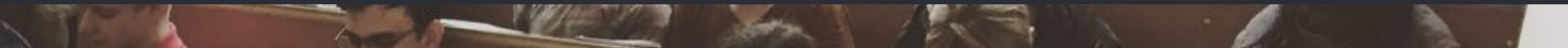
Data Mining in Action

Лекция 8. Практические рекомендации по валидации качества и работа с признаками



На прошлой лекции про оценку качества

- MAE
- RMSE
- MAPE
- SMAPE
- logloss

- Accuracy
- Precision
- Recall
- F-measure
- ROC-AUC

Немного мотивации: топ ошибок в индустрии

1. Постановка задачи отсутствует или неправильная (например, метрику вообще выбрали случайно)
2. A/B тест не проводится или не валиден
3. Утечка и переобучение

Субъективный топ причин

1. Безответственность: «и так сойдет»
2. Невнимательность, особенно в период «авралов»
3. Нехватка экспертизы: незнание, что вопросы, которые мы обсудим на этой лекции, существуют и важны

План

1. Пример выбора метрики
2. Анализ качества модели
3. Онлайн-качество
4. Извлечение признаков
5. Отбор признаков

1. Выбор метрики (пример: рекомендации)

Что можем делать

- Прогнозировать, какие товары будут куплены
- Максимизировать прибыль

Остается вопрос: какие прогнозы нужны и как их использовать, чтобы денег стало больше?

Максимизация количества покупок

Товар 1	Товар 2	Товар 3	Товар 4

Максимизация количества покупок

Товар 1	Товар 2	Товар 3	Товар 4
Вероятность:	p_1	p_2	p_3	p_4

Максимизация дохода

Товар 1	Товар 2	Товар 3	Товар 4

Вероятность:	p_1	p_2	p_3	p_4
Цена:	c_1	c_2	c_3	c_4

Максимизация дохода

Puma
Ветровка
3 490 руб.

Crocs
Сланцы
1 990 руб.

Tony-p
Слипоны
~~1 999 руб.~~ 1 590 руб.

Champion
Брюки спортивные
~~3 599 руб.~~ 1 970 руб.

Вероятность:	0.05	0.02	0.015	0.009
Цена:	3490	1990	1590	1970

Максимизация прибыли

Puma
Ветровка
3 490 руб.

Crocs
Сланцы
1 990 руб.

Tony-p
Слипоны
~~1 999 руб.~~ 1 590 руб.

Champion
Брюки спортивные
~~3 599 руб.~~ 1 970 руб.

Вероятность:	0.05	0.02	0.015	0.009
Цена:	3490	1990	1590	1970
Маржинальность	0.1	0.4	0.4	0.2

Мини-задача

Как изменится построение модели, если нам нужно максимизировать количество просмотренных пользователем товаров?

Точность (Precision@k)

Рекомендованные товары
Синяя футболка
Красная футболка
Кроссовки
Кепка
Зеленая футболка

Купленные товары
Красная футболка
Кеды
Кепка

k – количество рекомендаций

$$\text{Precision@k} = \frac{\text{купленное из рекомендованного}}{k}$$

AveragePrecision@k - усредненный по сессиям Precision@k

Полнота (Recall@k)

Рекомендованные товары
Синяя футболка
Красная футболка
Кроссовки
Кепка
Зеленая футболка

Купленные товары
Красная футболка
Кеды
Кепка

k – количество рекомендаций

$$\text{Recall@k} = \frac{\text{купленное из рекомендованного}}{\text{количество покупок}}$$

AverageRecall@k - усредненный по сессиям Recall@k

Взвешенный ценами recall@k

Рекомендованные товары
Синяя футболка – 1000р
Красная футболка – 1200р
Кроссовки – 3500р
Кепка – 900р
Зеленая футболка – 800р

Купленные товары
Красная футболка – 1200р
Кеды – 3000р
Кепка – 900р

$$\text{Взвешенный ценами Recall@k} = \frac{\text{стоимость купленного из рекомендованного}}{\text{стоимость покупок}}$$

AverageRecall@k - усредненный по сессиям Recall@k

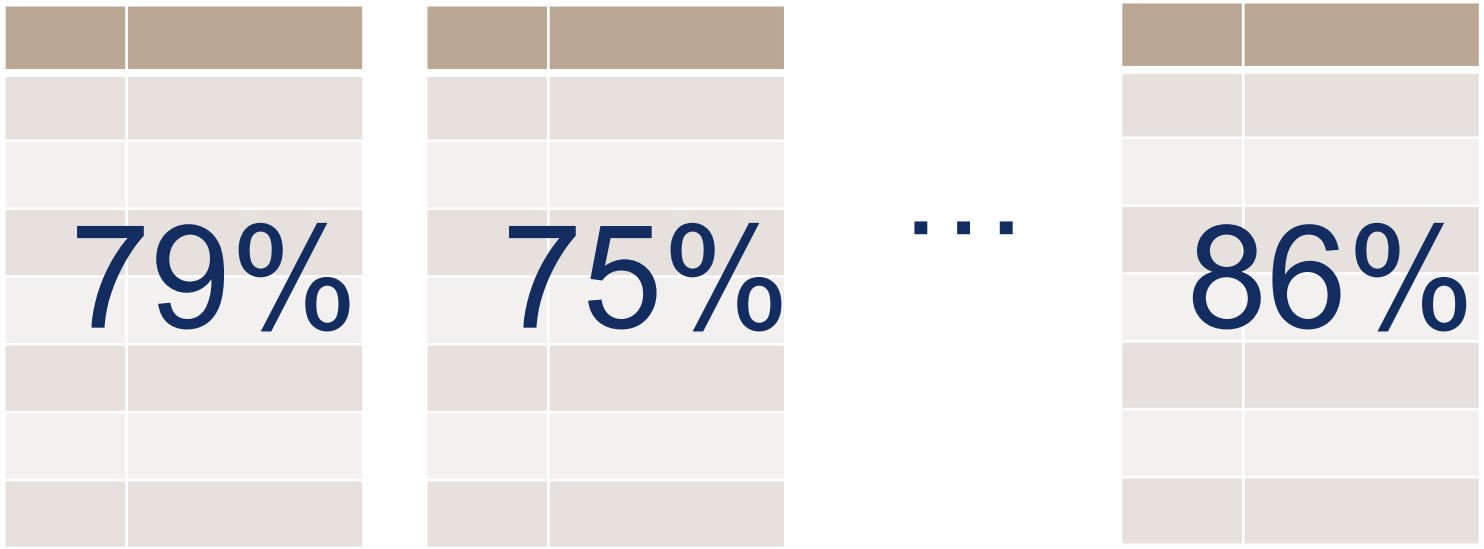
Качество классификации против качества рекомендаций

Пример – 2 решения для прогноза купит/не купит товар:

	Алгоритм 1	Алгоритм 2
AUC классификатора	0.52	0.85
Recall@5	0.72	0.71

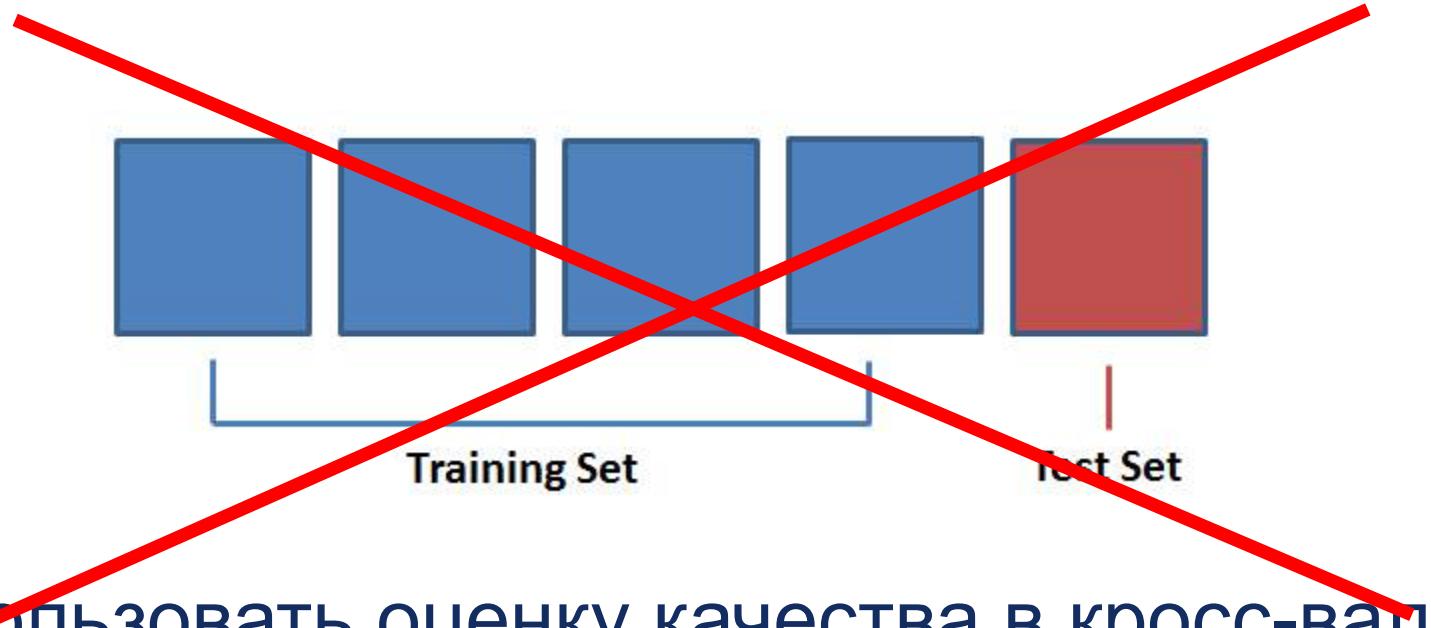
2. Анализ работы модели на исторических данных

Проблема: разброс на разных данных



Усреднение качества в CV

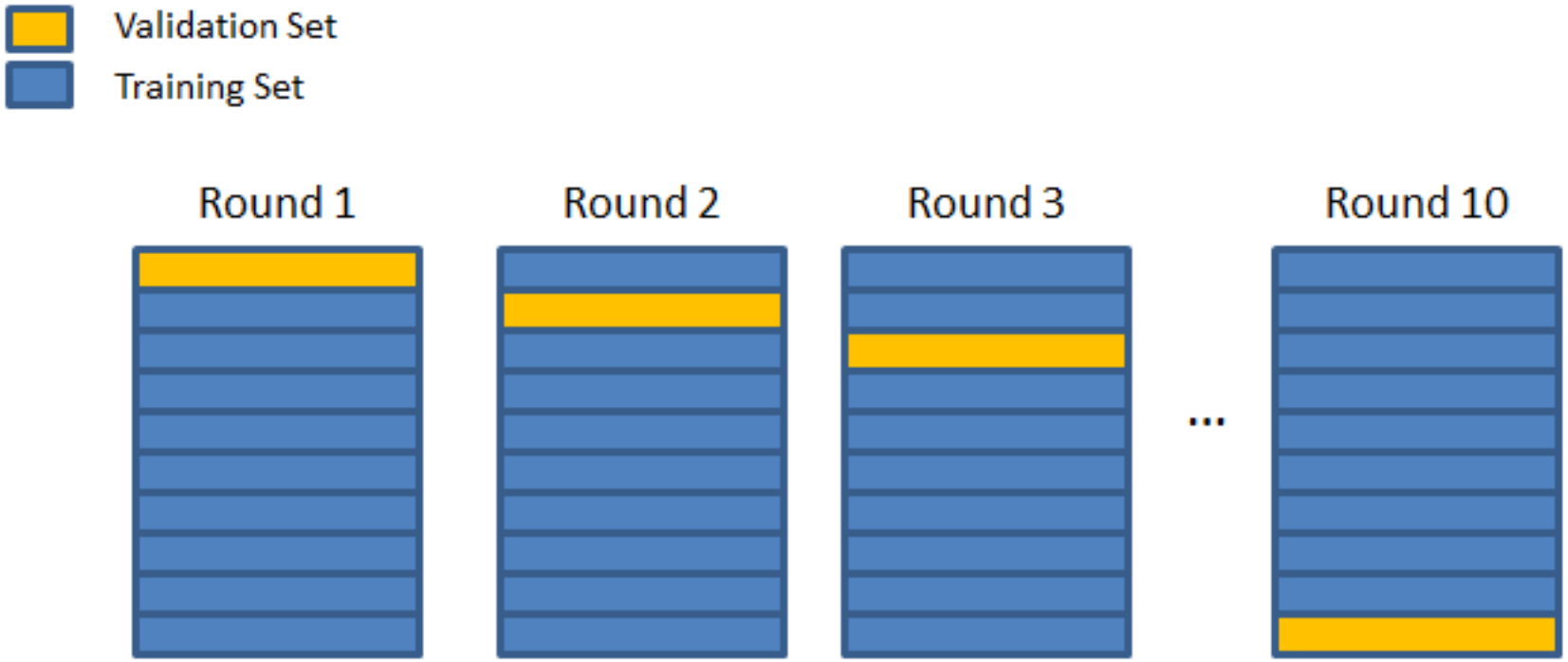
Если есть проблема со стабильностью модели, точно нужно избегать оценок на одном фиксированном датасете



Нужно использовать оценку качества в кросс-валидации

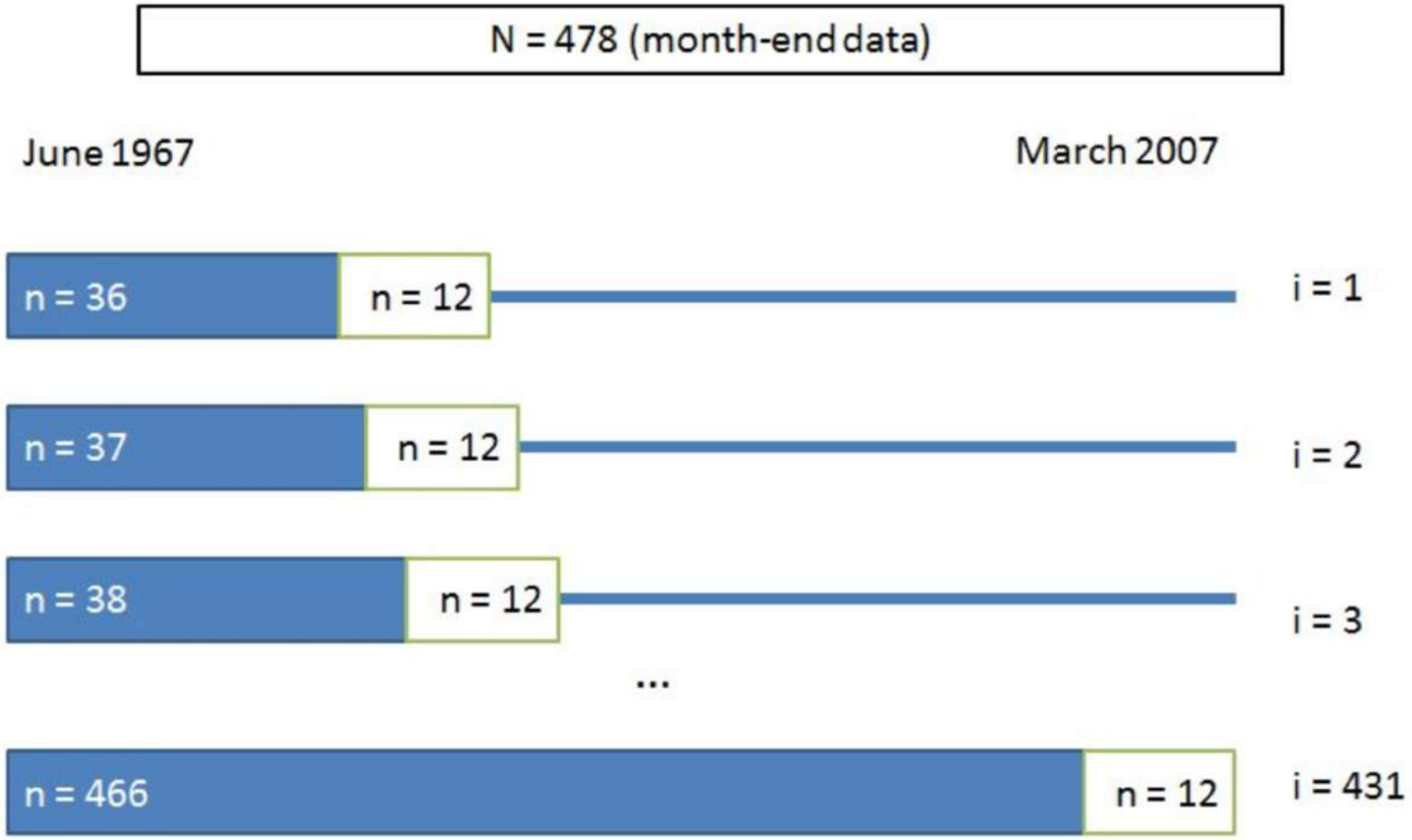
Кросс-валидация

K-Fold cross validation:

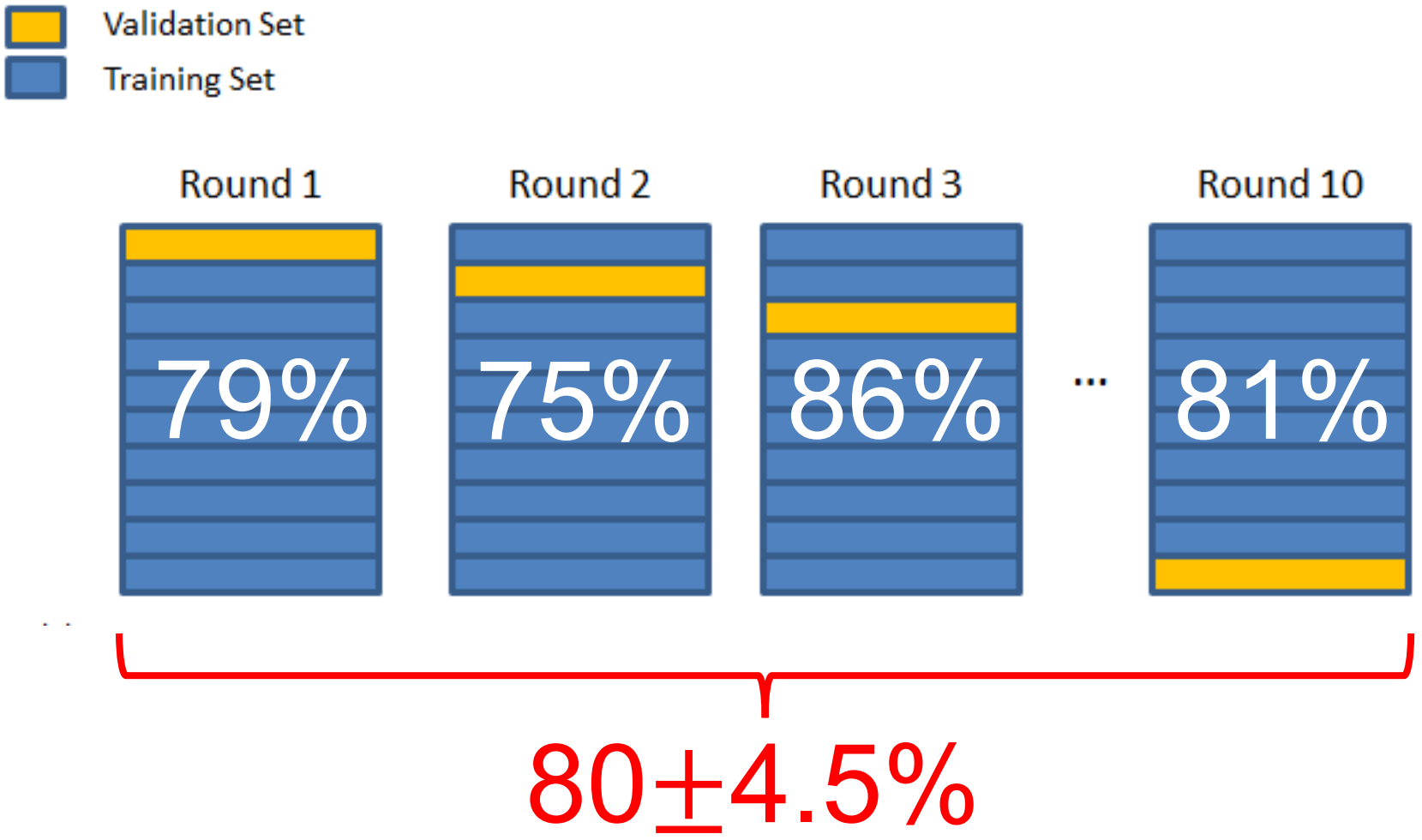


На ка

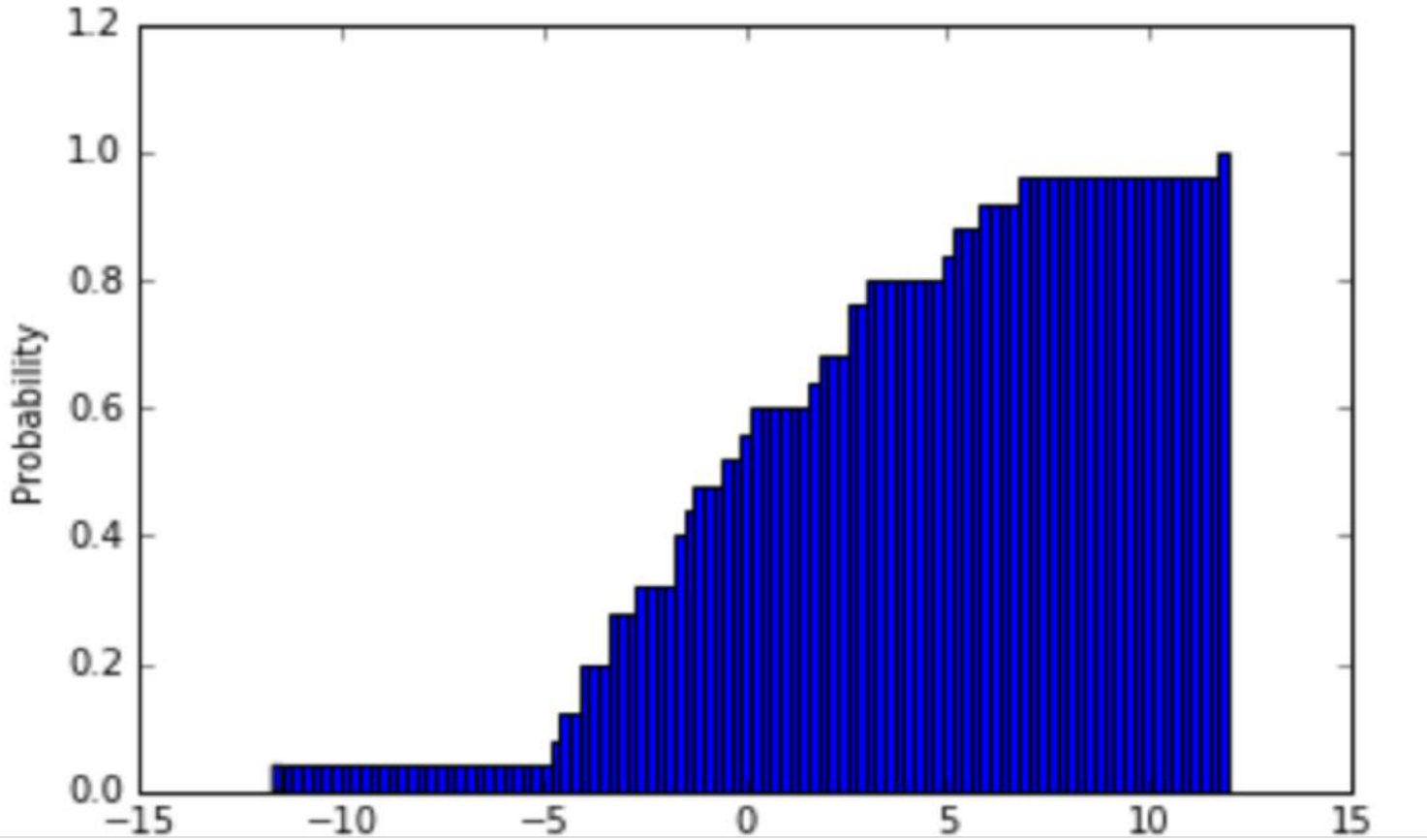
Предупреждение: будьте осторожны с CV



Учет разброса и распределения в CV



Учет разброса и распределения в CV



Анализ топа важных признаков

На одном фолде:

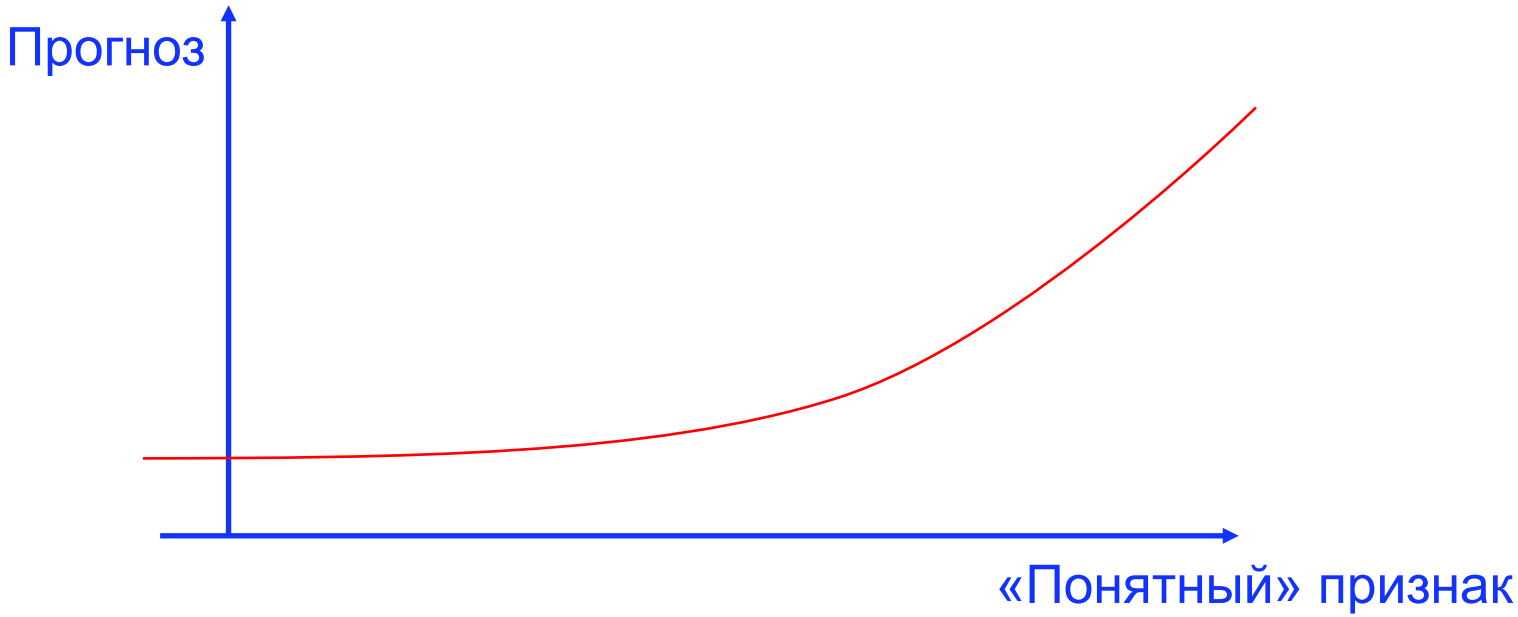
0.211268 Номер
0.147105 Ширина
0.128326 Вес
0.0954617 Параметр 1
0.0688576 Высота
0.057903 Параметр 2
0.0438185 Параметр 3
…

На другом:

0.285714 Номер
0.163265 Параметр 1
0.122449 Высота
0.102041 Параметр 4
0.0816327 Параметр 5
0.0816327 Вес
0.0612245 Параметр 2
…

Анализ зависимости от признаков

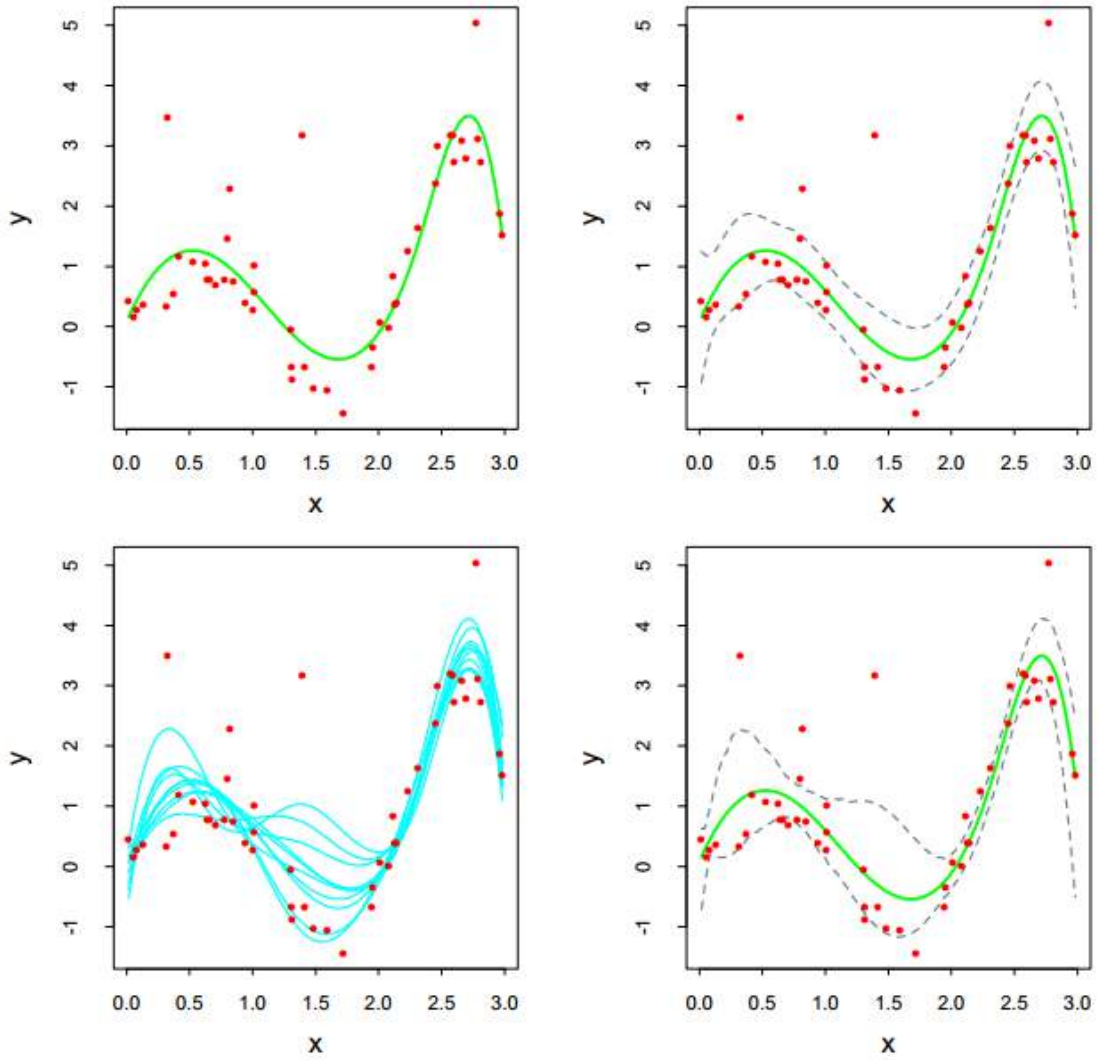
Если зависимость от каких-то признаков должна иметь понятный вид, можем поменять их (построить «искусственные» примеры) и посмотреть, как ведет себя прогноз



Уменьшение разброса

- Вариант 1: поиск допущенных ошибок
- Вариант 2: более устойчивые модели

Bagging



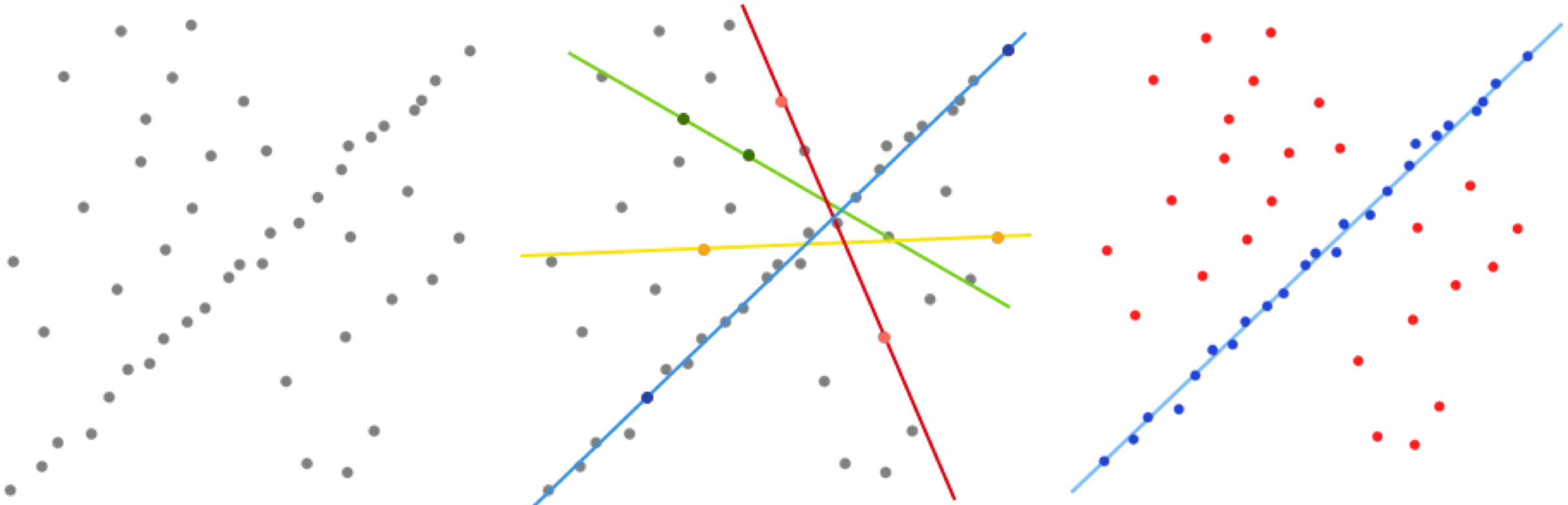
Бэггинг в sklearn.ensembles

- BaggingRegressor
- BaggingClassifier

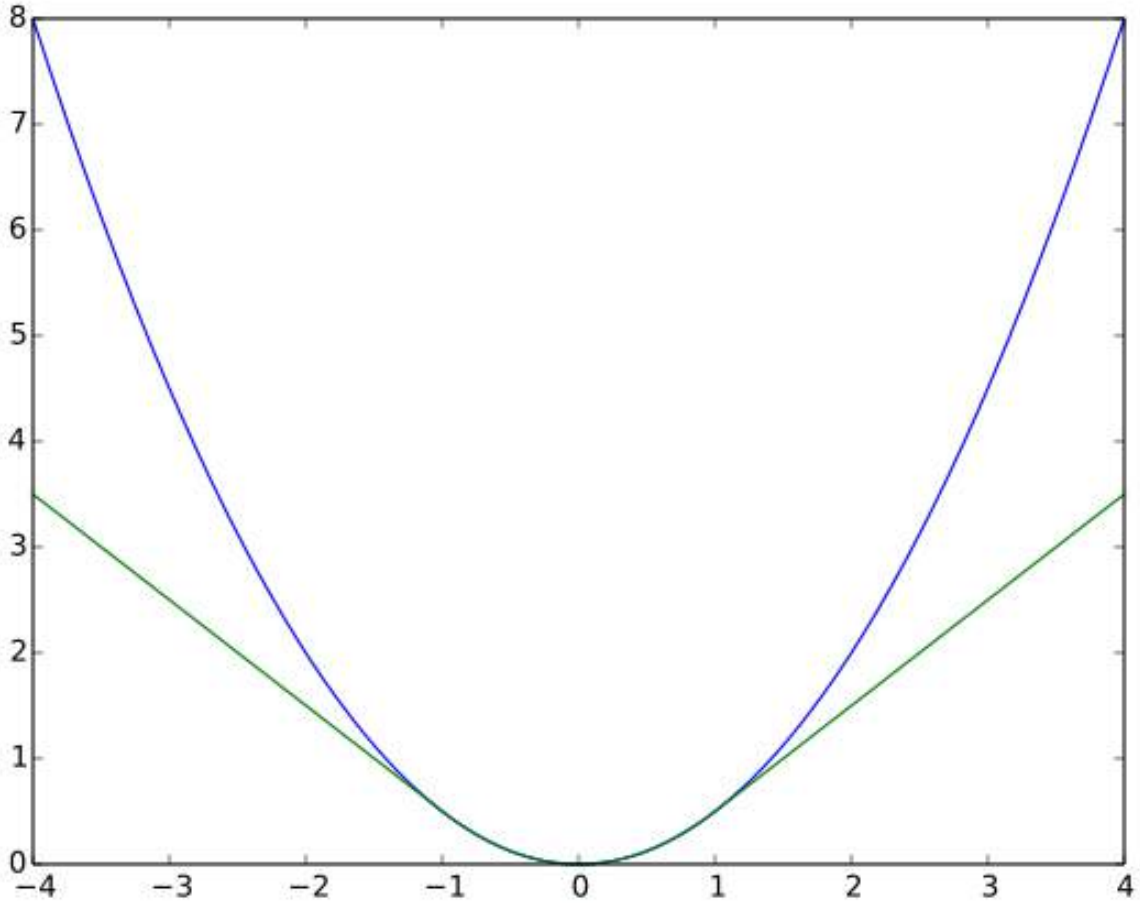
Робастные модели в sklearn.linear_model

- RANSACRegressor
- HuberRegressor
- Theil-Sen Regressor

RANSACRegressor



HuberRegressor



3. Онлайн эксперимент

Проблема

- Пока мы обсуждали качество на исторических данных
- Будет ли качество работы внедренной модели тем же?

Проблема

- Пока мы обсуждали качество на исторических данных
- Будет ли качество работы внедренной модели тем же?

Как правило, нет

A/B тестирование

Как измерить эффект от внедрения модели:

1. Разделить примеры, на которых применяем (например, пользователей) на две группы.
2. В одной группе использовать модель, в другой – нет
3. В конце измерить целевой показатель (продажи/конверсию/клики/что-то еще)

О чем поговорим сейчас

- Почему в продакшене качество бывает другим
- Как уменьшают это различие
- Как избежать ложных выводов из замеров качества в онлайне

Причины различия качества в оффлайне и онлайне

1. Переобучились, как следствие - на новых данных работаем плохо

Причины различия качества в оффлайне и онлайне

1. Переобучились, как следствие - на новых данных работаем плохо
2. Обучались не совсем на тех данных, на которых применяем (частный случай предыдущего)

Причины различия качества в оффлайне и онлайне

1. Переобучились, как следствие - на новых данных работаем плохо
2. Обучались не совсем на тех данных, на которых применяем (частный случай предыдущего)
3. В данных есть «утечка» (leak)

Причины различия качества в оффлайне и онлайне

1. Переобучились, как следствие - на новых данных работаем плохо
2. Обучались не совсем на тех данных, на которых применяем (частный случай предыдущего)
3. В данных есть «утечка» (leak)
4. Просто так «нарандомило»

Пример: есть ли приложение конкурентов

- Обучили модель на пользователях Android
- Надо применять для пользователей iOS

Пример: есть ли приложение конкурентов

- Обучили модель на пользователях Android
- Надо применять для пользователей iOS

Решение:

- Обучили на тех же признаках модель, определяющую Android или iOS у пользователя
- Те признаки, что в ней получились важными – не используем

Пример утечки 1

Задача:

Прогнозируем количество продаж в магазине на следующей неделе по данным предыдущих недель

Утечка (leak):

В признаки случайно добавили продажи и на той неделе, для которой прогнозируем (например, в продажах за последний месяц)

Пример утечки 2

Задача:

Прогнозируем по посещаемым человеком сайтам, наймут ли его в компанию

Утечка:

Профили пользователей взяты свежие, а не за тот день, когда кандидата из обучающей выборки еще не взяли в компанию и он еще не ходил на внутренние ресурсы

Онлайновая оценка качества

Как понять, какое качество в продакшене?

Онлайновая оценка качества

Как понять, какое качество в продакшене?

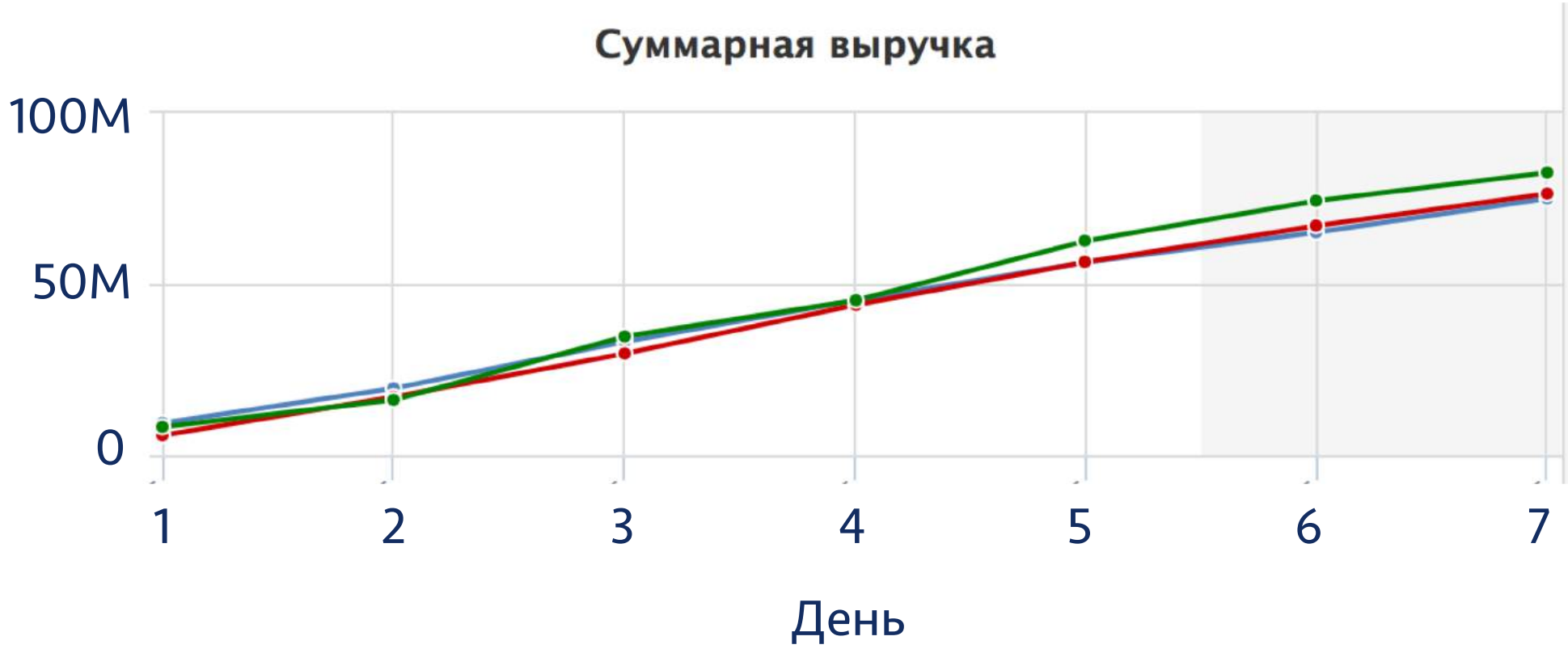
Идеи:

1. A/B тест
2. Оценка статзначимости результата

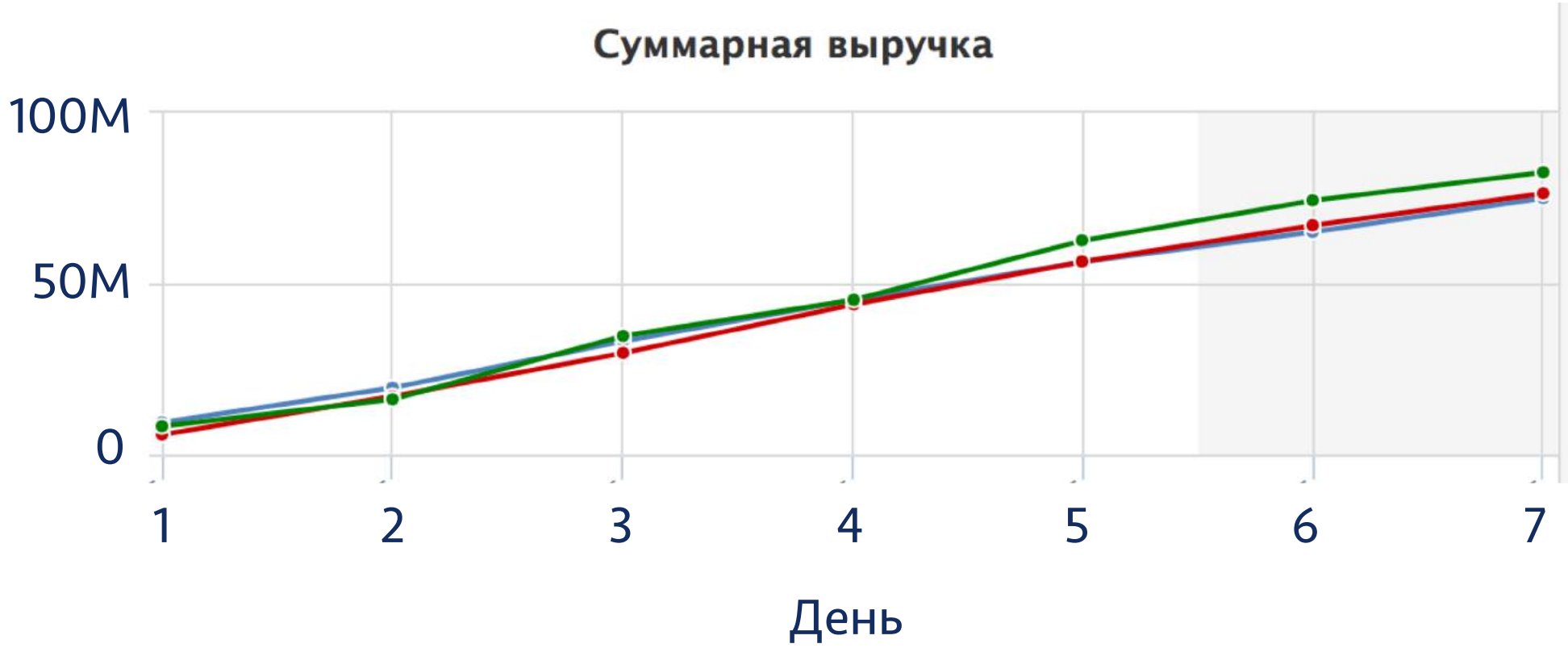
A/B тест

1. Случайным образом делим пользователей на равные группы
2. Измеряем целевые метрики (например, конверсию, количество заказов или доход) в каждой группе за длительный период времени
3. Получаем какое-то число для каждой группы
4. Что дальше?

Пример: A/B тесты и статистика

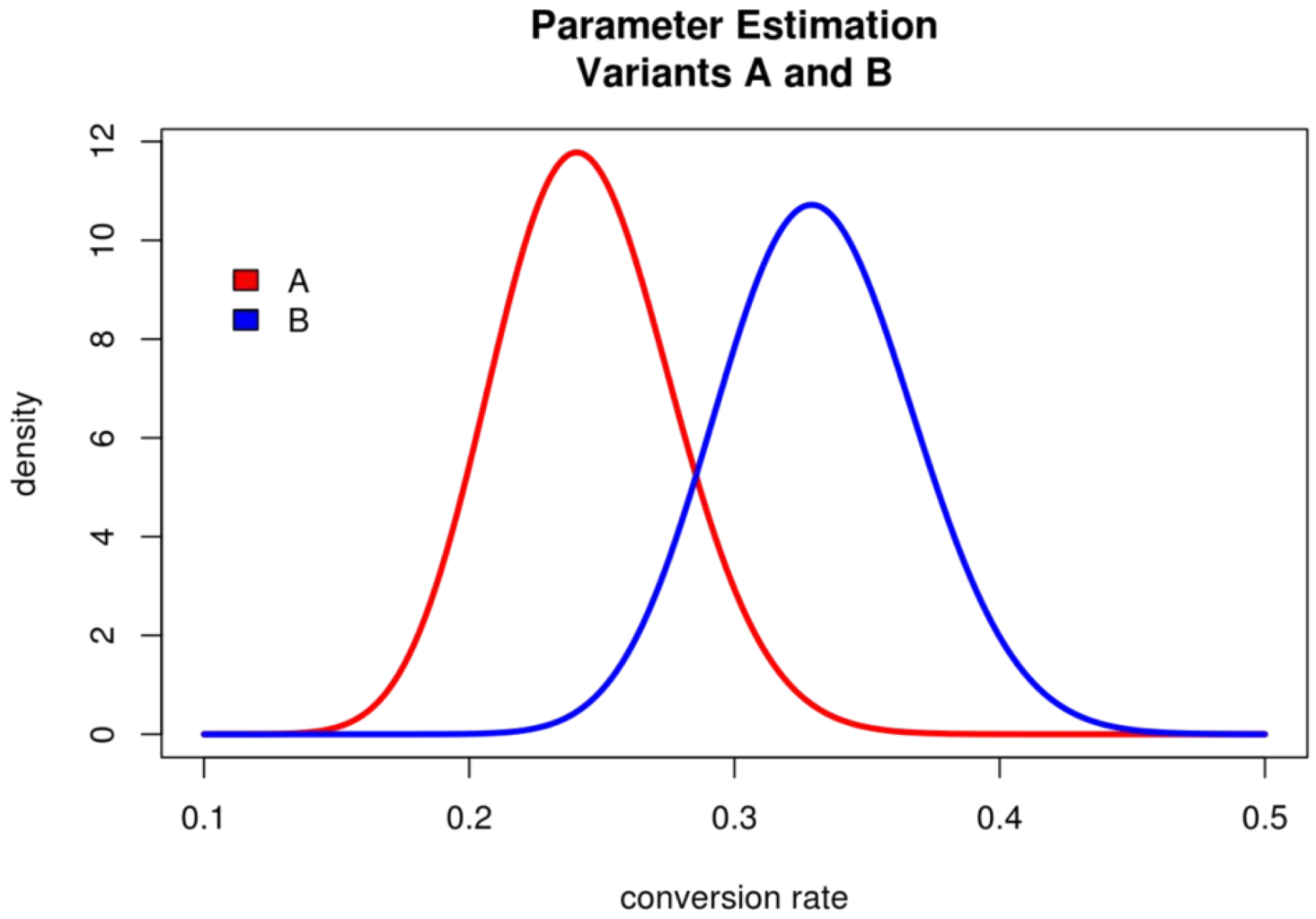


Пример: A/B тесты и статистика



Одна кривая отличается от других на 10%
Но разбиение на самом деле – случайное

Распределение результатов в группах



Проверка гипотез

Дано: значения, которые принимала случайная величина

Проверка гипотез

Дано: значения, которые принимала случайная величина

Нужно: выполнить некоторые операции с этими значениями, чтобы проверить наличие некоторого свойства у случайной величины (справедливость **статистической гипотезы**)

Проверка гипотез

Дано: значения, которые принимала случайная величина

Нужно: выполнить некоторые операции с этими значениями, чтобы проверить наличие некоторого свойства у случайной величины (справедливость **статистической гипотезы**)

Примеры гипотез: принадлежность к определенному семейству распределений, равенство матожидания нулю, равенство матожиданий у двух разных случайных величин

Пример гипотезы

Есть последовательность пользовательских сессий в которых произошел (1) или не произошел (0) заказ такси в группе A и в группе B:

Пример гипотезы

Есть последовательность пользовательских сессий в которых произошел (1) или не произошел (0) заказ такси в группе A и в группе B:

A: 0 1 0 0 0 0 1 0 0 1 0 0 1 ...

B: 0 0 0 1 0 0 0 1 0 0 0 1 0 ...

Пример гипотезы

Есть последовательность пользовательских сессий в которых произошел (1) или не произошел (0) заказ такси в группе A и в группе B:

A: 0 1 0 0 0 0 1 0 0 1 0 0 1 ...

B: 0 0 0 1 0 0 0 1 0 0 0 1 0 ...

Последовательность разностей:

0 1 0 -1 0 0 1 -1 0 1 0 -1 1

Пример гипотезы

Есть последовательность пользовательских сессий в которых произошел (1) или не произошел (0) заказ такси в группе A и в группе B:

A: 0 1 0 0 0 0 1 0 0 1 0 0 1 ...

B: 0 0 0 1 0 0 0 1 0 0 0 1 0 ...

Последовательность разностей:

0 1 0 -1 0 0 1 -1 0 1 0 -1 1

Посмотрим на эти числа как на значения случайной величины и проверим гипотезу, что ее матожидание равно нулю (что различие между группами A и B в среднем нулевое)

Статистические тесты

На входе: значения, которые принимала случайная величина (например, количество заказов в каждый день за последний месяц), **уровень значимости** (1%, 5%, 10%)

Статистические тесты

На входе: значения, которые принимала случайная величина (например, количество заказов в каждый день за последний месяц), **уровень значимости** (1%, 5%, 10%)

Что делаем: вычисляем некоторую величину и по ее значению принимаем или отвергаем гипотезу на заданном уровне значимости

Статистические тесты

На входе: значения, которые принимала случайная величина (например, количество заказов в каждый день за последний месяц), **уровень значимости** (1%, 5%, 10%)

Что делаем: вычисляем некоторую величину и по ее значению принимаем или отвергаем гипотезу на заданном уровне значимости

Примеры тестов:

- Тест Стьюдента
- Перестановочный тест
- Бутстреп

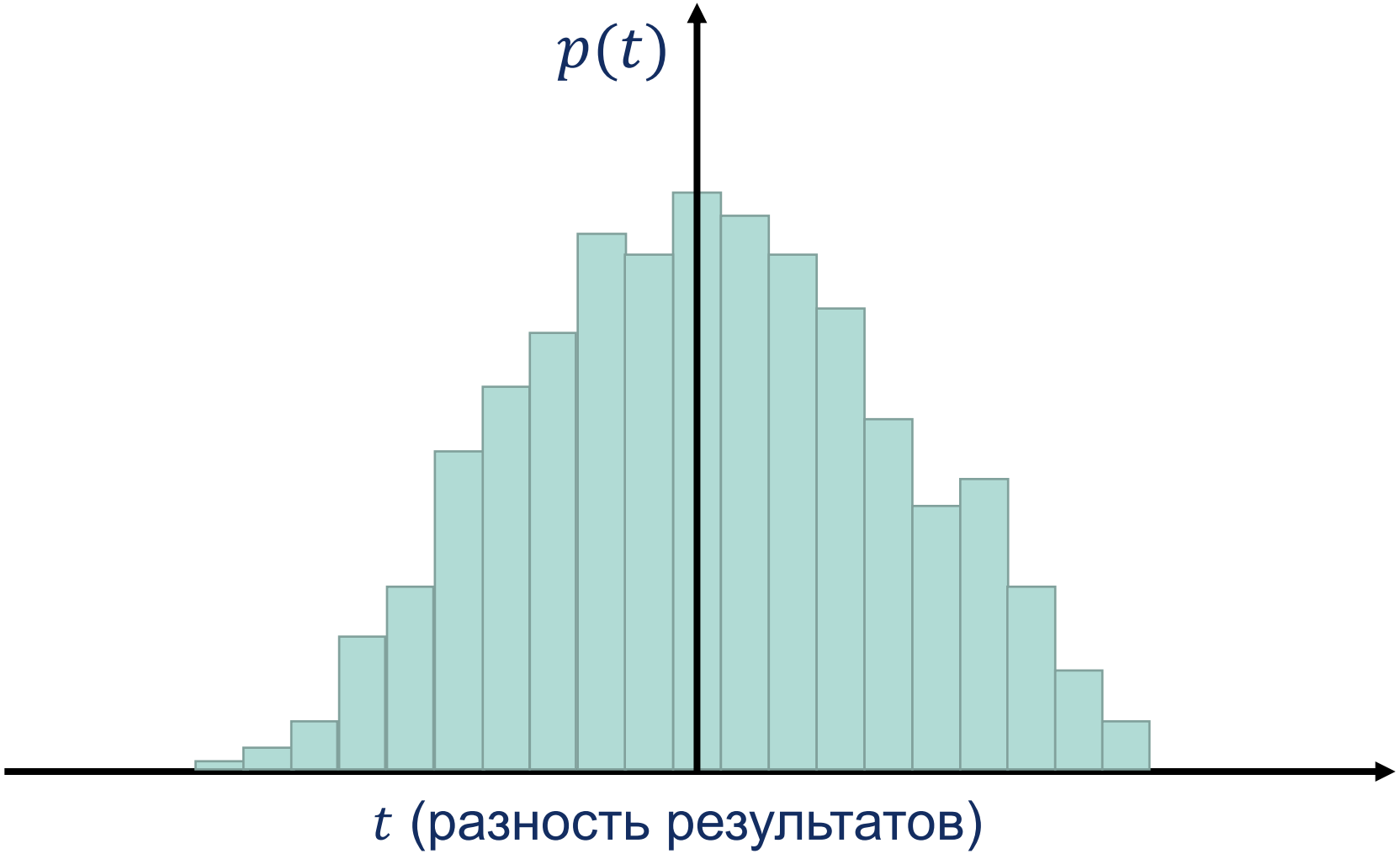
Чуть подробнее о происходящем в A/B тесте

- Пусть H_0- гипотеза, которую мы хотим отвергнуть: совпадение распределений результата в группе A и B (и, в частности, совпадение матожиданий)
- Обозначим возможное отклонение результатов в группах t, а то, которое фактически наблюдаем- T
- $P(t \geq T|H_0)$ – *достигаемый* уровень значимости
- Пусть 5% - уровень значимости
- Если $P(t \geq T|H_0) \leq 0.05$ – отвергаем H_0

Самый «простой» тест: бутстпреп

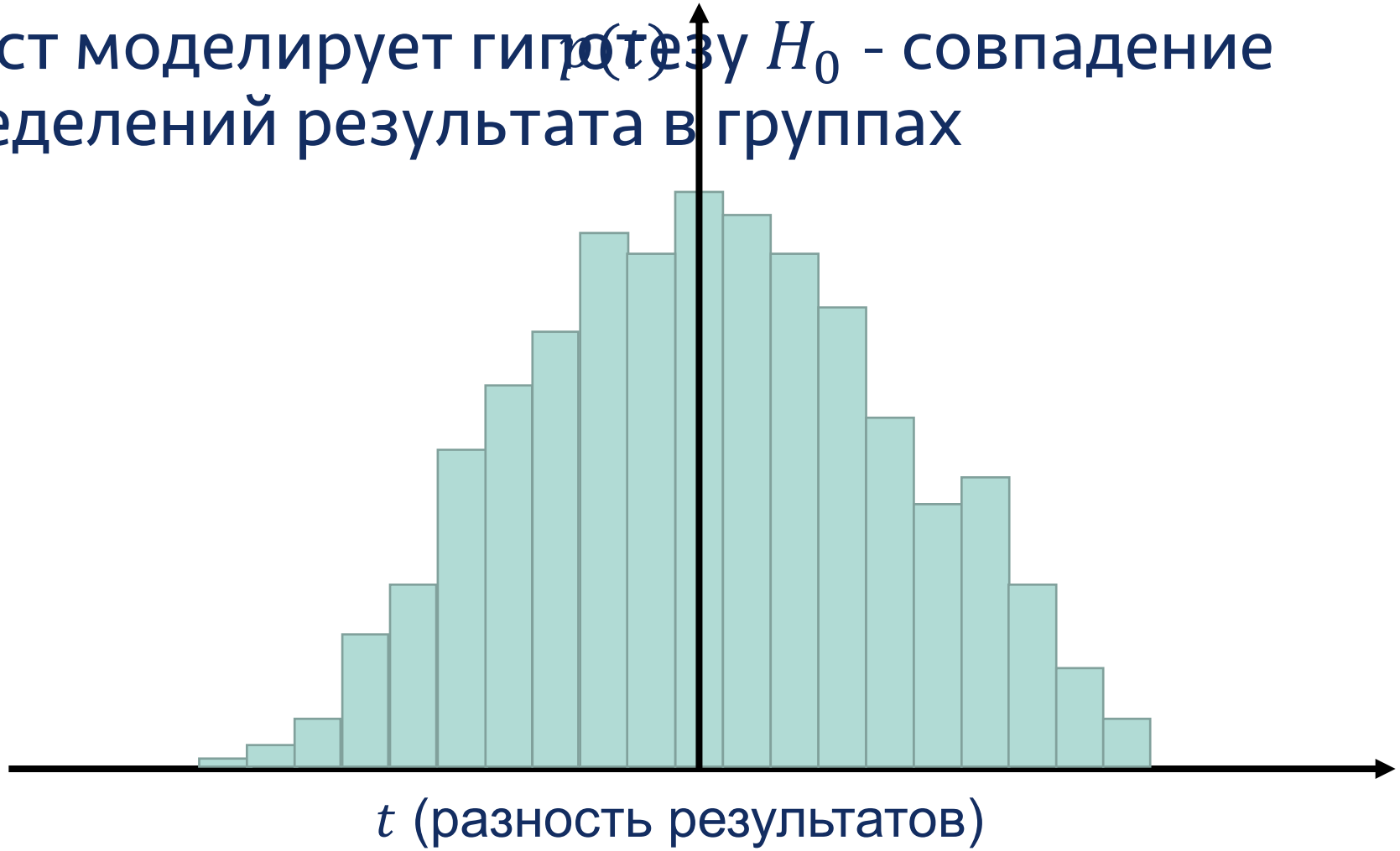
1. Имитируем A/A тест на исторических данных, N раз случайно разбив на две группы и посчитав результаты в каждой
2. Строим распределение разности результатов в группах
3. По этому распределению оцениваем вероятность получить в A/A тесте такую же разность как в A/B

Гистограмма распределения из A/A тестов

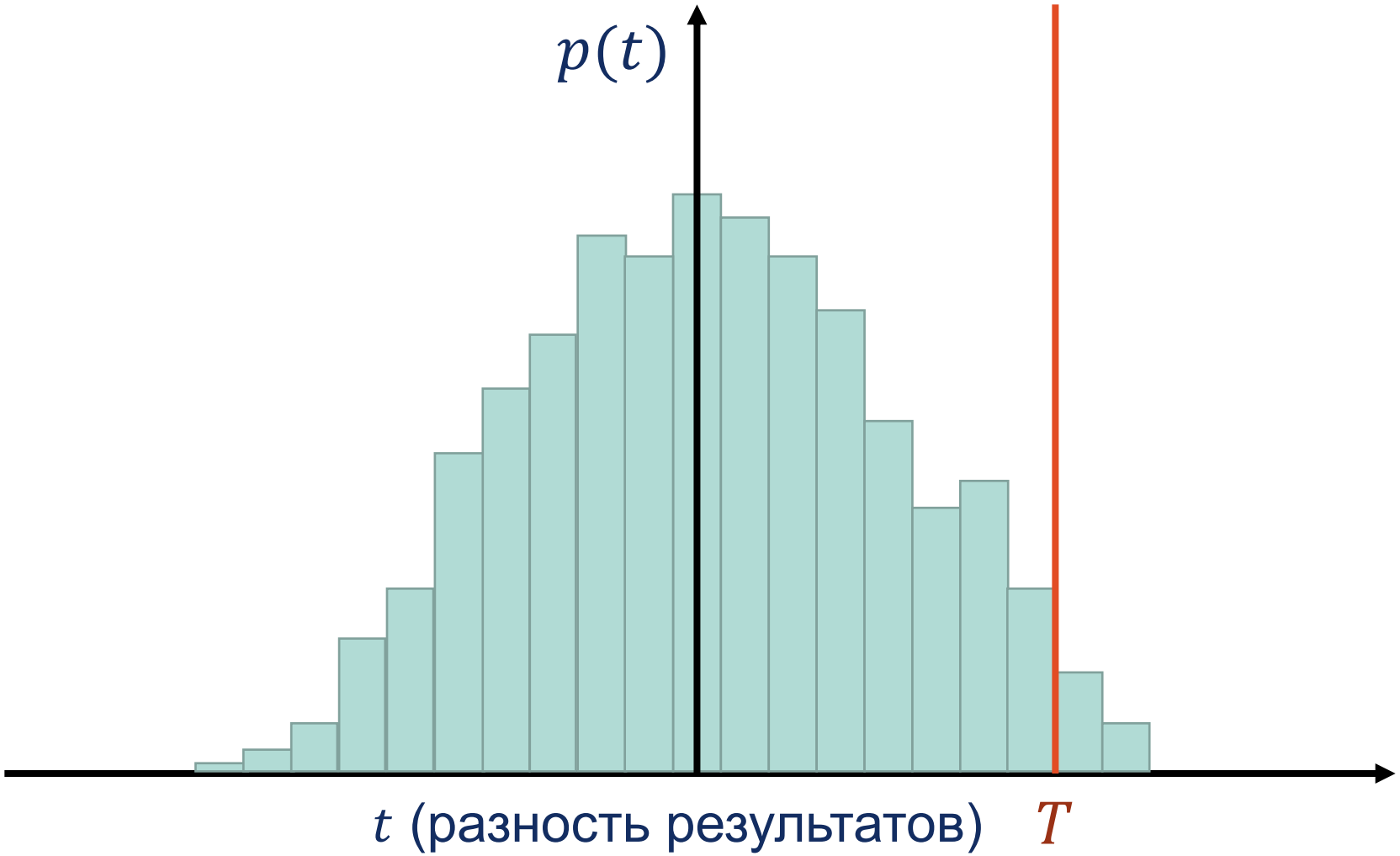


Гистограмма распределения из A/A тестов

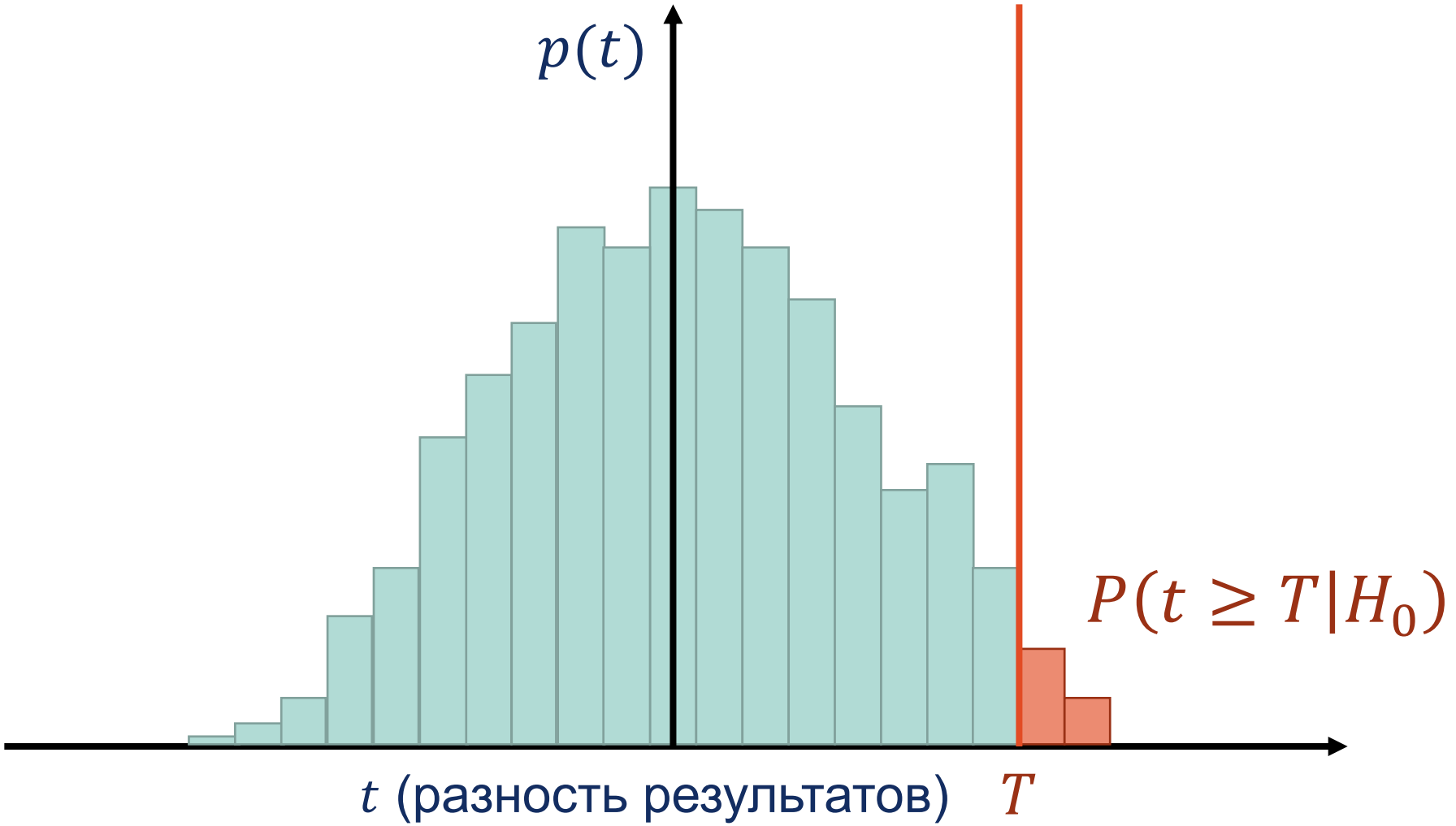
A/A тест моделирует гипотезу H_0 - совпадение распределений результата в группах



Разность из A/B теста на гистограмме

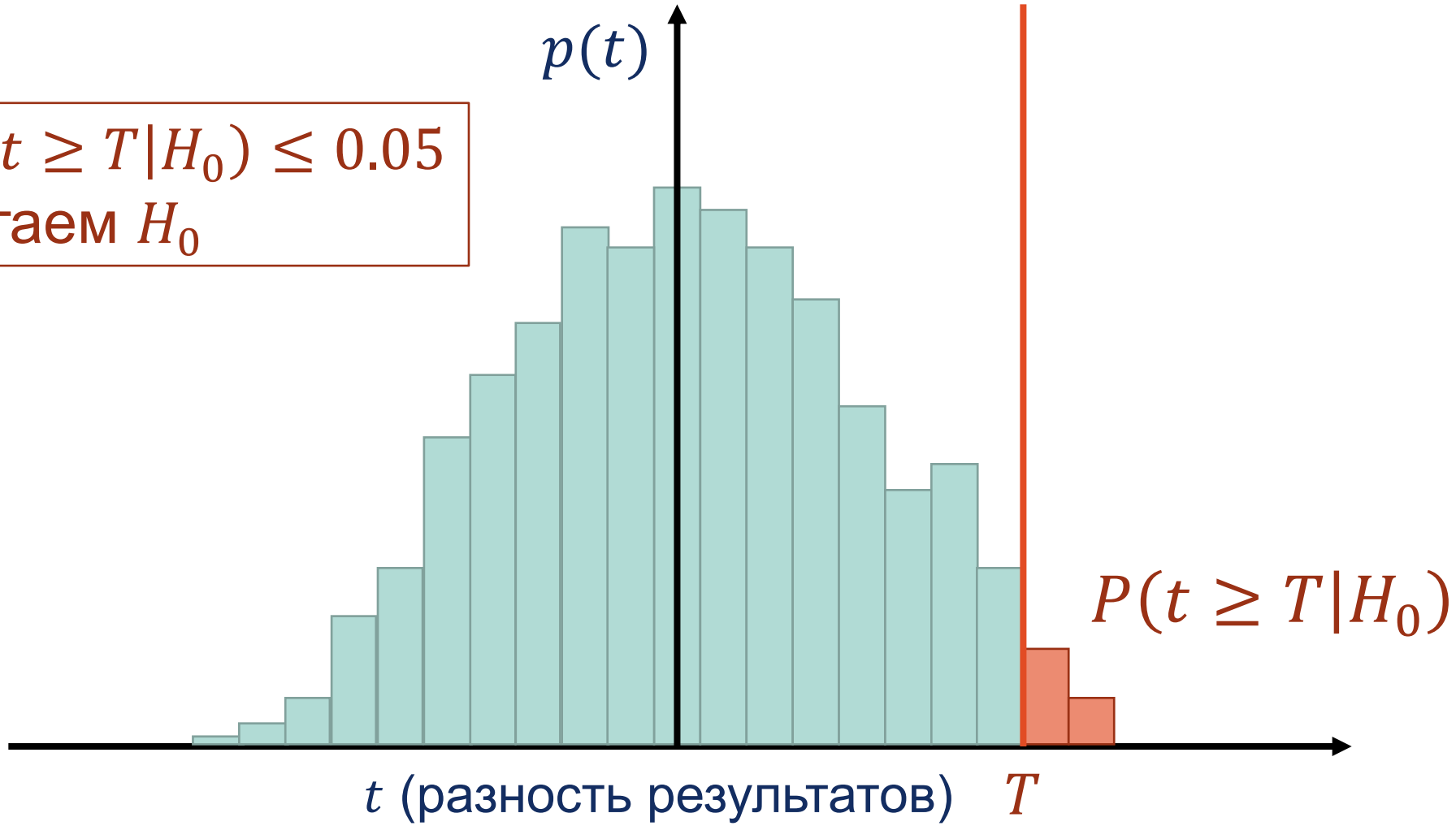


Вероятность не меньшего отклонения в A/A



Вероятность не меньшего отклонения в A/A

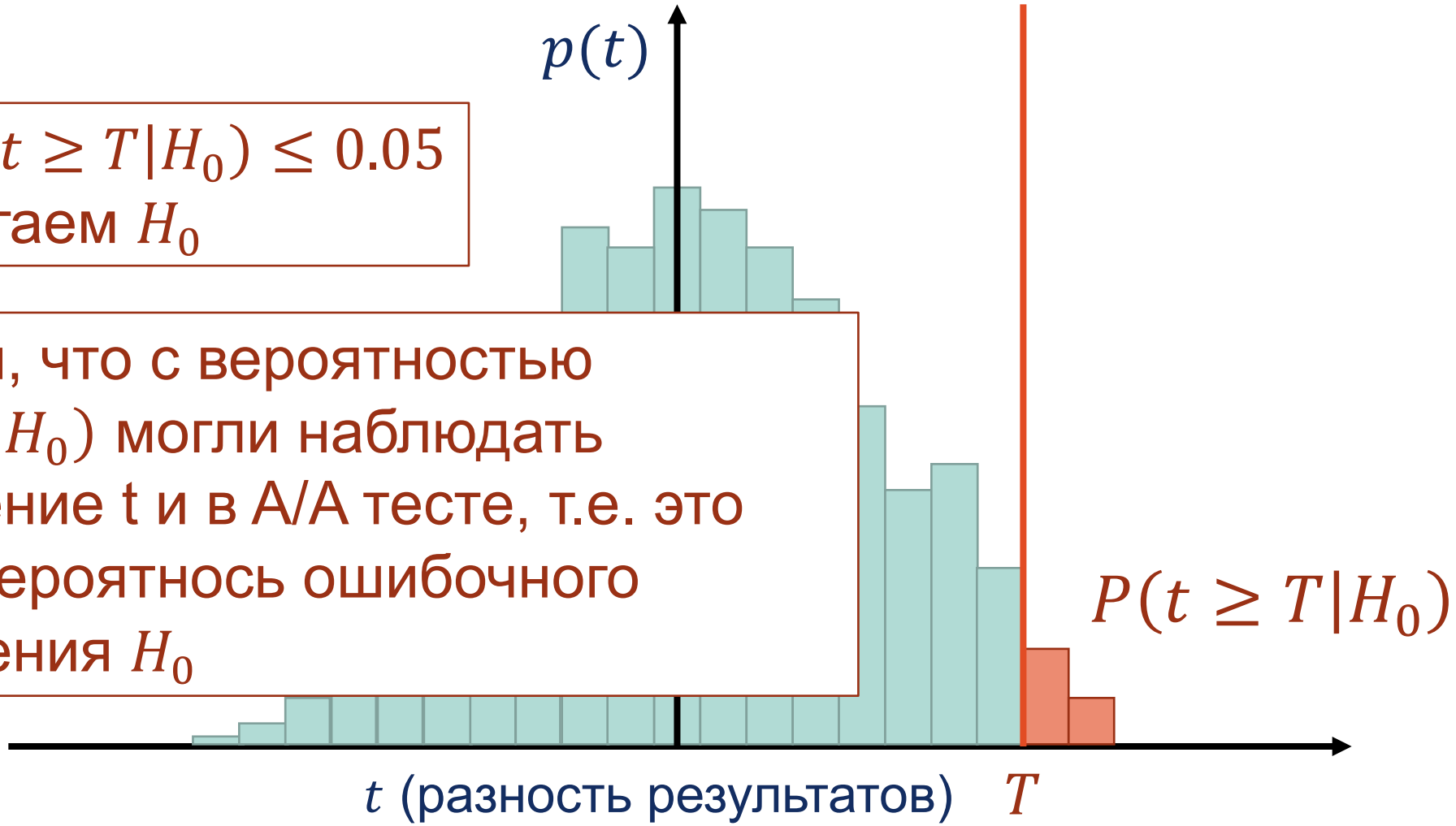
Если $P(t \geq T|H_0) \leq 0.05$
– отвергаем H_0



Вероятность не меньшего отклонения в A/A

Если $P(t \geq T|H_0) \leq 0.05$ – отвергаем H_0

Помним, что с вероятностью $P(t \geq T|H_0)$ могли наблюдать отклонение t и в A/A тесте, т.е. это еще и вероятнось ошибочного отвержения H_0



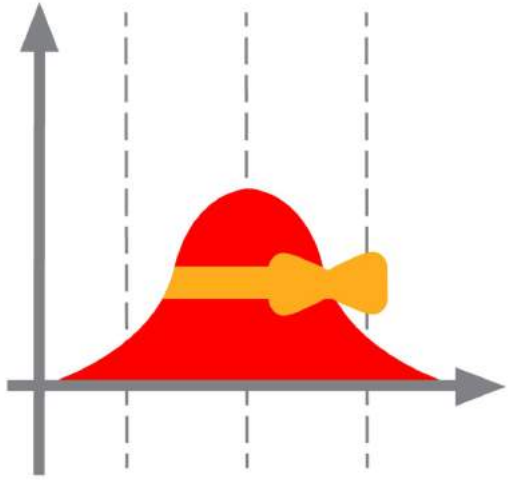
Подробнее о проверке гипотез

Лекция с весеннего потока DMIA 2018 года:

https://www.youtube.com/watch?v=YULMqwo7Tas&t=1021s

Подробнее о статистике в Data Science

Курс «Построение выводов по данным»:
https://www.coursera.org/learn/stats-for-data-analysis



Преподаватели и авторы курса:

Евгений Рябенко

Эмели Драль

История из практики: разбиение на группы

- Предложено аналитиками:
  - Брать hash от user_id
  - Смотреть на остаток от деления на 2

- Сделано:
  - Брать hash от user_id+user_email
  - Смотреть на остаток от деления на 2

История из практики: улучшение алгоритма

- Перед каждой выкаткой сравнивали качество новой версии алгоритма с предыдущей
- Сделали 15 последовательных версий
- Ради интереса решили посмотреть, насколько улучшился алгоритм по сравнению с первоначальным, и сделали A/B тест

История из практики: улучшение алгоритма

- Перед каждой выкаткой сравнивали качество новой версии алгоритма с предыдущей
- Сделали 15 последовательных версий
- Ради интереса решили посмотреть, насколько улучшился алгоритм по сравнению с первоначальным, и сделали A/B тест
- Первоначальный победил

Что не стоит делать при оценке статистической значимости

1. Постфактум подбирать такую метрику, по которой будет стат. значимый результат

Что не стоит делать при оценке статистической значимости

1. Постфактум подбирать такую метрику, по которой будет стат. значимый результат
2. Подбирать такой срез, в котором есть стат.значимый результат

Что не стоит делать при оценке статистической значимости

1. Постфактум подбирать такую метрику, по которой будет стат. значимый результат
2. Подбирать такой срез, в котором есть стат.значимый результат
3. Подбирать такой период времени, на котором есть стат.значимый результат

Что не стоит делать при оценке статистической значимости

1. Постфактум подбирать такую метрику, по которой будет стат. значимый результат
2. Подбирать такой срез, в котором есть стат.значимый результат
3. Подбирать такой период времени, на котором есть стат.значимый результат
4. Каждый день проверять, статзначим ли результат и останавливать тест, если да (частный случай предыдущего)

Резюме по A/B тестам

- Качество в онлайне и оффлайне обычно отличается
- Важно не допустить переобучение или утечку
- Нужно обязательно делать A/B тесты
- Нужно обязательно оценивать статзначимость
- Важно не делать ложных выводов по статистически незначимым результатам

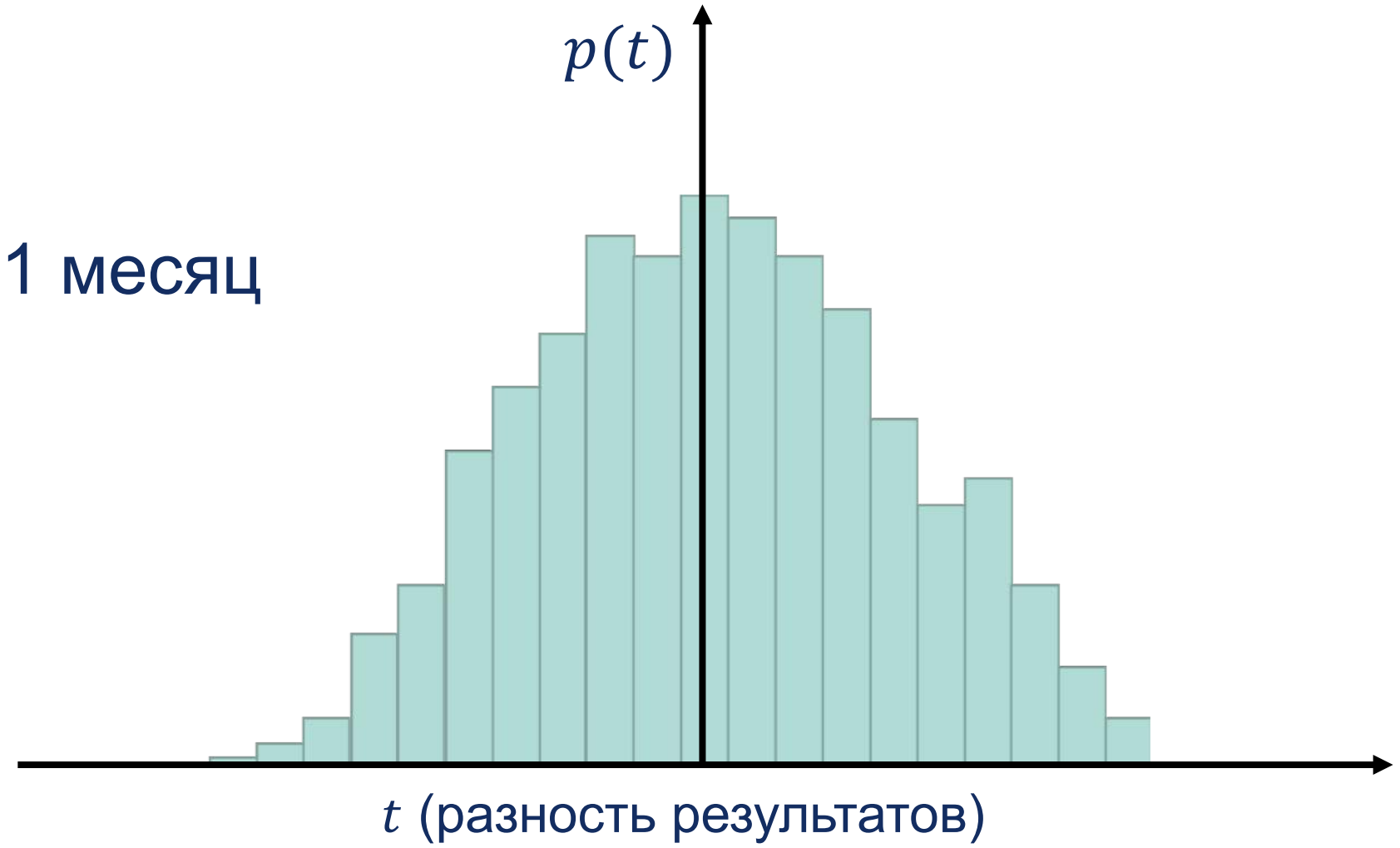
Планирование A/B теста

1. Решаете, насколько большое (по величине) улучшение метрики детектируете

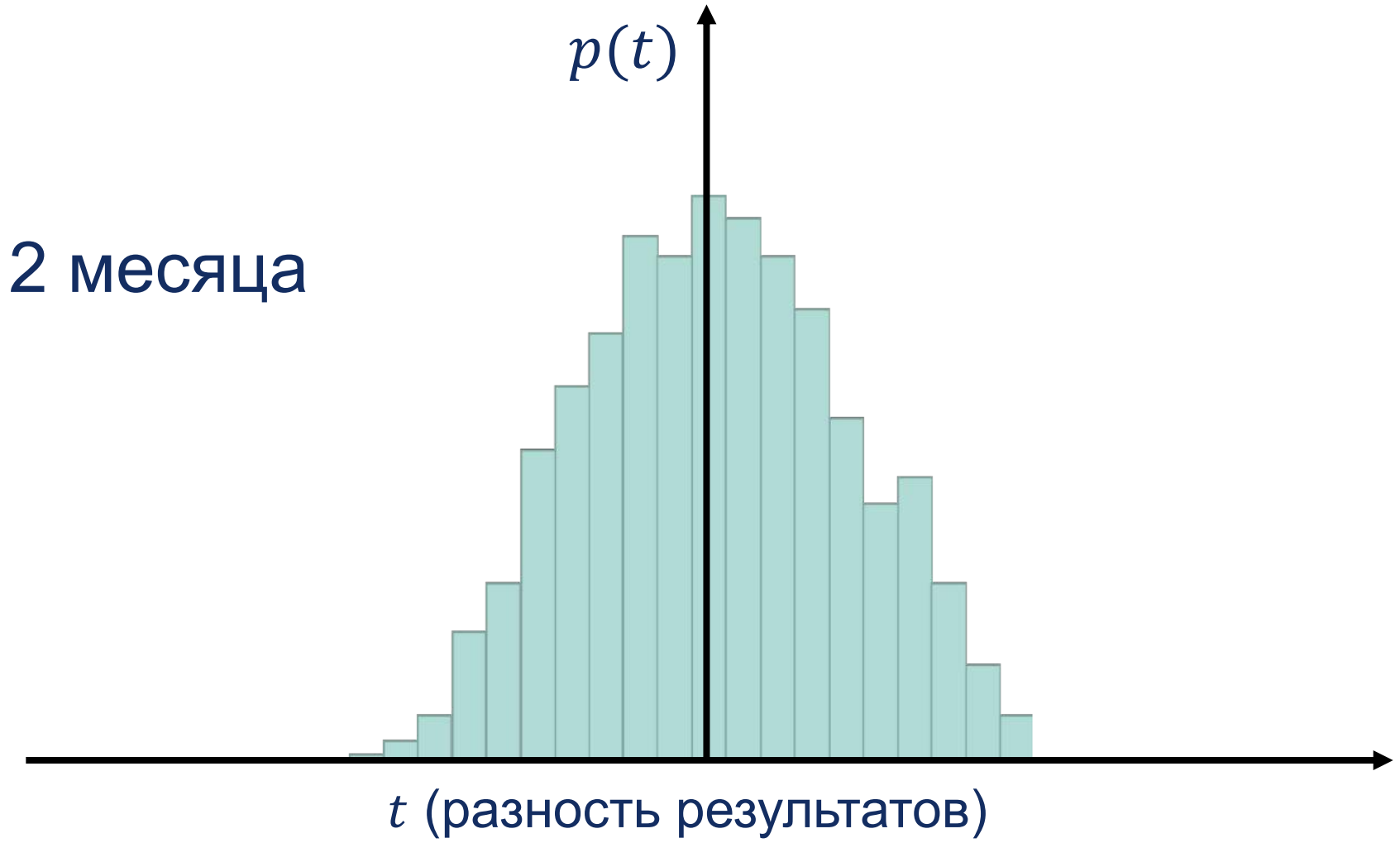
Планирование A/B теста

1. Решаете, насколько большое (по величине) улучшение метрики детектируете
2. По оценкам статзначимости (например, бутстрепом) на исторических данных понимаете, сколько данных нужно (=какая продолжительность у A/B теста), чтобы это улучшение алгоритма было статзначимым на нужном уровне значимости

Как подбирать длительность A/B теста

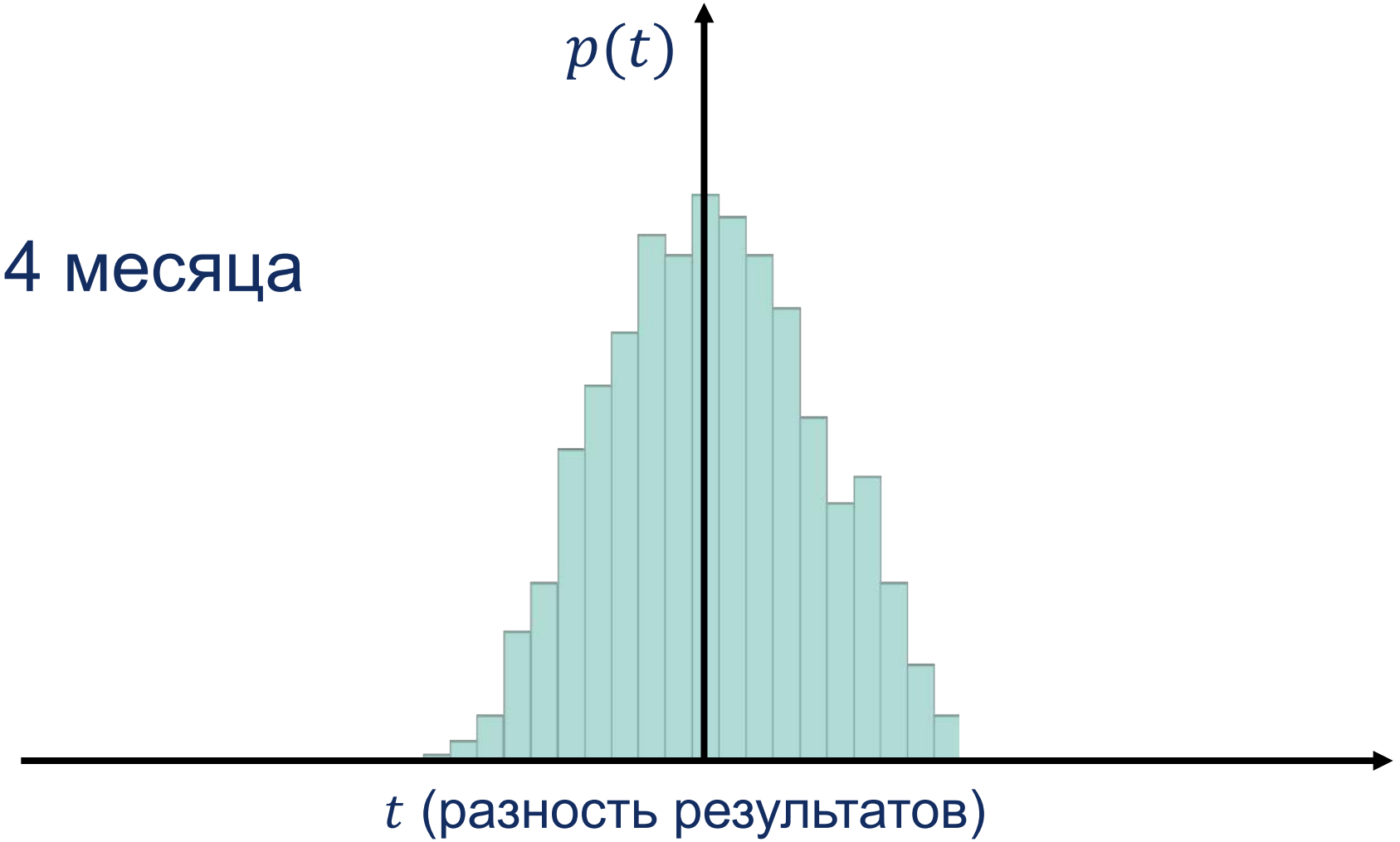


Как подбирать длительность A/B теста



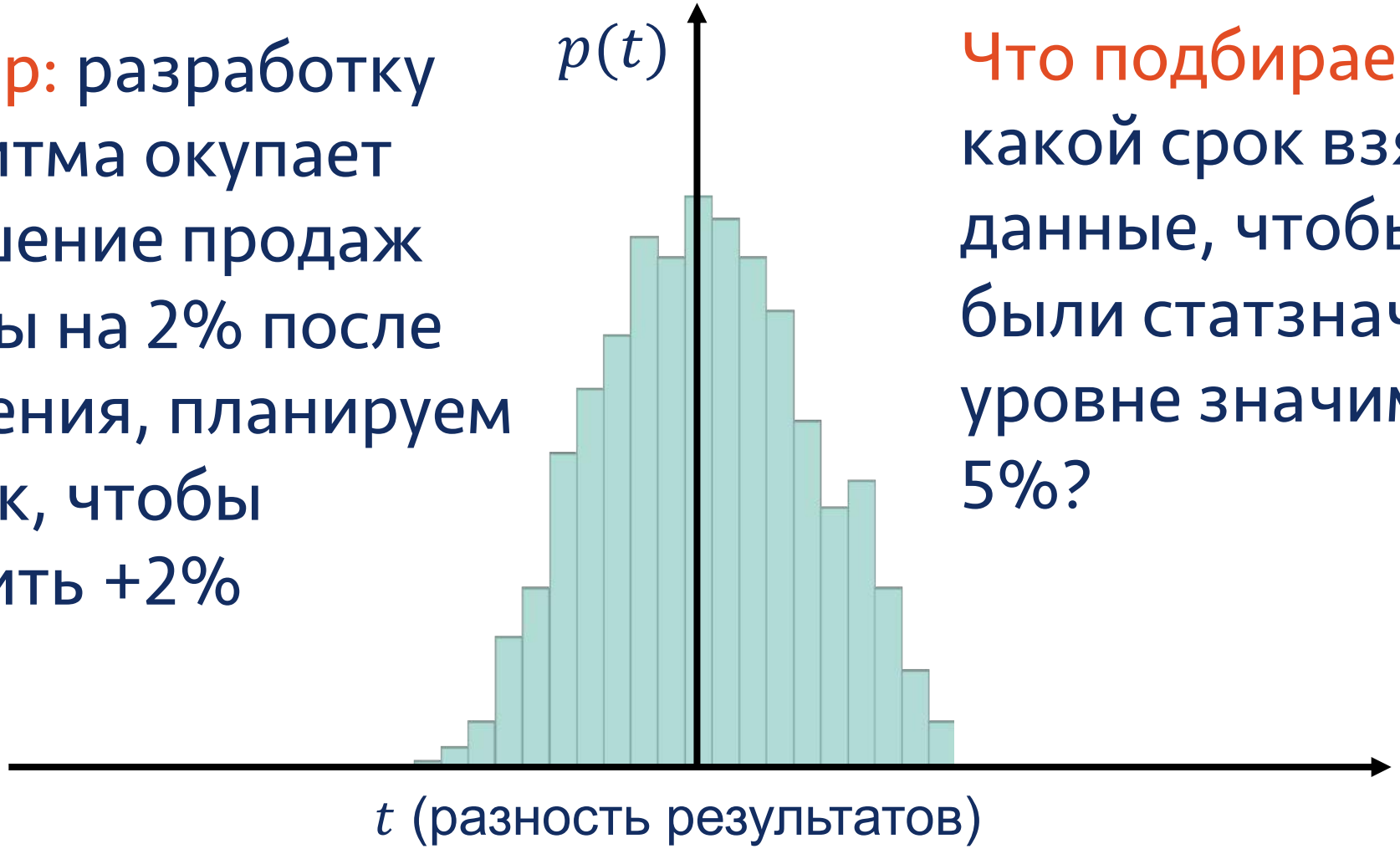
Как подбирать длительность A/B теста

4 месяца



Как подбирать длительность A/B теста

Пример: разработку алгоритма окупает повышение продаж хотя бы на 2% после внедрения, планируем A/B так, чтобы заметить +2%



Что подбираем: за какой срок взять данные, чтобы +2% были статзначимы на уровне значимости 5%?

Планирование A/B теста

1. Решаете, насколько большое (по величине) улучшение метрики детектируете
2. По оценкам статзначимости (например, бутстрепом) на исторических данных понимаете, сколько данных нужно (=какая продолжительность у A/B теста), чтобы это улучшение алгоритма было статзначимым на нужном уровне значимости
3. Помните про сезонность, округляете продолжительность теста хотя бы до недель

Планирование A/B теста

1. Решаете, насколько большое (по величине) улучшение метрики детектируете
2. По оценкам статзначимости (например, бутстрепом) на исторических данных понимаете, сколько данных нужно (=какая продолжительность у A/B теста), чтобы это улучшение алгоритма было статзначимым на нужном уровне значимости
3. Помните про сезонность, округляете продолжительность теста хотя бы до недель
4. Перед A/B тестом полезно провести A/A, чтобы проверить, настолько ли похожи результаты в группах, как на исторических данных, а возможно – даже проверить, не срабатывают ли ваши критерии в A/A тесте

Планирование A/B теста

1. Решаете, насколько большое (по величине) улучшение метрики детектируете
2. По оценкам статзначимости (например, бутстрепом) на исторических данных понимаете, сколько данных нужно (=какая продолжительность у A/B теста), чтобы это улучшение алгоритма было статзначимым на нужном уровне значимости
3. Помните про сезонность, округляете продолжительность теста хотя бы до недель
4. Перед A/B тестом полезно провести A/A, чтобы проверить, настолько ли похожи результаты в группах, как на исторических данных, а возможно – даже проверить, не срабатывают ли ваши критерии в A/A тесте

С учетом перезапусков из-за ошибок – фактические сроки могут быть еще в 2-3 раза больше

Резюме первой части лекции

1. Существует ряд стандартных метрик качества, которые допускают различные модификации
2. Важно выбрать релевантную задаче метрику
3. Полезно изучать стабильность обученной модели
4. Нужно оценивать качество после внедрения модели с помощью A/B теста
5. В A/B тесте обязательно нужно оценивать статзначимость и вообще планировать его так, чтобы ее можно было заметить

Напоминание: топ ошибок в индустрии

1. Постановка задачи отсутствует или неправильная (например, метрику вообще выбрали случайно)
2. A/B тест не проводится или не валиден
3. Утечка и переобучение

Субъективный топ причин

1. Безответственность: «и так сойдет»
2. Невнимательность, особенно в период «авралов»
3. Нехватка экспертизы: незнание, что вопросы, которые мы обсуждали на этой лекции, существуют и важны

4.Извлечение и простые преобразования признаков

Виды признаков

Какие бывают признаки:

1. Числовые
2. Порядковые
3. Категориальные
4. Даты и время
5. Координаты

Даты и время

1. Количество прошедших секунд
   например, с 00:00:00 UTC, 1 January 1970

2. Использование периодичности
   a. номер дня в году, в месяце, в неделе
   b. час, минута, секунда

3. Время до/после важных событий
   Например, количество дней, оставшихся до ближайшего праздника

Координаты

1. Повороты системы координат на 45 градусов, 22.5 градусов, etc
2. Добавление расстояний до:
    a. Других объектов из выборки
    b. Центров кластеров
    c. Инфраструктурных зданий - магазинов, школ, больниц

Категориальные признаки (строки)

Из колонок “name”, “ticket”, “cabin” можно сгенерировать новые признаки

	A	B	C	D	E	F	G	H	I	J	K
1	survived	pclass	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked
2	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
3	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
4	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
5	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
6	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
7	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
8	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
9	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
10	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
11	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
12	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
13	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S

Категориальные признаки

Бинаризация

feature
a
b
c
b

feature == a	feature == b	feature == c
1		
	1	
		1
	1	

Категориальные признаки

Hashing trick

feature
a
b
c
b

feature == a or feature == c	feature == b
1	
	1
1	
	1

Категориальные признаки

№ склада	Город	…	Продано вина (ящиков)
2343	Москва	…	56000
185	Самара	…	10500
121	Ростов	…	11300
…	…	…	…

Категориальные признаки

№ склада	В среднем продают в городе	…	Продано вина (ящиков)
2343	59000	…	56000
185	11200	…	10500
121	12100	…	11300
…	…	…	…

Метапризнаки

Использование ответов других алгоритмов

	xgb_prediction	knn_prediction	svm_prediction	target
train	0.192	0.293	0.122	0
train	0.789	0.890	0.670	1
test	0.542	0.310	0.173	?

Осторожно с переобучением: используйте KFold, LOO

Feature engineering

- Выделение признаков (feature extraction) – генерация признаков по известным данным
- Отбор признаков (feature selection) – ранжирование признаков по «полезности» и выкидывание наименее полезных (или даже наоборот «вредных»)
- Преобразование признаков (feature transform) – создание новых признаков на основе имеющихся

Генерация признаков

Для решения задачи нужно использовать разные типы данных

Пример: задача рекомендации музыки

1. Музыкальные треки
2. Тексты песен
3. Плейлисты

Проблема: нужно преобразовать к одному формату - матрице “объекты-признаки”

Пример 1: текстовые признаки

- Dataset: 20news_groups
- Электронные письма, разбитые по 20 темам (классам)
- Попробуем придумать классификатор, который различает две темы:
- auto и politics.mideast

Пример письма 1:

Извлечение текстовых признаков

From: carl_f_hoffman@cup.portal.com
Newsgroups: rec.autos
Subject: 1993 Infiniti G20
Message-ID: <78834@cup.portal.com>
Date: Mon, 5 Apr 93 07:36:47 PDT
Organization: The Portal System (TM)
Lines: 26

I am thinking about getting an Infiniti G20. In consumer reports it is ranked high in many catagories including highest in reliability index for compact cars.
Mitsubushi Galant was second followed by Honda Accord).

A couple of things though:
1) In looking around I have yet to see anyone driving this
car. I see lots of Honda's and Toyota's.
2) There is a special deal where I can get an Infinity G20, fully loaded, at dealer cost (I have check this out and the numbers match up). They are doing this because they are releasing and update mid-1993 version (includes dual air-bags) and want to get rid of their old 1993's.

I guess my question is: Is this a good deal?
Also, Can anyone give me any feedback on Infiniti?

Thanks,
Carl Hoffman

P.S.
The other cars that I have test driven and which are in the running are:
Mitsubishi Galant, Honda Accord, and Toyota Camary

Извлечение текстовых признаков

Пример письма 2:

From: Bob.Waldrop@f418.n104.z1.fidonet.org (Bob Waldrop)
Subject: Celebrate Liberty! 1993
Message-ID: <1993Apr5.201336.16132@dsd.es.com>
Followup-To: talk.politics.misc

Announcing. . . Announcing. . . Announcing. . .Announcing. . .

CELEBRATE LIBERTY!
1993 LIBERTARIAN PARTY NATIONAL CONVENTION
AND POLITICAL EXPO

THE MARRIOTT HOTEL AND THE SALT PALACE
SALT LAKE CITY, UTAH

INCLUDES INFORMATION ON DELEGATE DEALS!
(Back by Popular Demand!)

The convention will be held at the Salt Palace Convention Center and the Marriott Hotel, Salt Lake City, Utah. The business sessions, Karl Hess Institute, and Political Expo are at the Salt Palace; breakfasts, parties, and banquet are at the Marriott Hotel.

Marriott Hotel room rates are $79.00 night, plus 10.5% tax ($87.17 total). This rate is good for one to four persons room occupancy. Double is one or two beds; 3 or 4 people is 2 beds. You can make your reservations direct with the hotel (801-531-0800), or you can purchase your room

Текстовые признаки: bag-of-words

aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
...	
Zaire	0

Простой классификатор текстов

Взвешивание частот слов в текстах

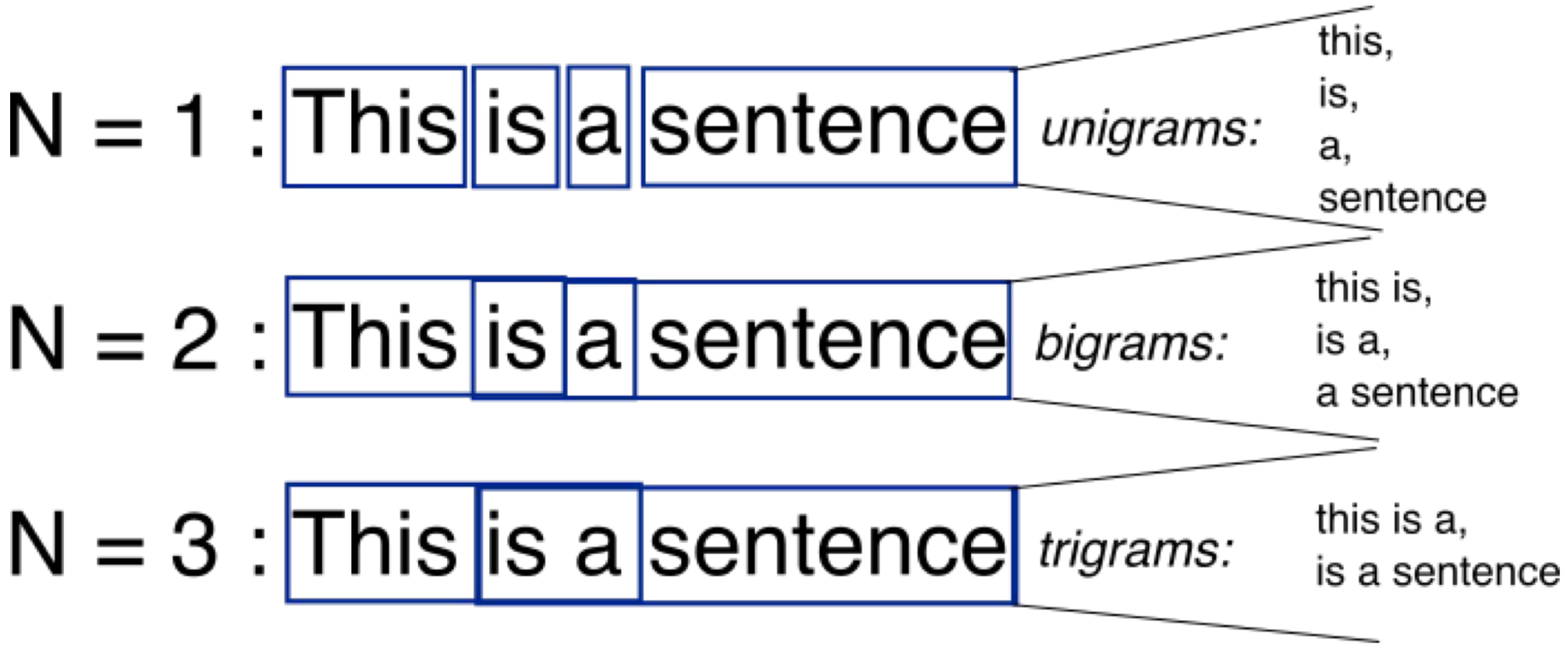
Term Frequency

$$\mathrm{tf}_{\mathrm{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

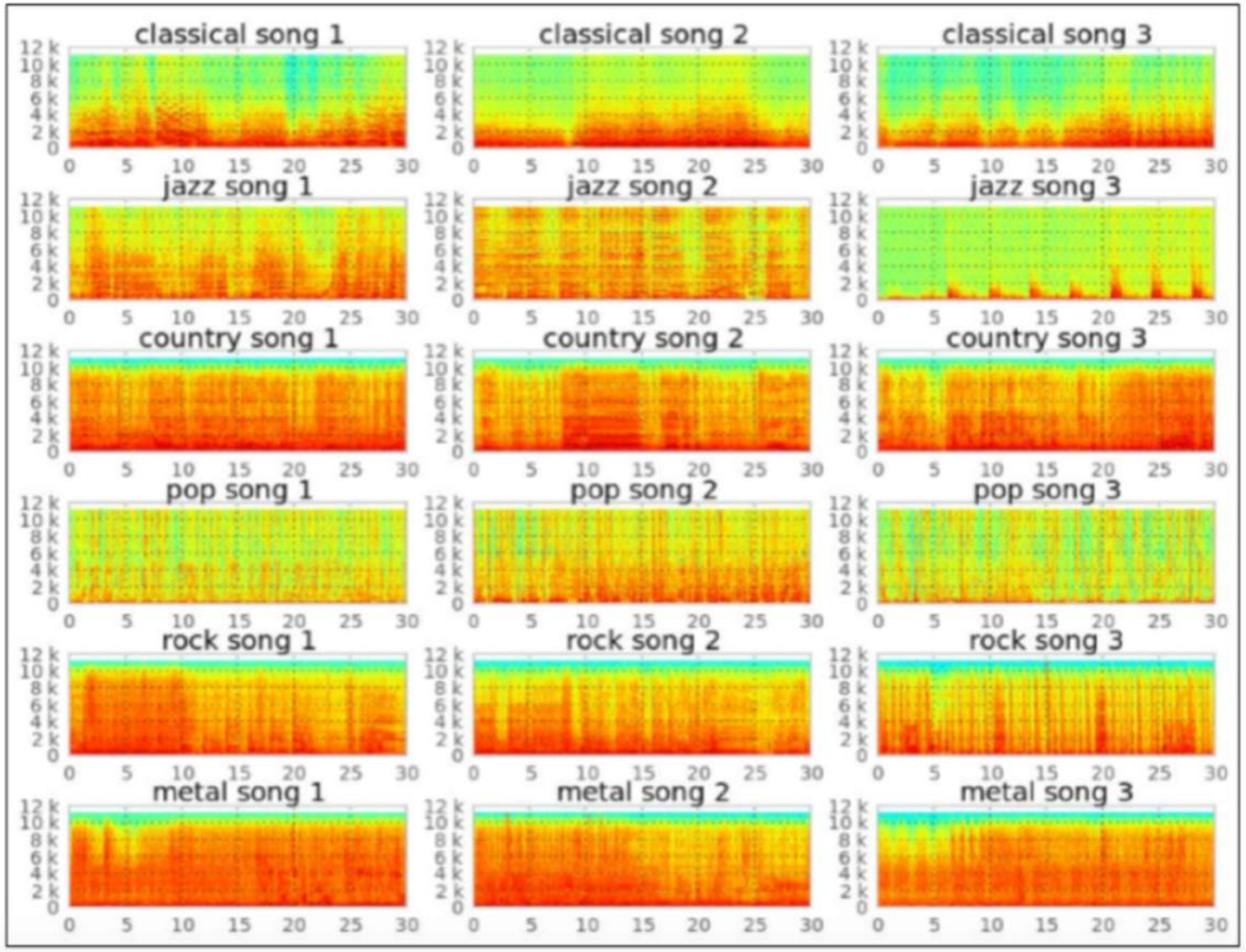
Inverse Document Frequency

$$\mathrm{idf}_{\mathrm{i}} = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

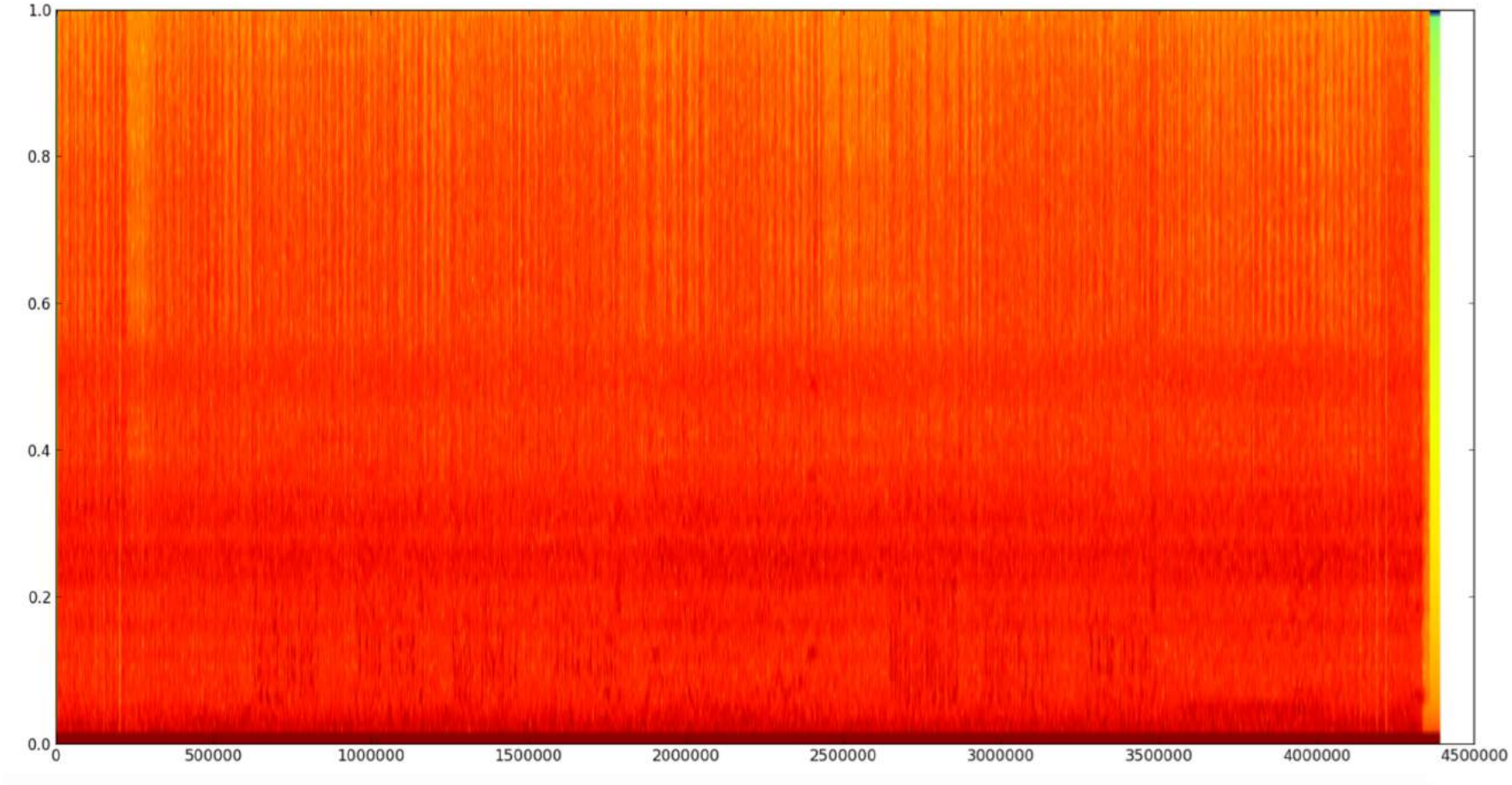
Частоты N-грамм



Пример 2: признаки аудиофайла

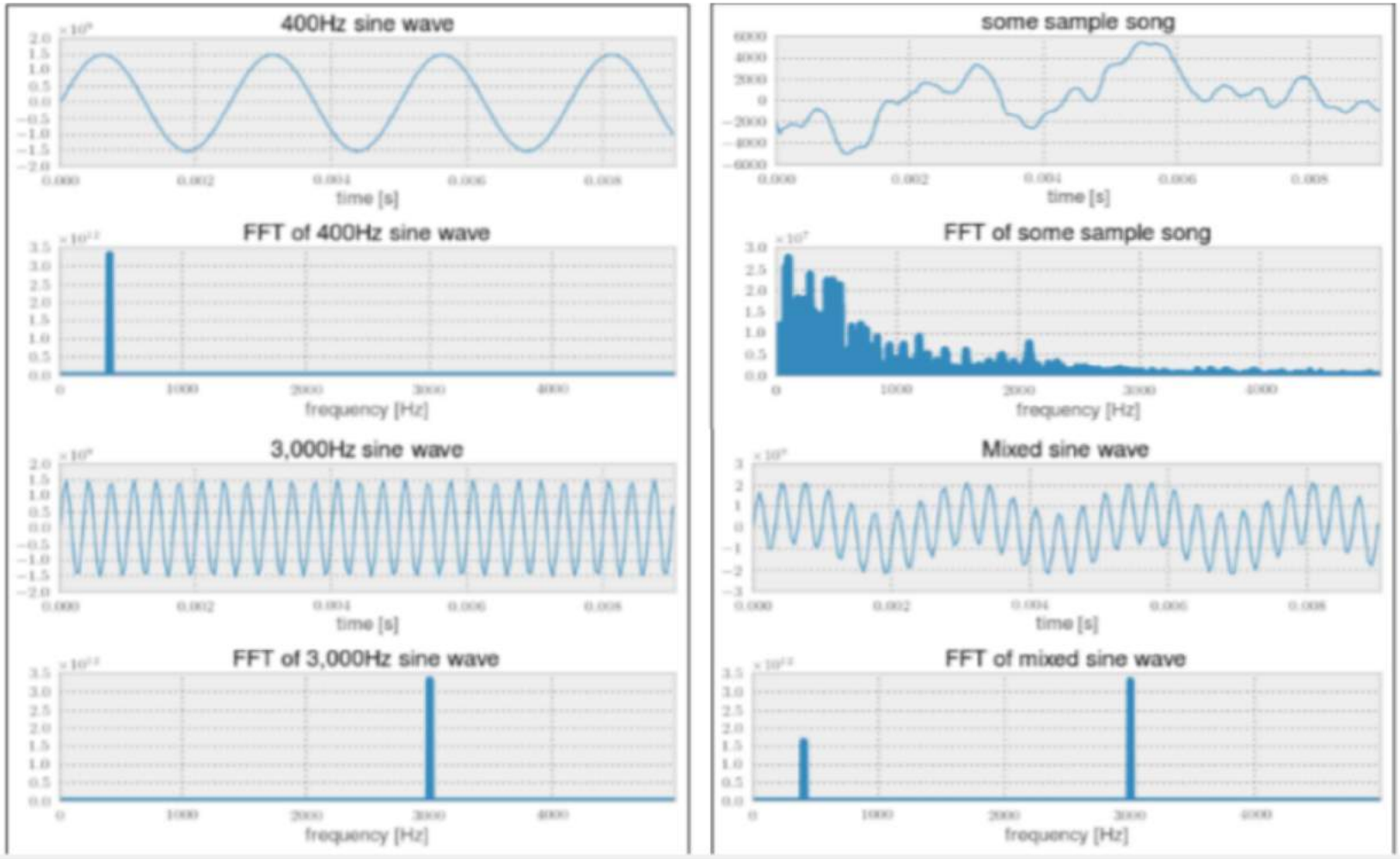


Пример 2: признаки аудиофайла



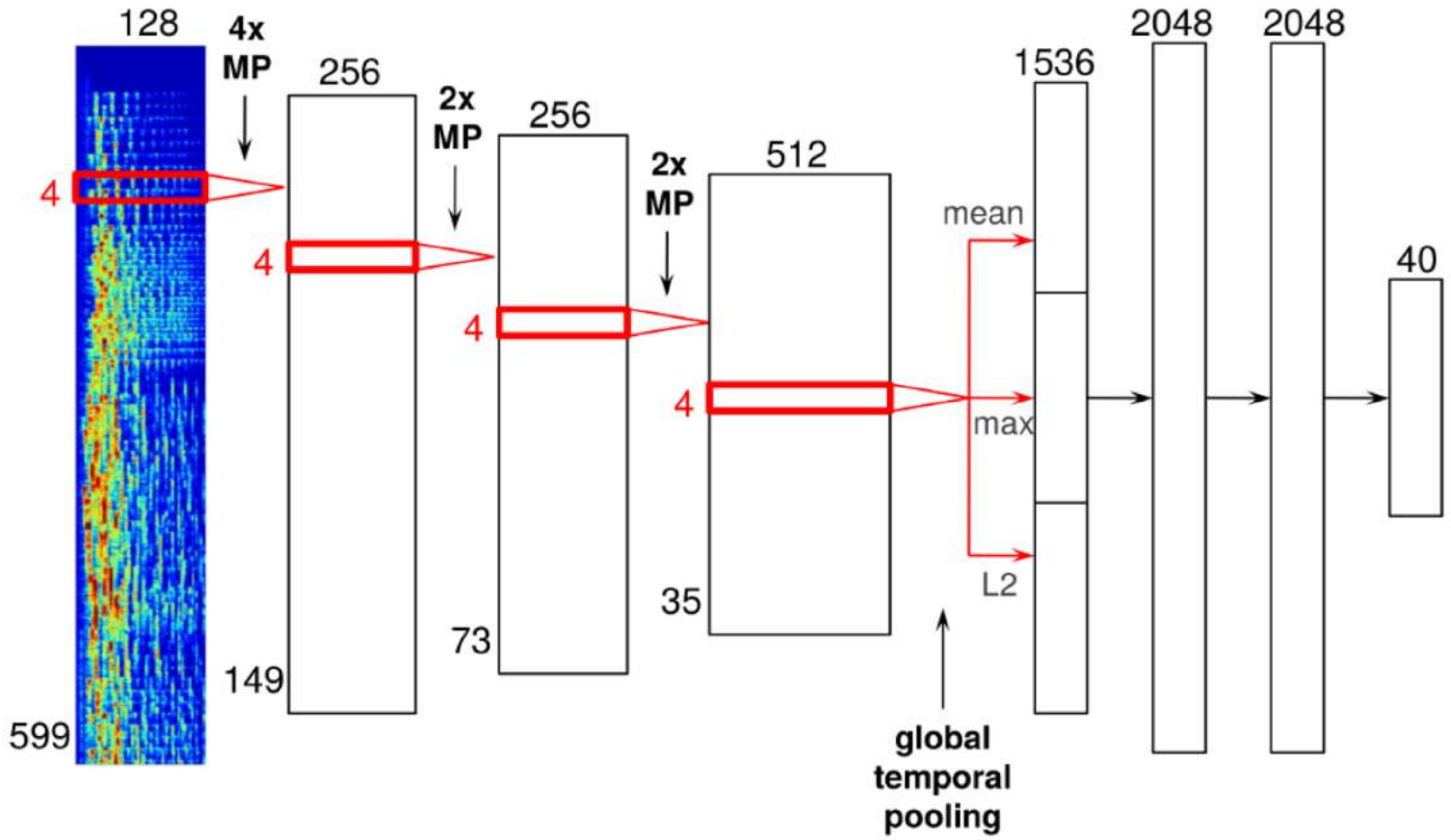
MFCC - преобразование Фурье логарифма спектра

Пример 2: признаки аудиофайла



Пример 2: признаки аудиофайла

Embeddings с помощью нейронных сетей:



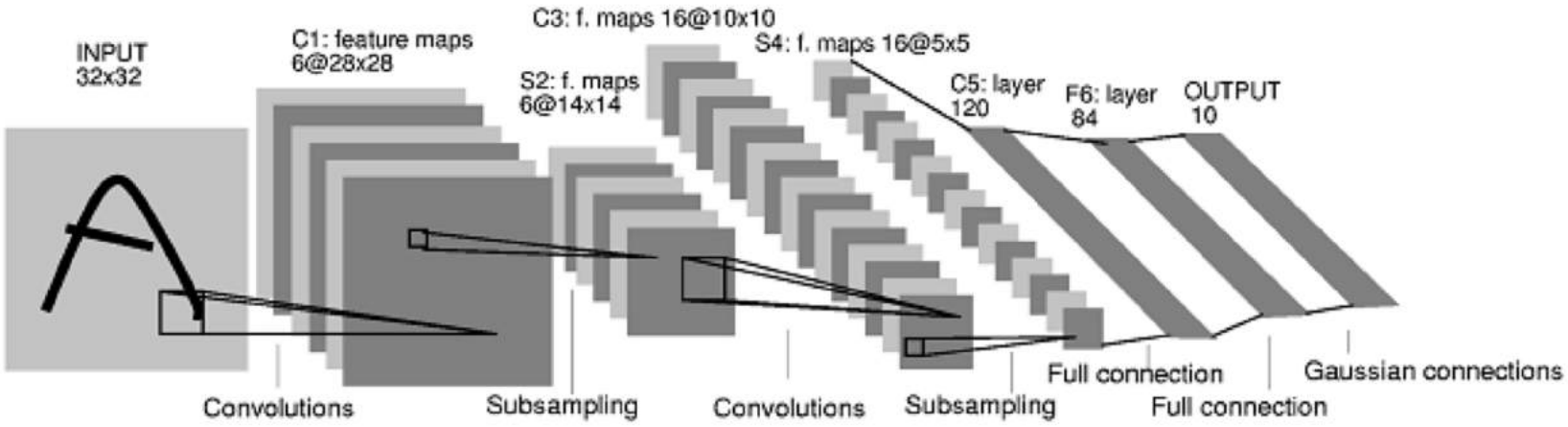
Пример 3: признаки изображения

Input image

Histogram of Oriented Gradients

Пример 3: признаки изображения

Выходы слоев из нейросети



Нормализация признаков

Центрирование и нормирование

1. На обучающей выборке:

$$f' = \frac{f - a}{b}$$

Например: a – среднее значение, b – дисперсия

Другой вариант: $b = \max\{f\} - \min\{f\}$

Иногда даже так: $b = \sqrt{\max\{f\} - \min\{f\}}$

Нормализация признаков

Центрирование и нормирование

1. На обучающей выборке:

$$f' = \frac{f - a}{b}$$

Например: a – среднее значение, b – дисперсия

Другой вариант: $b = \max\{f\} - \min\{f\}$

Иногда даже так: $b = \sqrt{\max\{f\} - \min\{f\}}$

2. На тестовой выборке – два варианта:
   a) Пересчитываем a и b
   b) Используем те же a и b

Нормализация признаков

Пересчитывать ли a и b

На тестовой выборке – два варианта:

a) Пересчитываем a и b
b) Используем те же a и b

Если пересчитываем – модель работает с тем же диапазоном значений. Это уместно, если важны значения признаков **относительно выборки**

Если используем те же – возможны выходы значения признака за диапазон из обучающей выборки. Это уместно, если в этих случаях модель должна **экстраполировать прогноз** и может это делать правильно.

Монотонные функции от признаков

Взвешивание корнем и логарифмом

Когда нужно сгладить большие значения признаков, допустимо заменить признак его корнем или логарифмом:

$$f' = \sqrt{f}$$

$$f' = \ln(1 + f)$$

Например, иногда этот прием применяется с частотами слов в тексте

Подгонка под более «нормальное» распределение

Преобразование Бокса-Кокса

Однопараметрический вариант:

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln y_i & \text{if } \lambda = 0, \end{cases}$$

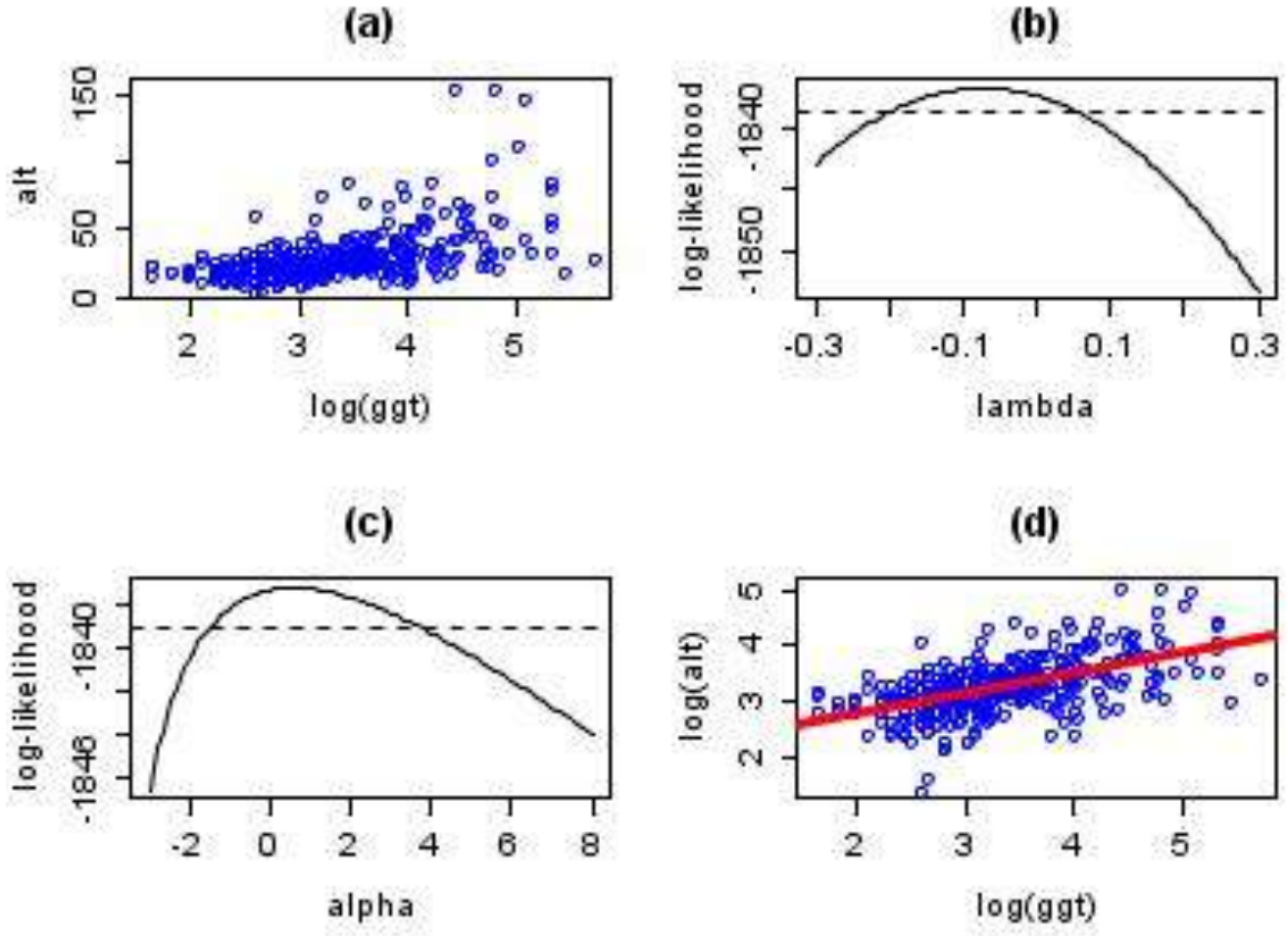
Двухпараметрический вариант:

$$y_i^{(\boldsymbol{\lambda})} = \begin{cases} \dfrac{(y_i + \lambda_2)^{\lambda_1} - 1}{\lambda_1} & \text{if } \lambda_1 \neq 0, \\ \ln(y_i + \lambda_2) & \text{if } \lambda_1 = 0, \end{cases}$$

Подгонка под более «нормальное» распределение

Преобразование Бокса-Кокса

Таргет тоже можно так преобразовывать:



Подгонка под более «нормальное» распределение

Преобразование Йео-Джонсона

Еще один вариант, который может работать с нулевыми и отрицательными значениями:

https://www.stat.umn.edu/arc/yjpower.pdf

5.Отбор признаков

Отбор признаков

1. Статистические методы
2. С помощью регуляризации L1
3. Жадный отбор
4. С помощью моделей

Отбор признаков по статистическим критериям

Пример: критерий хи-квадрат позволяет отобрать лучшие бинарные признаки для каждого класса

	Значение признака 1	**Значение признака 0**
Объект принадлежит классу	A	B
Объект не принадлежит классу	C	D

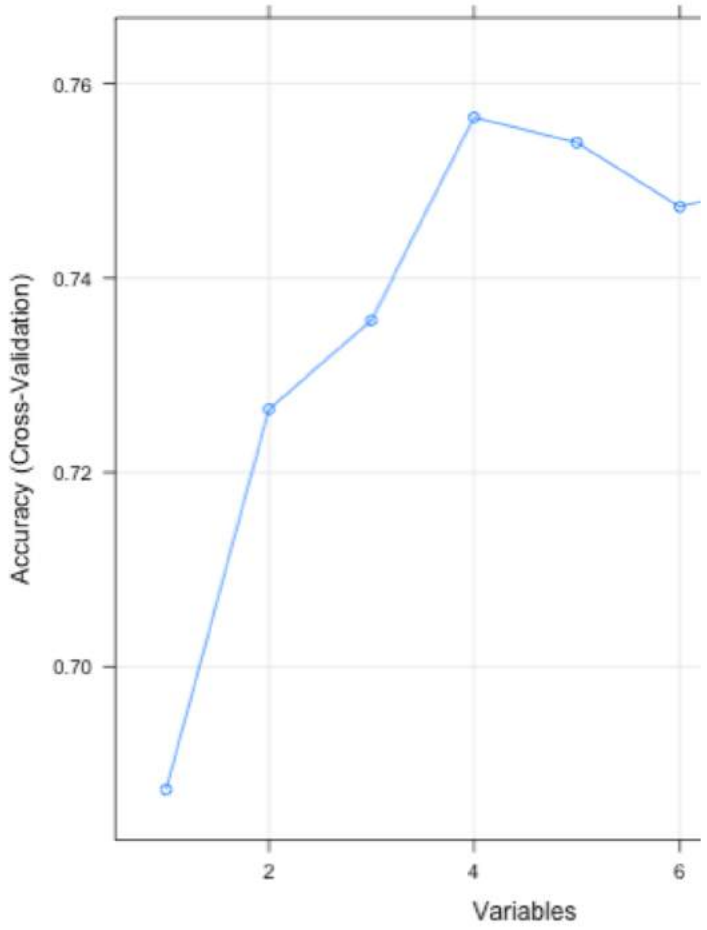
$$\chi^2 = \frac{N(AD - BC)^2}{(A + B)(A + C)(B + D)(C + D)}$$

Отбор признаков с помощью l1-регуляризации

$$\sum_{i=1}^{l} L(M_i) + \gamma \sum_{k=1}^{m} |w_k| \to min$$

Жадный отбор признаков

Чередование добавления и удаления признаков

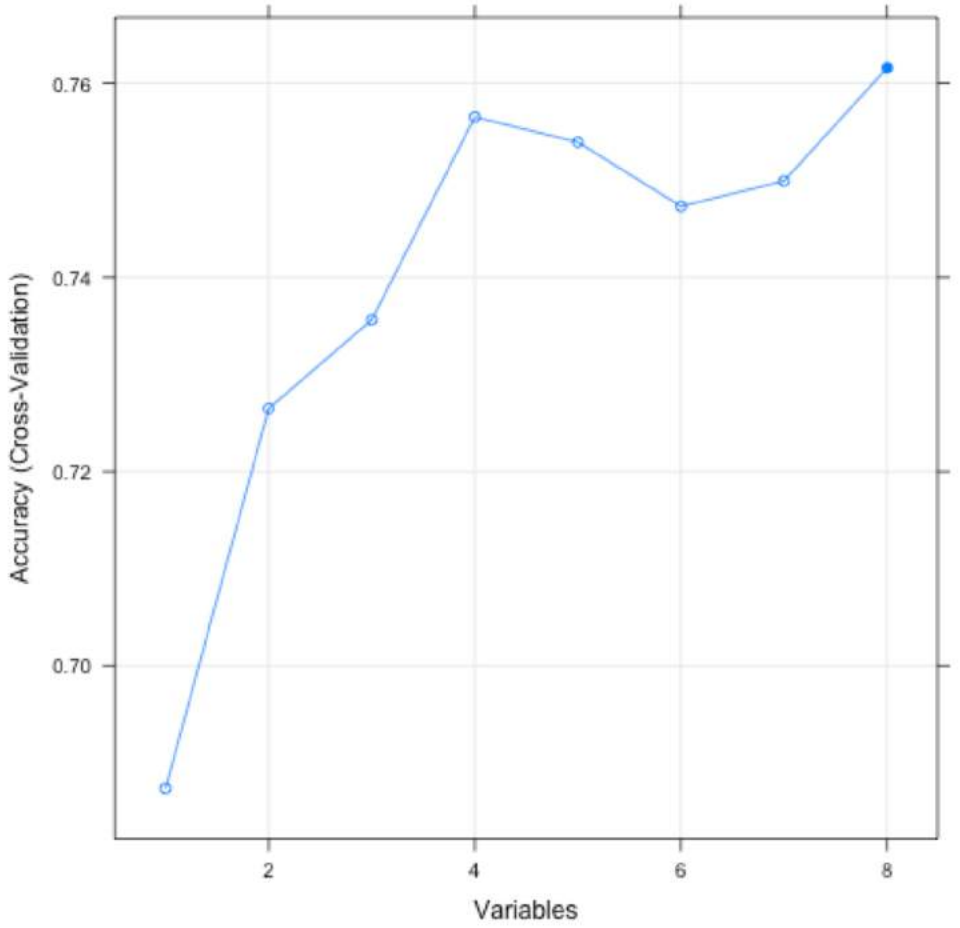


Жадный отбор признаков

Чередование добавления и удаления признаков

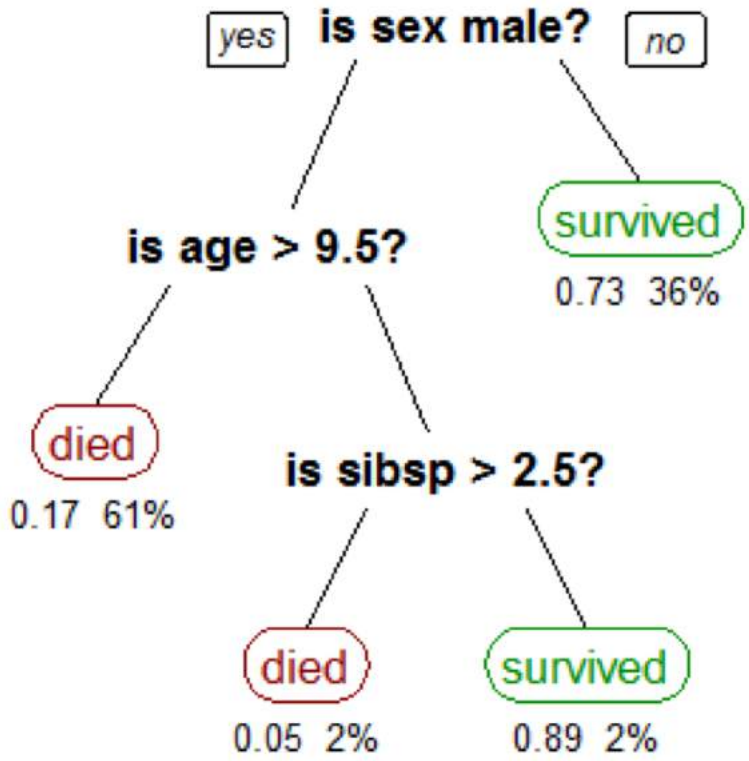
Этап добавления: добавляем лучшие признаки

Этап удаления: удаляем худшие признаки



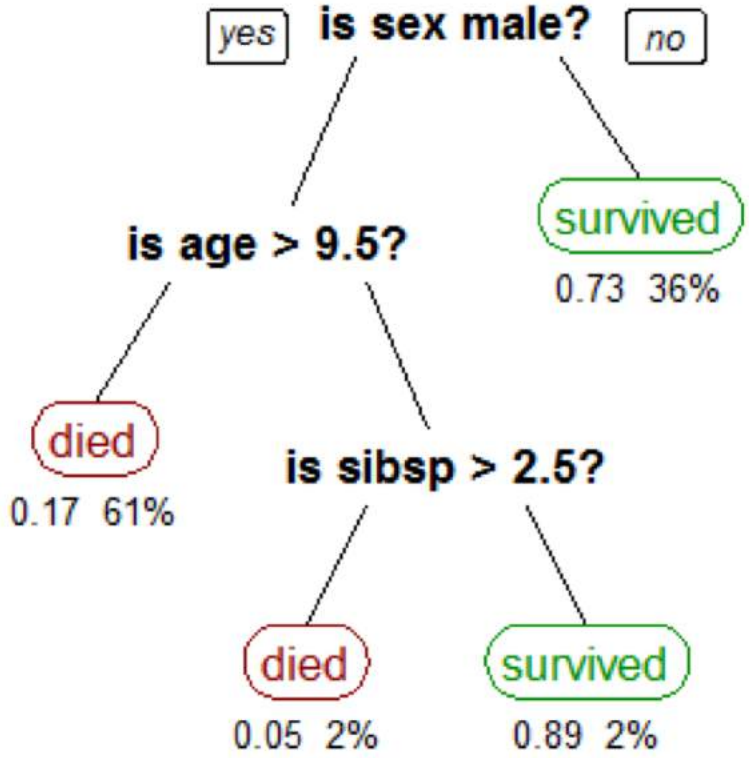
Отбор признаков с помощью моделей

Вопрос: как можно оценивать важность признака в решающих деревьях?



Отбор признаков с помощью моделей

Вопрос: как можно оценивать важность признака в решающих деревьях?



А в линейных моделях?

План

1. Пример выбора метрики
2. Анализ качества модели
3. Онлайн-качество
4. Извлечение признаков
5. Отбор признаков

Data Mining in Action

Лекция 8

Группа курса в Telegram:

https://t.me/joinchat/B1OlTk74nRV56Dp1TDJGNA