Машинное обучение:
валидация и тестирование моделей в production

МТС Тета
Эмели Драль

Basics

1. ML basics & tools
2. Валидация моделей по историческим данным
3. **Тестирование моделей в production**

Результат изучения: знаете стандартные виды обучения, понимаете логику работы базовых алгоритмов, можете валидировать модели

Валидация в production

1. Сложность и качество
2. Дополнительные свойства
3. Пилотный тест и АБ-тестирование

Качество vs Сложность

Качество
vs
Сложность

Quality vs Complexity

Quality: чем меньше ошибка модели, тем лучше

Complexity: чем проще модель, тем стабильнее она работает

Качество
vs
Сложность

Quality vs Complexity

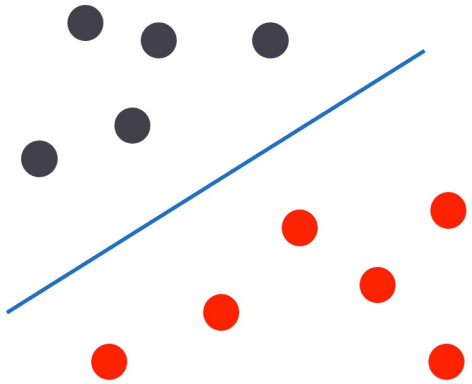
Quality: чем меньше ошибка модели, тем лучше

Complexity: чем проще модель, тем стабильнее она работает

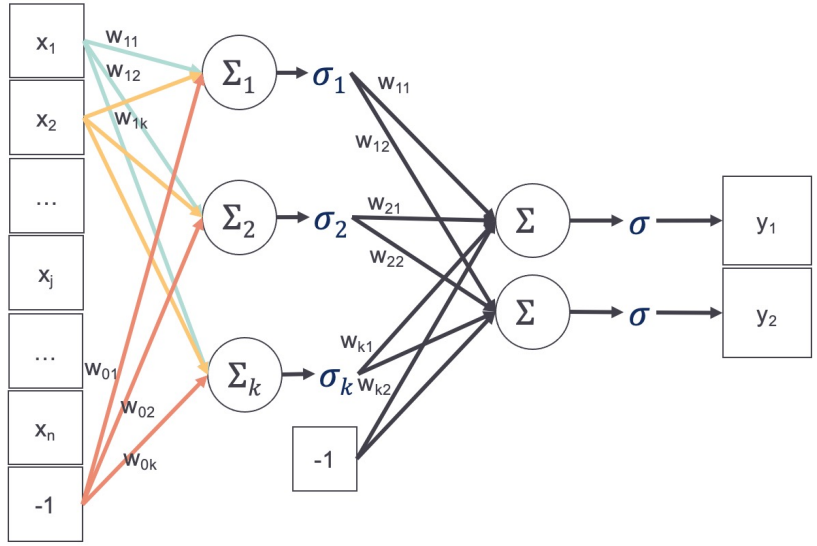
Часто, более сложные модели (или комбинации моделей) дают меньшую ошибку, но для использования в сервисе выбирают ближайший по качеству более простой аналог

Качество
vs
Сложность

Quality vs Complexity

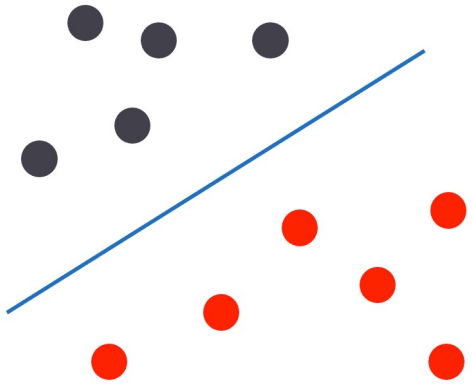


vs

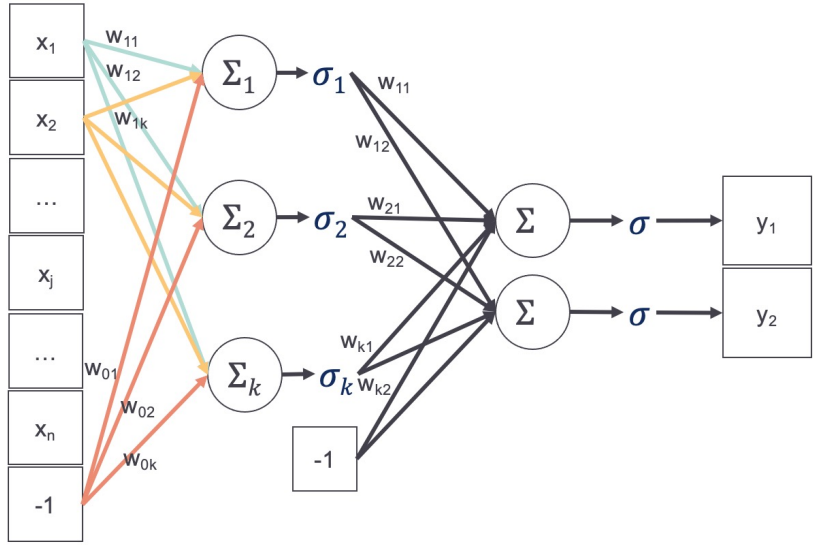


Качество
vs
Сложность

Quality vs Complexity



vs

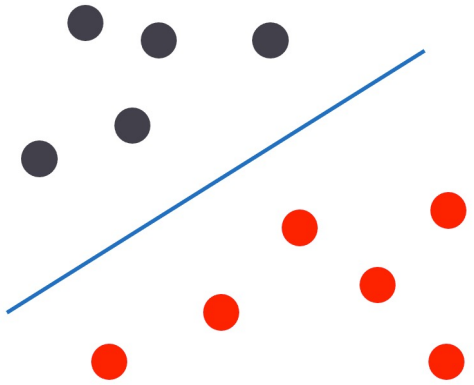


ROC AUC = 0,74

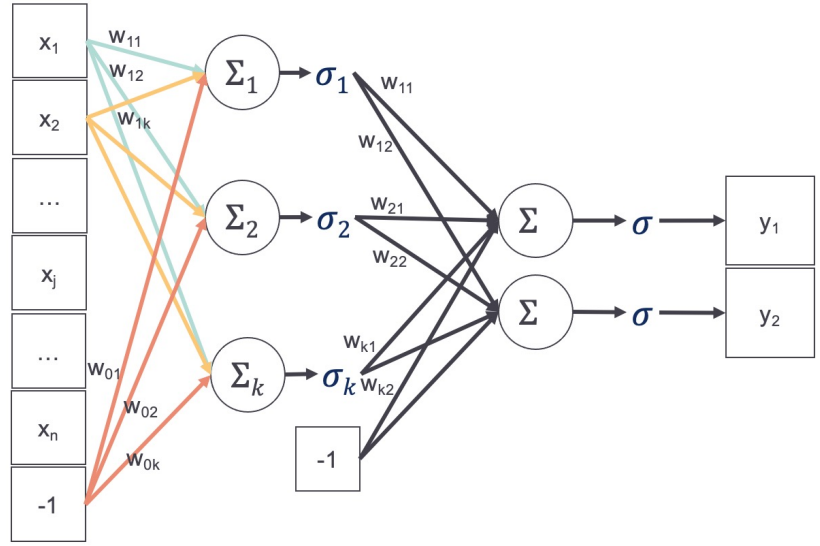
ROC AUC = 0,79

Качество
vs
Сложность

Quality vs Complexity



vs



ROC AUC = 0,74

ROC AUC = 0,79

- Связь качества модели и экономического эффекта: сколько нам стоит 0.05 ROC AUC?
- Готовы ли мы ради этого эффекта усложнить архитектуру для поддержки нейронных сетей?

Качество
vs
Сложность

Quality vs Complexity

Важно иметь несколько моделей-кандидатов разной сложности и понимать, какой прирост в качестве и эффекте дает усложнение модели

Качество
vs
Сложность

Модели-кандидаты

Важно иметь несколько моделей-кандидатов разной сложности и понимать, какой прирост в качестве и эффекте дает усложнение модели

Пример:

1. Constant model
2. Simple model with numeric features only
3. Complex model with numeric features only
4. Simple model with some feature engineering
5. Complex model with some feature engineering
6. Hybrid model

Качество
vs
Сложность

Constant model

1. Самый популярный класс в задаче классификации
2. Среднее или медиана (посчитанные по обучающей выборке!) в задаче регрессии
3. Last value (можно с учетом сезонности) в задаче прогнозирования
4. Most popular items для рекомендательной системы

Качество
vs
Сложность

Constant model

1. Самый популярный класс в задаче классификации
2. Среднее или медиана (посчитанные по обучающей выборке!) в задаче регрессии
3. Last value (можно с учетом сезонности) в задаче прогнозирования
4. Most popular items для рекомендательной системы

- для каждой задачи можно подобрать условно оптимальную константу
- это важный benchmark, позволяющий понять ценность решения

Качество
vs
Сложность

Constant model

В некоторых индустриях метрики качества даже устроены таким образом, чтобы оценивать относительный прирост качества модели.

Пример: задача прогнозирования оттока в телеком.

Метрика lift@k - во сколько раз ранжирование среди top k% абонентов согласно модели лучше случайного ранжирования?

$$lift@k = \frac{precision@k}{precision@all} = \frac{precision@k}{churn\ rate}$$

Качество
vs
Сложность

Simple model

1. Регрессия по одному или нескольким признакам
2. Дерево решений небольшой глубины
3. Метод ближайших соседей по нескольким признакам
4. Rule-based (часто, это текущее production решение)

Качество
vs
Сложность

Модель другого типа

Часто, текущее production решение не является моделью машинного обучения

1. Rule-based system
2. Математическая модель (аналитическая формула)
3. Физическая модель

Их не вполне справедливо считать простыми, но это также хороший benchmark

Качество
vs
Сложность

Complex model

Дальнейшее снижение ошибки модели возможно за счет:

- feature engineering
- более сложный алгоритм с большим количеством параметров
- комбинации более сложного алгоритма и feature engineering

Качество
vs
Сложность

Complex model

Дальнейшее снижение ошибки модели возможно за счет:

- feature engineering
- более сложный алгоритм с большим количеством параметров
- комбинации более сложного алгоритма и feature engineering

Полезно проанализировать остатки модели, чтобы оценить наличие оставшегося сигнала в данных;

Имеем смысл смотреть на feature importance добавленных признаков, особенно если их сложно рассчитывать

Качество
vs
Сложность

Hybrid model

Альтернативный способ снижения ошибки – использование комбинации из нескольких подходов к решению задачи.

Подходов очень много, например:

- Стандартный stacking
- Content based + collaborative filtering recommender system
- Бинарная классификация + регрессия для одного из классов
- Физико-химическая модель + ml модель
- Термодинамическая модель + ml модель

и пр.

Качество
vs
Сложность

Quality vs Complexity

Модель	Precision@10% (cv mean)
Constant model	0.08
Physical model	0.71
Linear model (num features)	0.61
GB (feature engineering)	0.76
Physical model + GB on residuals	0.82
Ideal model	0.9

Полезно дать рекомендацию о том, какую модель вы считаете оптимальной.

Качество
vs
Сложность

Quality vs Complexity

Модель	Precision@10% (cv mean)
Constant model	0.08
Physical model	**0.71**
Linear model (num features)	0.61
GB (feature engineering)	0.76
Physical model + GB on residuals	**0.82**
Ideal model	0.9

Полезно дать рекомендацию о том, какую модель вы считаете оптимальной.

Дополнительные свойства

Валидация модели

Качество модели

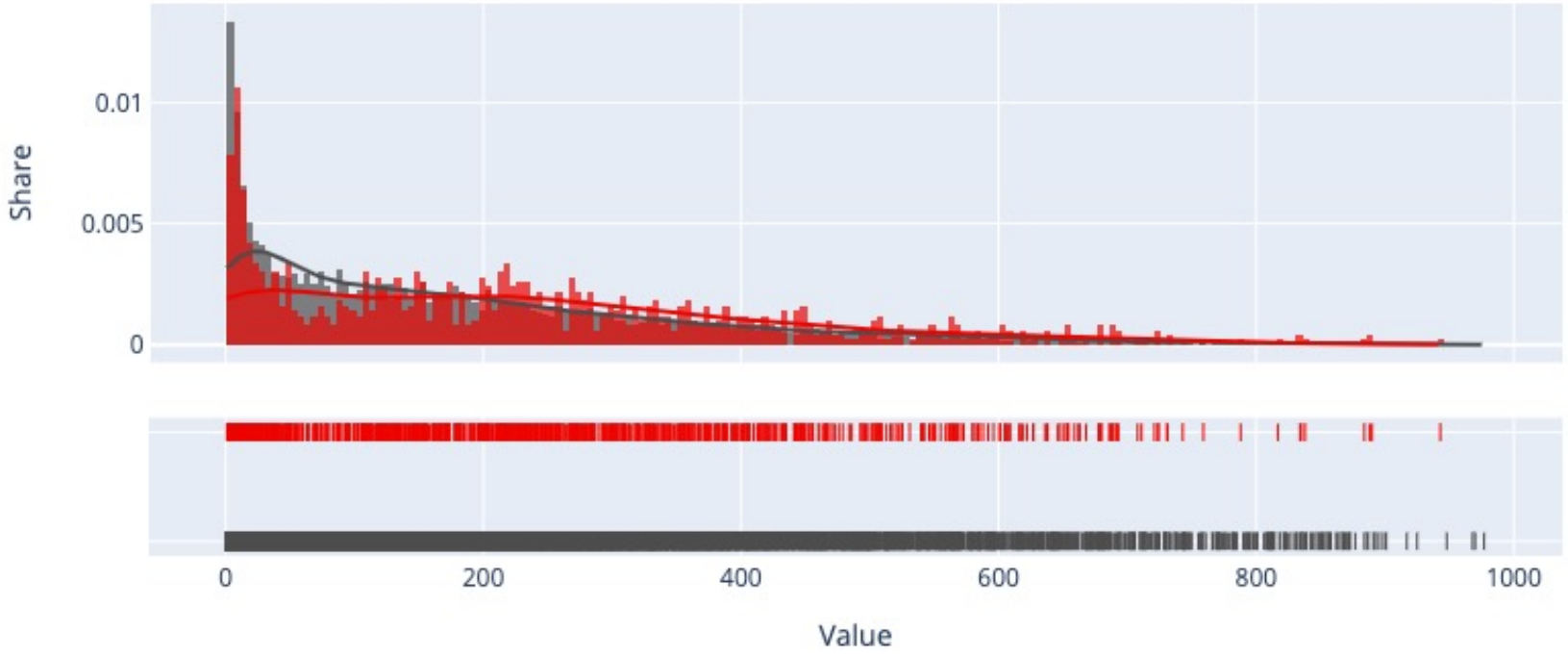
- Можно смотреть на несколько метрик, хотя оптимизируем всегда одну
- Интервальные оценки лучше точечных
- Cross-validation + hold-out test

Также, с помощью cross-validation можно оценить стабильность модели:

- меняется ли качество от фолда к фолду?
- меняется ли feature importance от фолда к фолду?

Валидация модели

Качество модели

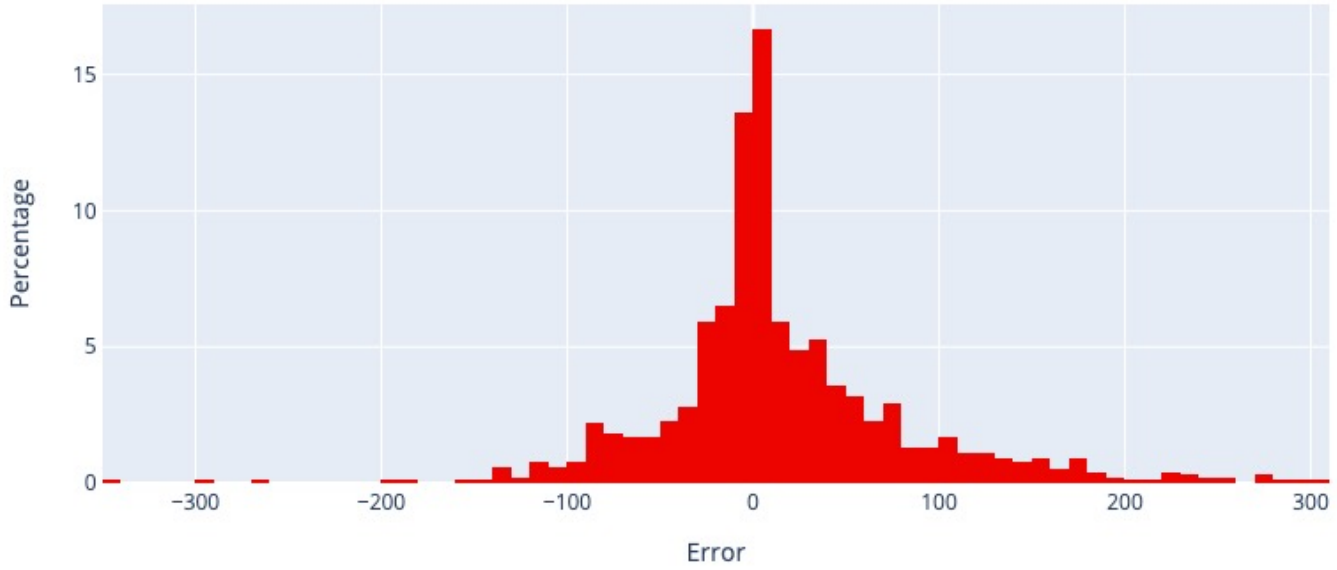


Дополнительно, имеет смысл сравнить:

- распределение target на обучении и отложенной выборке
- распределение model output на обучении и отложенной выборке

Валидация модели

Качество модели



Распределение ошибок поможет понять:

- склонна ли модель к недо/переоценке целевой функции
- остался ли сигнал в данных
- есть ли выбросы или сегменты с большей ошибкой

Валидация модели

Что ещё нужно оценить?

- Скорость устаревания модели
- Bias & fairness
- Интерпретация

Валидация модели

Скорость устаревания

Важная характеристика, на основе которой можно сделать вывод о необходимой частоте переобучения модели

Подход к оценке: (обучение, ошибка внутри ожидаемого интервала, ошибка за пределами интервала)

Быстрое устаревание:

Модель не устаревает:

Среднее устаревание:

Валидация модели

Валидация модели и предотвращение ошибок

Валидация модели

Избежание предвзятости

Amazon scraps secret AI recruiting tool that showed bias against women

In effect, Amazon's system taught itself that male candidates were preferable. It penalized resumes that included the word "women's," as in "women's chess club captain." And it downgraded graduates of two all-women's colleges, according to people familiar with the matter. They did not specify the names of the schools.

https://www.technologyreview.com/2019/10/25/132184/a-biased-medical-algorithm-favored-white-people-for-healthcare-programs/

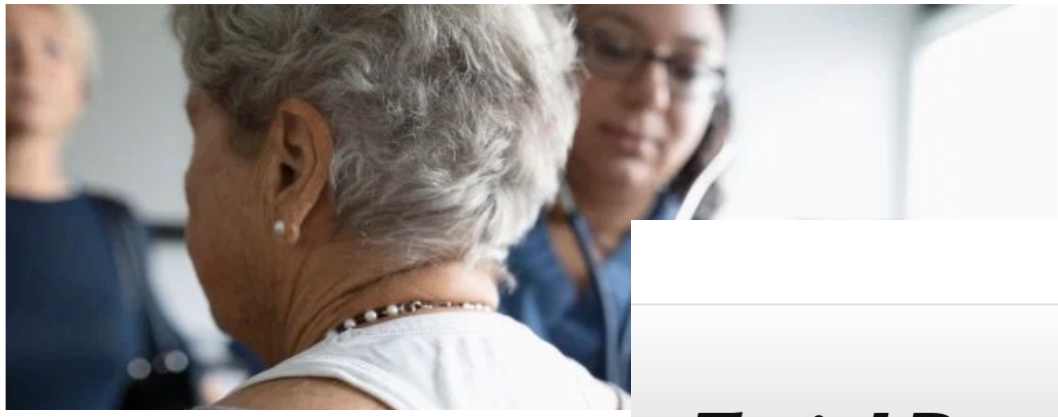
Валидация модели

Избежание предвзятости

MIT Technology Review

Artificial intelligence Oct 25

A biased medical algorithm favored white people for health-care programs



The New York Times

Facial Recognition Is Accurate, if You're a White Guy

https://www.technologyreview.com/2019/10/25/132184/a-biased-medical-algorithm-favored-white-people-for-healthcare-programs/
https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html

Валидация модели

Доверие

More CEOs (84%) ‘agree’ that AI-based decisions need to be explainable than that AI is good for society (79%).

Mark J. Girouard, an employment attorney at Nilan Johnson Lewis, says one of his clients was vetting a company selling a resume screening tool, but didn’t want to make the decision until they knew what the algorithm was prioritizing in a person’s CV.

After an audit of the algorithm, the resume screening company found that the algorithm found two factors to be most indicative of job performance: their name was Jared, and whether they played high school lacrosse. Girouard’s client did not use the tool.

https://qz.com/1427621/companies-are-on-the-hook-if-their-hiring-algorithms-are-biased
https://hackernoon.com/dogs-wolves-data-science-and-why-machines-must-learn-like-humans-do-41c43bc7f982
https://www.pwc.com/mu/pwc-22nd-annual-global-ceo-survey-mu.pdf

Валидация модели

Регуляторные требования

Европа – GDPR

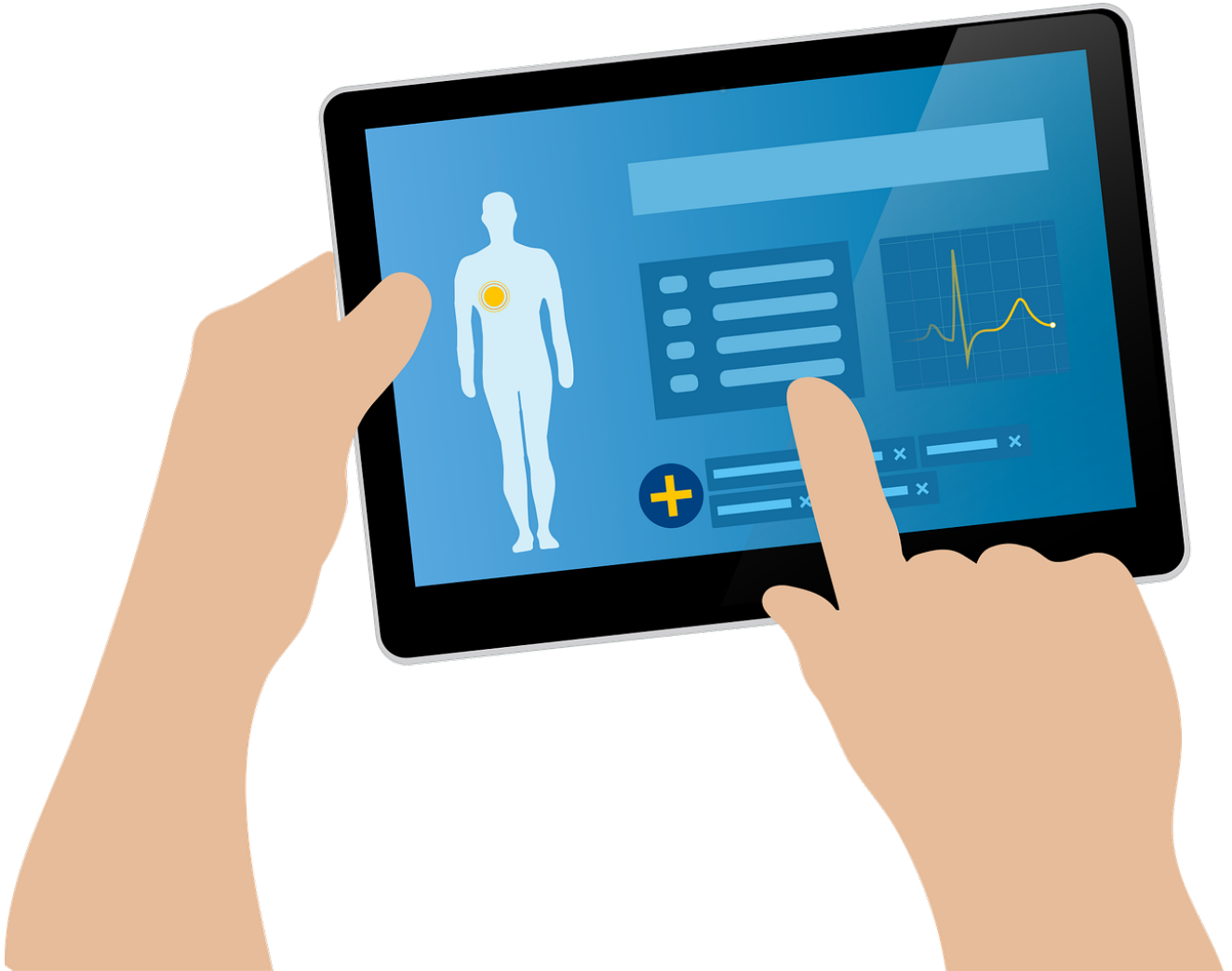
The right to access meaningful information about the logic involved, as well as the significance and the envisaged consequences of automated decision-making”

США - Equal Credit Opportunity Act

Statement of reasons for adverse action, must be specific and indicate the principal reason(s) for the adverse action

Валидация модели

Интерпретация модели



Валидация модели

Дополнительные свойства

Такие характеристики модели, как:

- калибровка
- качество в топе прогнозов
- ошибка в разрезе выбранных сегментов
- линейность по выбранным признакам

и пр.

Тестирование в production

Тестирование в production

Тестирование в production

Варианты тестирования в боевых условиях:

- Пилотный тест
- АБ-тестирование

Тестирование в production

Пилотный тест

- Изменение применяется в течение ограниченного промежутка времени
- Изменение может быть применено на небольшой группе пользователей/объектов
- Сравниваются значения метрик до и во время пилота
- Оценка проверяется на практическую и статистическую значимость

Тестирование в production

Пилотный тест

Пилот:

- Сравнивают значения метрик до и во время пилота
- Часто для большей достоверности сравнивают также значения метрик во время и после пилота

Тестирование в production

Пилотный тест

В чем ограничения данного метода тестирования?

Тестирование в production

Пилотный тест

В чем ограничения данного метода тестирования?

- Сложно изолировать влияние изменения от влияния внешних факторов
- В частности влияние сезонных факторов, трендов
- Требуется длительное время для проведение серии экспериментов

Тестирование в production

АБ-тестирование

Что если усложнить пилотный тест?

Тестирование в production

АБ-тестирование

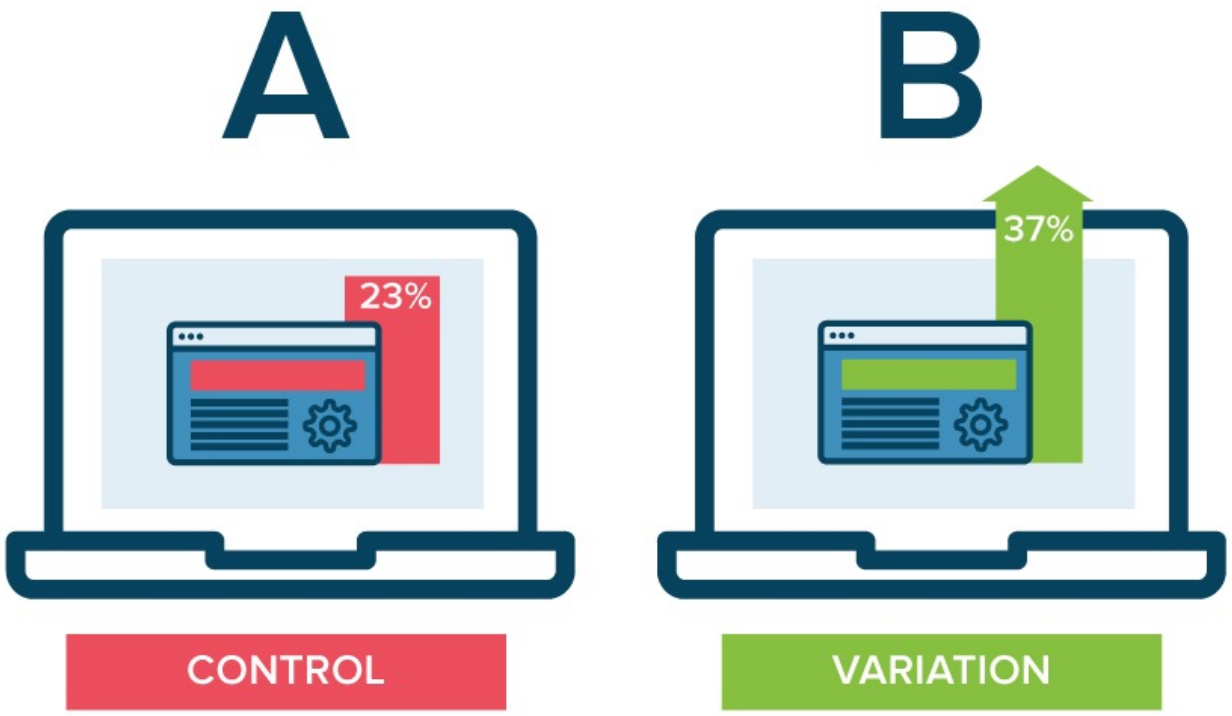
Что если усложнить пилотный тест?

АБ-тестирование:

- Пользователи/объекты делятся на контрольную и тестовую группы (сегменты, “сплиты”) – А и Б
- Изменение производится только в одной из групп (группа Б)
- Единовременно тестируется только одно изменение
- Оцениваются отличия между контрольной и тестовой группой
- Оценка проверяется на практическую и статистическую значимость

Тестирование в production

АБ-тестирование



В чем преимущества и ограничения АБ-тестирования по сравнению с пилотным тестом?

Тестирование в production

АБ-тестирование

Преимущества АБ-тестирования

- Легче изолировать влияние от влияние внешних факторов
- Возможно отделить влияние сезонных факторов, трендов
- Возможно проведение серии экспериментов одновременно (если данных достаточно)

Ограничения АБ-тестирования

- Требуется разбиение на А и Б группы, пригодное для сравнения
- Может потребоваться длительное время для достижения значимого результата
- Риск совершения ошибок I и II рода: ложная детекция результата или пропуск значимого результата при неудачном дизайне эксперимента

Этапы АБ-тестирования

Этапы АБ-тестирования

Тест прошел успешно, если:

1. зафиксировали эффект, когда он есть
2. не детектировали ложный эффект
3. смогли атрибуцировать эффект
4. провели тест оптимальным количеством ресурсов, в том числе максимально быстро

Этапы АБ-тестирования

Этапы АБ-тестирования

1. Подготовительный этап
2. Проведение тестирования
3. Оценка результатов тестирования

Этапы АБ-тестирования

Подготовительный этап

Цель: разработать дизайн эксперимента

Дизайн эксперимента – детальный ответ на вопрос:

Как будет проводиться эксперимент и оцениваться его результаты?

Этапы АБ-тестирования

Подготовительный этап

Цель: разработать дизайн эксперимента

На самом деле вопросов довольно много:

- Какой эффект (метрика, размер) вы ожидаете получить?
- Какой уровень значимости (степень уверенности) вас устроит?
- Какой стат. критерий подойдет, и какова его мощность?
- Каким способом выделяются тестовая и контрольная группы?
- Какого размера они должны быть?
- Сколько времени должен длиться тест?
- Можно ли оптимизировать тест, сделав его быстрее или сократив размер группы?

Этапы АБ-тестирования

Ожидаемый эффект

Оценка эффекта:

- Экспертная оценка
- По данным (формула для расчета, тест по историческим данным)

Этапы АБ-тестирования

Разбиение на группы

Если данных очень много – можно разбить случайно

Что если данных меньше?

- Стратификация по статическим (пол, возраст, регион и пр.) и историческим признаками (активность, коммерческий сегмент и пр.)
- Если стратификация не была сделала – отличия можно учесть при расчета эффекта

Этапы АБ-тестирования

Разбиение на группы

Если данных очень много данных – можно разбить случайно

Что если данных совсем мало?

Этапы АБ-тестирования

Разбиение на группы

Если данных очень много – можно разбить случайно

Что если данных мало и они не независимы?

- Преобразование для получения выборки из независимых наблюдений (перешли от пользователей к сессиям, от сессий – к транзакциям)

Этапы АБ-тестирования

Размер групп и длительность теста

Оценивается достаточное количество событий – N

N можно получить различными сочетаниями длительности теста (во времени) и размером групп (в объектах)

При выборе длительности и размера групп следует учитывать дополнительные ограничения, например, сезонность

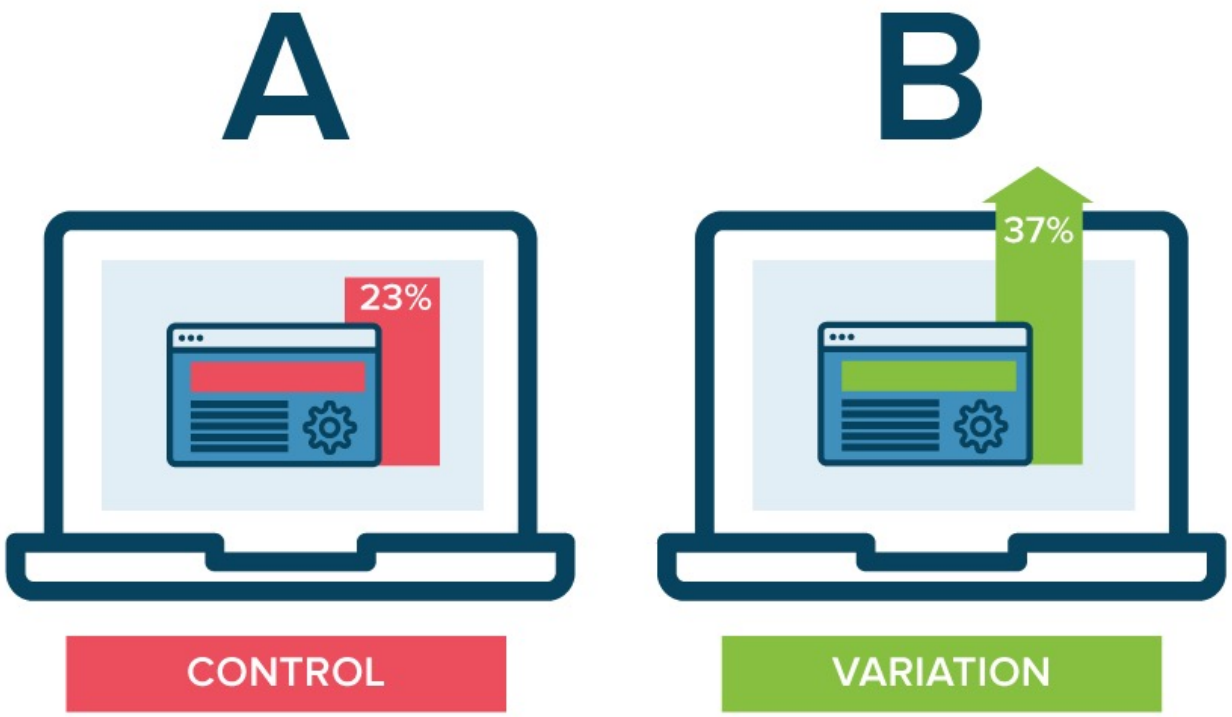
Этапы АБ-тестирования

Дизайн эксперимента

- Какой эффект (метрика, размер) вы ожидаете получить? экспертно или расчет по данным
- Каким способом выделяются тестовая и контрольная группы? экспертно с валиацией по данным
- Какой уровень значимости (степень уверенности) вас устроит? экспертно
- Какой стат. критерий подойдет, и какова его мощность? статистический расчет по данным
- Какого размера должны быть группы ? статистический расчет по данным
- Сколько времени должен длиться тест? статистический расчет по данным
- Можно ли оптимизировать тест, сделав его быстрее или сократив размер группы? опыт и чтение статей=)

Этапы АБ-тестирования

Корректность дизайна

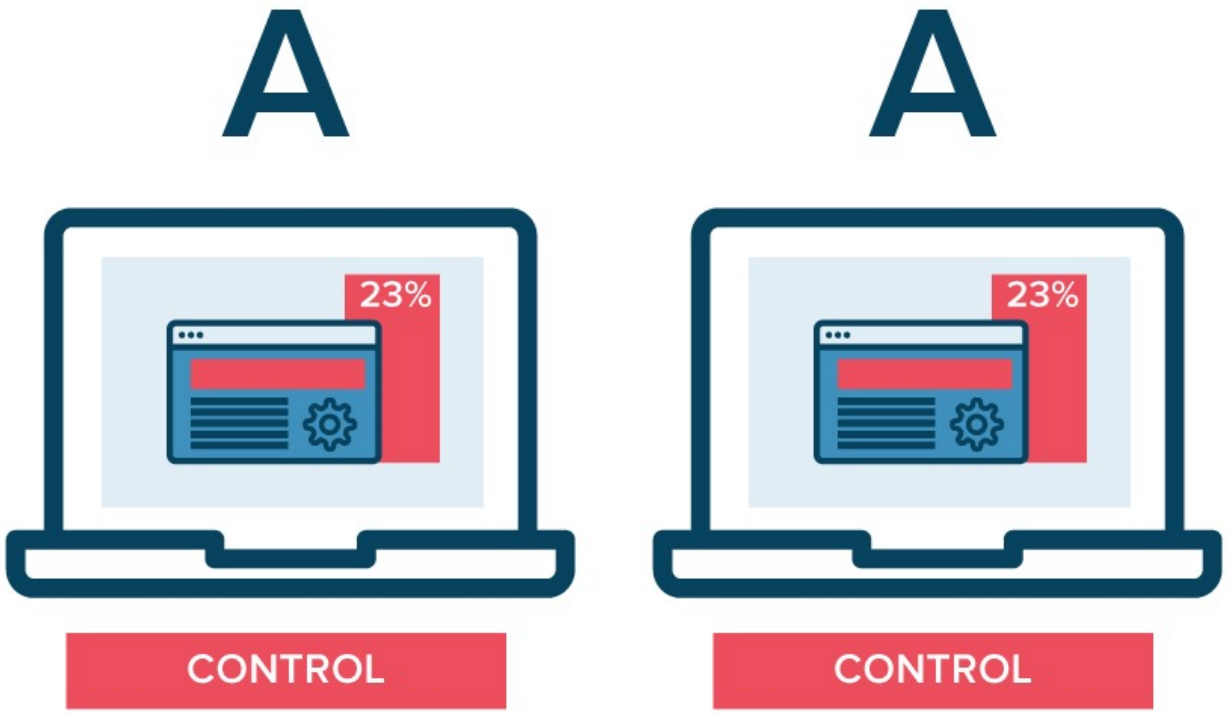


Как часто мы ловим эффект, когда его нет?

https://www.optimizely.com/optimization-glossary/ab-testing/

Этапы АБ-тестирования

AA-тестирование



https://www.optimizely.com/optimization-glossary/ab-testing/

Этапы АБ-тестирования

АА-тестирование

Можно реализовать в двух режимах:

- Офлайн: по историческим данным
- Онлайн: на работающем сервисе

Этапы АБ-тестирования

АА-тестирование

Реализация офлайн:

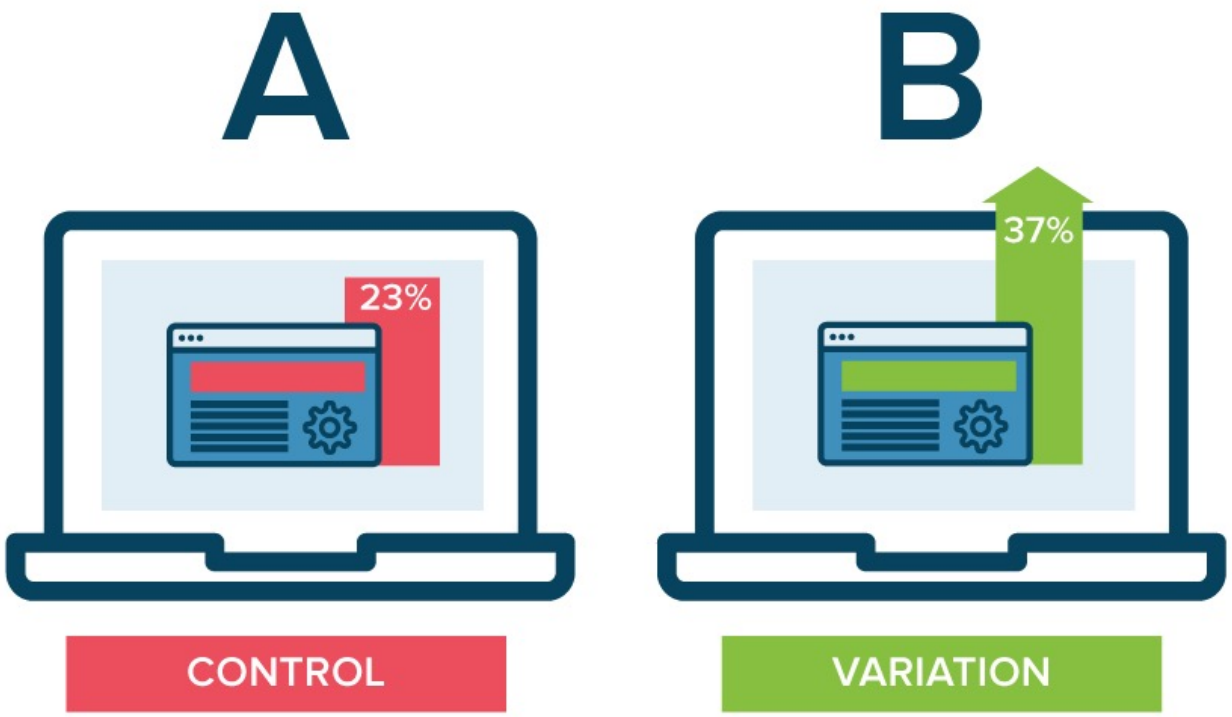
- Делим исторические данные на тест и контроль, считаем метрики и делаем стат. тест
- Можно протестировать множество вариантов очень быстро
- Не требует разработки

Реализация онлайн:

- В АБ-тесте демонстрируем одно и тоже в обоих сегментах
- Тестируем не только дизайн, но и инфраструктуру

Этапы АБ-тестирования

Корректность дизайна



Как часто мы пропускаем эффект, когда его есть?

https://www.optimizely.com/optimization-glossary/ab-testing/

Этапы АБ-тестирования

Синтетические резульаты

Идея:

- Эмулируем наличие эффекта
- Считаем метрики и делаем стат. тест

Этапы АБ-тестирования

Синтетические результаты

Идея:

- Эмулируем наличие эффекта
- Считаем метрики и делаем стат. тест

Реализация:

1. Добавим шум и оценим результаты
2. Добавим эффект и оценим результаты

Этапы АБ-тестирования

Проведение тестирования

Цель: корректный запуск и мониторинг в процессе теста

Корректный запуск складывается из корректного дизайна и правильно работающей инфраструктуры для тестирования

Этапы АБ-тестирования

Проведение тестирования

Цель: корректный запуск и мониторинг в процессе теста

После проверки дизайна тестирования и инфраструктуры, важно наладить мониторинг теста и настроить оповещения в случае возникновения проблем.

Этапы АБ-тестирования

Проведение тестирования

Цель: корректный запуск и мониторинг в процессе теста

В рамках мониторинга мы следим за “здоровьем “ тестирования:

- сколько объектов (пользователей) попадает в тест и контроль?
- соотношение по регионам
- время ответа

и др.

Этапы АБ-тестирования

Оценка результатов

Цель: верно оценить эффект и определить следующие шаги

Связь оценки результатов АБ тестирования и проверки стат. гипотезы:

1. Оценка эффекта в рамках теста – оценка параметра генеральной совокупности по выборке
2. Оценка должна быть надежной – статистическая значимость в тесте
3. Зафиксировали ложный эффект – сделали ошибку 1 рода
4. Пропустили эффект – сделали ошибку 2 рода
5. Оценка размера эффекта – практическая значимость, стоит сделать интервальные оценки

Этапы АБ-тестирования

Оценка результатов

Цель: верно оценить эффект и определить следующие шаги

- В рамках АБ теста важно не только проверить наличие эффекта, но и оценить его размер. То есть важны и статистическая и практическая значимости
- Выбор удачного дизайна эксперимента (в том числе его длительность, H0 и H1, критерий) важен для оценки статистической и практической значимостей

Этапы АБ-тестирования

Оценка результатов

Цель: верно оценить эффект и определить следующие шаги

Практическая / Статистическая	Есть	Нет
Есть	принимаем гипотезу	не внедряем изменение
Нет	продолжаем тест	отклоняем гипотезу

Этапы АБ-тестирования

Оценка результатов

Цель: верно оценить эффект и определить следующие шаги

Всё равно сомневаетесь?

- Обратный эксперимент - способ проверить наличие долгосрочного эффекта от вашего изменения.

- Реализация: внедряя изменение в продакшн оставьте небольшой сегмент пользователей со старой версией продукта/сервиса. Проведите АБ тестирование.

Этапы АБ-тестирования

	Подготовительный этап дизайн тестирования	Проведение теста не ошибиться	Оценка результатов верно оценить эффект
1	Зафиксировать критерий успешности	Реализовать разделение на корректные группы (возможно, с проверкой в виде АА тестирования)	Пост-обработка данных: фильтрация выбросов, ошибок и пр.
2	Оценить потенциальный эффект, выбрать уровень значимости	Настроить технический мониторинг сервиса	Оценка статистической значимости результатов
3	Подобрать критерий, достаточно мощный при фиксированном уровне значимости	Настроить мониторинг метрик ”здоровья” тестирования	Оценка практической значимости результатов
4	Выбрать стратегию выделения групп, рассчитать необходимый размер групп и длительность тестов	Настроить алерты	Решение о дальнейших шагах: отказ, внедрение, обратный эксперимент
5	Оценить возможности для оптимизации (ускорения) тестирования: стратификация, прогнозирование поведения пользователей, CUPED и пр.		

Машинное обучение:
валидация моделей по историческим данным

Спасибо!

Эмели Драль