

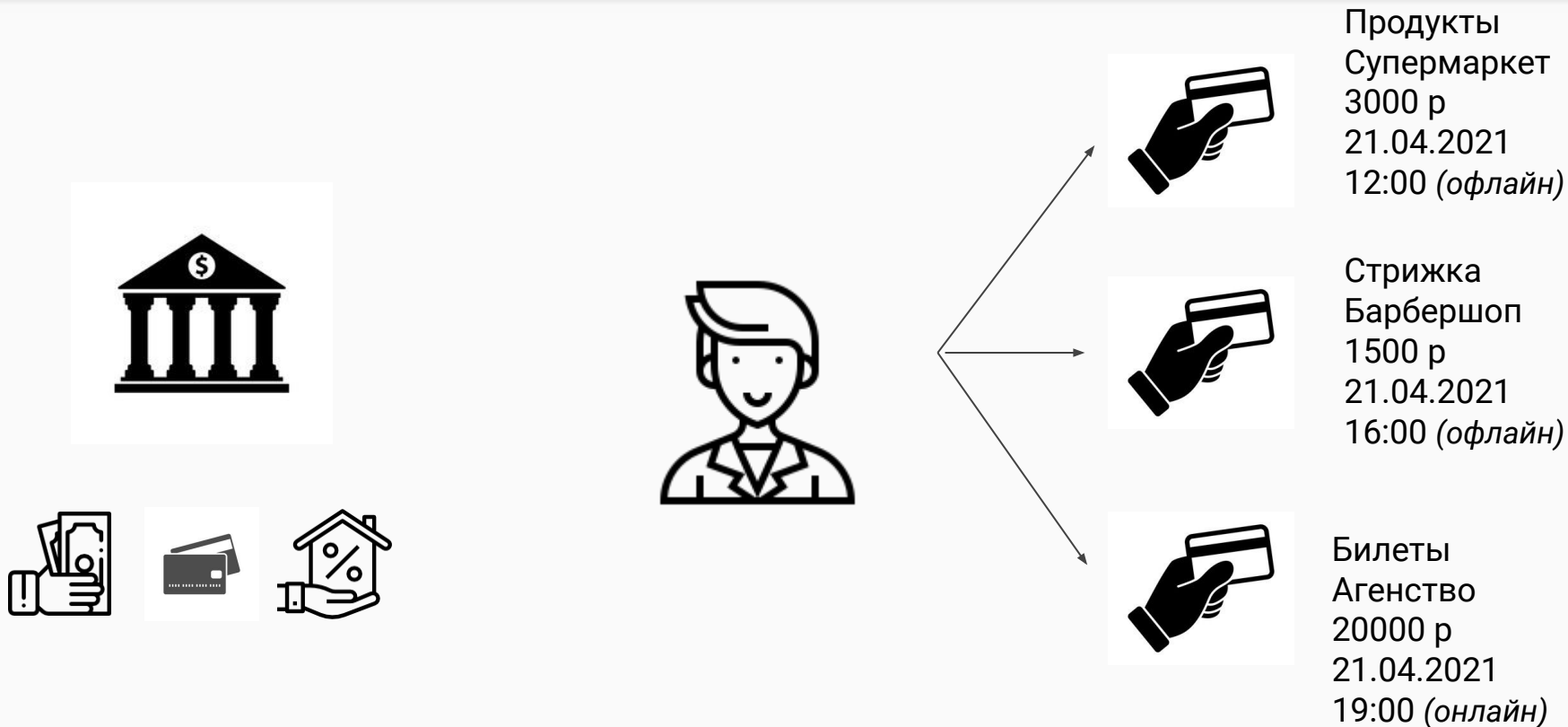
Решение задачи кредитного скоринга на основе градиентного бустинга на данных карточных транзакций клиента

дипломная работа Соколова Александра

Содержание

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

Постановка задачи

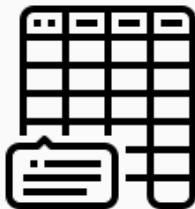


1. войти в число медалистов соревнования - AUC ROC на привате не ниже 0.7616840
2. реализовать прототип оптимального по размеру и качеству ML-решения в виде web приложения на heroku
3. использовать для оценки моделей метрики f1, AUC PRC, а также матрицы ошибок и другие метрики пытаюсь максимально приблизить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

Описание датасета



1 500 000 клиентов



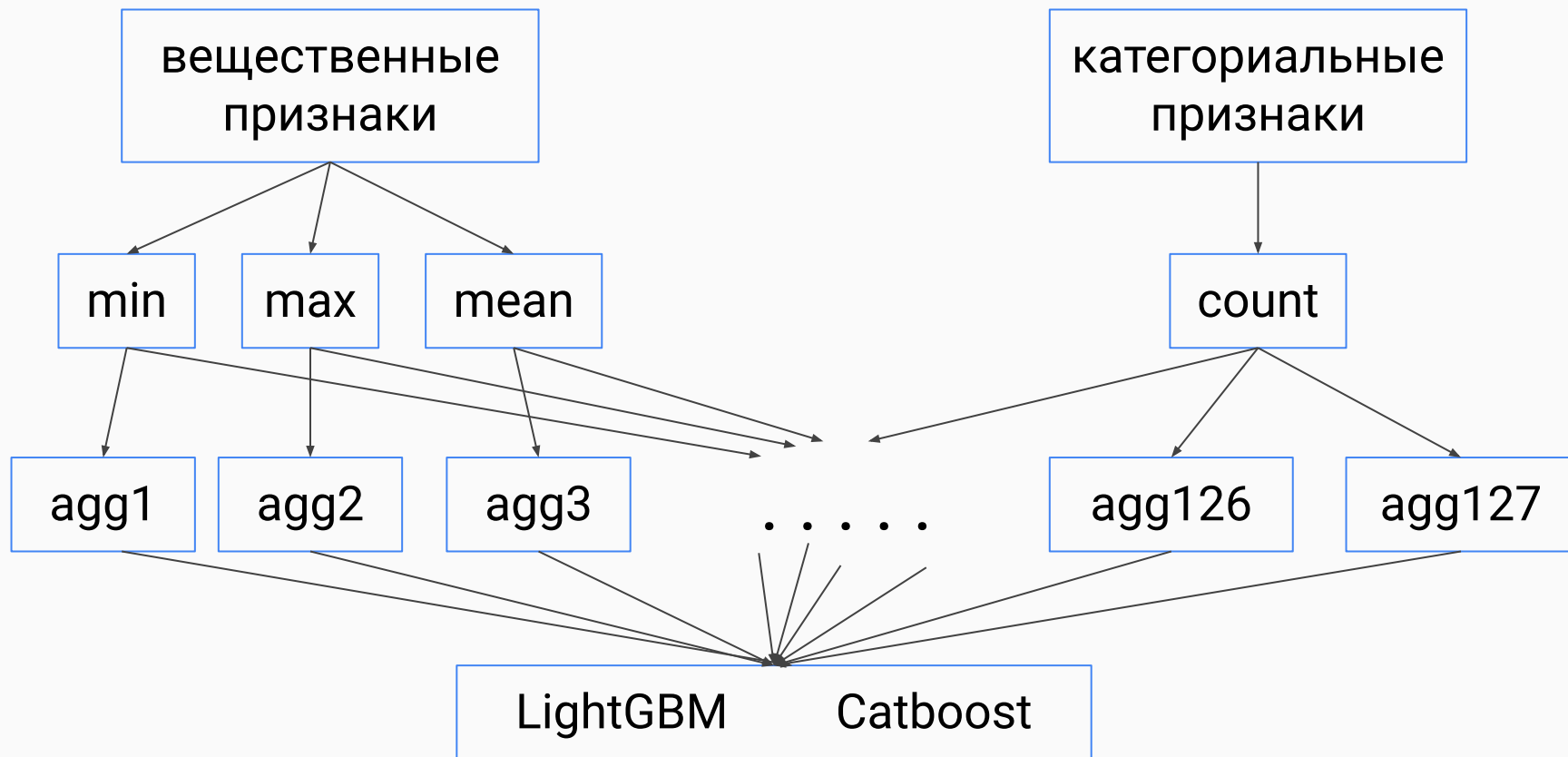
450 000 000 строк



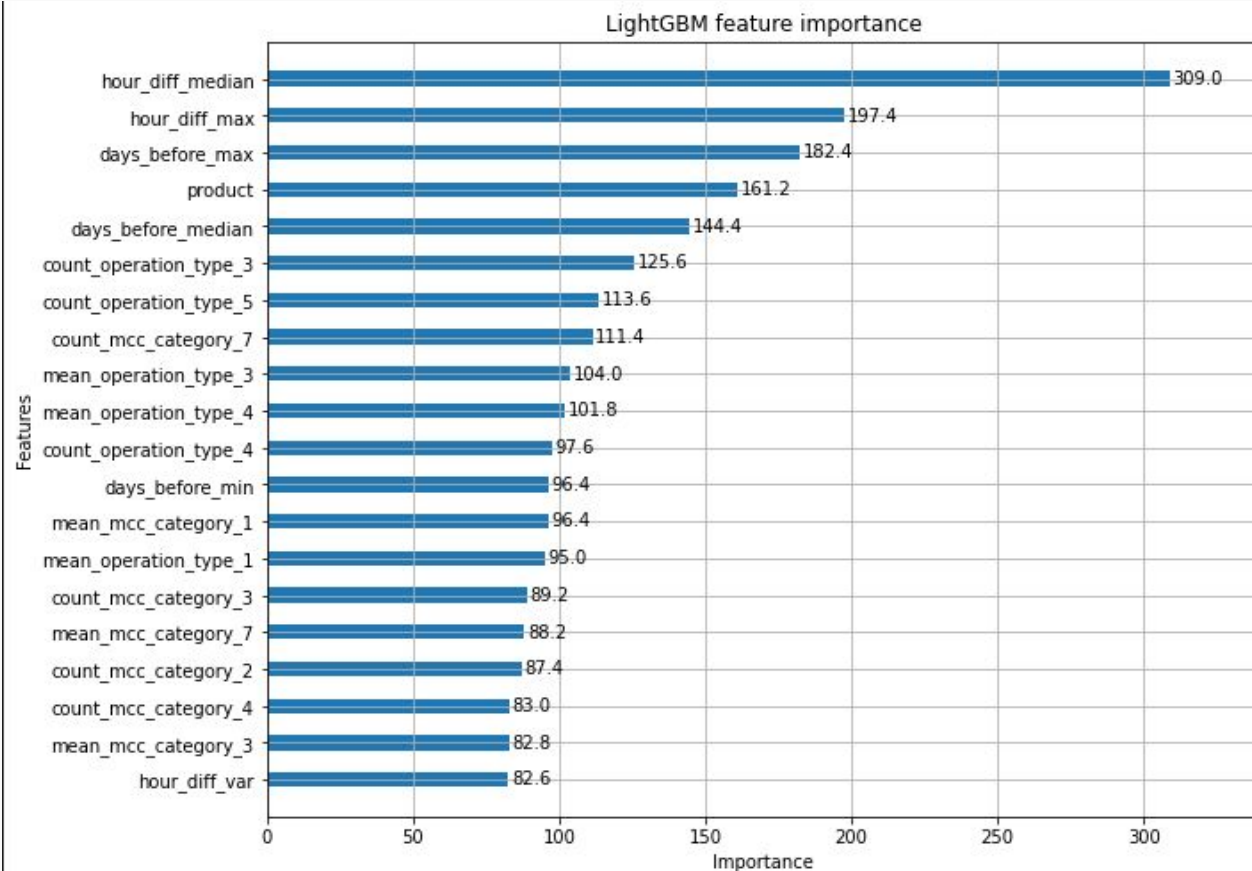
6 ГБ в формате parquet



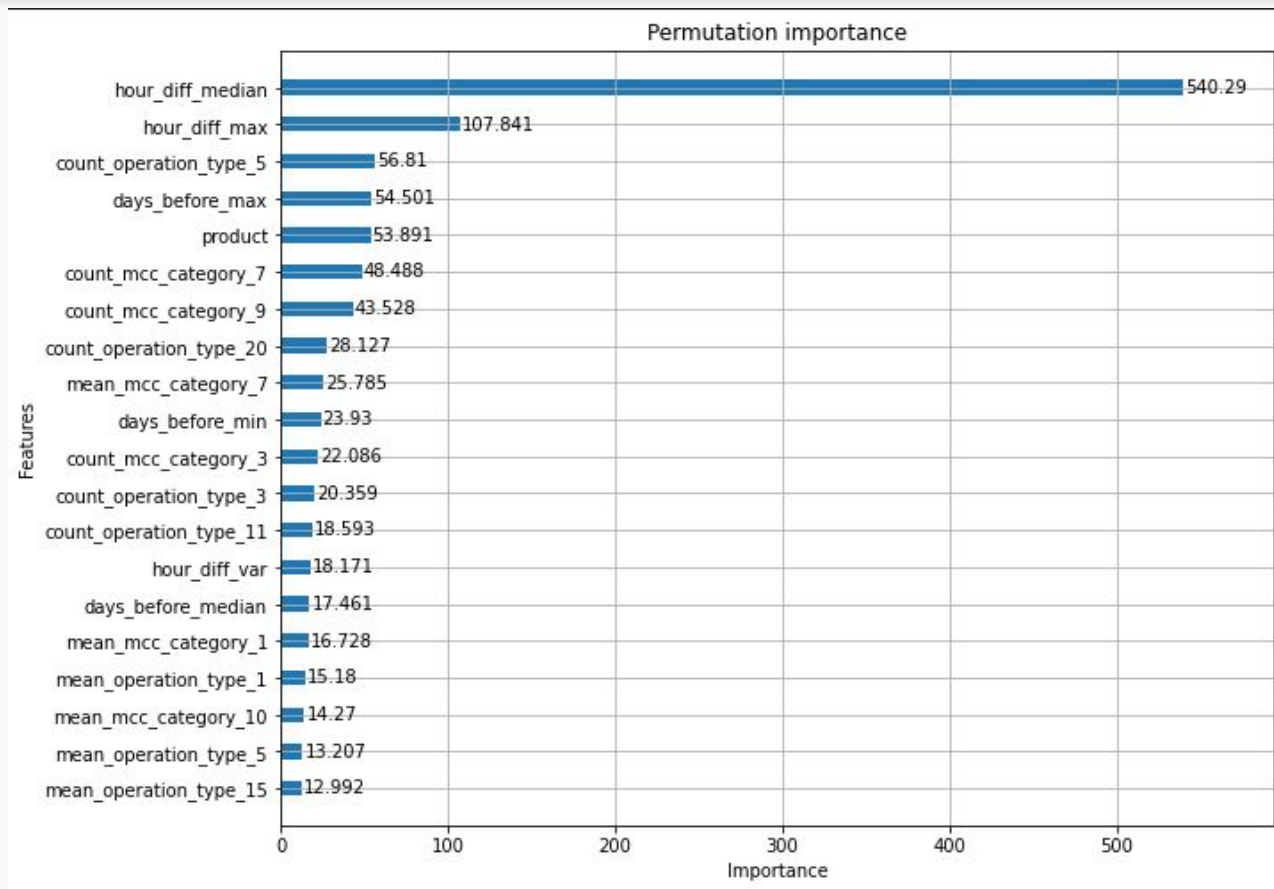
тестовая выборка смещена по времени



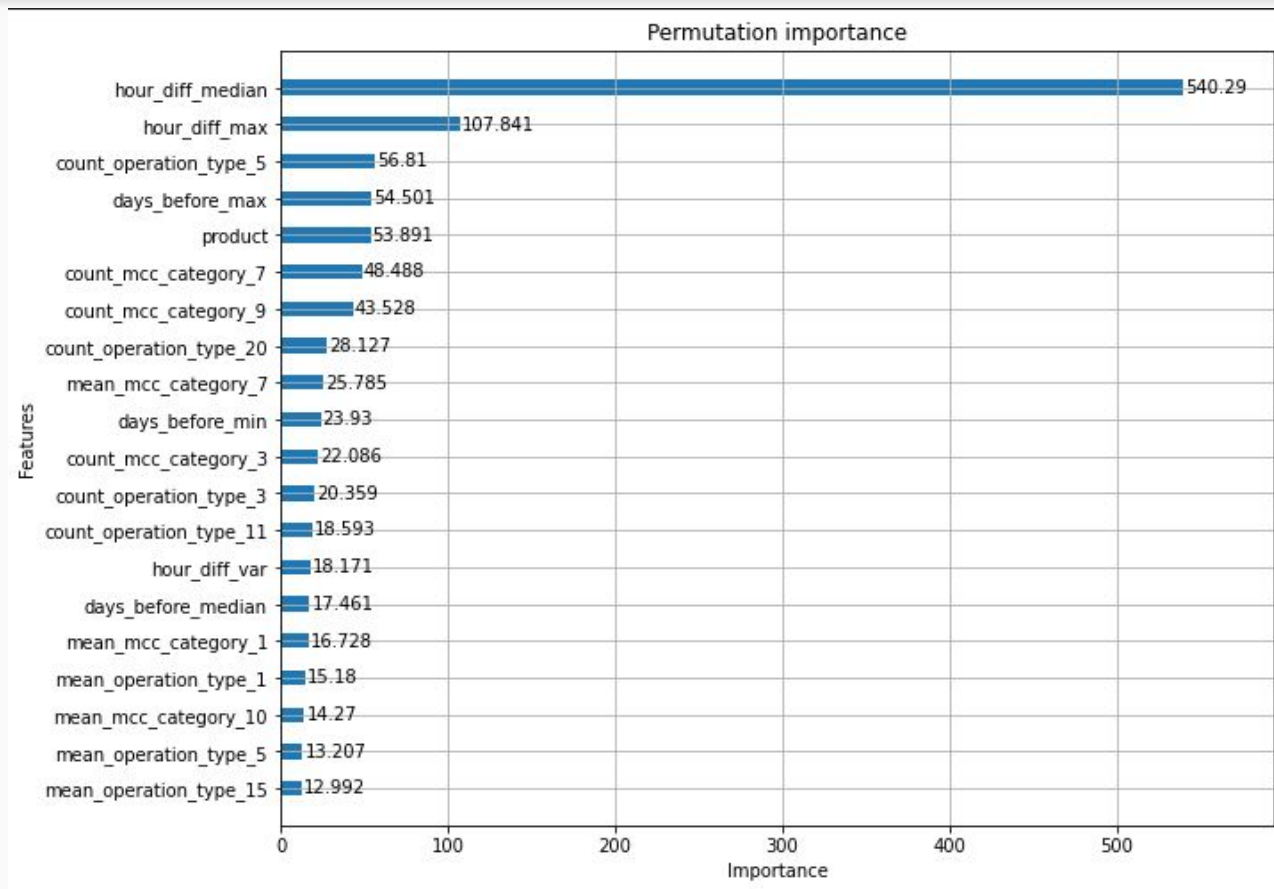
Отбор признаков (LightGBM feature importance)



Отбор признаков (Feature permutation importance)



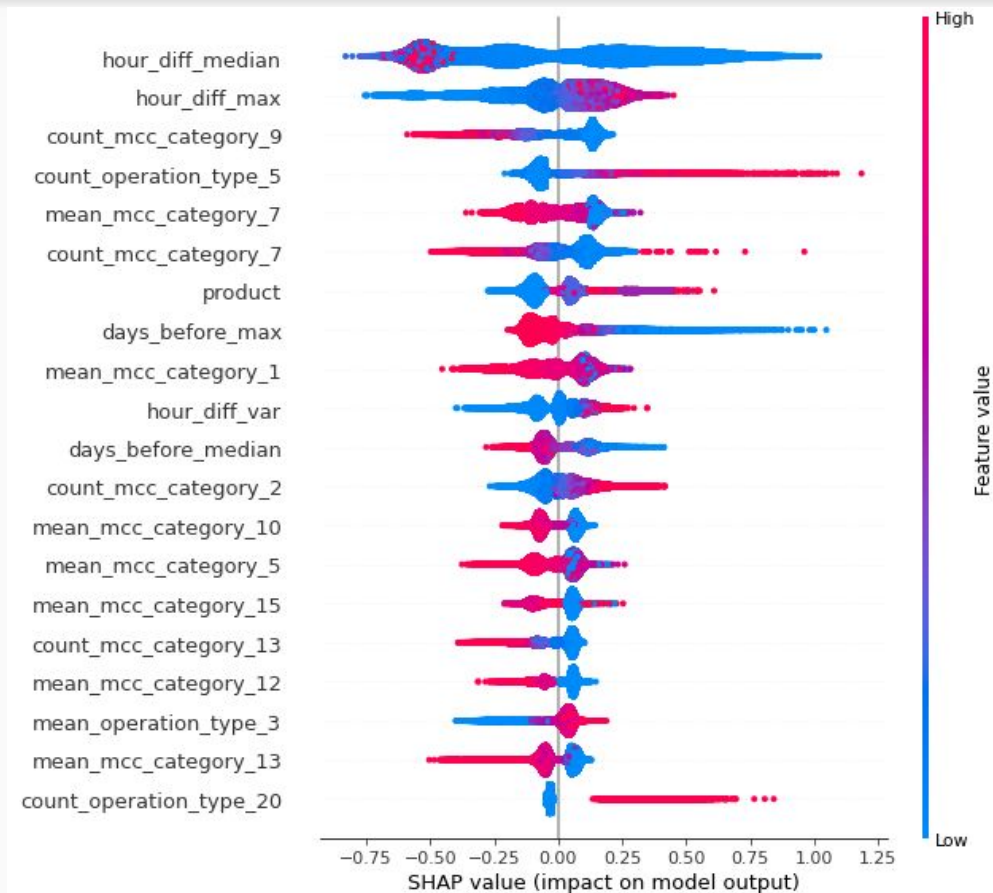
Отбор признаков (Target permutation importance)



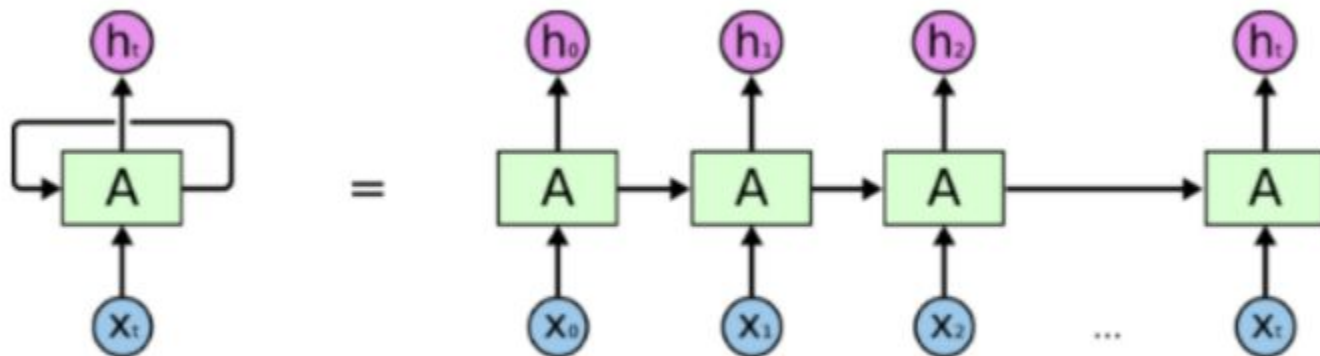
Промежуточные результаты по ML

	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	58	0.796	0.769	0.735
LightGBM	target permutation	60	0.792	0.765	0.733

Интерпретация прогнозов модели (shap)

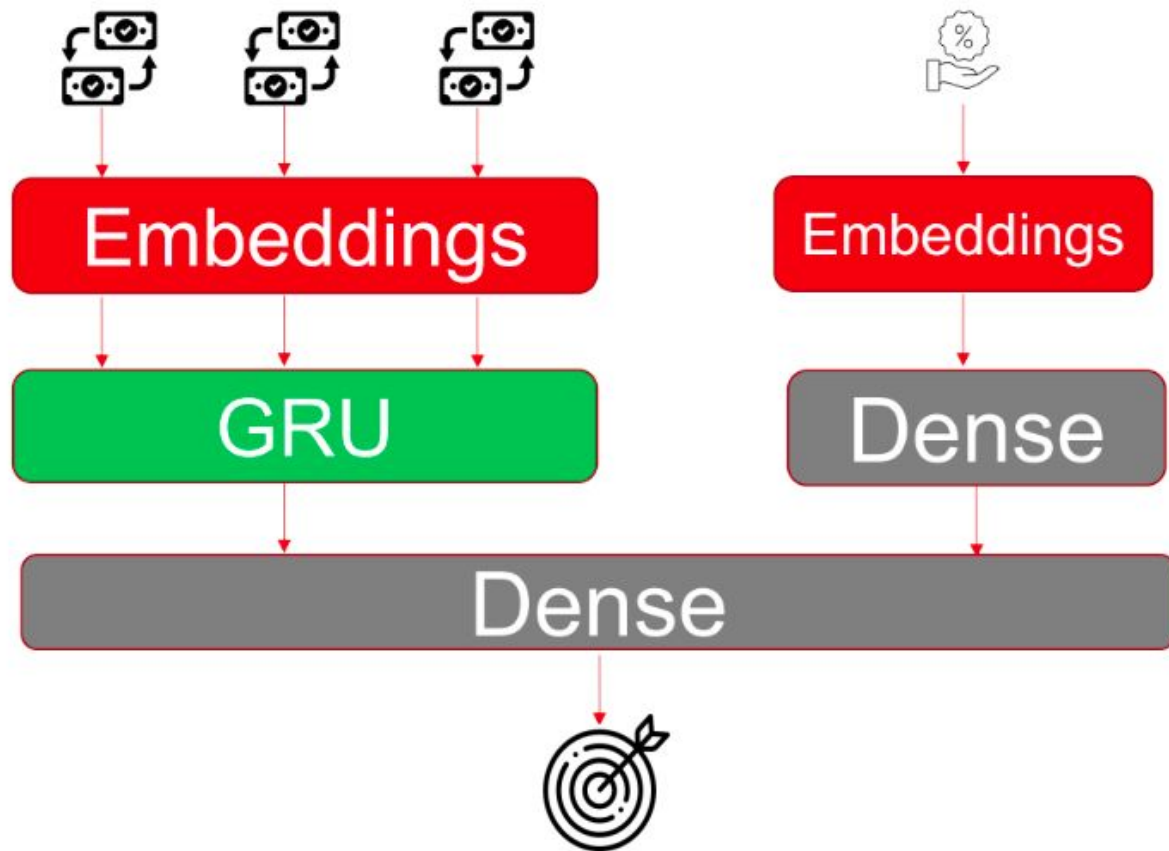


Рекуррентная нейронная сеть

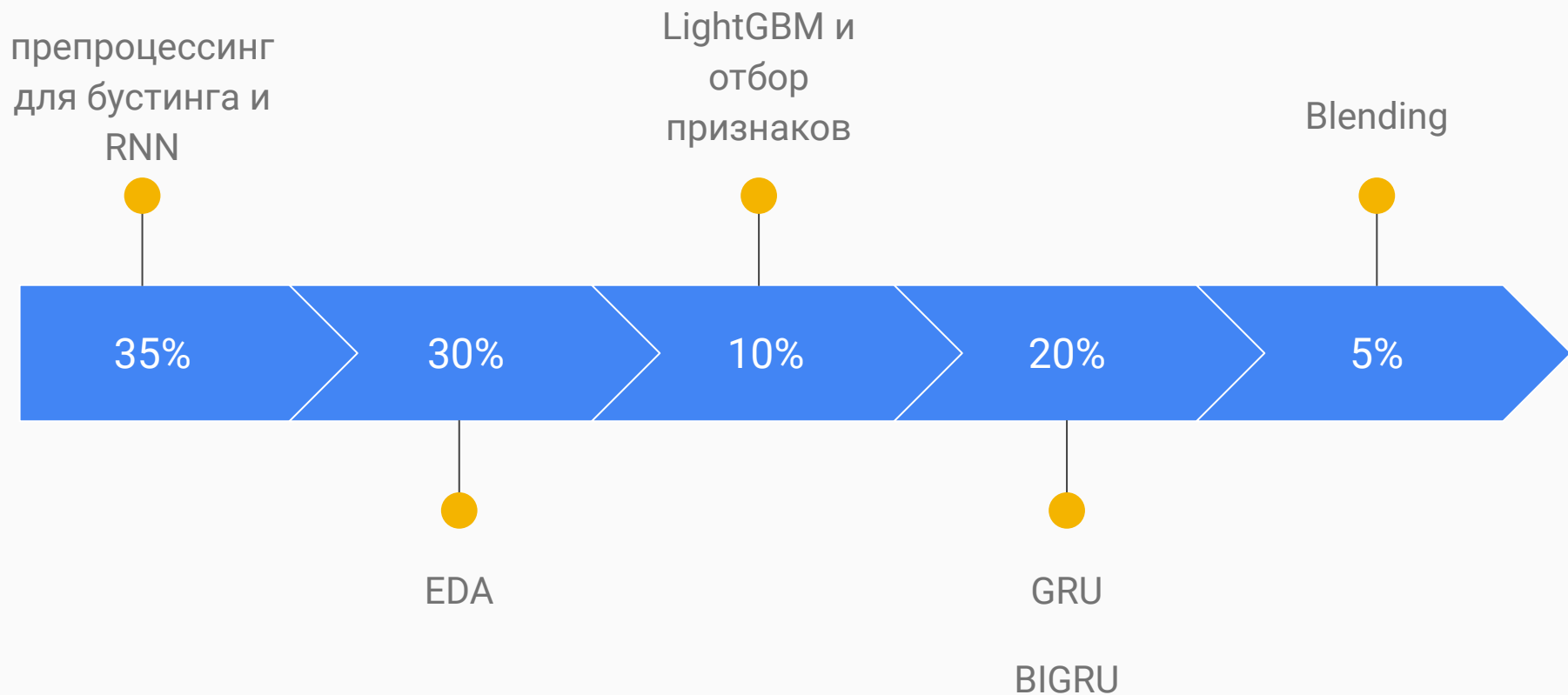


An unrolled recurrent neural network.

Архитектура рекуррентной нейронной сети



Тайминг



Результаты



- score = 0.7660770



- прототип



- визуализация сравнения
моделей на расширенных
метриках

1. Разобраться как применять `shap` для интерпретации модели обученной на нескольких фолдах
2. Попробовать вместо блендинга использовать прогнозы нейросети как отдельные признаки и на них построить модели ML

Спасибо за внимание