

ПРОГНОЗИРОВАНИЕ ОТТОКА

ДЛЯ РОЗНИЧНОГО МАГАЗИНА

СОДЕРЖАНИЕ:

- Подготовка данных:
 - ▶ Разбиваем выборку на датасеты
 - ▶ Формируем признаки
- Анализ данных и определение оттока
- Построение модели и интерпретация результатов

КРАТКОЕ РЕЗЮМЕ

- ▶ Идея заключается в том, что обучать модель лучше на данных, сгруппированных по транзакциям.
- ▶ В течение жизненного цикла клиента паттерны его поведения могут меняться. Поэтому отток - это не характеристика клиента, а событие. Оно может быть вызвано определенными триггерами (неудачная покупка), либо трендом (увеличение интервала между последними покупками).
- ▶ Задача происходит из постановки: "С какой вероятностью вернется клиент после текущей транзакции, учитывая историю клиента на момент транзакции?".
- ▶ Если вероятность возвращения снижается, то клиента надо дополнительно стимулировать.

ПЛАН ДЕЙСТВИЙ

- ▶ Разные клиенты имеют разные привычки и частоту покупок. **Сгруппируем похожих клиентов**, выясним распределение дней до следующей покупки в кластерах.
- ▶ Чем больше дней прошло с момента покупки, тем меньше вероятность возвращения клиента. Возьмем **для каждого кластера пороговое значение в днях** - после невозвращения клиента в течение этого времени с момента последней покупки будем считать, что произошел отток.
- ▶ Можно экспериментировать: порог выбрать в качестве **N-% квантиля распределения дней до следующей покупки**. Чем больше N, тем больше точность модели, но меньше практическая польза.
- ▶ Дав определение оттоку, обучим модель прогнозировать его вероятность.



ПОДГОТОВКА ДАННЫХ

Описание признаков и их
агрегирования

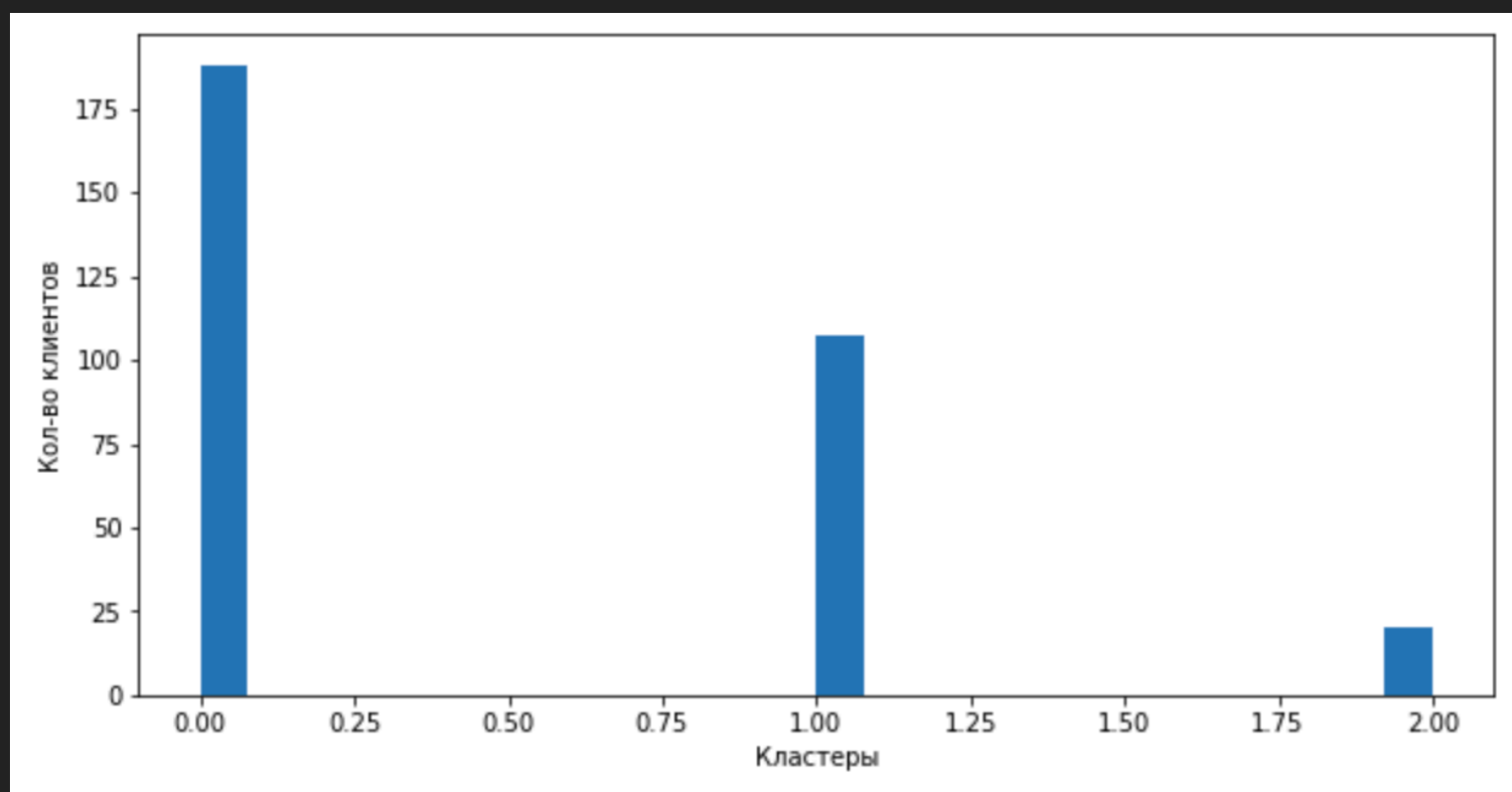
СОБЕРЕМ ДАТАСЕТЫ

- ▶ Выборка для обучения кластеризации - с начала 2017 года до июля 2017 года.
- ▶ Выборка для прогнозирования оттока - с августа 2018 года до конца 2018 года:
 - ▶ На ней же предсказывается кластер согласно накопленной на дату транзакции истории по клиенту.
 - ▶ Выборка делится на обучающую и отложенную. На отложенной оценивается ошибка.
- ▶ Будем использовать только данные по клиентам, у которых было 3 и более покупок.

КЛАСТЕРИЗАЦИЯ (ДАННЫЕ ДО ИЮЛЯ 2018)

БЕРЕМ НАКОПЛЕННУЮ ПО КЛИЕНТУ ИСТОРИЮ НА ДАТУ ПОСЛЕДНЕЙ ПОКУПКИ:

- Сколько у клиента было транзакций / транзакций определенного товара / транзакций товаров определенной категории
- С какой средней периодичностью в днях клиент ранее совершал покупки / покупки определенных товаров / покупки товаров определенной категории



ОСОБЕННОСТИ
КЛАСТЕРИЗАЦИИ:

- 3 кластера
- Кластеризация KMeans

Количество клиентов в кластерах

ПРОГНОЗИРОВАНИЕ ОТТОКА (ДААННЫЕ С АВГУСТА 2018)

БЕРЕМ НАКОПЛЕННУЮ ПО КЛИЕНТУ ИСТОРИЮ НА ДАТУ КОНКРЕТНОЙ ПОКУПКИ:

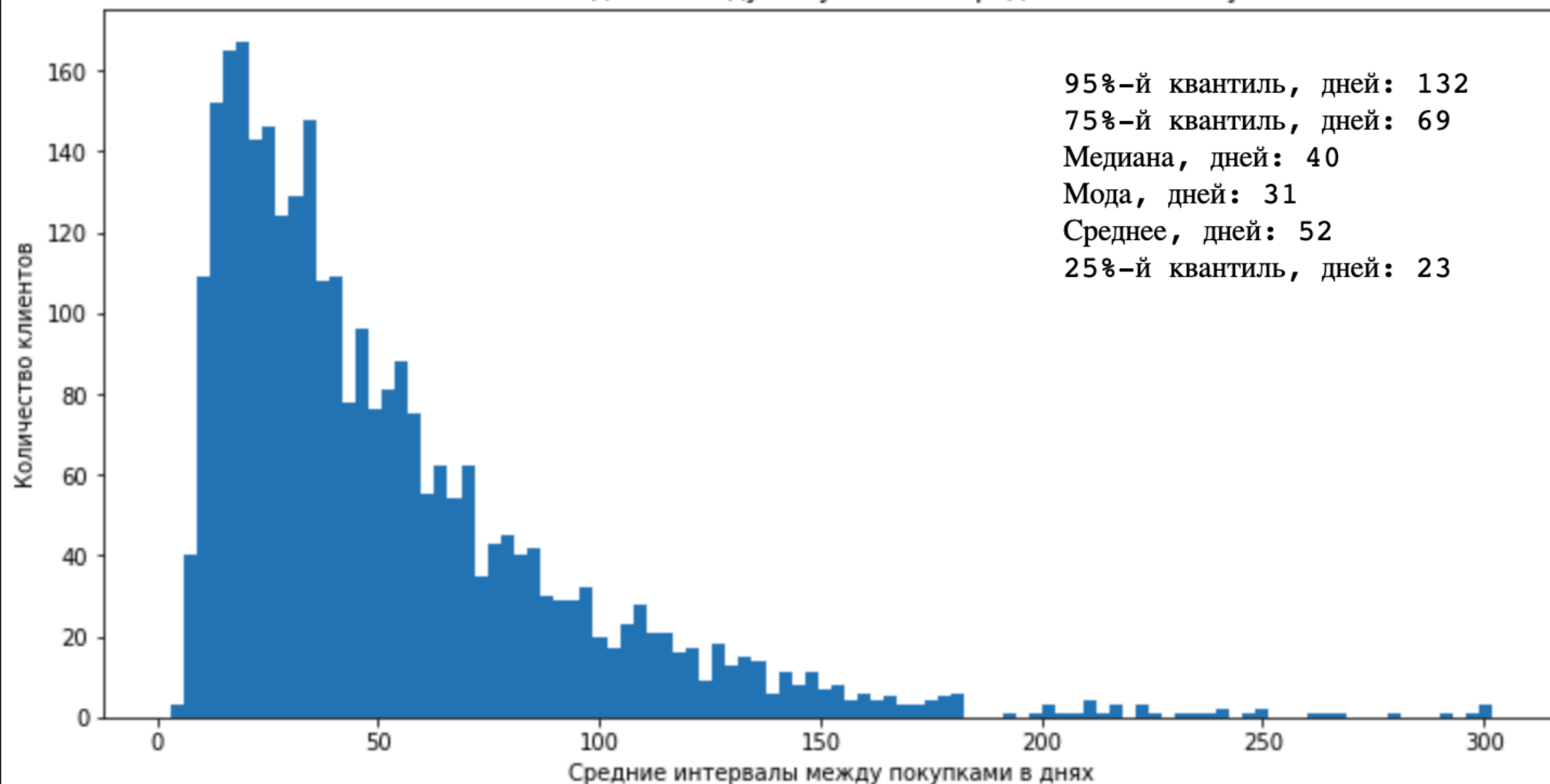
- Сколько у клиента ранее было транзакций / транзакций определенного товара / транзакций товаров определенной категории
- С какой средней периодичностью в днях клиент ранее совершал покупки / покупки определенного товара / покупки товаров определенной категории
- Разница в днях между последней и предпоследней транзакцией / между 2-й и 3-й транзакциями с конца / 3-й и 4-й транзакциями
- В текущей и предыдущих транзакциях, какая средняя стоимость товаров в корзине / средний чек корзины / среднее количество товаров в корзине
- Кластеры по клиентам
- Кол-во дней до следующей покупки, либо до текущего числа



АНАЛИЗ ДАННЫХ

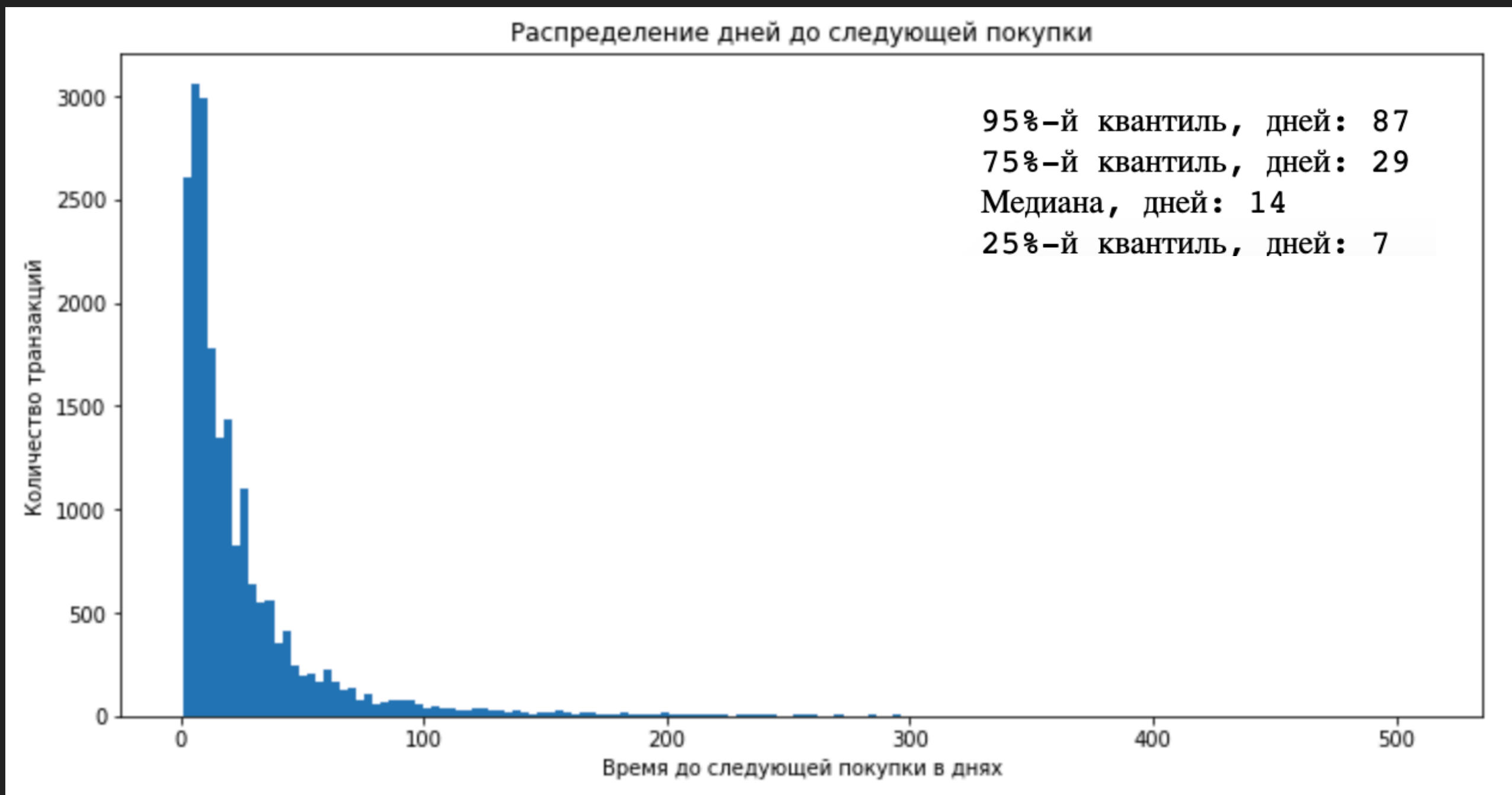
РАСПРЕДЕЛЕНИЕ СРЕДНИХ ИНТЕРВАЛОВ МЕЖДУ ПОКУПКАМИ

Количество дней между покупками, в среднем по клиенту



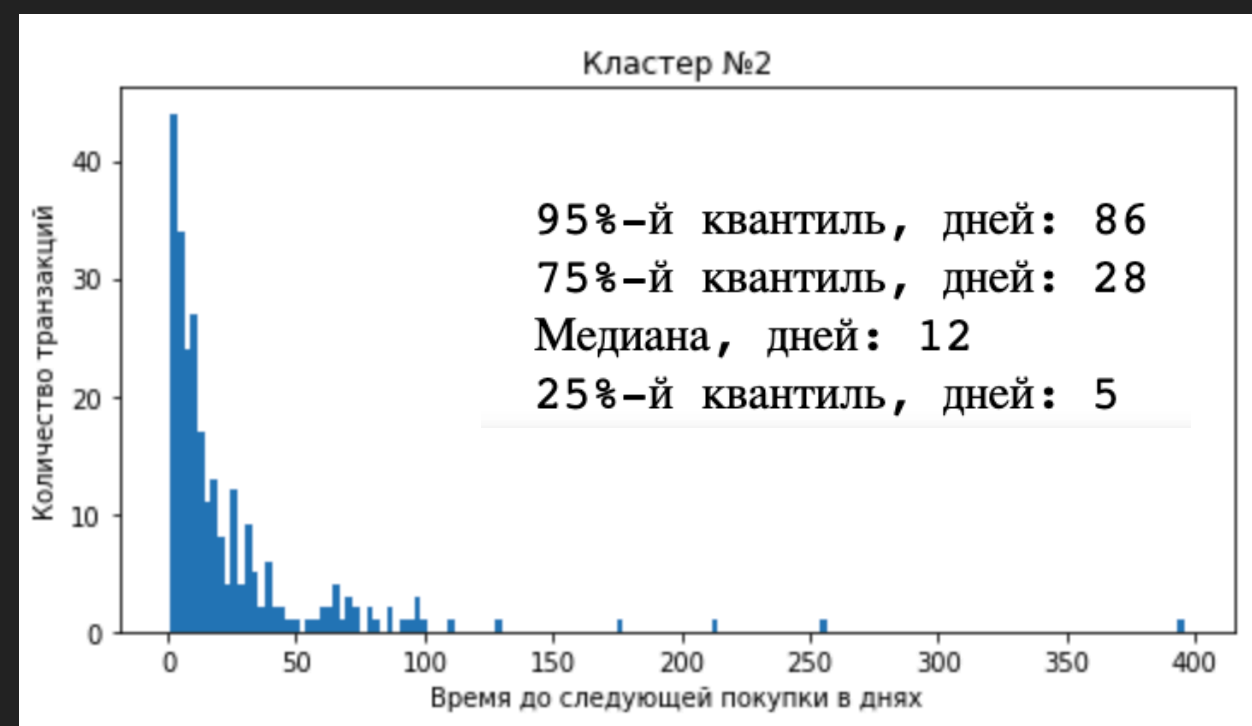
Наибольшее число клиентов в среднем возвращается через 31 день, половина клиентов возвращается в среднем в течение 40 дней. Средний интервал составляет 52 дня.

РАСПРЕДЕЛЕНИЕ ИНТЕРВАЛОВ ДО БУДУЩЕЙ ПОКУПКИ



В половине случаев клиенты возвращаются в течение 14 дней. Почти всегда (95% случаев) клиенты возвращались в течение 87 дней.

РАСПРЕДЕЛЕНИЕ ИНТЕРВАЛОВ ДО БУДУЩЕЙ ПОКУПКИ В КЛАСТЕРАХ



Распределение в кластерах значительно отличаются от общей выборки.

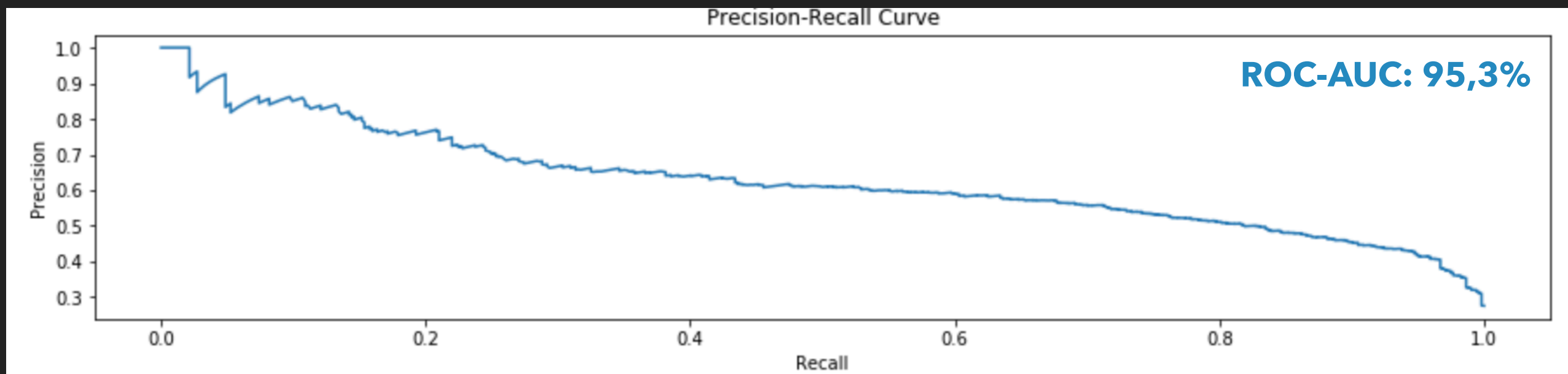


ПОСТРОЕНИЕ МОДЕЛЕЙ

КАКИЕ МОДЕЛИ БЫЛИ ПОСТРОЕНЫ:

- **Случайный лес:**
 - Кол-во деревьев 3000
 - Кол-во признаков, из которых выбирается лучший для разбиения, вычисляется как корень из числа признаков
 - Сбалансированная по классам целевой переменной выборка
 - ▶ Дополнительный вес для транзакций, из числа последних для клиента

КРИВАЯ ТОЧНОСТИ / ПОЛНОТЫ



Для порога вероятности 5%:

- Precision = 34,7%
- recall = 98,6%

	predict_0	predict_1
true_0	5396	185
true_1	7	270

Confusion matrix

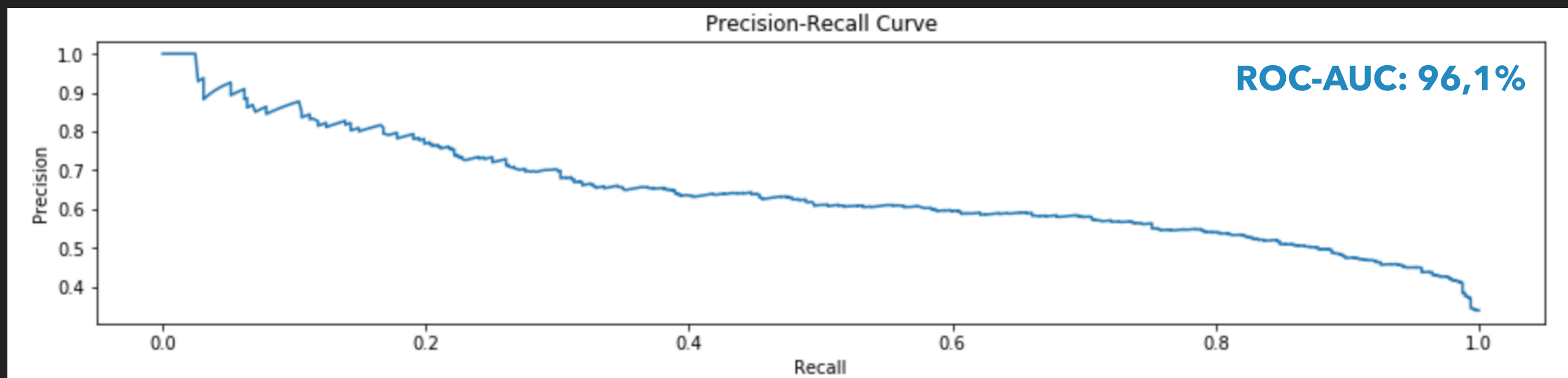
Для порога вероятности 30%:

- Precision = 53,1%
- recall = 75,1%

	predict_0	predict_1
true_0	4575	340
true_1	128	386

Confusion matrix

КРИВАЯ ТОЧНОСТИ / ПОЛНОТЫ



Для порога вероятности 5%:

- ▶ Precision = 35,5%
- ▶ recall = 99,3%

	predict_0	predict_1
true_0	4075	871
true_1	3	480

Confusion matrix

Для порога вероятности 30%:

- ▶ Precision = 53,5%
- ▶ recall = 81%

	predict_0	predict_1
true_0	4606	340
true_1	92	391

Confusion matrix

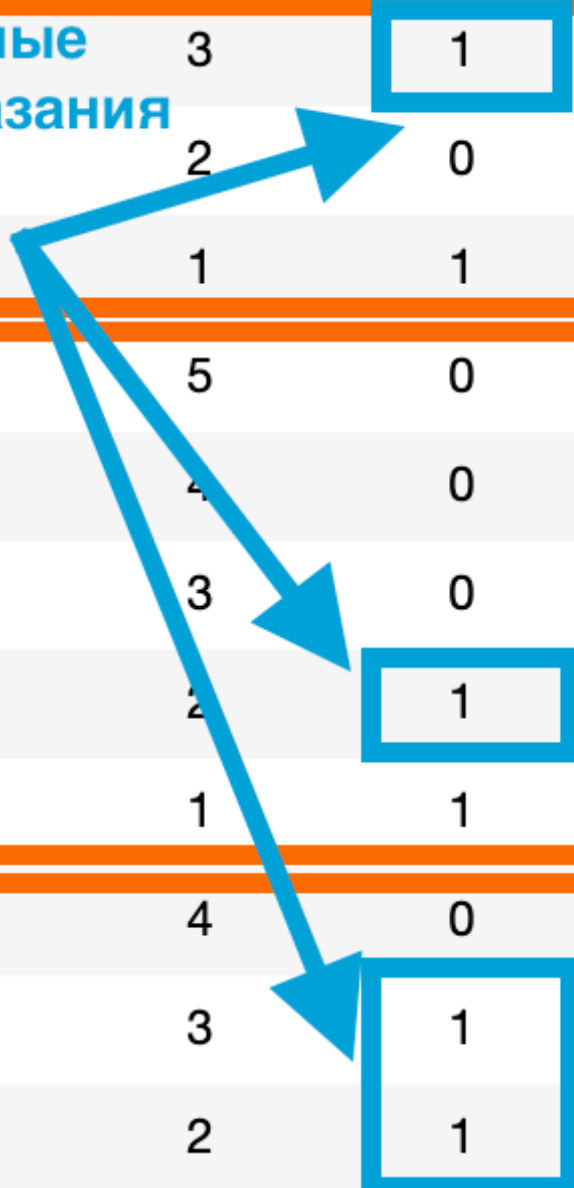
ВЫВОДЫ

- В дальнейшем предполагается делать прогноз по оттоку следующим образом:
 - ▶ С необходимой периодичностью делается срез по последним транзакциям всех клиентов. В качестве признака **future_interval** берется количество дней с момента последней транзакции по текущее число.
 - ▶ С помощью модели прогнозируется вероятность оттока.
 - ▶ В зависимости от вероятности принимаются те или иные меры по маркетинговому стимулированию клиента.

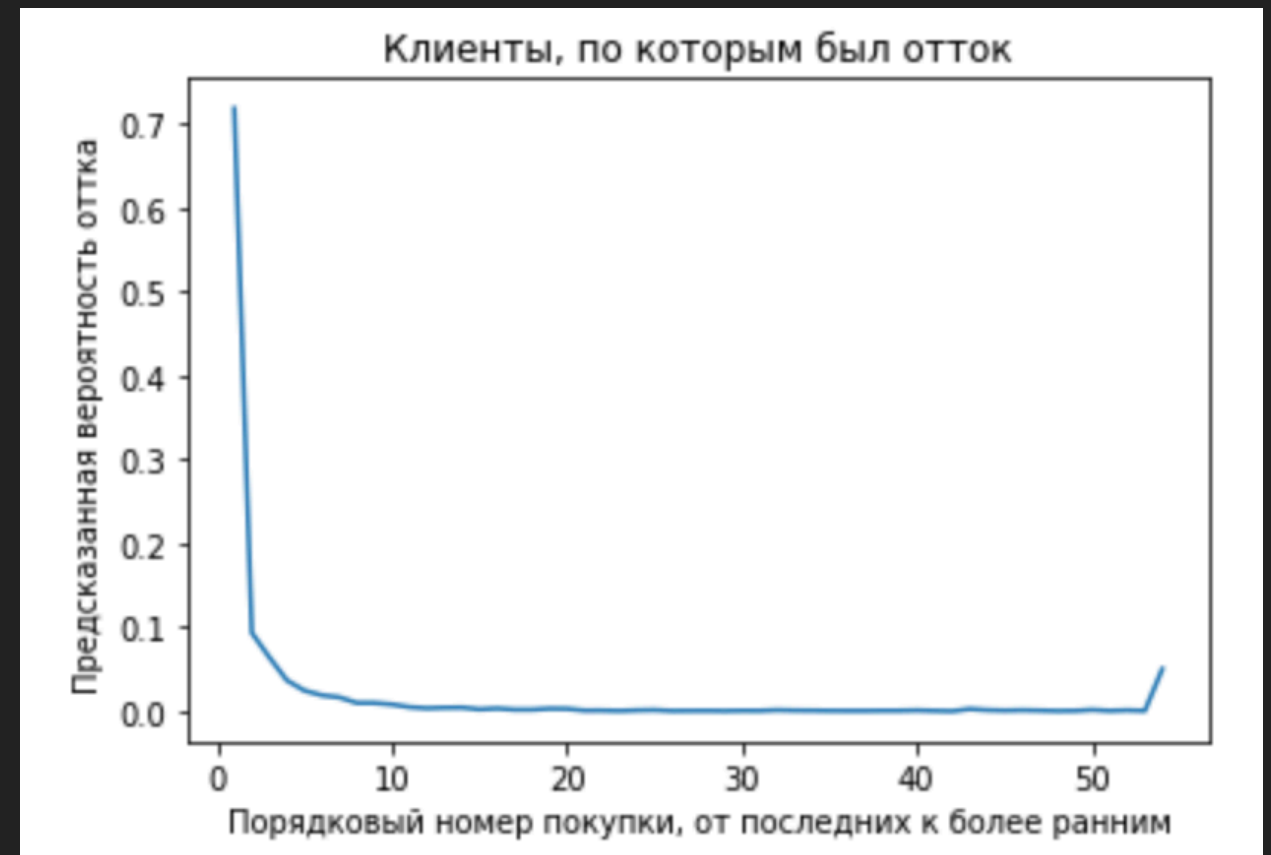
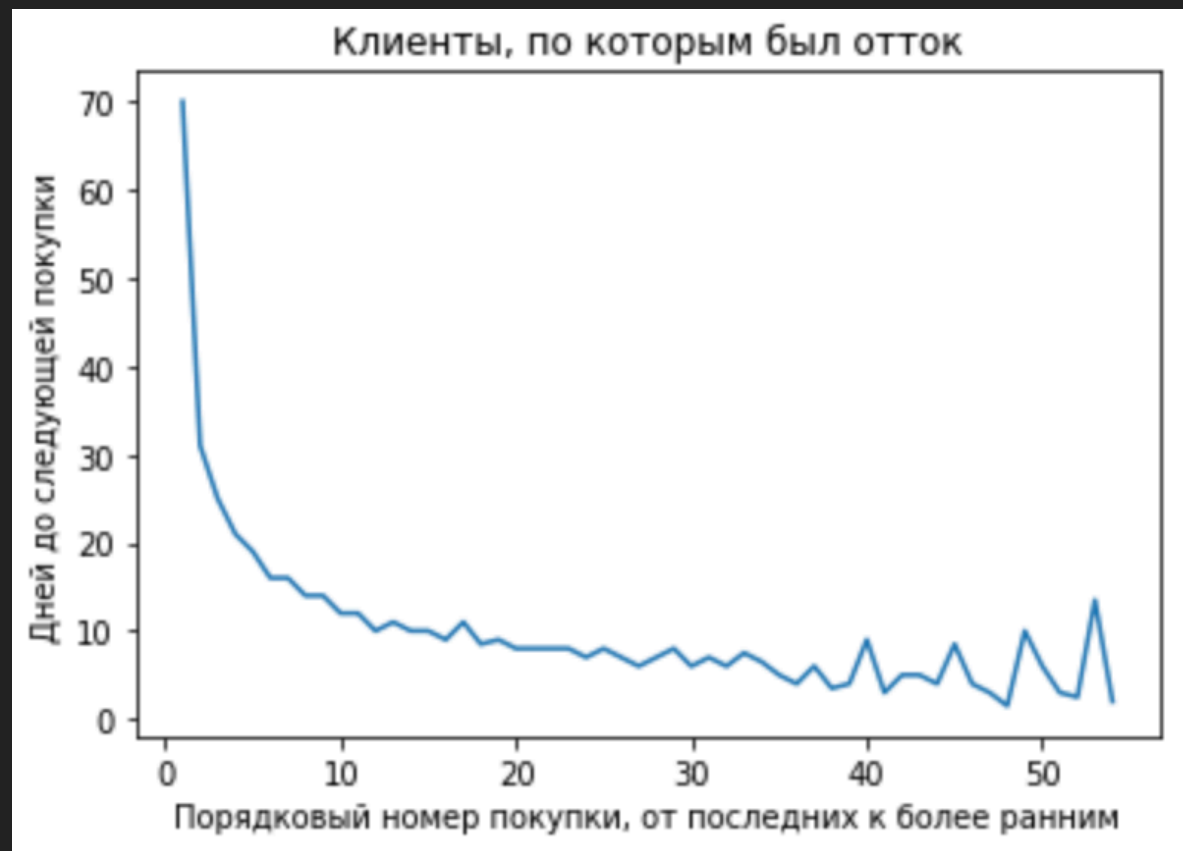
ВЫВОДЫ

	client_ID	future_interval	last_trans_number	y_predict	y_predict_proba	y_true
92	42921	85.0	3	1	0.105333	0
93	42921	12.0	2	0	0.001667	0
94	42921	43.0	1	1	0.793667	1
52	34008	44.0	5	0	0.001667	0
53	34008	110.0	4	0	0.044000	0
54	34008	2.0	3	0	0.004000	0
55	34008	34.0	2	1	0.132333	0
56	34008	89.0	1	1	0.835000	1
37	19509	1.0	4	0	0.009667	0
38	19509	55.0	3	1	0.094667	0
39	19509	63.0	2	1	0.319333	0
40	19509	193.0	1	1	0.798000	1

Ложные предсказания ?



ВЫВОДЫ



- ▶ На первом графике видно, что в среднем отток происходит не сразу, а ему предшествует тренд, выраженный в увеличении интервалов между покупками.
- ▶ На втором графике можно увидеть, что даже ложно предсказанная нами высокая вероятность оттока была в большинстве случаев "предвестником" реального оттока.