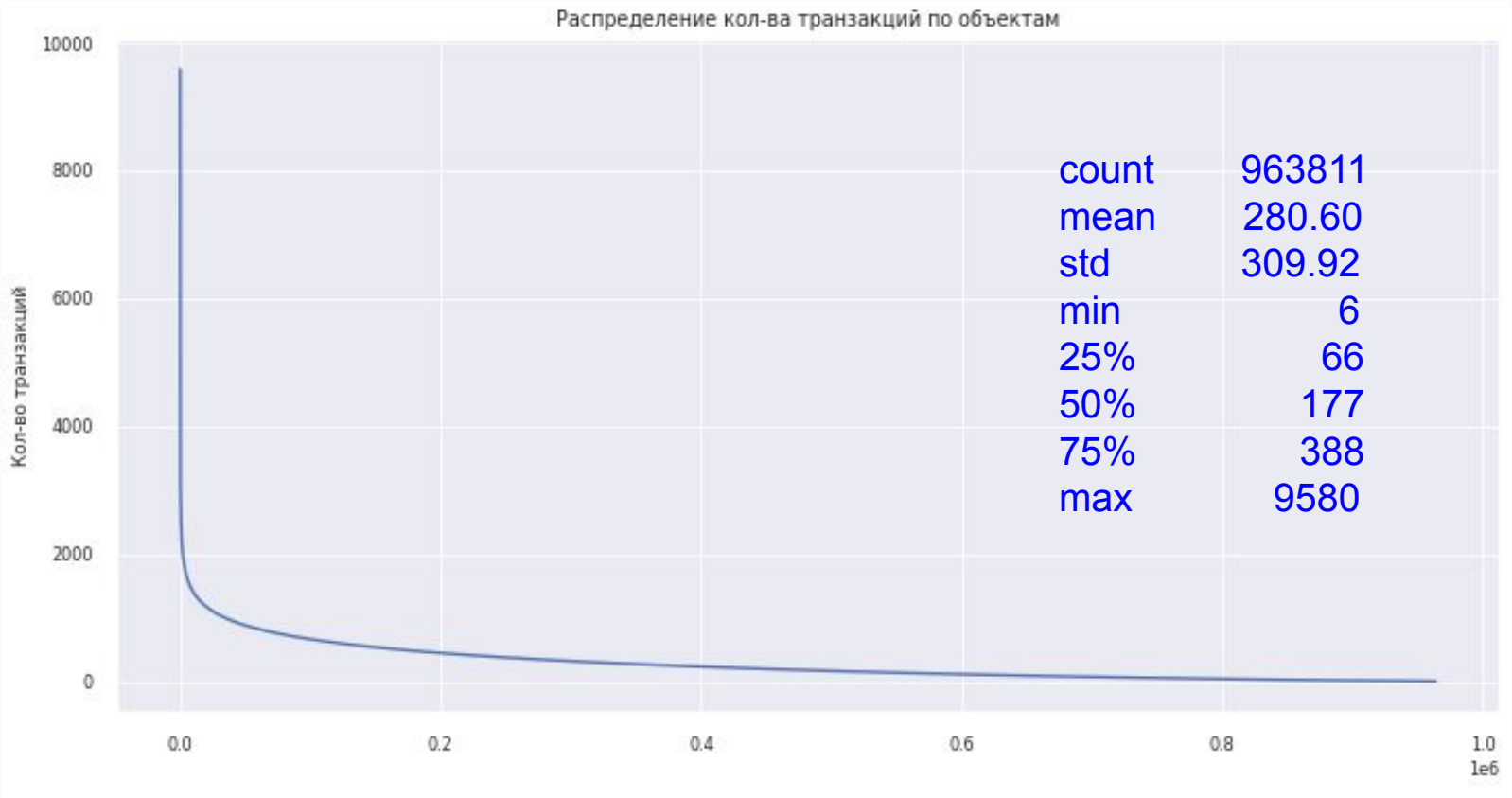
EDA по признаку ‘product’ (кредитный продукт)

Нулевой и первый кредитные продукты занимают около 80% всех кредитов



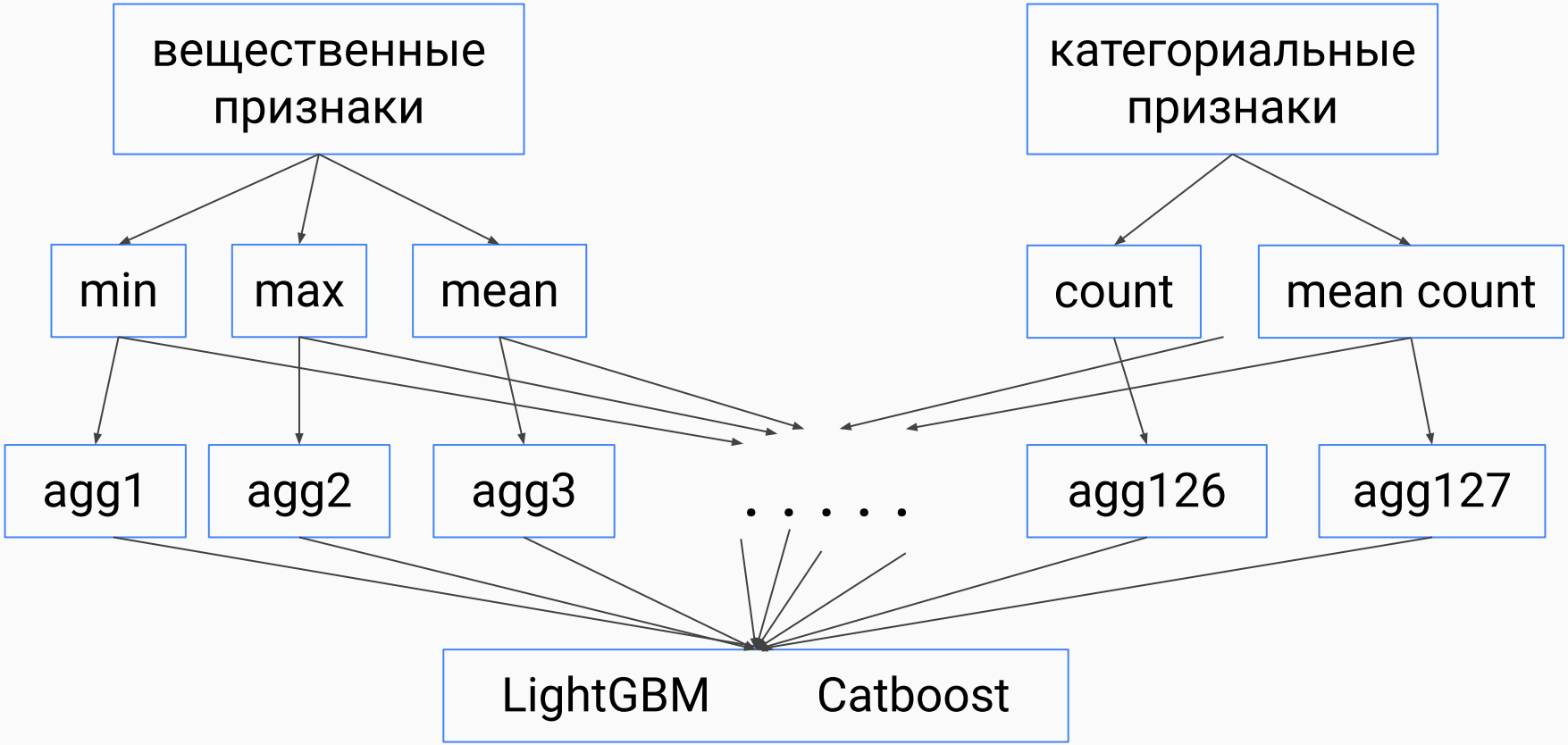
Процент дефолта по продукту 2 (7.1%) почти в три раза выше среднего значения по продуктам (2.7%)

EDA по количеству транзакций на один объект

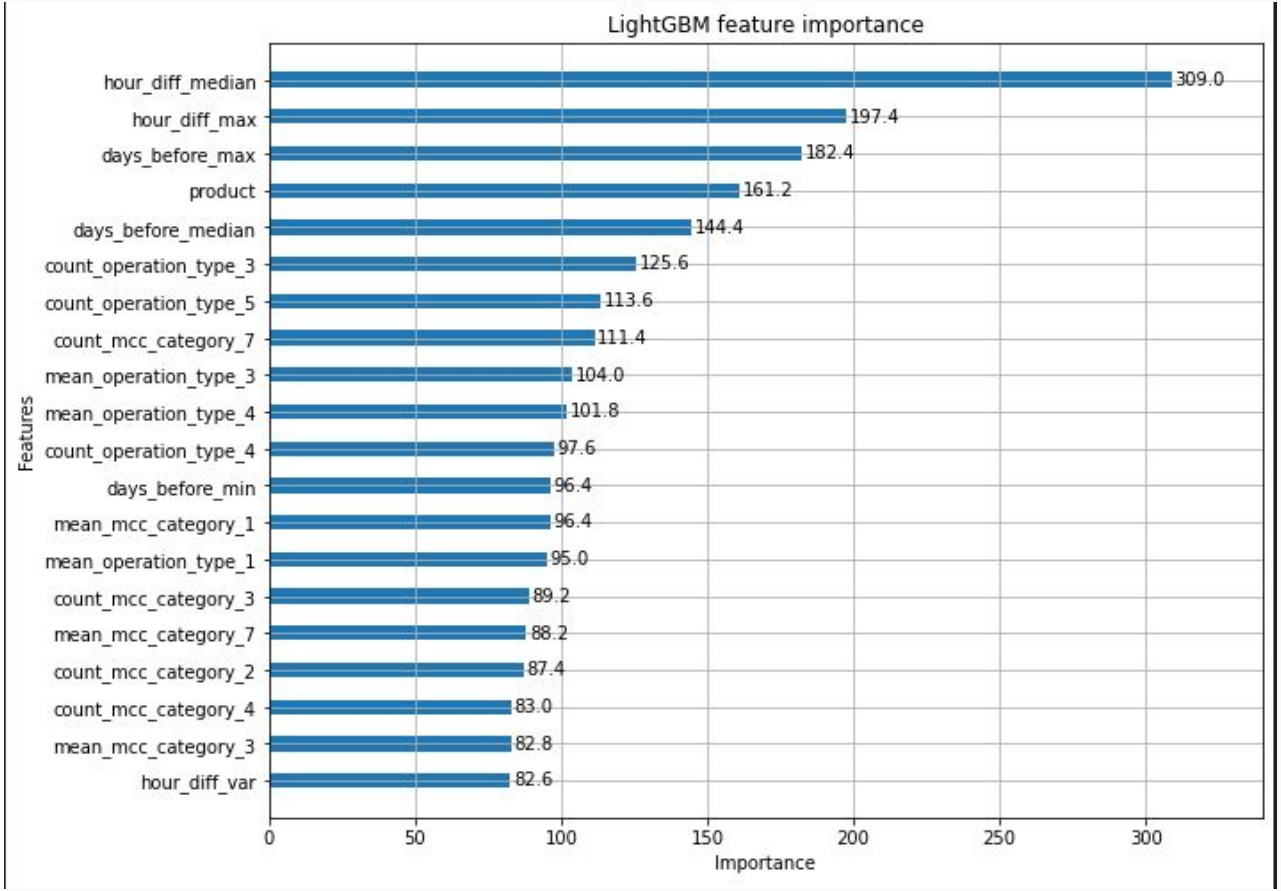


count	963811
mean	280.60
std	309.92
min	6
25%	66
50%	177
75%	388
max	9580

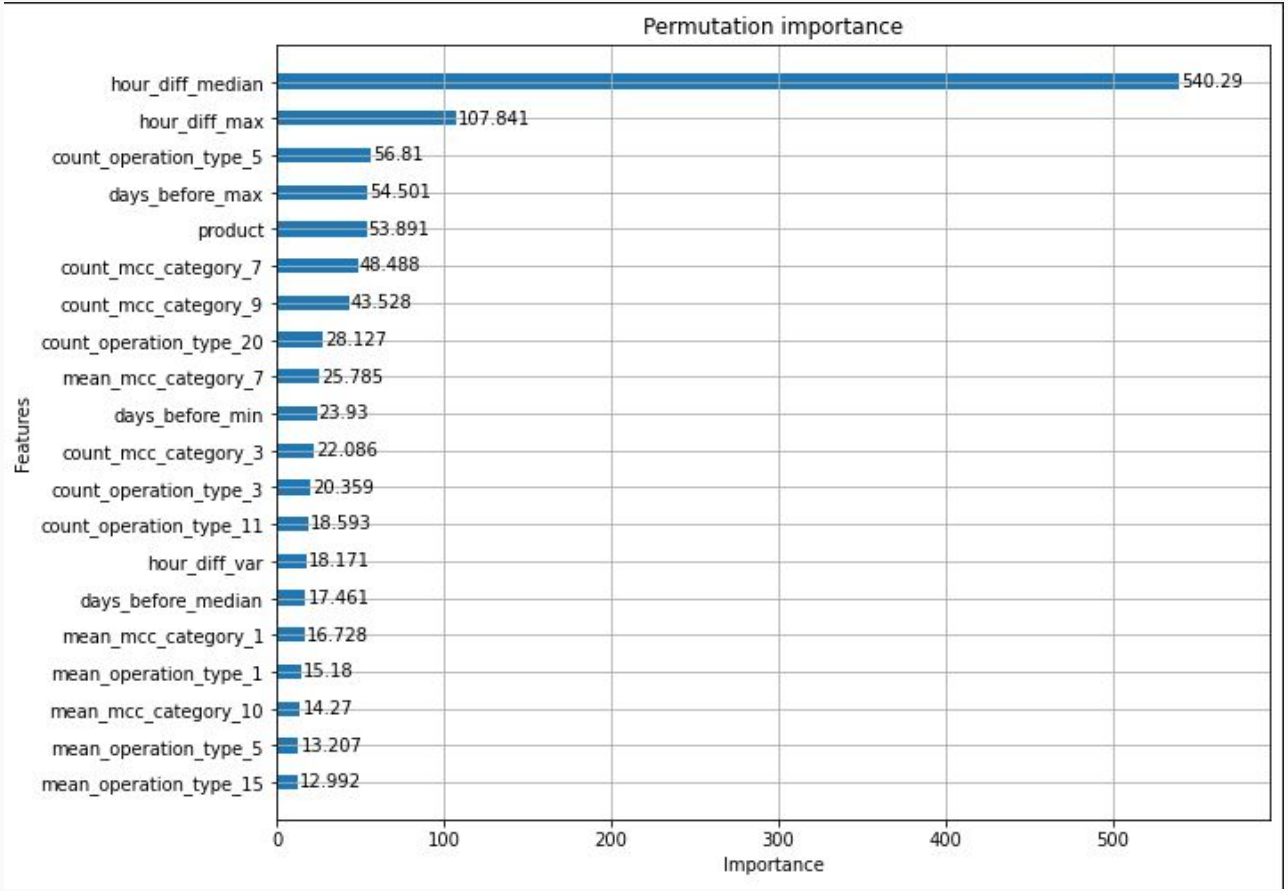
Препроцессинг для ML



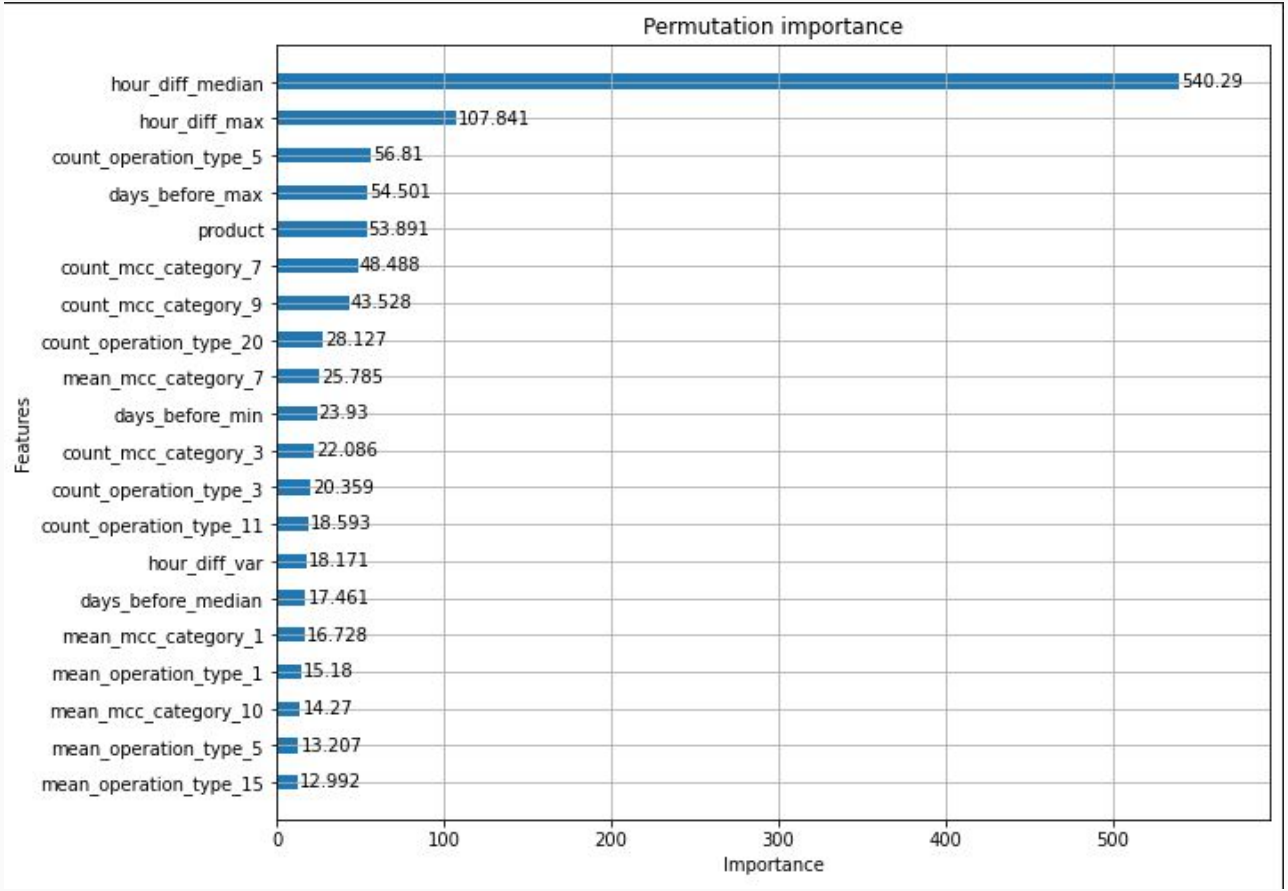
Отбор признаков (LightGBM feature importance)



Отбор признаков (Feature permutation importance)



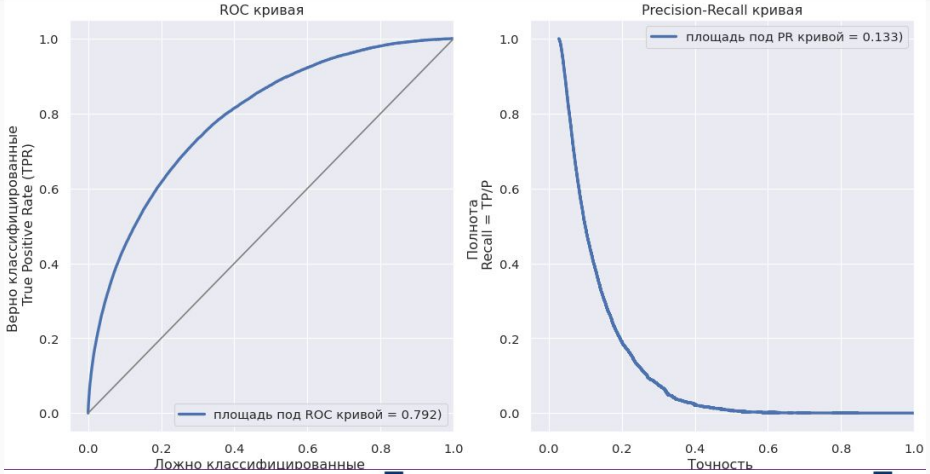
Отбор признаков (Target permutation importance)

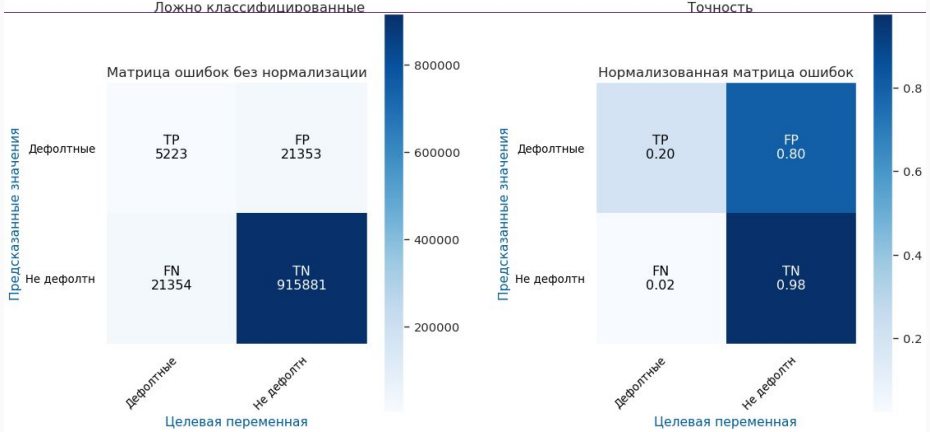


Промежуточные результаты по ML

	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

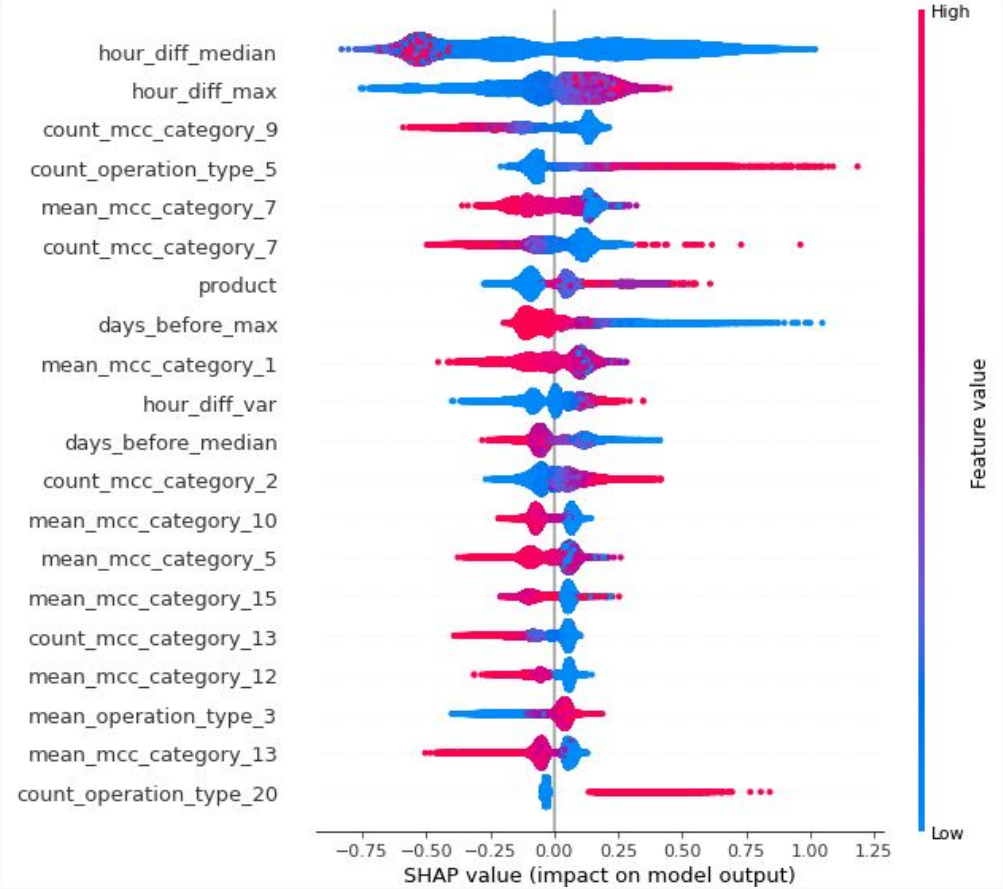
Анализ моделей с помощью расширенного списка метрик ML



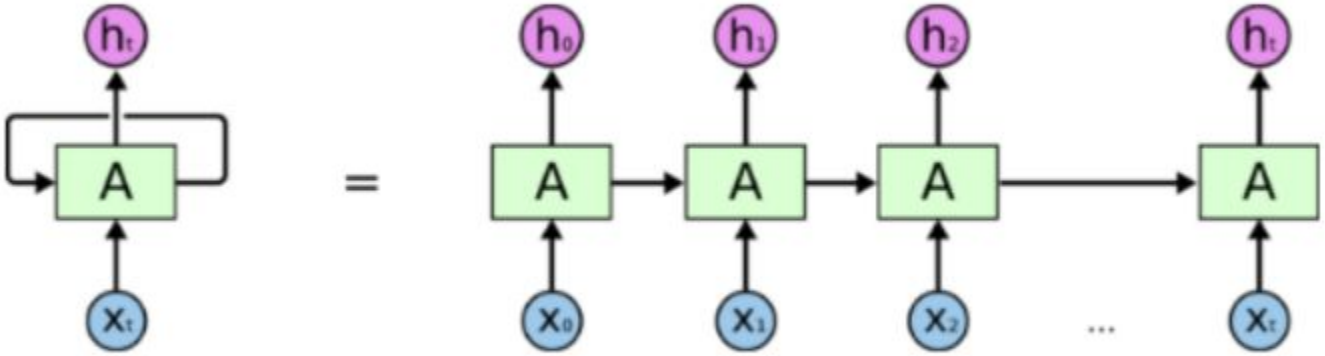


	model1	model2	model3	model4
Кол-во признаков	127	64	59	52
accuracy	0.9560	0.9558	0.9556	0.9556
balanced accuracy	0.5900	0.5885	0.5868	0.5864
precision	0.2026	0.1998	0.1965	0.1957
recall	0.2026	0.1998	0.1965	0.1957
Ошибка II рода	0.7973	0.8001	0.8034	0.8042
f1_score	0.2026	0.1998	0.1965	0.1957
roc_auc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

Интерпретация прогнозов модели (shap)

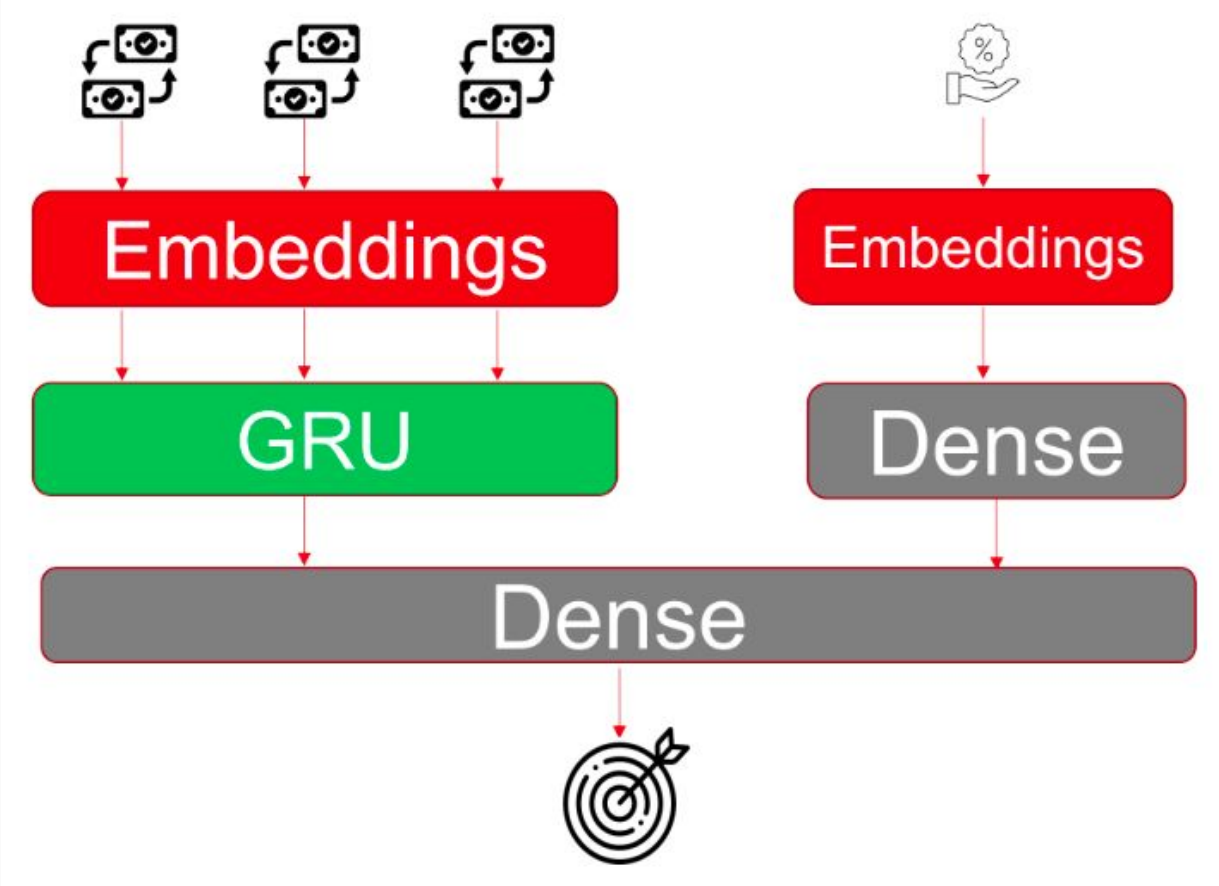


Рекуррентная нейронная сеть

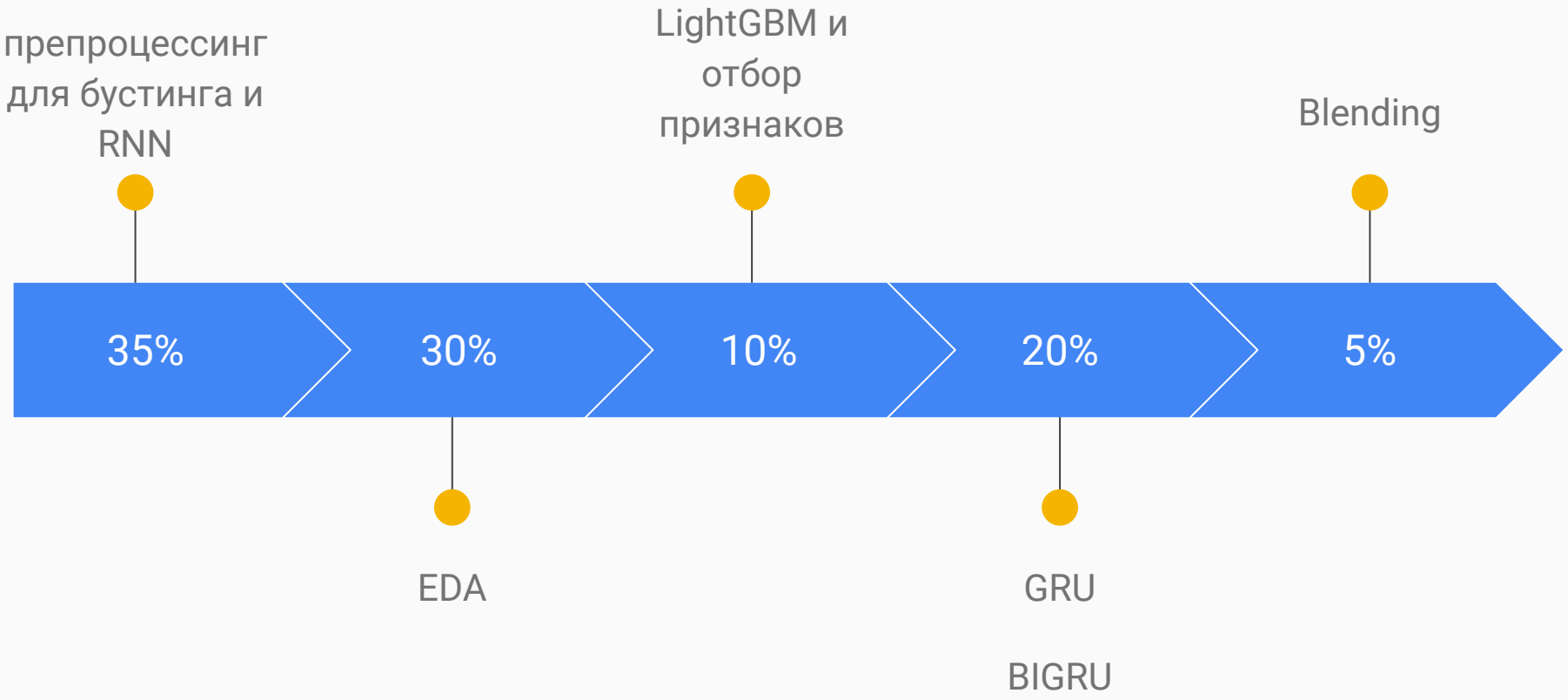


An unrolled recurrent neural network.

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