Data Mining in Action

Лекция 5. Оценка качества

Партнеры курса

МИСиС
Национальный исследовательский технологический университет

misis.ru

jet.su

На прошлой лекции

- Решающие деревья
- Ансамбли деревьев
- Общие идеи построения ансамблей
- Извлечение и простые преобразования признаков
- Отбор признаков

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. Валидация в задачах регрессии
2. Валидация при классификации
3. Пример выбора метрики
4. Стабильность модели
5. Онлайн-эксперимент

1. Валидация в задачах регрессии

Функционал ошибки (loss)

- MAE
- RMSE
- MAPE
- SMAPE
- logloss

MEAN AVERAGE ERROR

- Отклонение прогноза от исходного значения
- Усредненное по всем наблюдениям

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n} |y_j - \hat{y}_j|$$

ROOT MEAN SQUARED ERROR

- Корень из среднего квадратичного отклонения прогноза от исходного значения
- Сильнее штрафует за бОльшие по модулю отклонения

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

MEAN AVERAGE PERCENTAGE ERROR

- Ошибка прогнозирования оценивается в процентах

$$\mathrm{M} = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

SYMMETRIC MEAN AVERAGE PERCENTAGE ERROR

- Ошибка оценивается в процентах

$$\text{SMAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \frac{|F_t - A_t|}{(|A_t| + |F_t|)/2}$$

SYMMETRIC MEAN AVERAGE PERCENTAGE ERROR

- Ошибка оценивается в процентах

$$\mathrm{SMAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \frac{|F_t - A_t|}{(|A_t| + |F_t|)/2}$$

$$\mathrm{SMAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \frac{|F_t - A_t|}{|A_t| + |F_t|}$$

SYMMETRIC MEAN AVERAGE PERCENTAGE ERROR

- По-разному штрафует за перепрогнозирование и недопрогнозирование

- Перепрогнозирование:

$A_t = 100$, $F_t = 110$ ~ SMAPE = 4.76%

- Недопрогнозирование:

$A_t = 100$, $F_t = 90$ ~ SMAPE = 5.26%

Log Loss

- Логарифмическая ошибка
- Хорошо оценивает вероятность

$$\text{LogLoss} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right]$$

Почему Log Loss так выглядит

Пусть $p_i = P(y_i = 1|x_i)$, тогда $1 - p_i = P(y_i = 0|x_i)$

Почему Log Loss так выглядит

Пусть $p_i = P(y_i = 1|x_i)$, тогда $1 - p_i = P(y_i = 0|x_i)$

Теперь заметим, что выражение ${p_i}^{y_i}(1 - p_i)^{(1-y_i)}$ - просто запись вероятности того класса, к которому x_i фактически принадлежит

Почему Log Loss так выглядит

Пусть $p_i = P(y_i = 1|x_i)$, тогда $1 - p_i = P(y_i = 0|x_i)$

Теперь заметим, что выражение ${p_i}^{y_i}(1 - p_i)^{(1-y_i)}$ - просто запись вероятности того класса, к которому x_i фактически принадлежит

Произведение вероятностей фактических классов объектов из выборки – правдоподобие выборки:

$$\prod_{i=1}^{n} {p_i}^{y_i}(1 - p_i)^{(1-y_i)}$$

Почему Log Loss так выглядит

Пусть $p_i = P(y_i = 1|x_i)$, тогда $1 - p_i = P(y_i = 0|x_i)$

Теперь заметим, что выражение ${p_i}^{y_i}(1 - p_i)^{(1-y_i)}$ - просто запись вероятности того класса, к которому x_i фактически принадлежит

Произведение вероятностей фактических классов объектов из выборки – правдоподобие выборки:

$$\prod_{i=1}^{n} {p_i}^{y_i}(1 - p_i)^{(1-y_i)}$$

Если взять логарифм и умножить на -1 – получим log loss

Log Loss константного прогноза

Рассмотрим выборку из n объектов с одинаковыми векторами признаков x, на pn из которых таргет равен 1, а на остальных – 0.

Log Loss константного прогноза

Рассмотрим выборку из n объектов с одинаковыми векторами признаков x, на pn из которых таргет равен 1, а на остальных – 0.

Пусть $a(x) = c$, тогда log loss минимален при:

$$\left(\sum_{i=1}^{n} y_i \ln c + (1 - y_i) \ln(1 - c) \right)'_c = 0$$

Log Loss константного прогноза

$$\left(\sum_{i=1}^{n} y_i \ln c + (1 - y_i)\ln(1 - c)\right)'_c = 0$$

$$\frac{pn}{c} - \frac{n - pn}{1 - c} = 0$$

$$pn - cpn = cn - cpn$$

$$pn = cn$$

$$c = p$$

История про MAE вместо log loss

- Заказчик очень хотел, чтобы алгоритм оценивал вероятности в задаче бинарной классификации
- Немного знал про функции потерь
- Просил решать задачу регрессии на ответах 0 и 1 оптимизируя MAE, думал ответы будут между 0 и 1
- Ответы получились только 0 и 1

Упражнение

1. Показать, что если вместо log loss оптимизировать MAE в задаче с ответами 0 и 1, прогноз алгоритма будет округляться к 0 или к 1

2. Показать, что константный прогноз в регрессии, оптимизирующий MSE – среднее значение таргетов

2. Валидация в задаче классификации

Метрики качества

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

Accuracy

Доля правильных ответов при классификации

Accuracy

Доля правильных ответов при классификации

target: 1 0 1 0 0 0 0 1 0 0

Accuracy

Доля правильных ответов при классификации

target: 1 0 1 0 0 0 0 1 0 0

predicted: 0 0 1 0 0 0 0 1 1 0

Accuracy

Доля правильных ответов при классификации

target: 1 0 1 0 0 0 0 1 0 0

predicted: 0 0 1 0 0 0 0 1 1 0

Accuracy

Доля правильных ответов при классификации

target: 1 0 1 0 0 0 0 1 0 0

predicted: 0 0 1 0 0 0 0 1 1 0

accuracy = 8/10 = 0.8

Метрики качества

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

Precision & Recall

- Precision – точность
- Recall - полнота

Сбитые самолеты

Сбитые самолеты

$y = (0\ 0\ 0\ 0\ 1\ 0\ 1\ 1\ 0\ 1)$

$\hat{y} = (0\ 1\ 1\ 0\ 1\ 0\ 0\ 1\ 0\ 1)$

Precision

Precision – точность выстрелов:

$$\frac{\text{Количество сбитых самолётов}}{\text{Количество выстрелов}}$$

$y = (0\ 0\ 0\ 0\ 1\ 0\ 1\ 1\ 0\ 1)$

$\hat{y} = (0\ 1\ 1\ 0\ 1\ 0\ 0\ 1\ 0\ 1)$

Recall

Recall – «полнота» сбивания самолетов:

$$\frac{\text{Количество сбитых самолётов}}{\text{Общее количество самолётов}}$$

$y = (0\ 0\ 0\ 0\ 1\ 0\ 1\ 1\ 0\ 1)$

$\hat{y} = (0\ 1\ 1\ 0\ 1\ 0\ 0\ 1\ 0\ 1)$

Обычно объясняется так:

		Actual Class	
		Yes	No
Predicted Class	Yes	**T**rue **P**ositive	**F**alse **P**ositive
	No	**F**alse **N**egative	**T**rue **N**egative

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

F-measure (F-score, F1)

- Среднее гармоническое между precision и recall:

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

- Значение F-measure ближе к меньшему из precision и recall

Метрики качества

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

ROC-AUC

- Применяется для оценки «вероятностной» классификации*
- «Качество» ранжирования объектов по вероятности принадлежности к целевому классу
- Доля «правильно» отранжированных пар
- Вероятность встретить объект целевого класса раньше, чем объект нецелевого класса

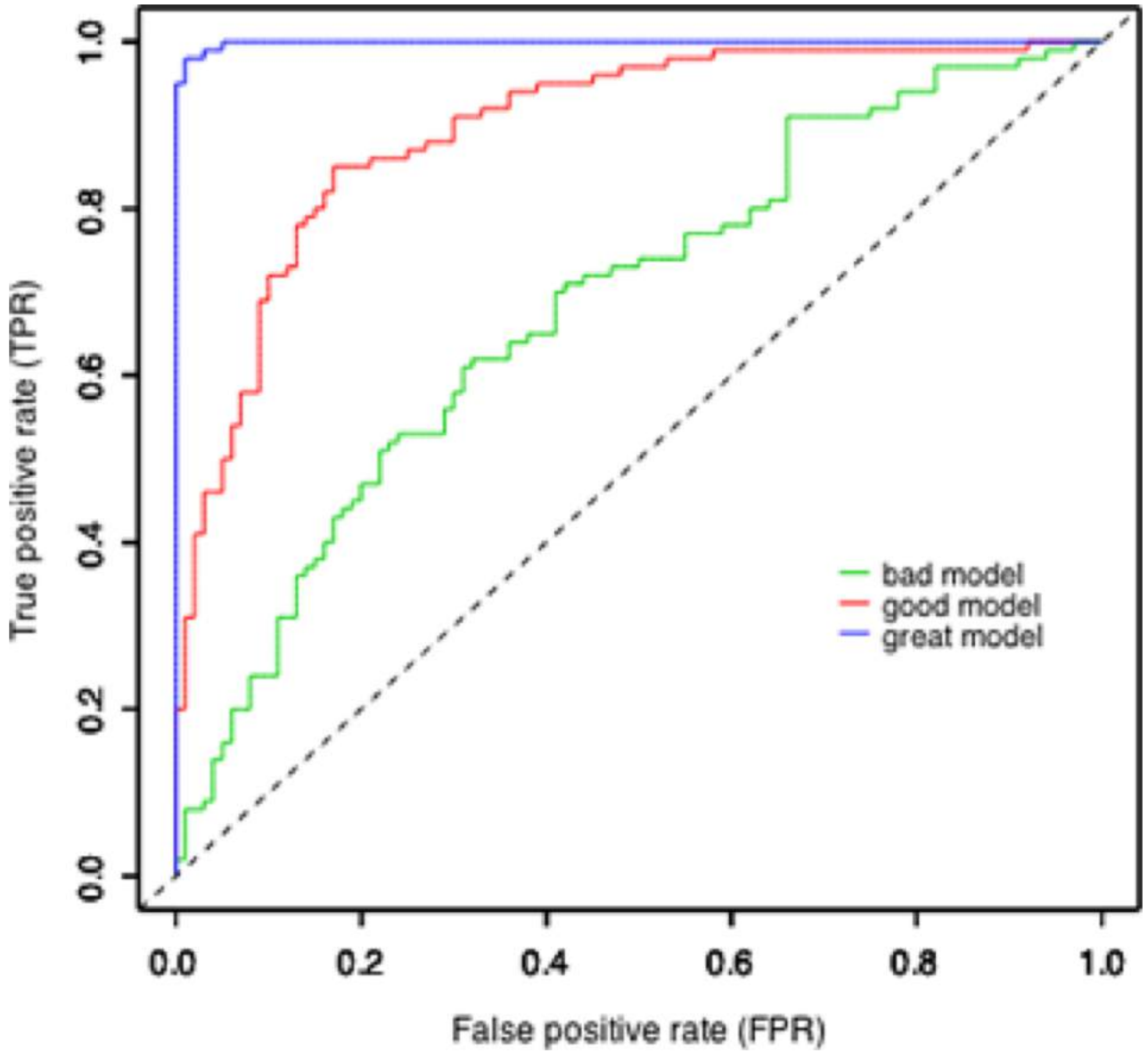
ROC

		Actual Class	
		Yes	No
Predicted Class	Yes	**T**rue **P**ositive	**F**alse **P**ositive
	No	**F**alse **N**egative	**T**rue **N**egative

$$TPR = \frac{True\ positives}{True\ positives + False\ negatives}$$

$$FPR = \frac{False\ positives}{False\ positives + True\ negatives}.$$

ROC


True positive rate (TPR)
False positive rate (FPR)
0.0
0.2
0.4
0.6
0.8
1.0
bad model
good model
great model

ROC

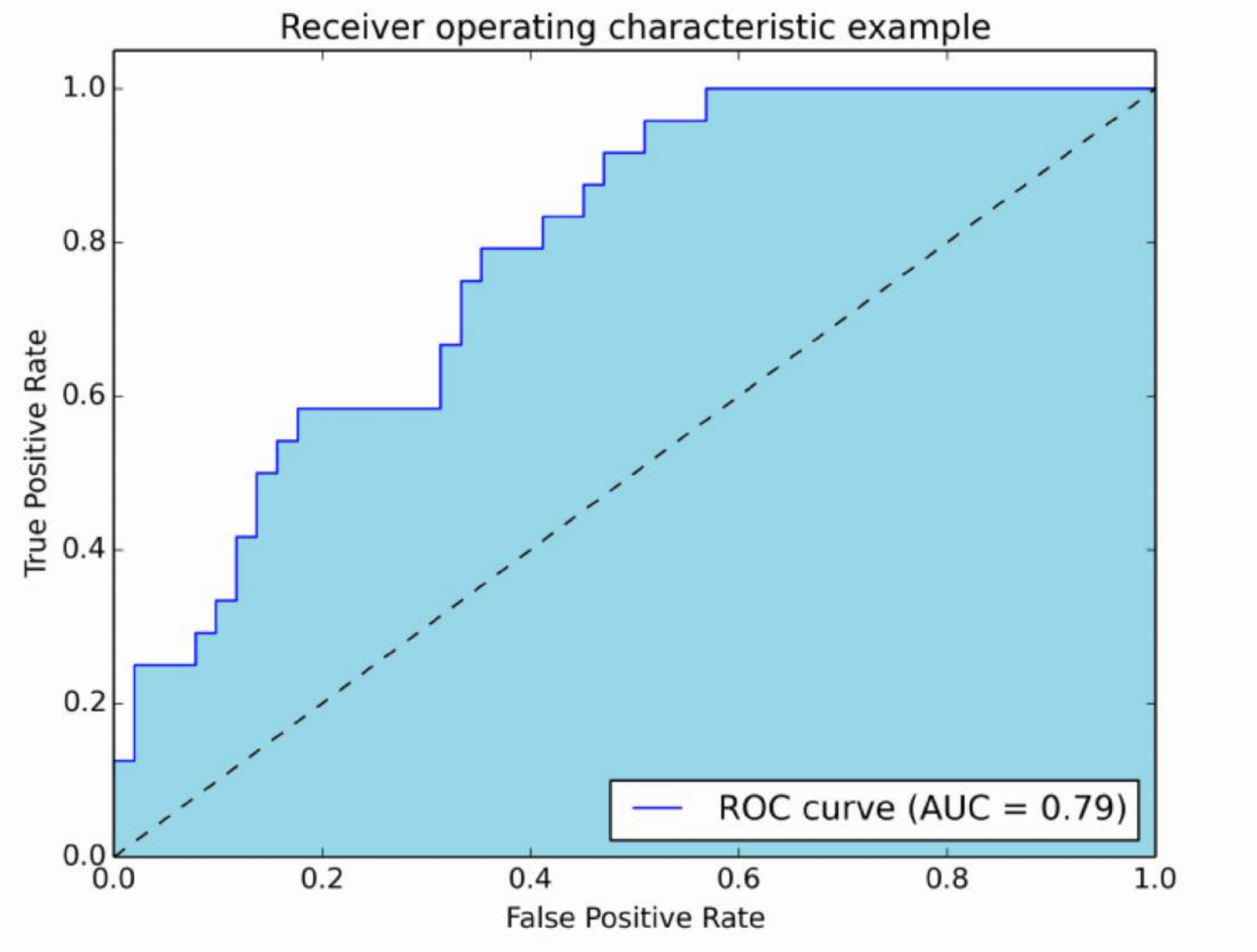
- Как оценить кривую численно?

ROC-AUC

- Как оценить кривую численно?

- **Измерить площадь под кривой** – area under the curve!

ROC-AUC

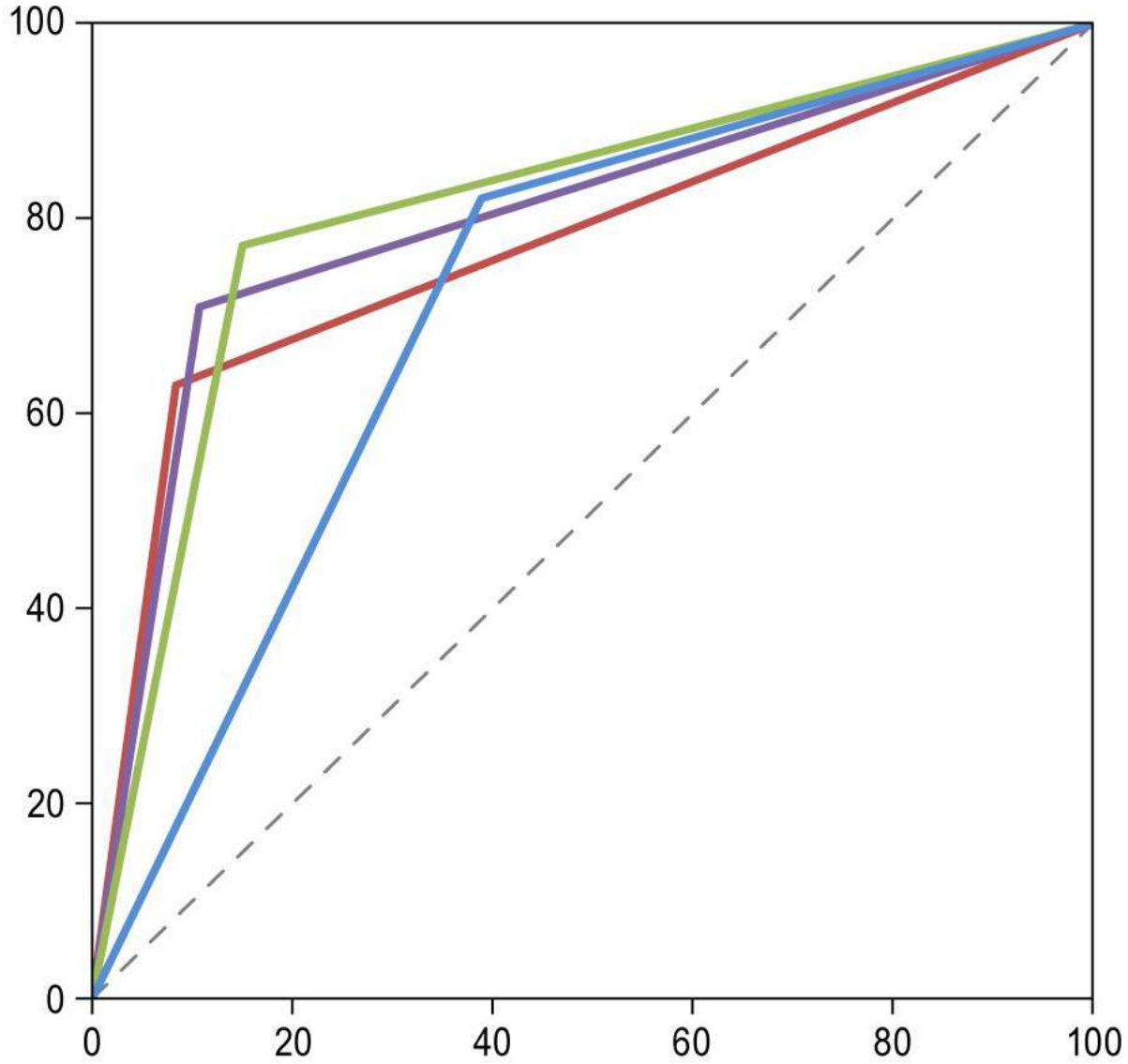

Receiver operating characteristic example
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (AUC = 0.79)

ROC-AUC по-простому

Рассмотрим всевозможные пары объектов из выборки. ROC-AUC – доля тех пар, которые алгоритм отранжировал правильно.

История про ROC-AUC по 0 и 1

История про ROC-AUC по 0 и 1



Упражнение

- Показать, что треугольный ROC-AUC для константного ответа равен 0.5
- Показать, что треугольный ROC-AUC для случайного ответа 0 или 1 (с любой вероятностью ответа 1) – тоже равен 0.5
- Показать, что обычный ROC-AUC для случайных ответов из равномерного распределения на [0, 1] равен 0.5

Дополнительные материалы

Рассказ про ROC-AUC в блоге Александра Дьяконова:

https://dyakonov.org/2017/07/28/auc-roc-площадь-под-кривой-ошибок/

Подумайте, почему мы вводили ROC-AUC **не** с помощью движения по сетке вправо и вверх (подсказка: ответ кроется в шаге с сортировкой)

3. 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

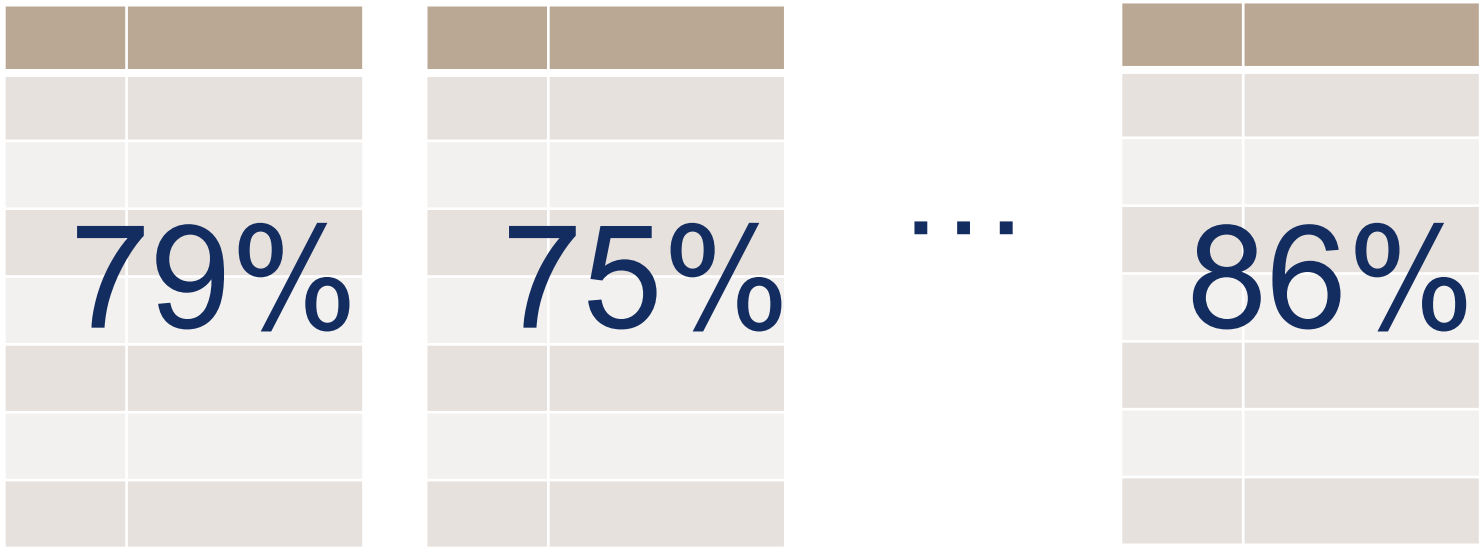
Качество классификации против качества рекомендаций

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

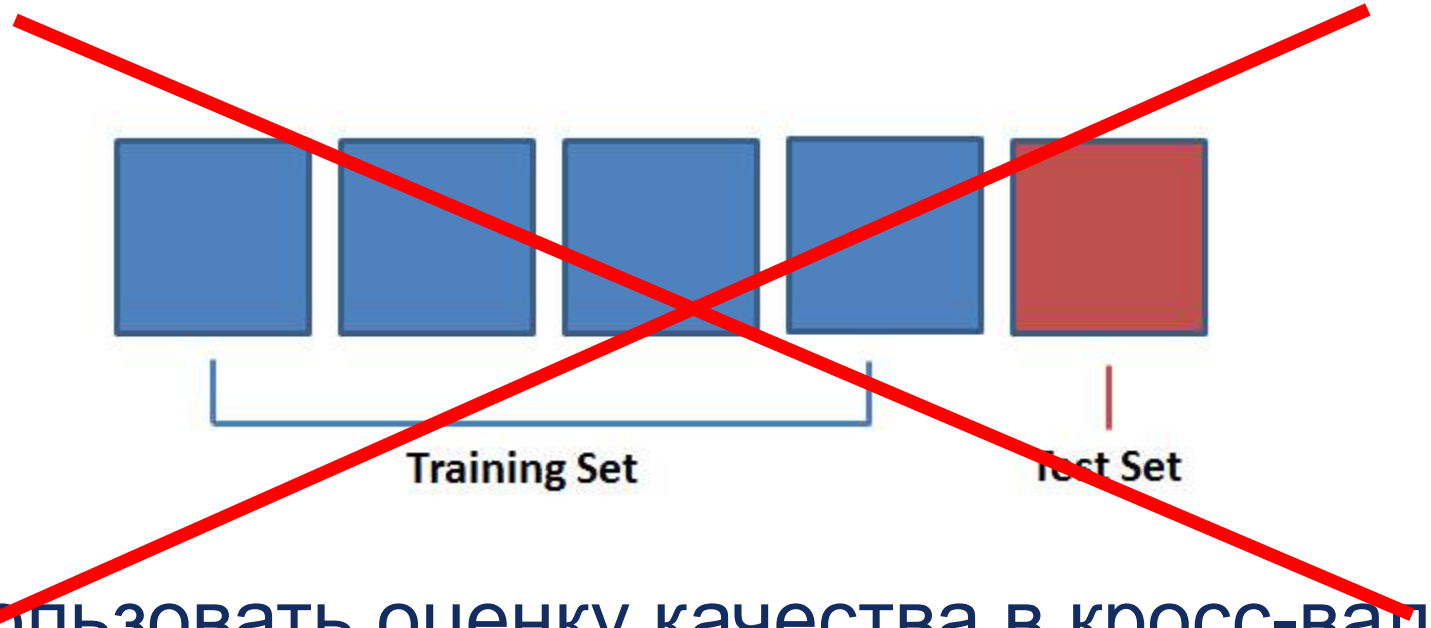
4. Стабильность моделей

Проблема: разброс на разных данных



Шаг 1: усреднение качества в CV

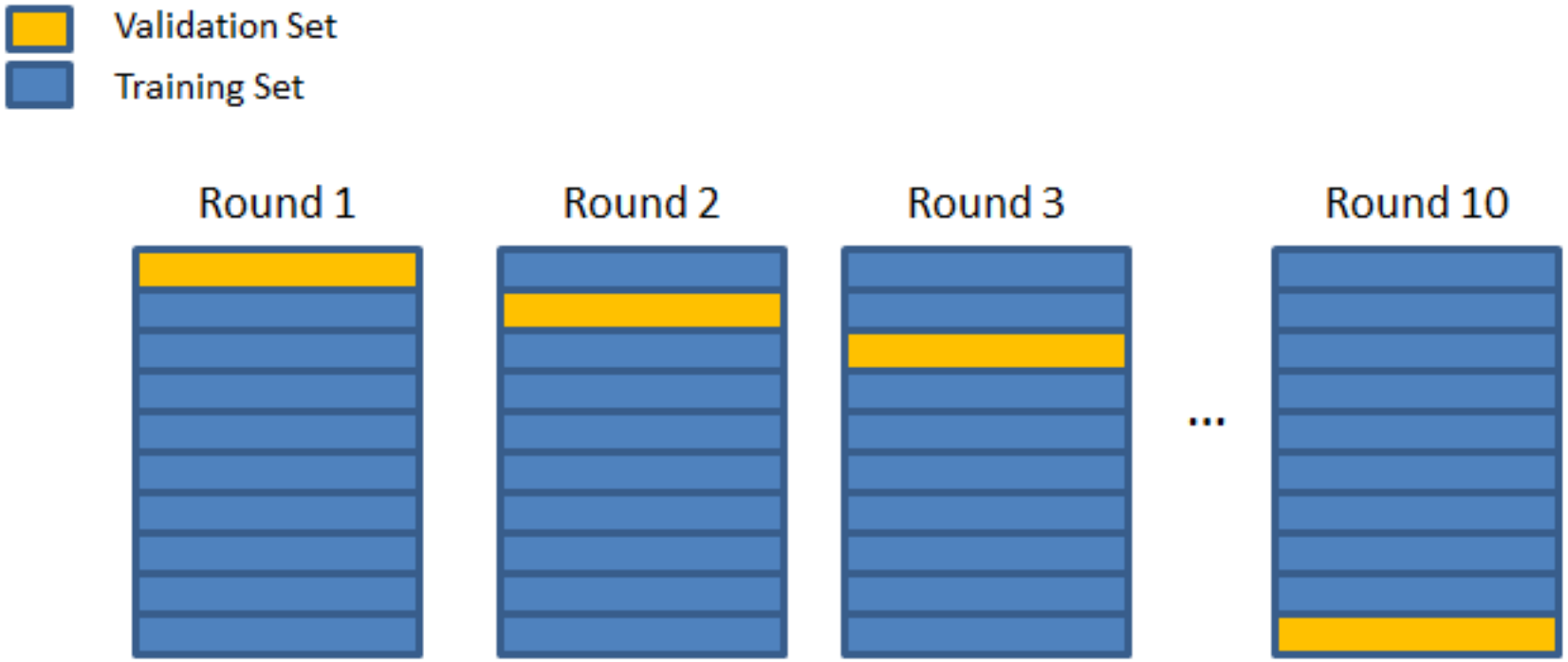
Если есть проблема со стабильностью модели, точно нужно избегать оценок на одном фиксированном датасете



Нужно использовать оценку качества в кросс-валидации

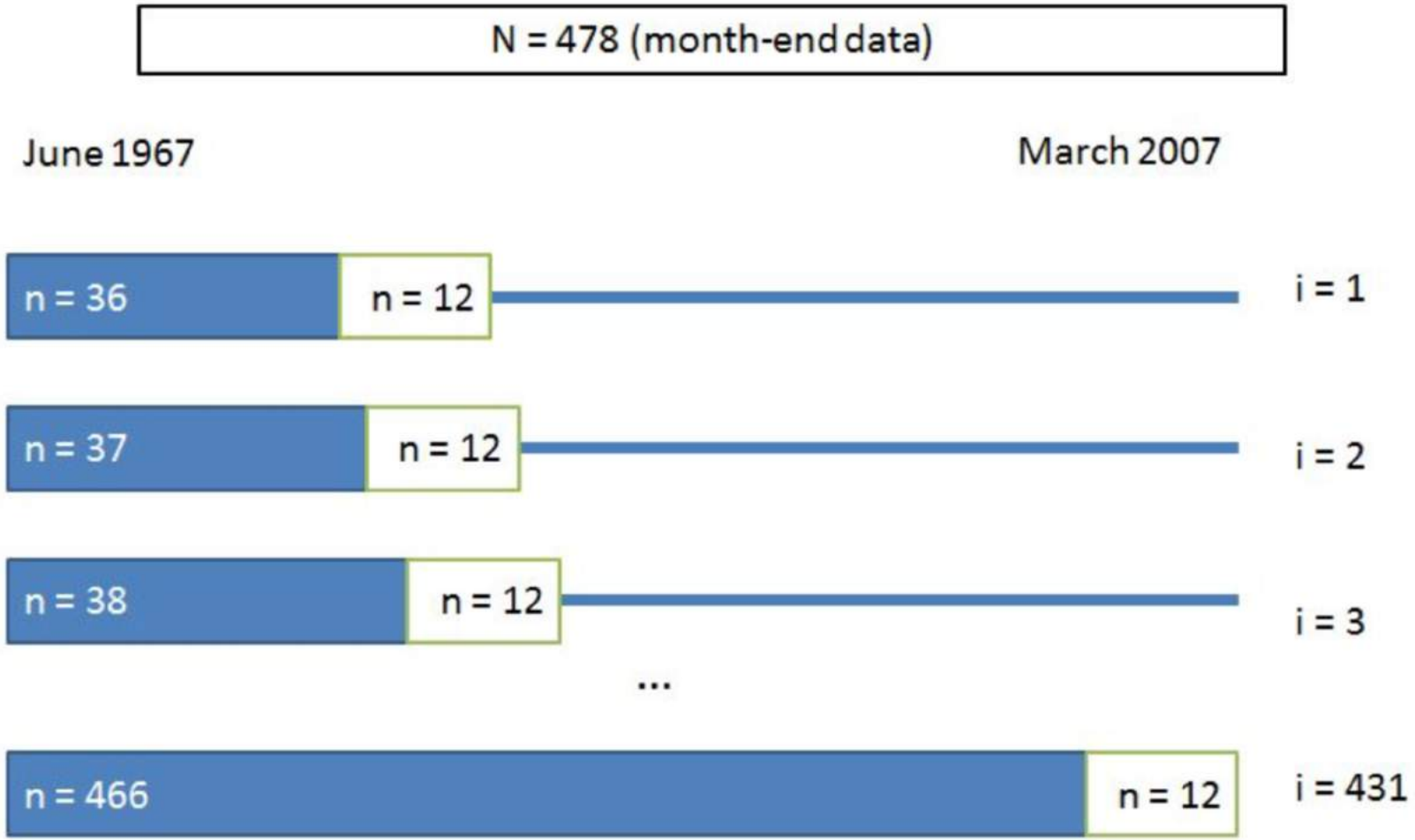
Кросс-валидация

K-Fold cross validation:

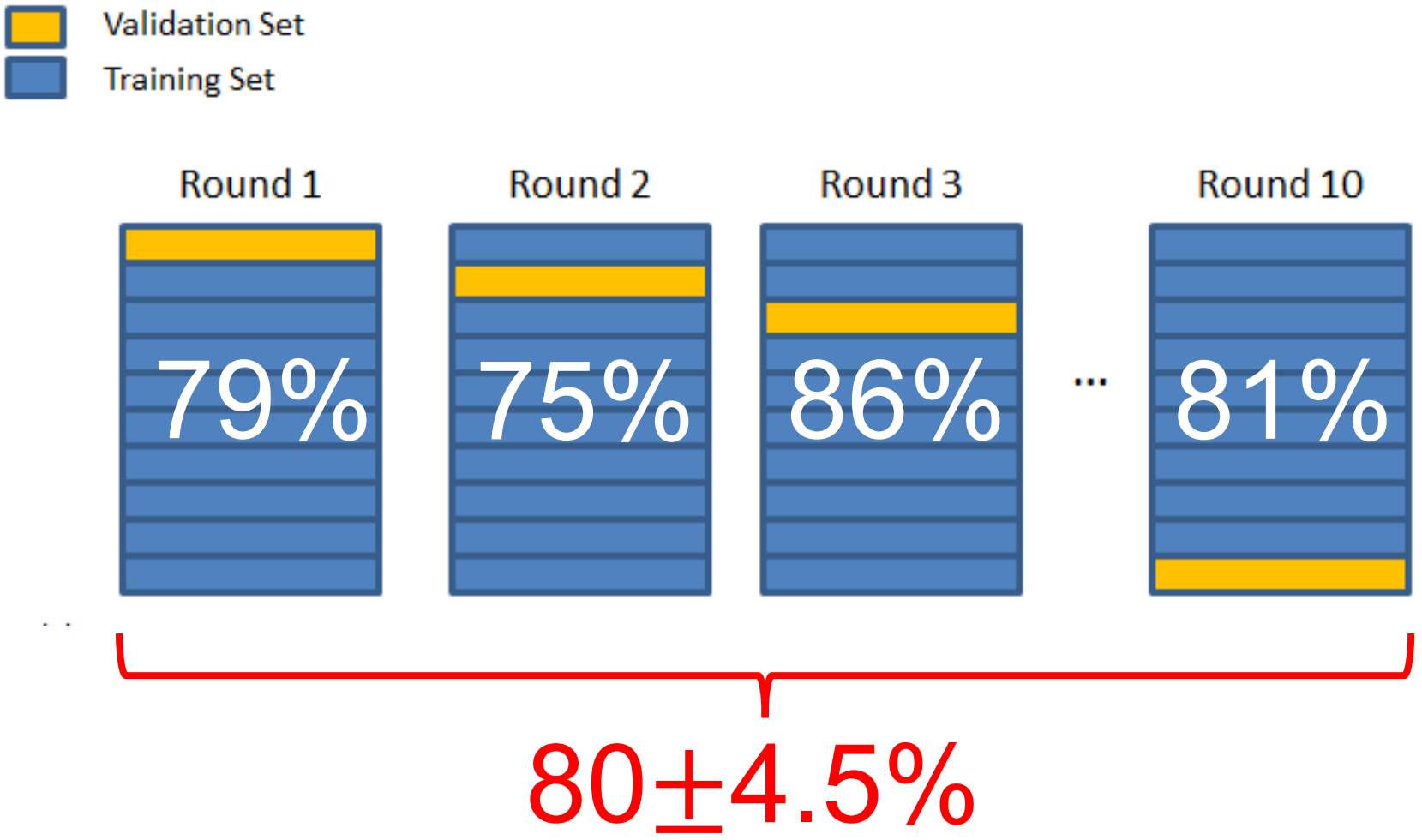


На ка

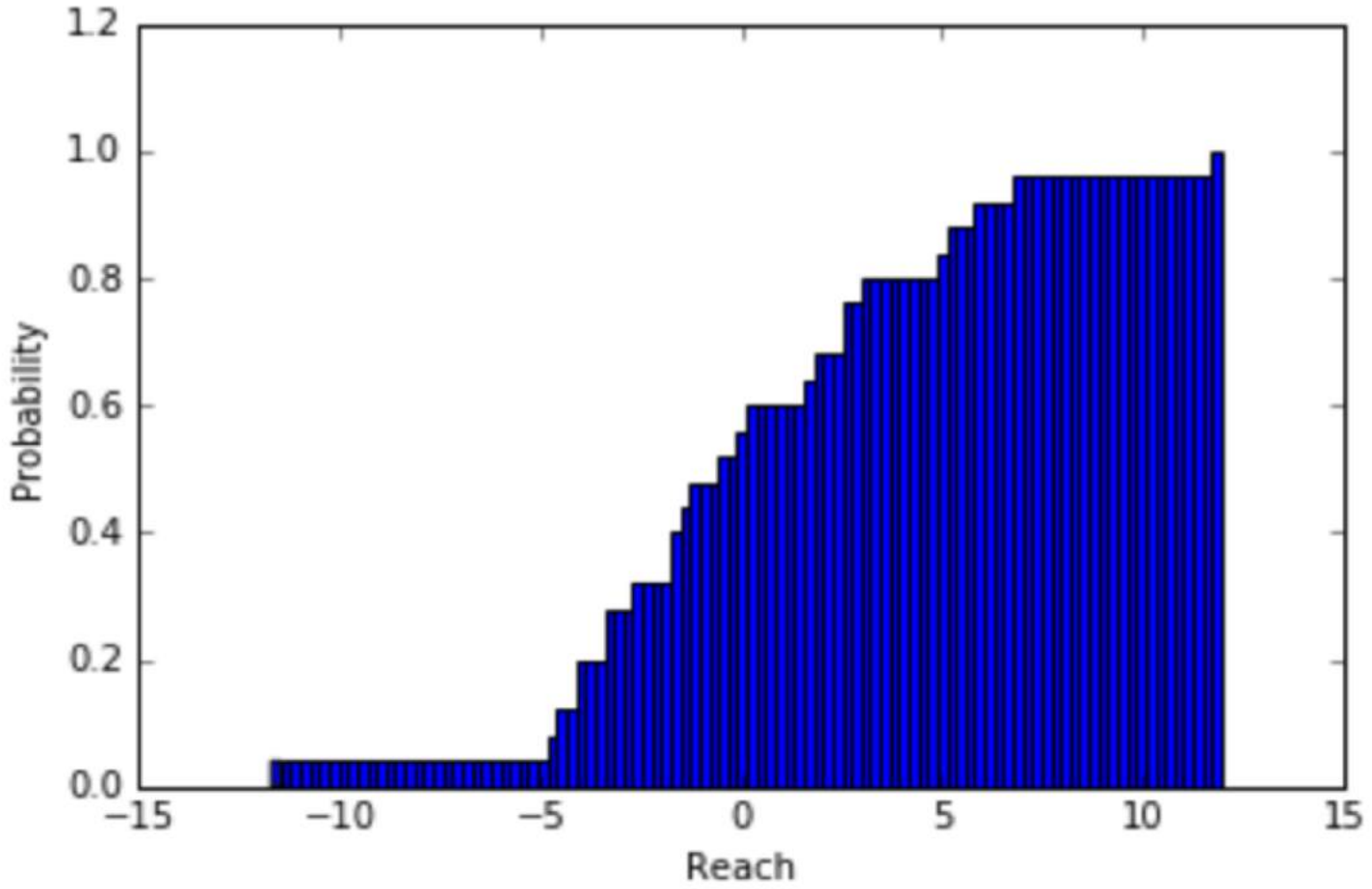
Предупреждение: будьте осторожны с CV



Шаг 2: учет разброса и распределения в CV



Шаг 2: учет разброса и распределения в CV



Шаг 3: анализ топа важных признаков

На одном фолде:

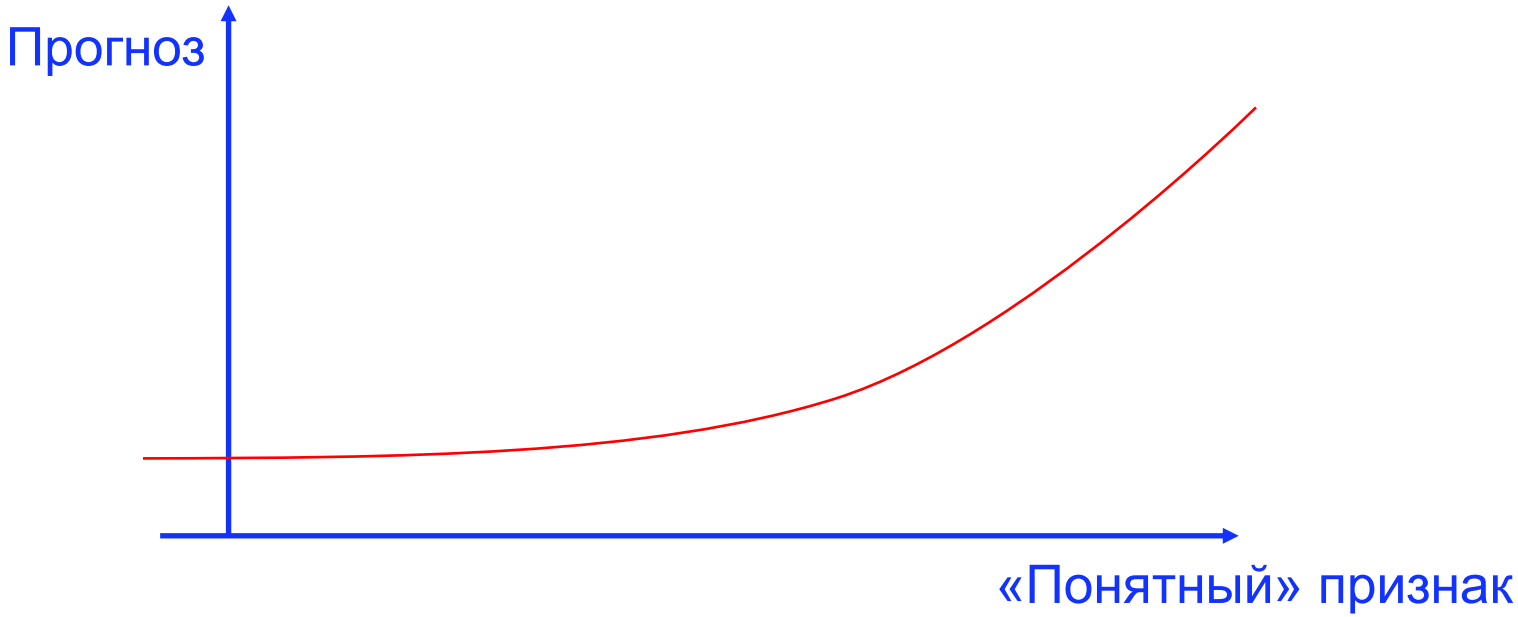
0.211268 Номер
0.147105 Ширина
0.128326 Вес
0.0954617 Параметр 1
0.0688576 Высота
0.057903 Параметр 2
0.0438185 Параметр 3
…

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
…

Шаг 4: Анализ зависимости от признаков

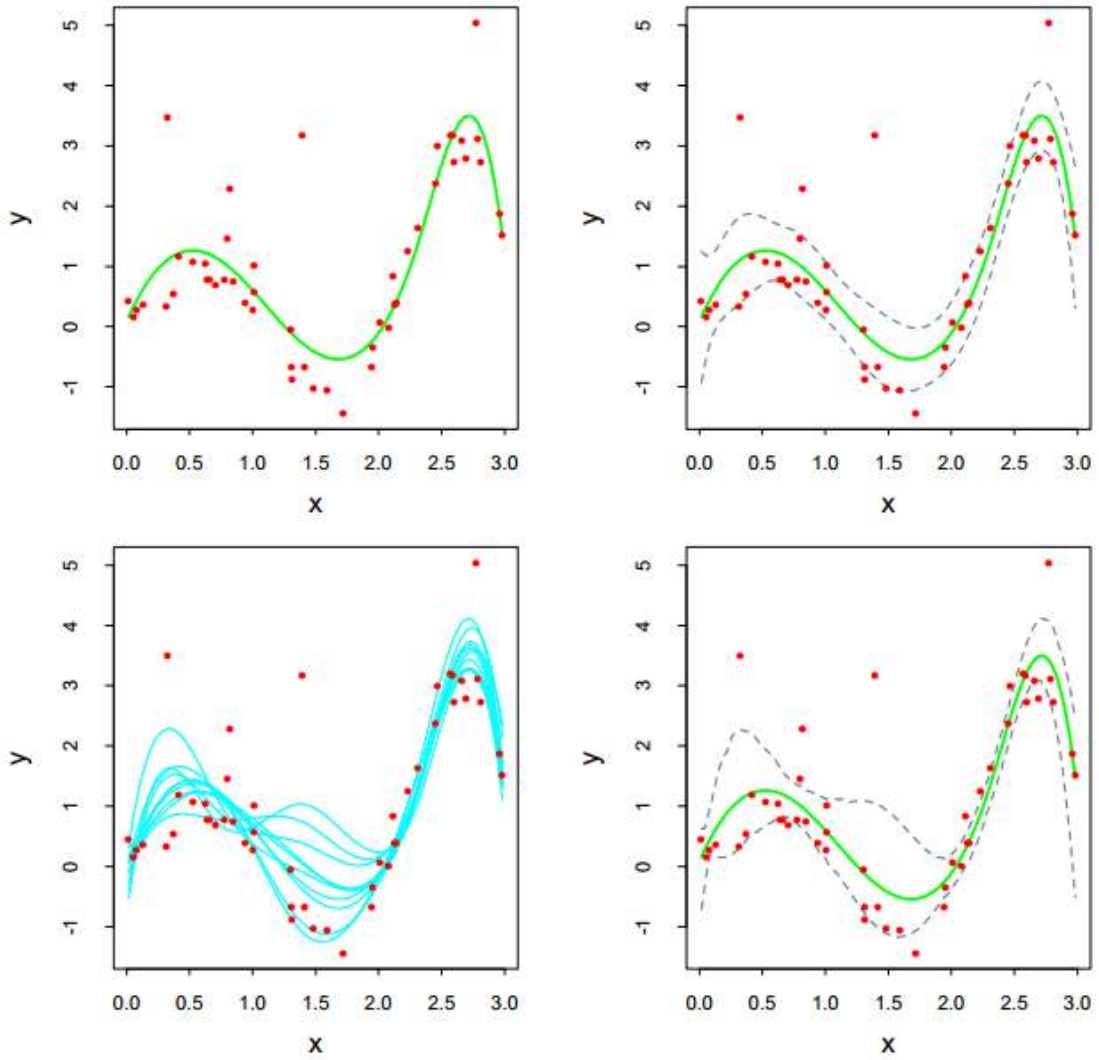
Если зависимость от каких-то признаков должна иметь понятный вид, можем поменять их (построить «искусственные» примеры) и посмотреть, как ведет себя прогноз



Шаг 5: Уменьшение разброса

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

Bagging



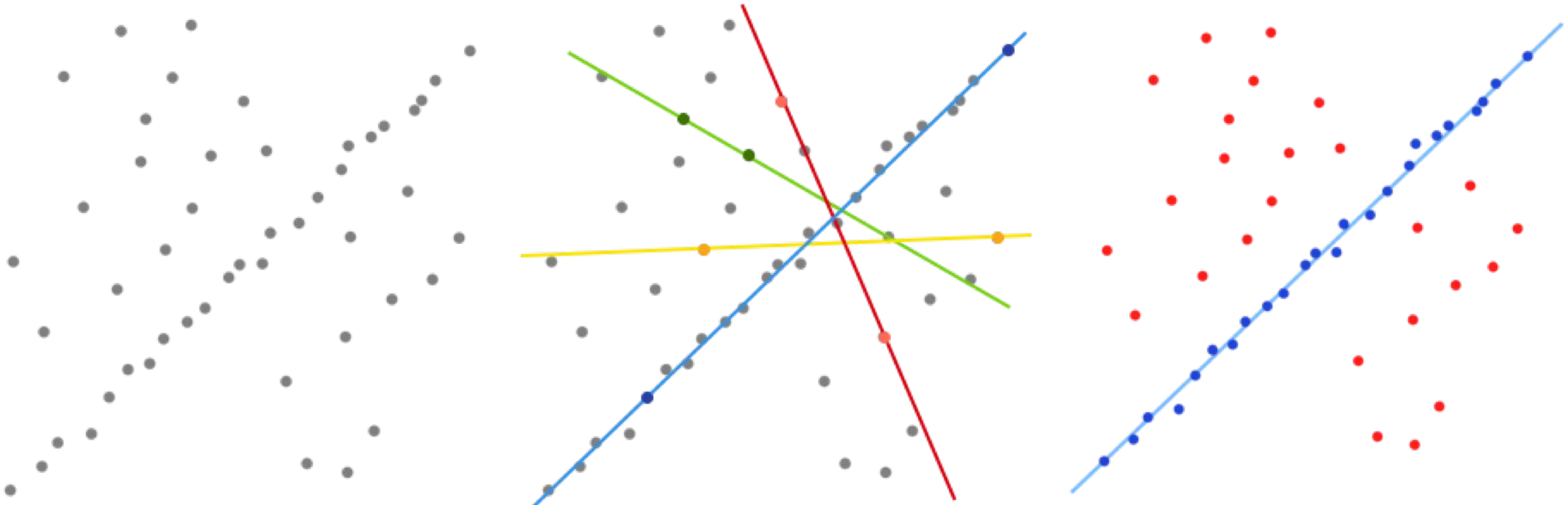
Бэггинг в sklearn.ensembles

- BaggingRegressor
- BaggingClassifier

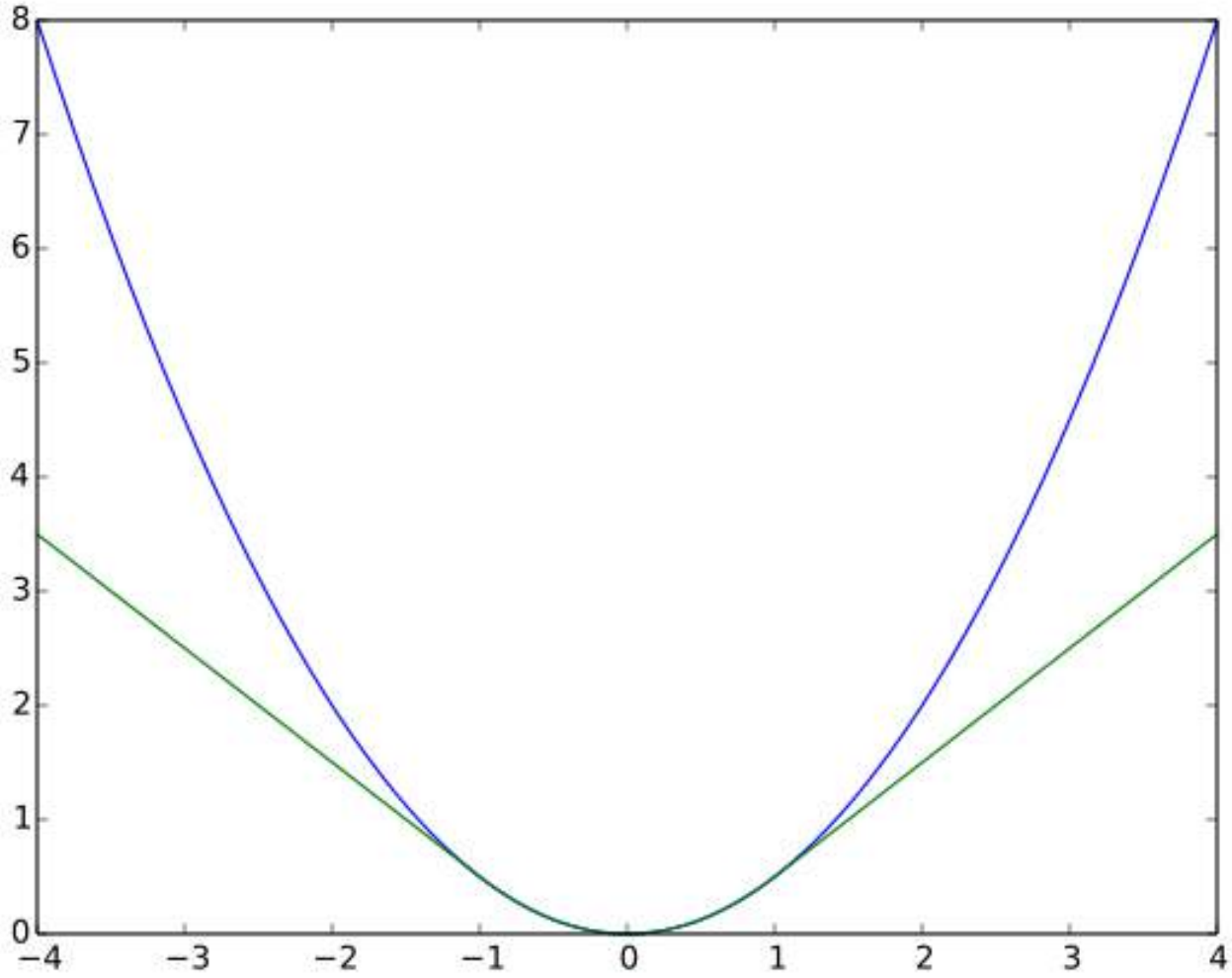
Робастные модели в sklearn.linear_model

- RANSACRegressor
- HuberRegressor
- Theil-Sen Regressor

RANSACRegressor



HuberRegressor



5. 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

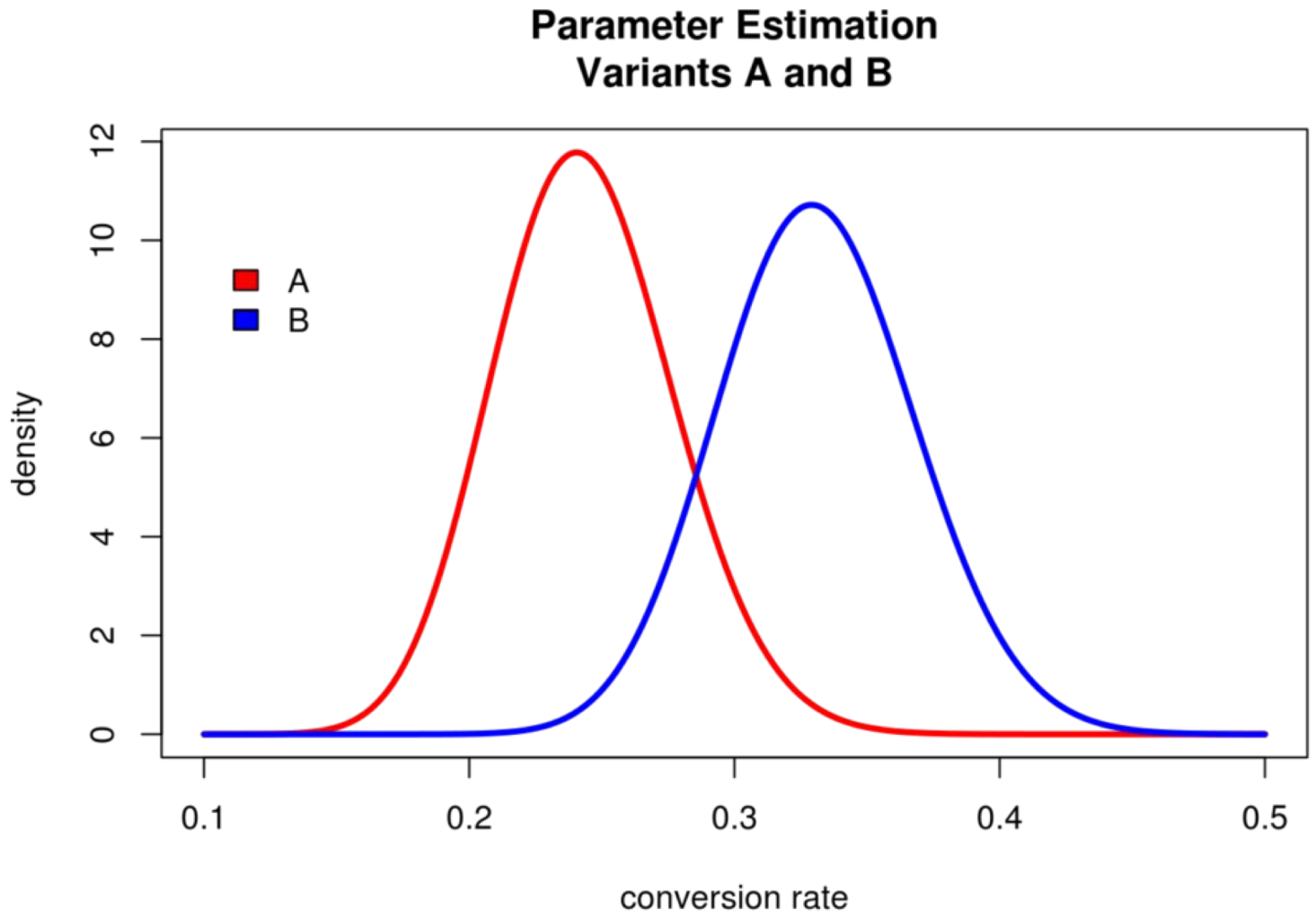
1. Случайным образом делим пользователей на равные группы
2. Измеряем целевые метрики (например, конверсию, количество заказов или доход) в каждой группе за длительный период времени
3. Получаем какое-то число для каждой группы
4. Что дальше?

Статистическая значимость: пример

Статистическая значимость: пример

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

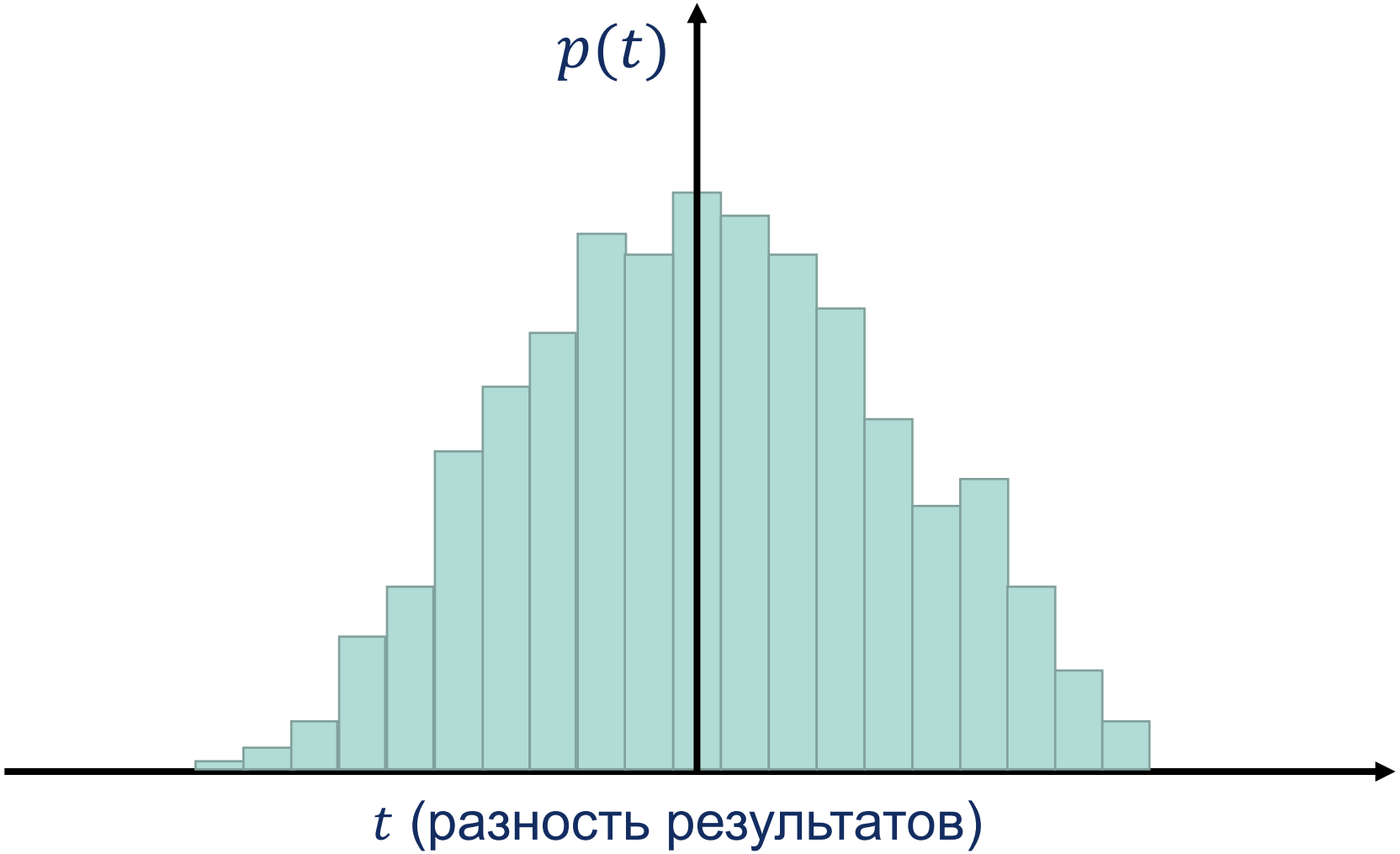
Чуть подробнее о происходящем в A/B тесте

- 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

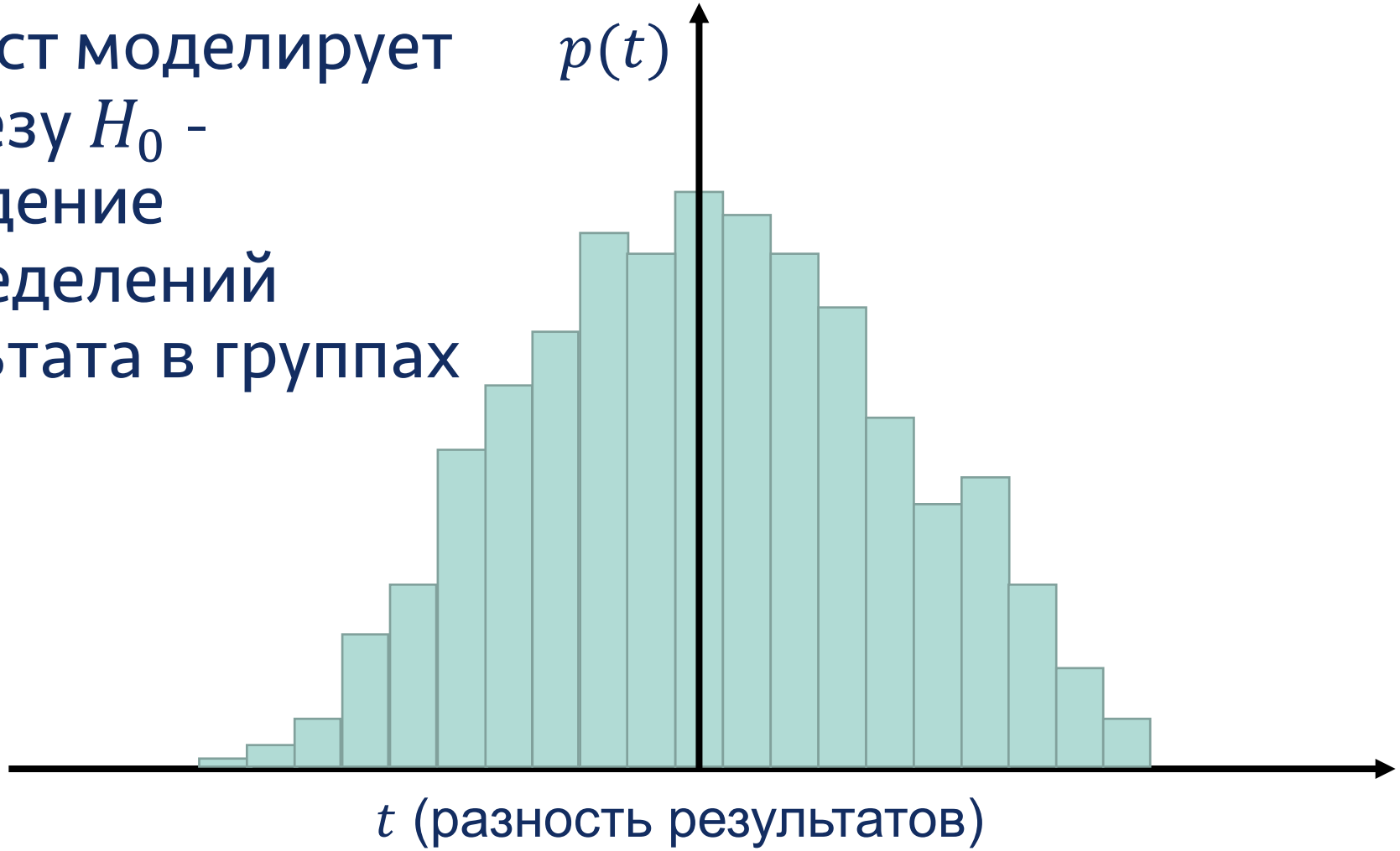
1. Имитируем A/A тест на исторических данных, N раз случайно разбив на две группы и посчитав результаты в каждой
2. Строим распределение разности результатов в группах
3. По этому распределению оцениваем вероятность получить в A/A тесте такую же разность как в A/B

Гистограмма распределения из A/A тестов

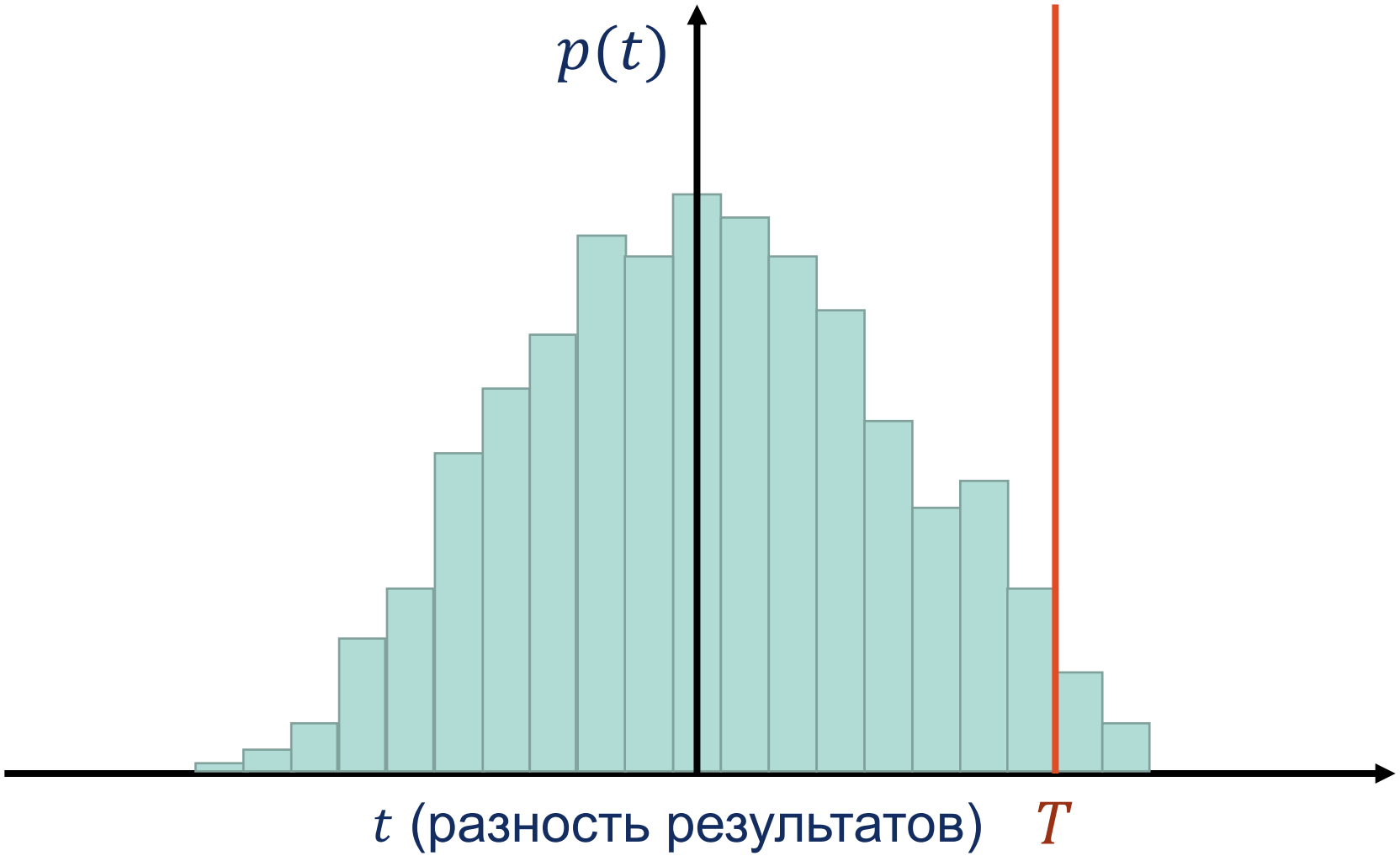


Гистограмма распределения из A/A тестов

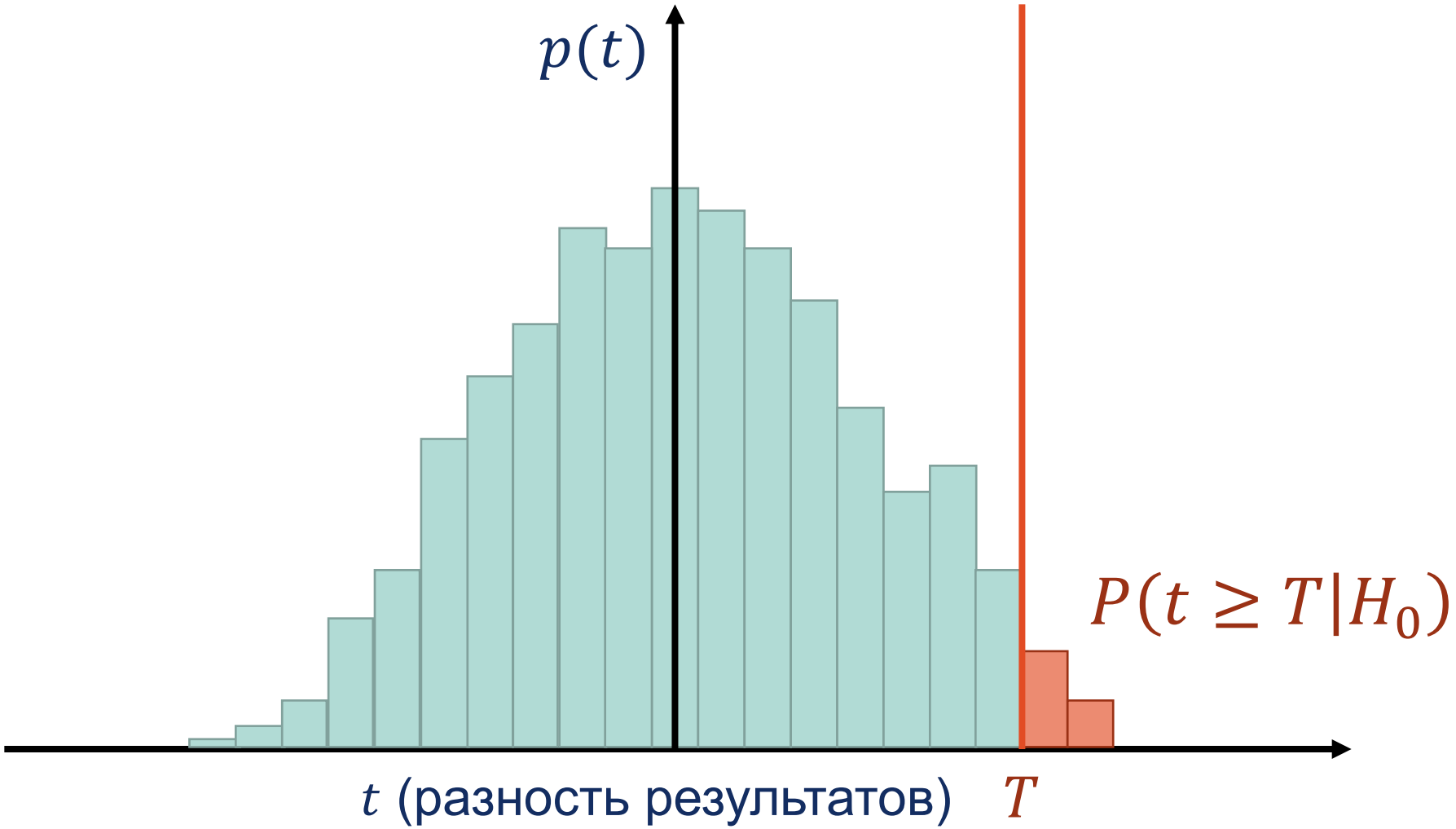
A/A тест моделирует гипотезу H_0 - совпадение распределений результата в группах



Разность из A/B теста на гистограмме

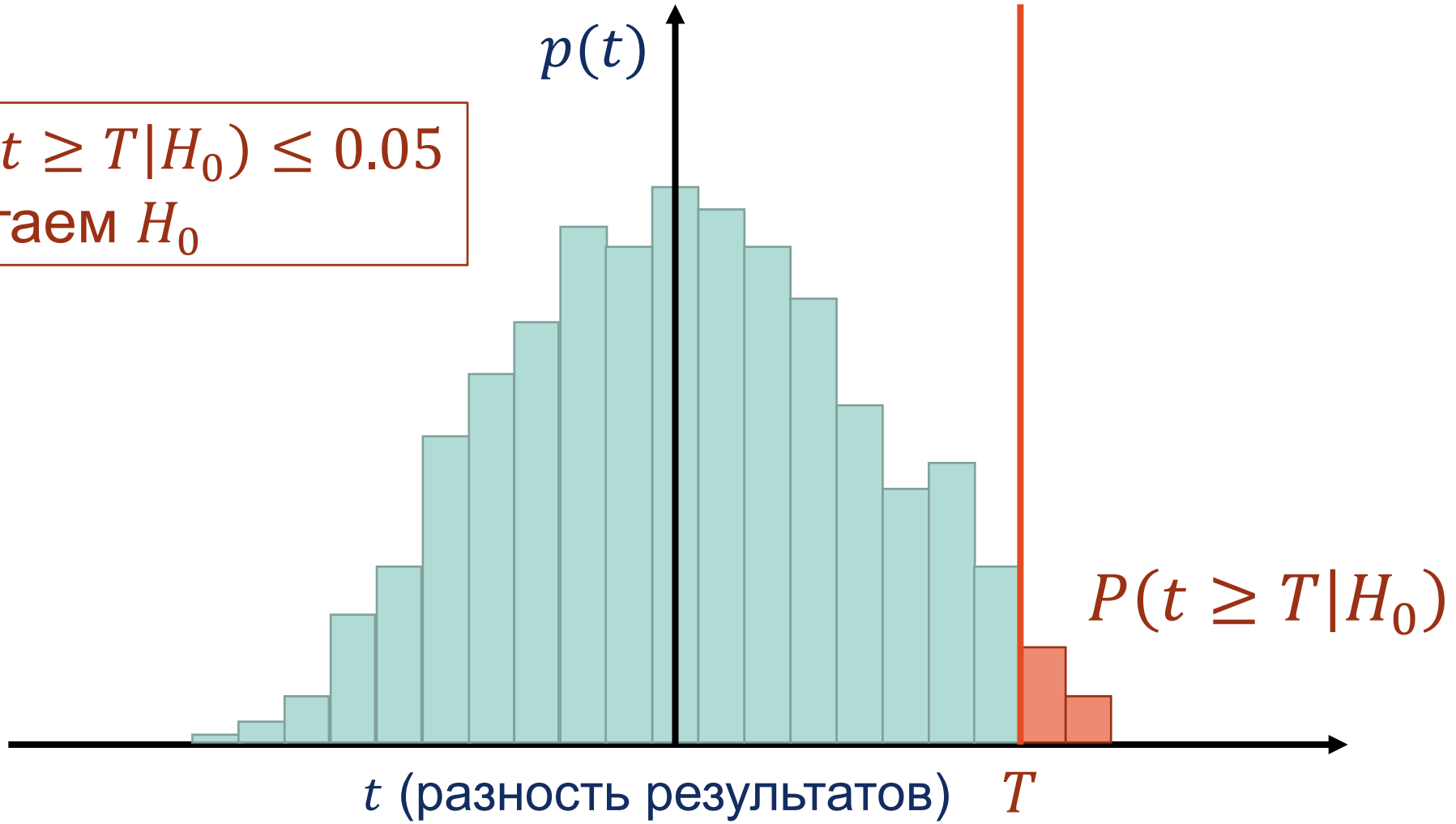


Вероятность не меньшего отклонения в A/A



Вероятность не меньшего отклонения в A/A

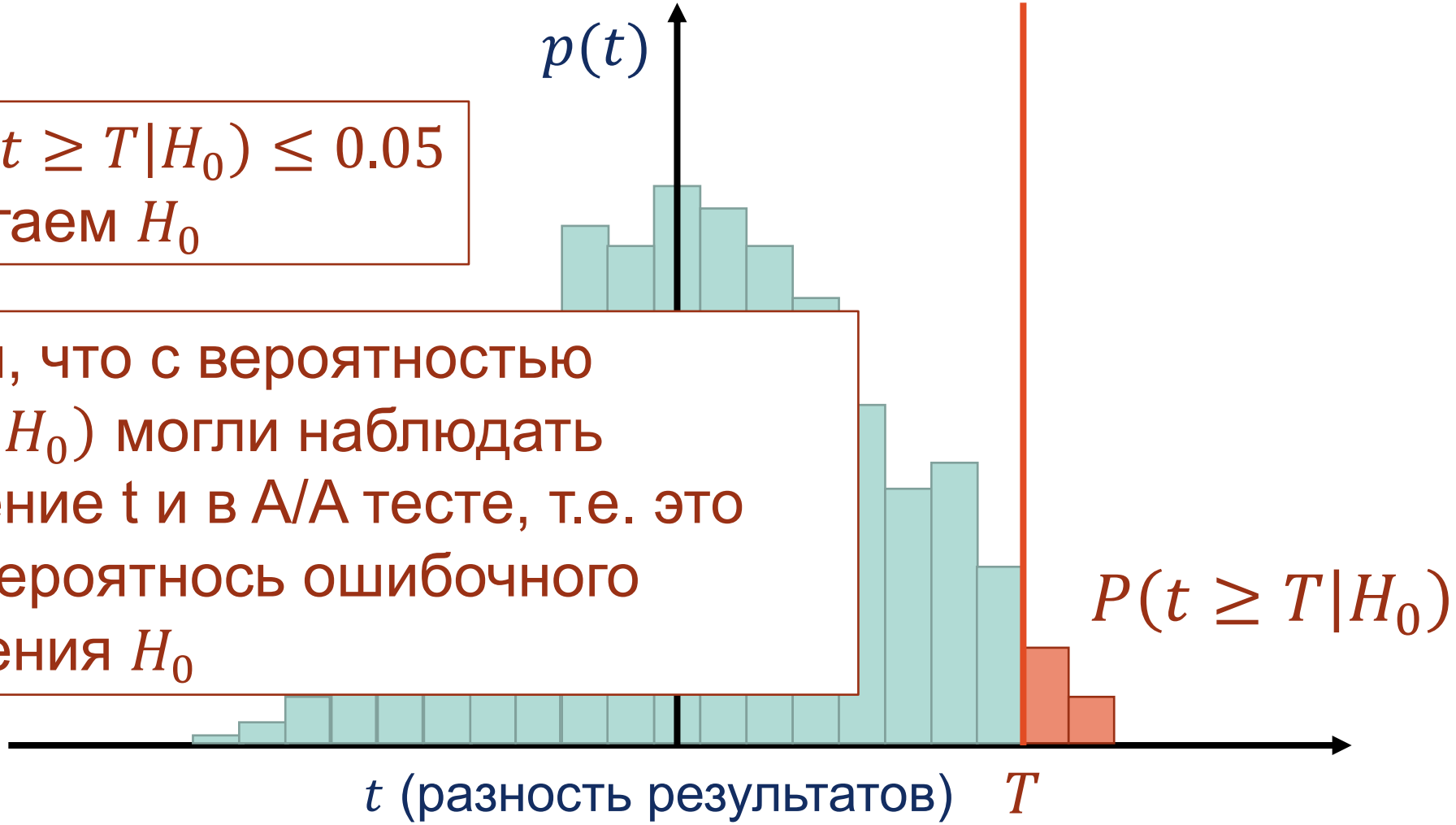
Если $P(t \geq T|H_0) \leq 0.05$
– отвергаем H_0



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

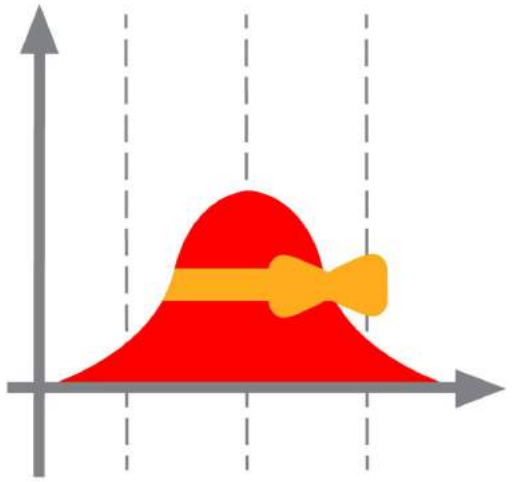


Подробнее о проверке гипотез

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

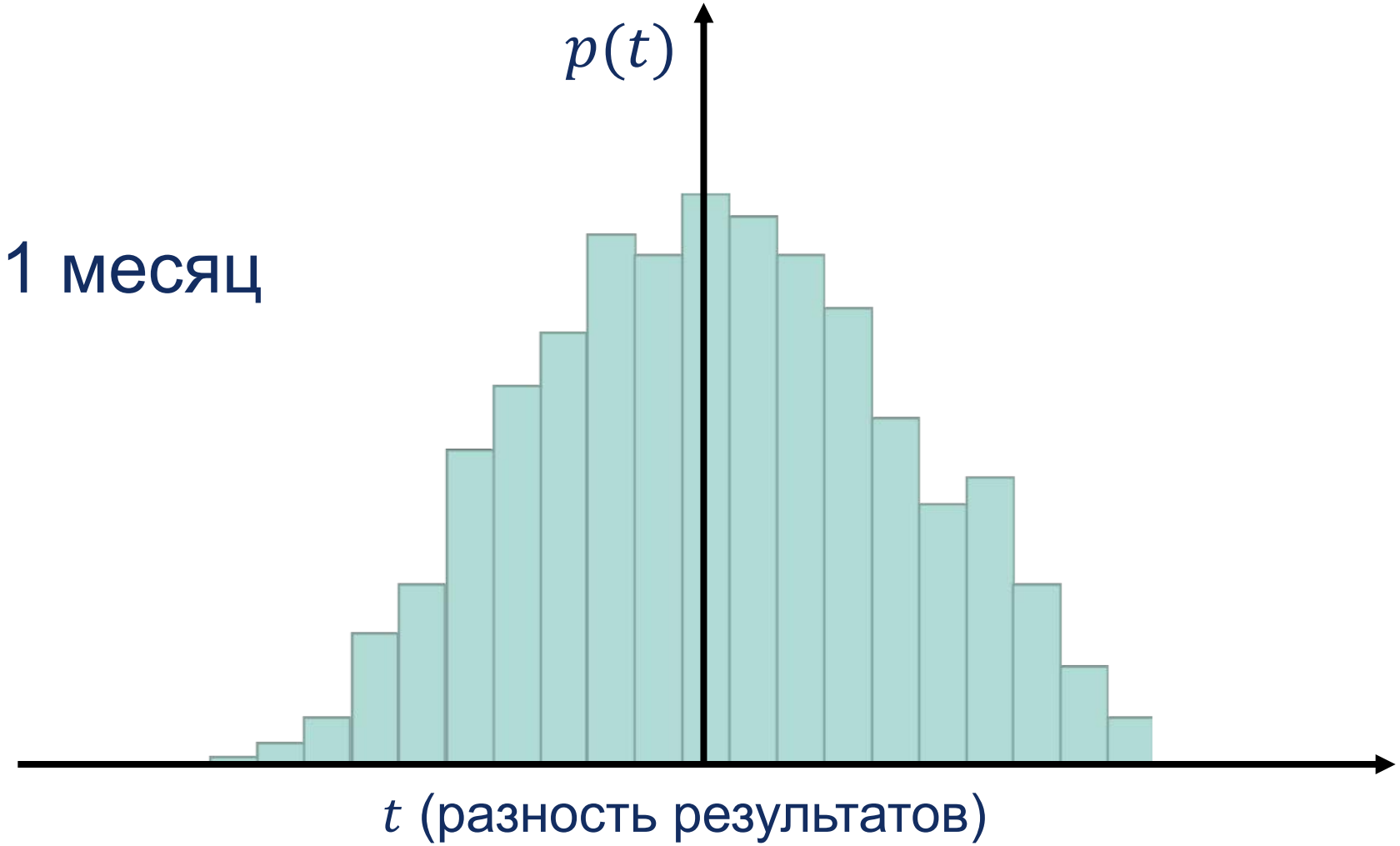
Планирование A/B теста

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

Планирование A/B теста

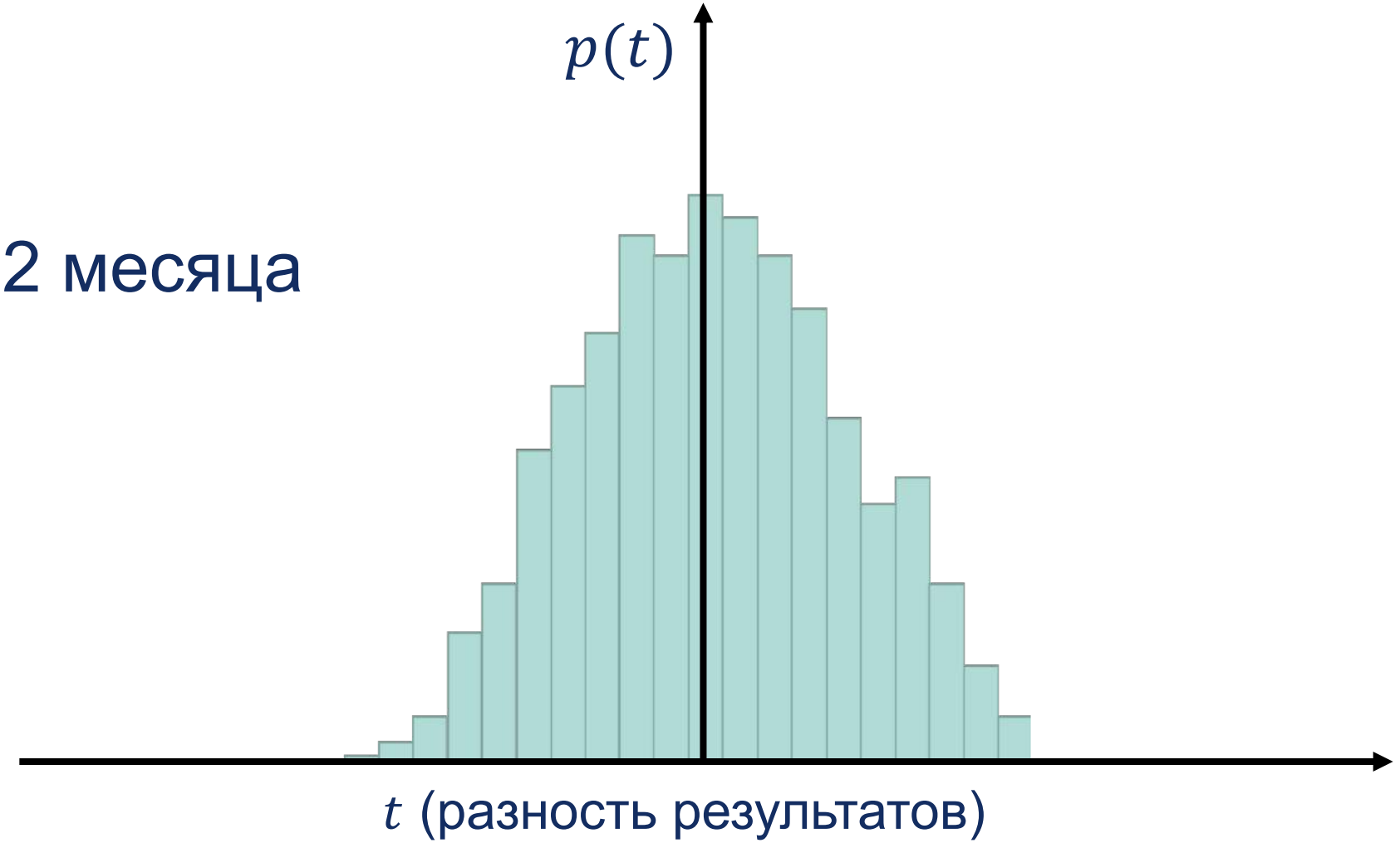
1. Решаете, насколько большое (по величине) улучшение метрики детектируете
2. По оценкам статзначимости (например, бутстрепом) на исторических данных понимаете, сколько данных нужно (=какая продолжительность у A/B теста), чтобы это улучшение алгоритма было статзначимым на нужном уровне значимости

Как подбирать длительность A/B теста



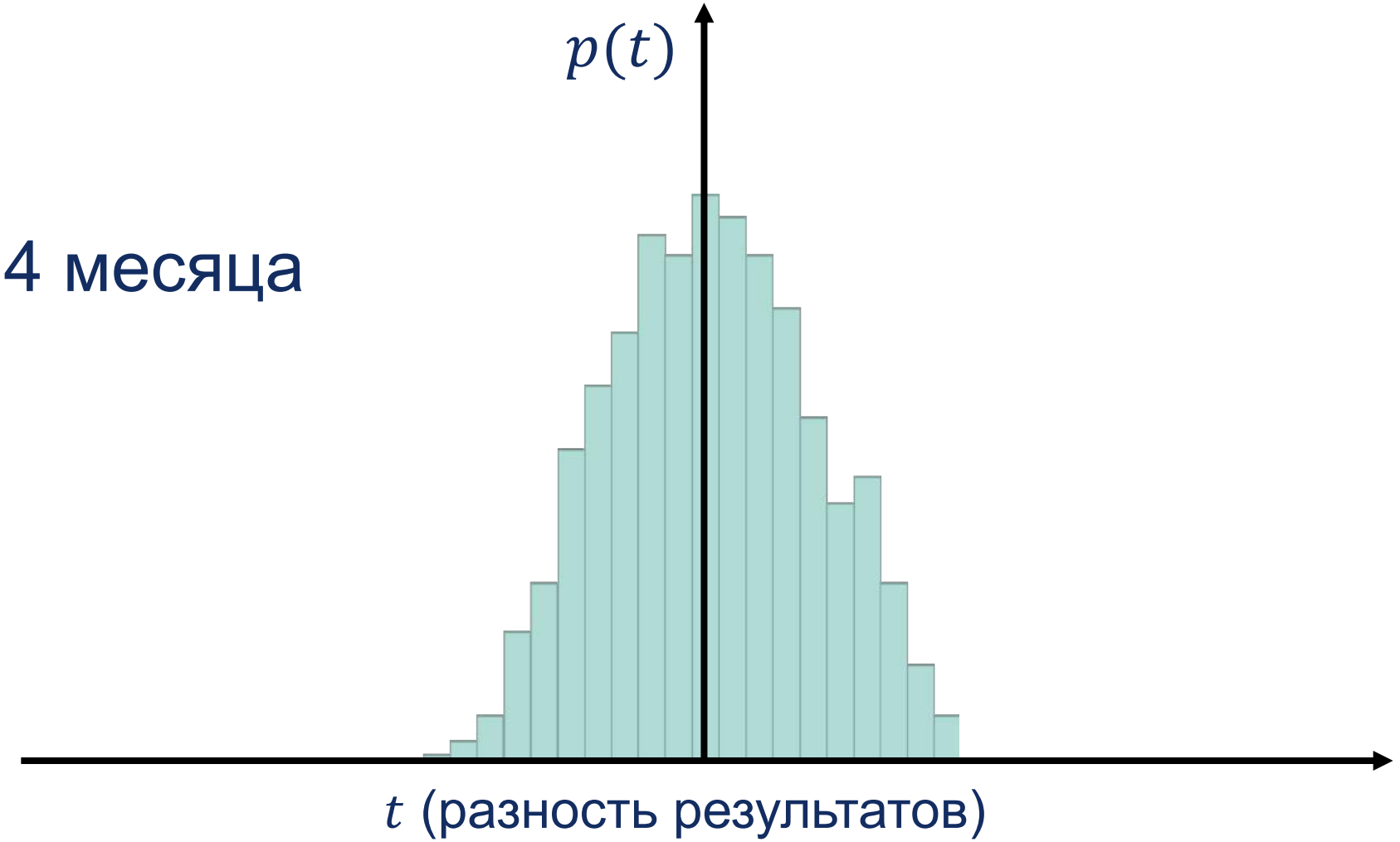
Как подбирать длительность A/B теста

2 месяца



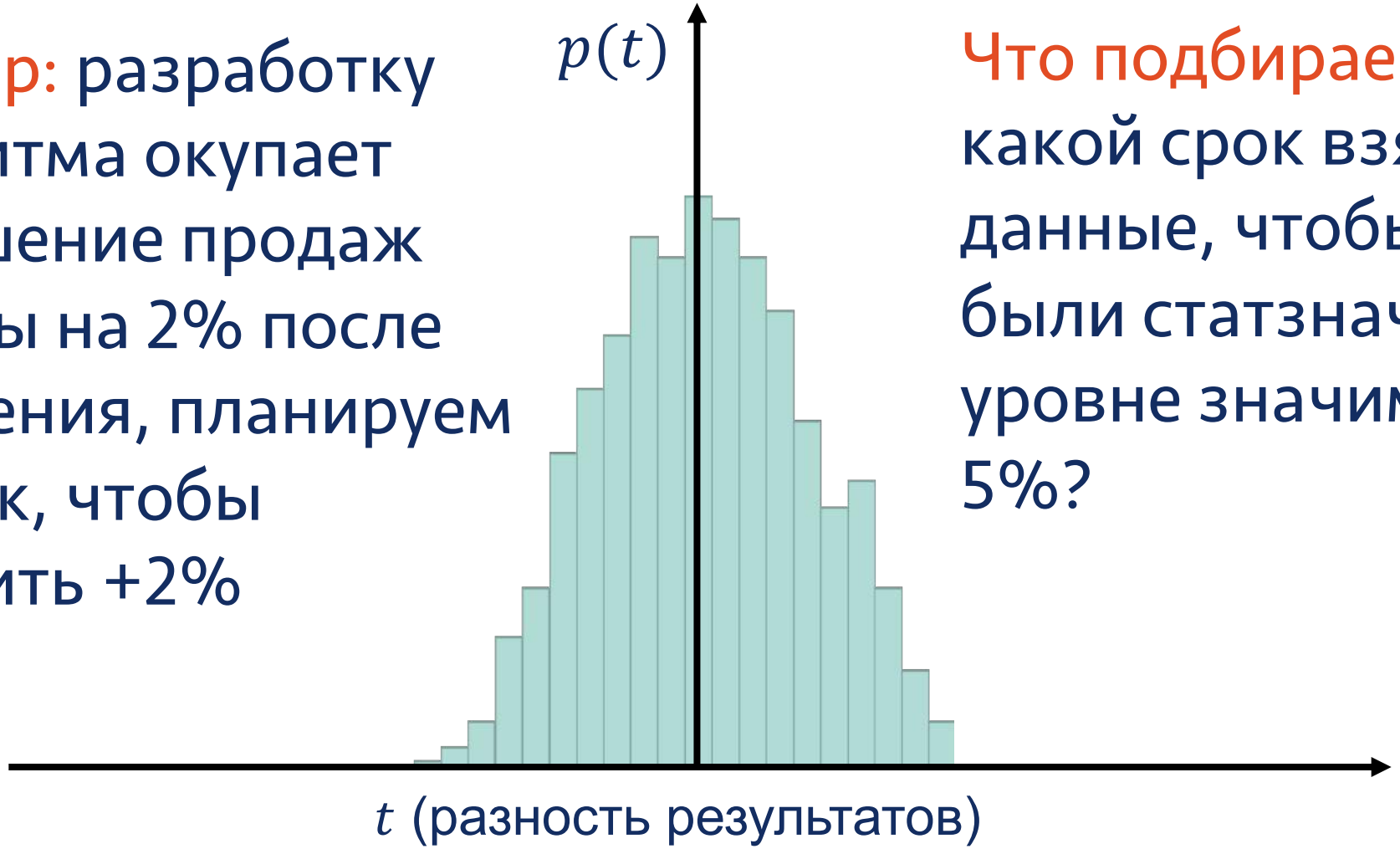
Как подбирать длительность A/B теста

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

План

1. Валидация в задачах регрессии
2. Валидация при классификации
3. Пример выбора метрики
4. Стабильность модели
5. Онлайн-эксперимент

Резюме по всей лекции

1. Существует множество стандартных метрик качества, которые допускают небольшие 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

Для справки: 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

Data Mining in Action

Лекция 5

Группа курса в Telegram:

 https://t.me/joinchat/B1OlTk74nRV56Dp1TDJGNA