


Алгоритмы машинного обучения: как применять и рассчитать эффект

Арустамов Алексей

Loginom Company

Цифровая трансформация — это процесс интеграции цифровых технологий во все аспекты деятельности организаций:

- Бизнес-процессы
 - Продукты и услуги
 - Культура производства
 - Коммуникации
- 

Основные цели цифровой трансформации



Исследование Forbes Insights и Hitachi Data Systems , опрос 573 топ-менеджеров компаний из разных стран мира



Машинное обучение –
ключевой компонент
цифровой трансформации
для гибкого и быстрого
принятия оптимальных
решений с минимальным
участием людей.

Препятствия к цифровой трансформации

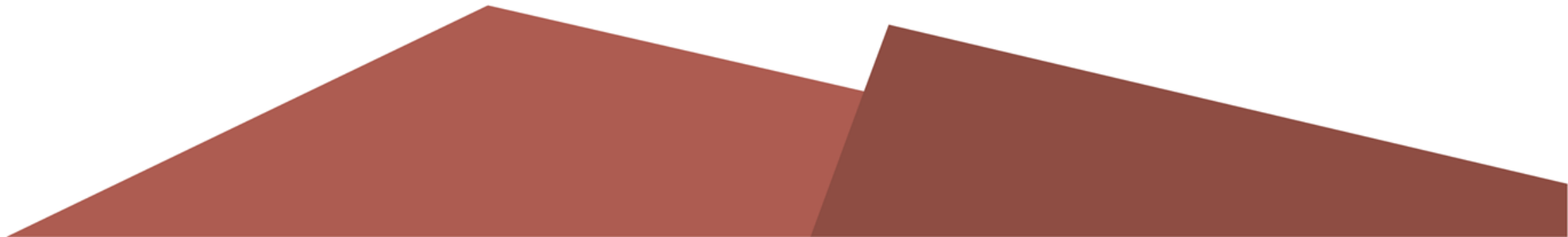


Исследование Forbes Insights и Hitachi Data Systems , опрос 573 топ-менеджеров компаний из разных стран мира



Возможность обосновать
эффективность в деньгах
значительно повышает
готовность реально
применять машинное
обучение на практике


Пример: оценка вероятности
оттока клиентов –
доказательство окупаемости

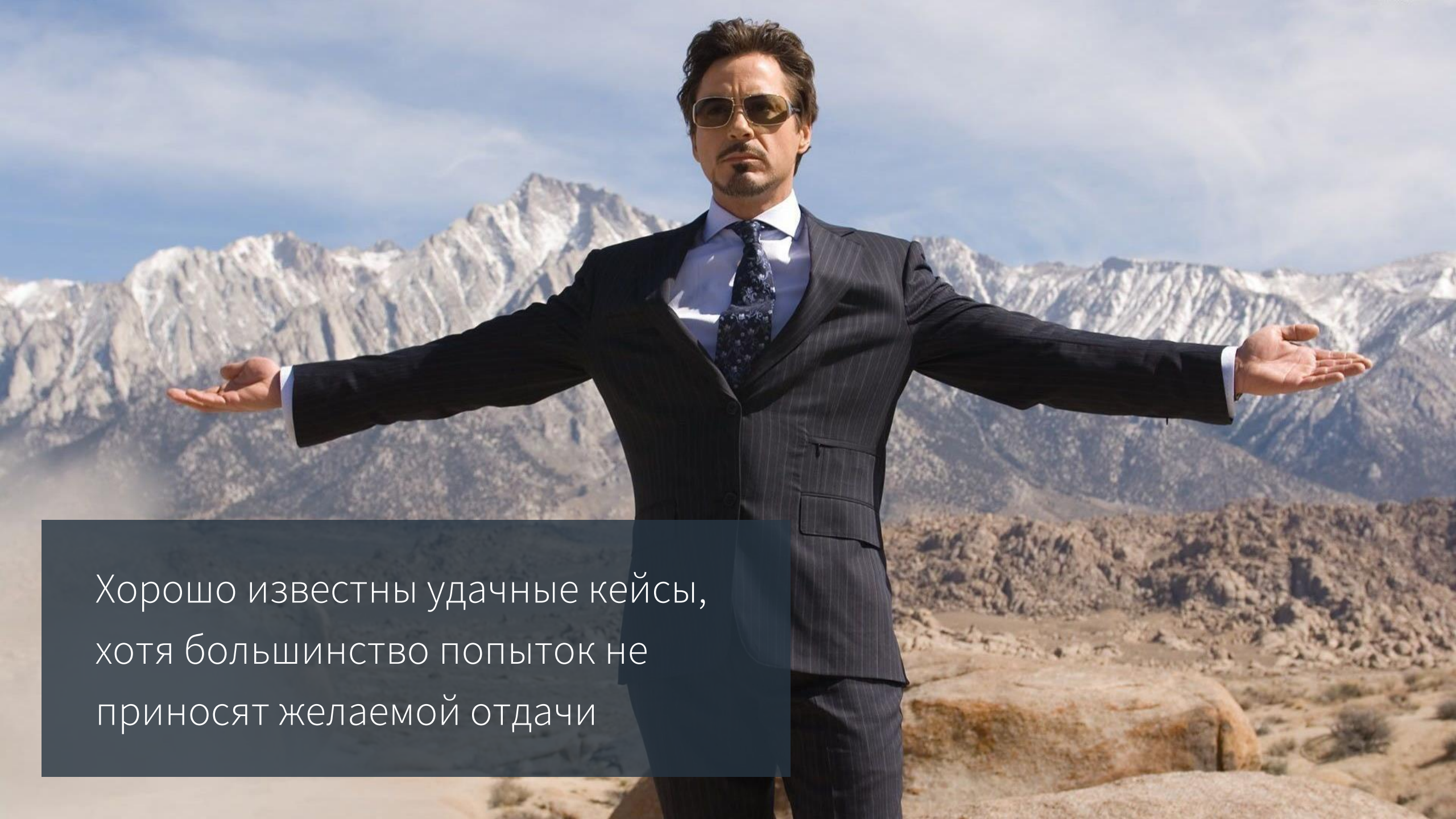


Вариант 1: Удачные кейсы

Логика: Компания «Х» благодаря машинному обучению снизила отток клиентов на 20% – таков эффект систем противодействия оттоку.

Проблема: Игнорирование систематической ошибки выжившего.






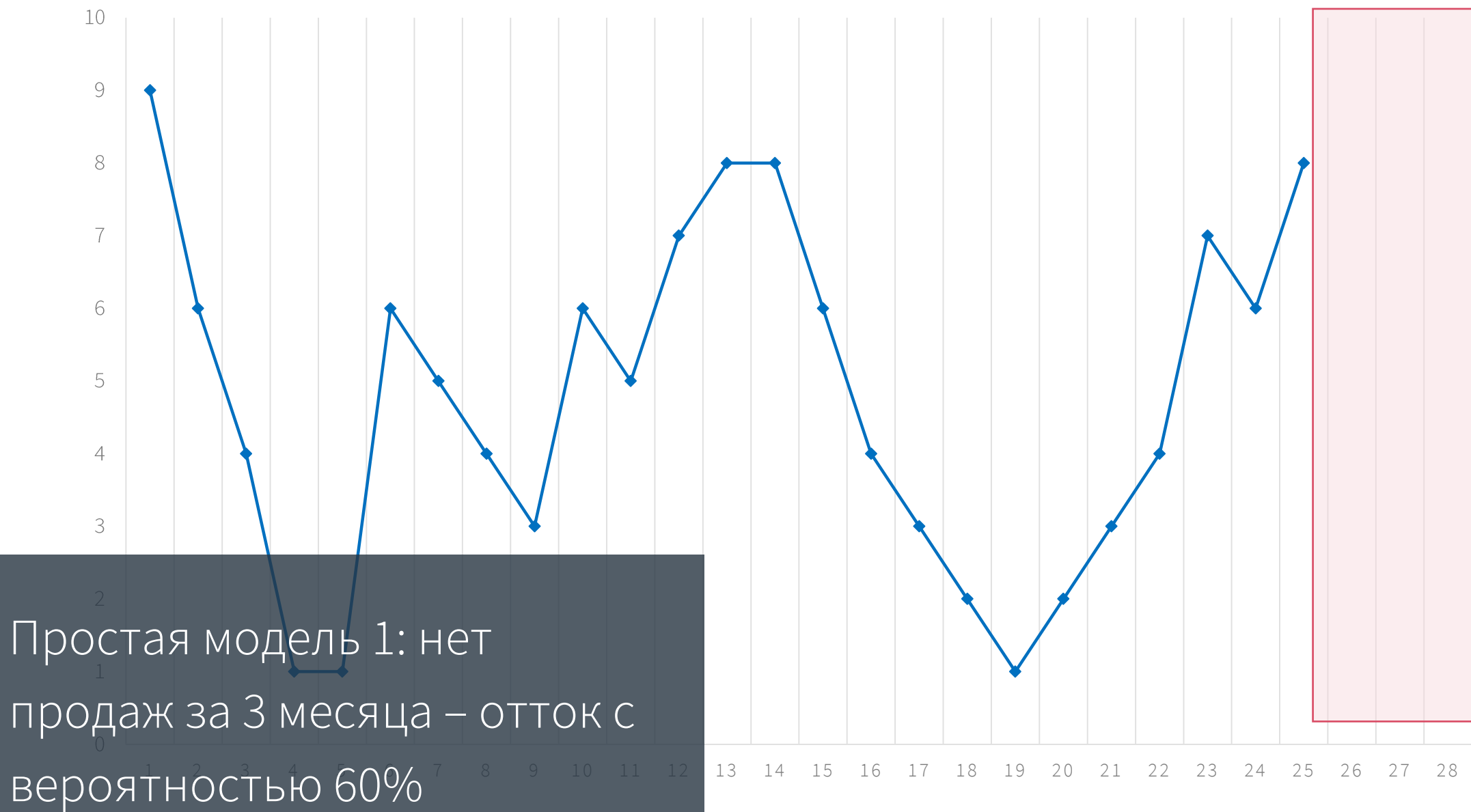
Хорошо известны удачные кейсы,
хотя большинство попыток не
приносят желаемой отдачи

Вариант 2: Сравнение с нулем

Логика: Машинное обучение обнаружило 80% склонных к оттоку, все доходы от удержанных – эффект от применения модели.

Проблема: Игнорирование простых методов предсказания оттока клиентов.







Отток клиентов по месяцам



Вариант 3: A/B тестирование

Логика: Создаем 2 группы одну оставляем без изменений, а для второй применяем ML модель, разница между ними – эффект от модели машинного обучения.

Проблема: нужен работающий процесс со встроенными инструментами A/B тестирования.

Вариант 4: Сравнение с тривиальной моделью

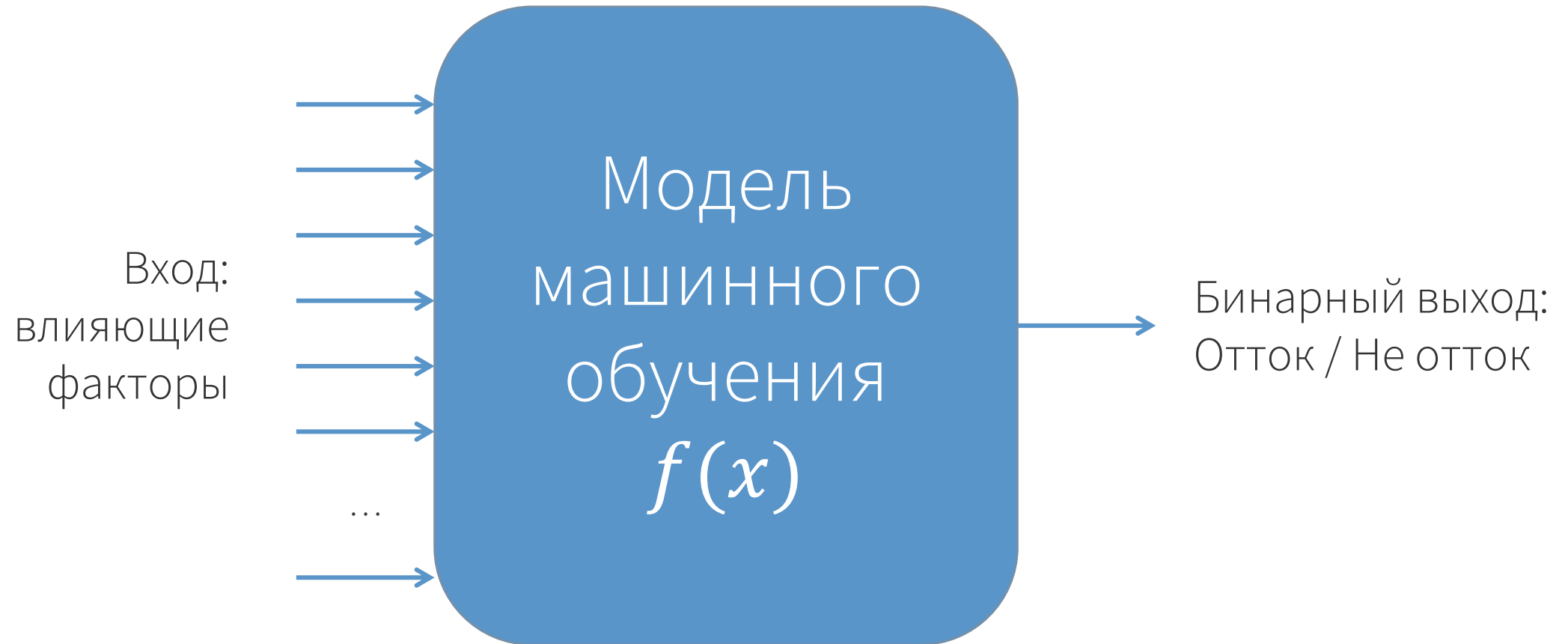
Логика: На одном наборе данных строим тривиальную модель и модель с машинным обучением, разница между ними – эффект от модели машинного обучения.

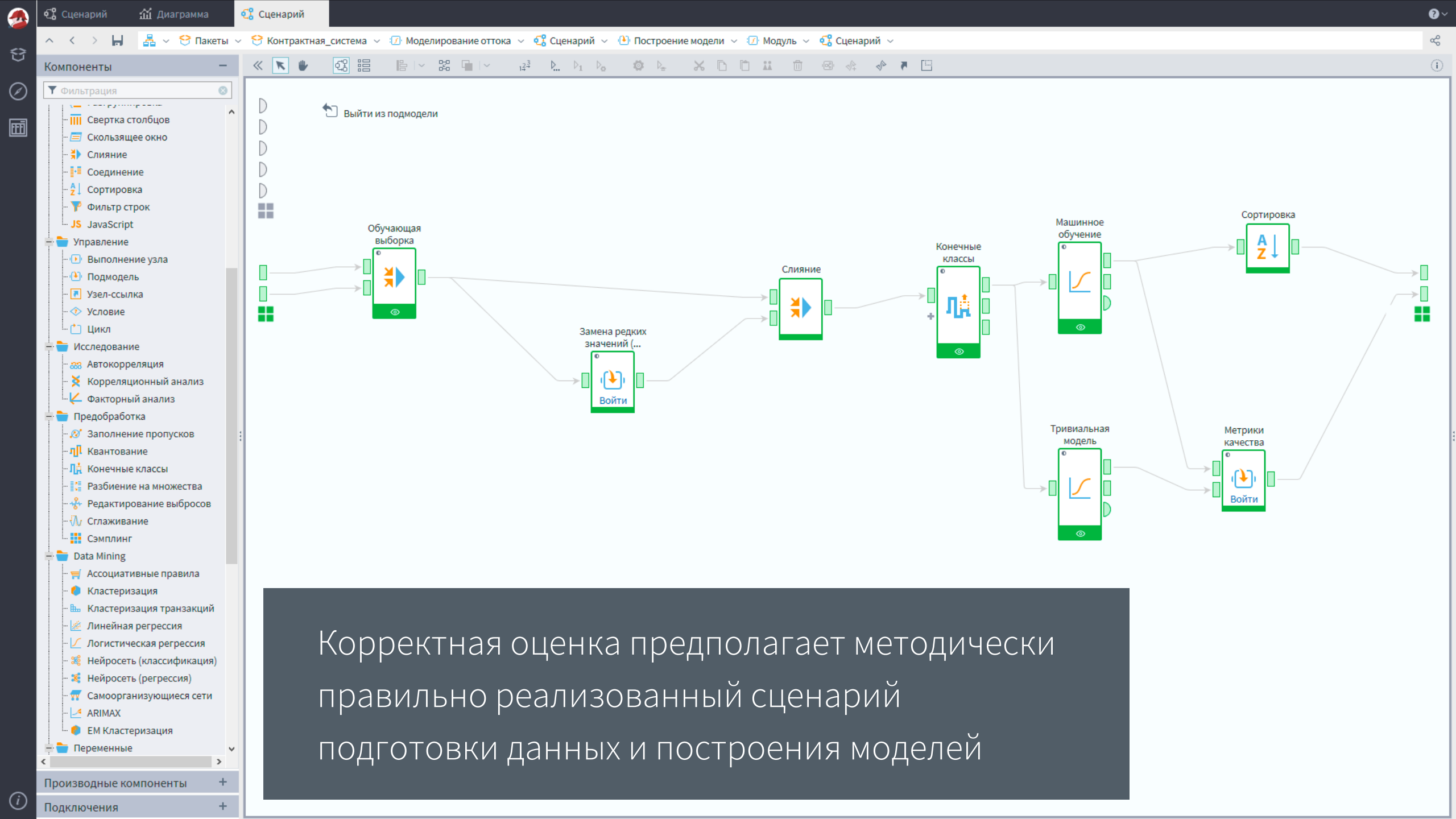
Проблема: эффект ML модели зависит от «тривиальности» простой модели.

Кейс: Сравнение с тривиальной моделью



Задача бинарной классификации





Корректная оценка предполагает методически правильно реализованный сценарий подготовки данных и построения моделей

Тривиальная модель

1. Простое бизнес-правило
2. Использование одного значимого фактора
3. Минимум внешней информации

VS

Сложная модель

1. Применение сложных алгоритмов
2. Использование множества факторов
3. Обогащение данных перед построением модели



demo

Сравнение ROC-кривых



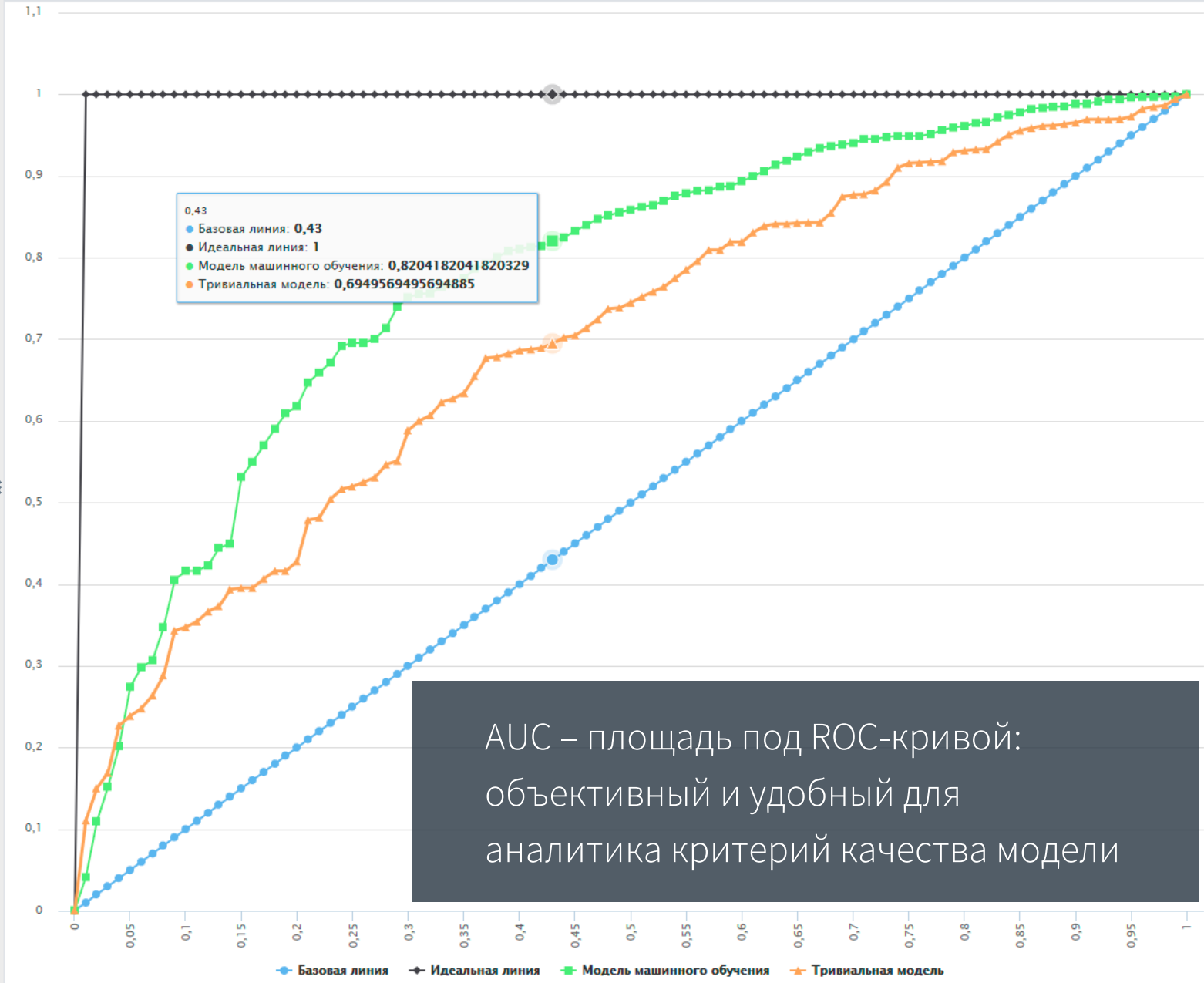
Пакеты lsm_demo Модуль1 Сценарий Сравнение скоринговых карт Модуль Сценарий Визуализация Визуализаторы Сравнение ROC-кривых

Линии TPR - Сумма

Модель оттока + Факты

FRP Базова... Идеал... Модел... Триви...

0	0,00	0,00	0,00	0,00
0,01	0,01	1,00	0,04	0,11
0,02	0,02	1,00	0,11	0,15
0,03	0,03	1,00	0,15	0,17
0,04	0,04	1,00	0,20	0,23
0,05	0,05	1,00	0,27	0,24
0,06	0,06	1,00	0,30	0,25
0,07	0,07	1,00	0,31	0,26
0,08	0,08	1,00	0,35	0,29
0,09	0,09	1,00	0,41	0,34
0,1	0,10	1,00	0,42	0,35
0,11	0,11	1,00	0,42	0,35
0,12	0,12	1,00	0,42	0,37
0,13	0,13	1,00	0,44	0,37
0,14	0,14	1,00	0,45	0,39
0,15	0,15	1,00	0,53	0,40
0,16	0,16	1,00	0,55	0,40
0,17	0,17	1,00	0,57	0,41
0,18	0,18	1,00	0,59	0,42
0,19	0,19	1,00	0,61	0,42
0,2	0,20	1,00	0,62	0,43
0,21	0,21	1,00	0,65	0,48
0,22	0,22	1,00	0,66	0,48
0,23	0,23	1,00	0,67	0,50
0,24	0,24	1,00	0,69	0,52
0,25	0,25	1,00	0,70	0,52
0,26	0,26	1,00	0,70	0,53
0,27	0,27	1,00	0,70	0,53
0,28	0,28	1,00	0,71	0,55
0,29	0,29	1,00	0,74	0,55
0,3	0,30	1,00	0,75	0,59



Площадь под ROC-кривой

Достоинство: стандартный и общепринятый в машинном обучении критерий сравнения нескольких моделей.

Недостаток: нет простого способа оценки разницы моделей в деньгах.

Ошибка первого и второго рода

Тривиальная модель

	Факт		
Модель	Отток	Не отток	Итого
Отток	6 512 (63.9%)	856 (1.1%)	7 368
Не отток	3 667 (36.1%)	78 903 (98.9%)	82 570
Итого	10 179	79 759	89 938

Сложная модель

	Факт		
Модель	Отток	Не отток	Итого
Отток	7 025 (69.0%)	343 (0.4%)	7 368
Не отток	3 154 (31.0%)	79 416 (99.6%)	82 570
Итого	10 179	79 759	89 938

Ошибка первого и второго рода

Достоинство: стандартный критерий сравнения нескольких моделей, позволяющий учесть цену ошибки.

Недостаток: сложно цену ошибки напрямую свести к финансовым потерям.

Процент распознанных событий

После построения модели для каждого клиента рассчитывается вероятность оттока, записи сортируются по убыванию вероятности и строится кумулятивный график угаданных моделью случаев оттока.



Выбор диаграммы

- ☐ ROC-кривая
- ☐ PR-кривая
- ☐ Базовые показатели
- ☐ Диаграмма точности
- ☐ Диаграмма равновесия
- ☒ % распознанных событий
- ☐ Диаграмма роста
- ☐ Диаграмма отклика
- ☐ Диаграмма выигрыша

☒ Кумулятивная

10 диапазонов

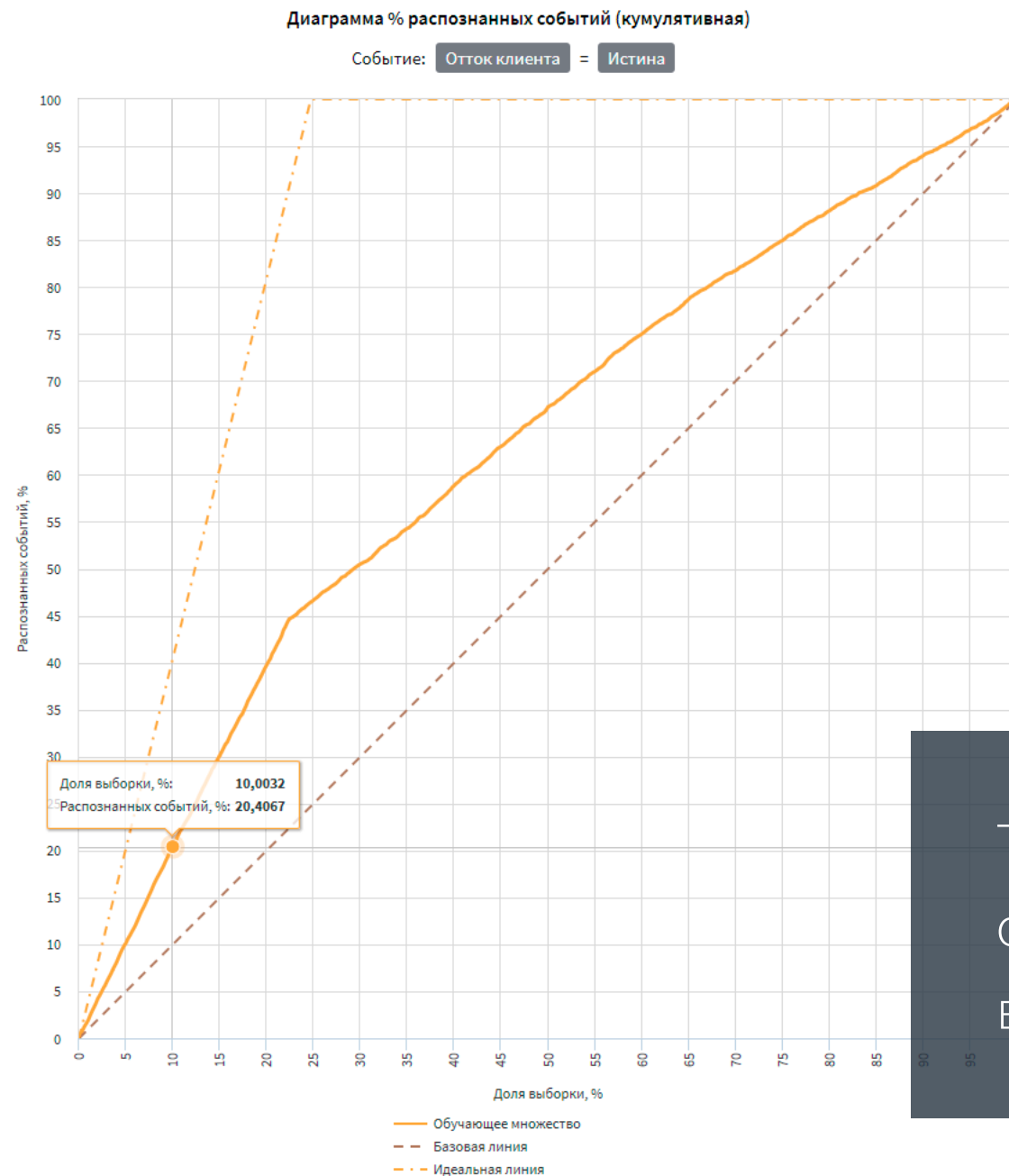
Множества

☒ Обучающее ☒ Тестовое

Порог отсечения

Из настроек узла

Значение порога:



Оценки классификации

Показатель	Множества	
	Обучающее	Тестовое
Оценки классификатора		
AUC ROC	0,6662	
AUC PR	0,4132	
Коэффициент Джини	0,3324	
KS	29,5452	
Порог отсечения: Из настроек узла		
Значение	0,5000	
TPR (Чувствительность)	0,0000	
TNR (Специфичность)	1,0000	
FPR (1-Специфичность)	0,0000	
PPV	0,4929	
F1 Score	0,0000	
MCC	0,0000	

Матрицы ошибок

Классифицировано	Фактически		Итого
	Событие	Не-событие	
Обучающее	5 410	16 483	
Событие	0	0	0
Не-событие	5 410	16 483	21 893
Тестовое			
Событие			
Не-событие			

Распознано

Обучающее	16 483/21 893
-----------	---------------

Тривиальная модель, учет
одного фактора: 10%
выборки = 20% оттока



Выбор диаграммы

- ☐ ROC-кривая
- ☐ PR-кривая
- ☐ Базовые показатели
- ☐ Диаграмма точности
- ☐ Диаграмма равновесия
- ☒ % распознанных событий
- ☐ Диаграмма роста
- ☐ Диаграмма отклика
- ☐ Диаграмма выигрыша

☒ Кумулятивная

10 диапазонов

Множества

☒ Обучающее ☒ Тестовое

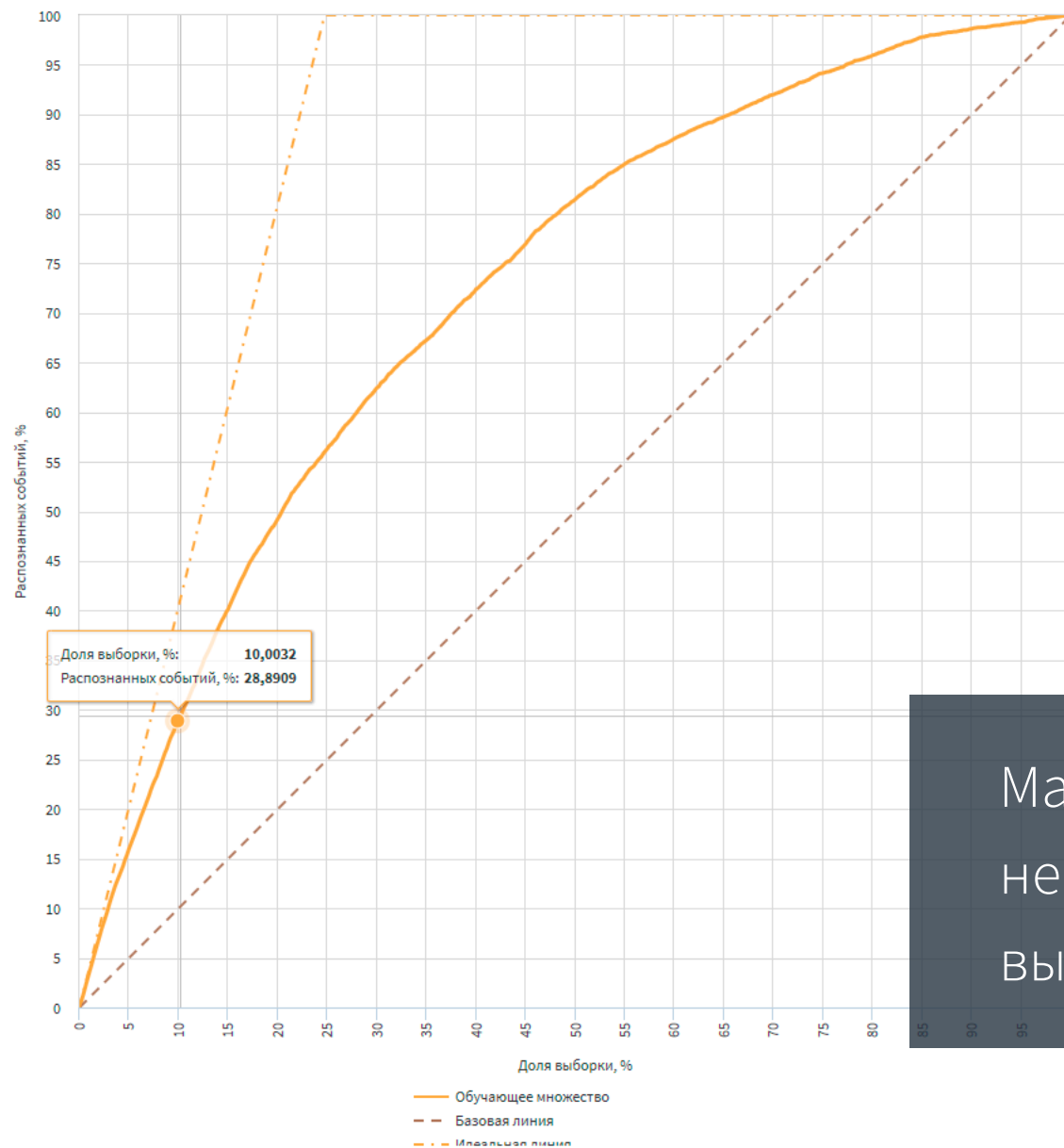
Порог отсечения

Баланс (TPR = TNR)

Значение порога:

Диаграмма % распознанных событий (кумулятивная)

Событие: Отток клиента = Истина



Оценки классификации

Показатель	Множества	
	Обучающее	Тестовое
Оценки классификатора		
AUC ROC	0,7943	
AUC PR	0,5905	
Коэффициент Джини	0,5887	
KS	43,3823	
Порог отсеивания: Баланс (TPR = TNR)		
Значение	0,2464	
TPR (Чувствительность)	0,7146	
TNR (Специфичность)	0,7146	
FPR (1-Специфичность)	0,2854	
PPV	0,4511	
F1 Score	0,5531	
MCC	0,3793	

Матрицы ошибок

Классифицировано	Фактически		Итого
	Событие	Не-событие	
Обучающее	5 410	16 483	
Событие	3 866	4 704	8 570
Не-событие	1 544	11 779	13 323
Тестовое			
Событие			
Не-событие			

Распознано

Обучающее	15 645/21 893
-----------	---------------

Машинное обучение, учет
нескольких факторов : 10%
выборки = 28% оттока

Показатель	Значение
Активных клиентов	21 893
Средний отток	960
Процент оттока	4.38%
Средний доход на год (руб.)	50 000
Потери от оттока	48 000 000
Стоимость удержания одного клиента	1 000
На какой процент клиентов будем воздействовать	10.00%
Процент удержания после воздействий	50.00%
Тривиальная модель	
Распознали склонных к оттоку	20.40%
На сколько клиентов надо воздействовать	2 190
Стоимость мероприятий по удержанию	2 190 000
Сколько удержим (в клиентах)	98
Сколько удержим (в рублях)	4 896 000
Выгода от удержания	2 706 000
Машинное обучение	
Распознали склонных к оттоку	28.90%
На сколько клиентов надо воздействовать	2 190
Стоимость мероприятий по удержанию	2 190 000
Сколько удержим (в клиентах)	139
Сколько удержим (в рублях)	6 936 000
Выгода от удержания	4 746 000
Выгода от ML модели	2 040 000

Процент распознанных событий

Достоинство: можно свести к понятной бизнесу методике оценке эффекта до начала применения модели.

Недостаток: требуются определенные допущения для расчета эффекта.

Описанные методики позволяют, хоть и с некоторыми допущениями, доказать окупаемость инвестиций в машинное обучение повышают вероятность их реального применения для решения практических задач.



loginom.ru

