Лекция №6
МатСтат в Бизнесе

Даниил Потапов

Руководитель Лаборатории Искусственного Интеллекта

РСХБ

План лекции

- АБ тестирование
- Causal inference
- Валидация ML моделей

План лекции

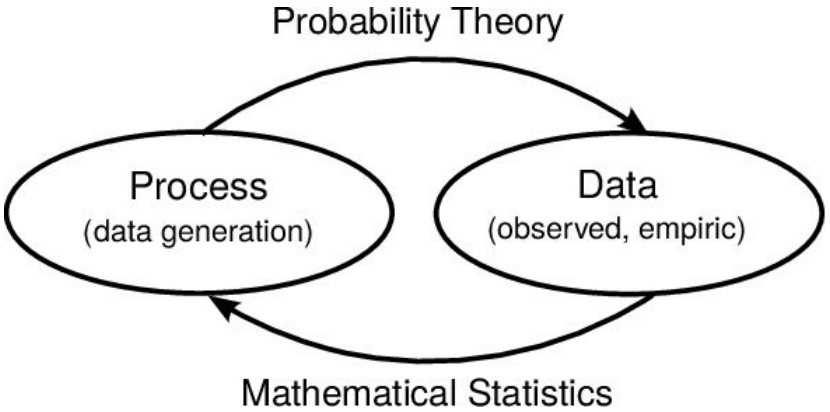
ИТМО

- АБ тестирование
- Causal inference
- Валидация ML моделей

ТерВер vs МатСтат

ИТМО

- Мир вокруг нас порождает данные мириадами различных процессов. Механизмы порождения изучаются ***теорией вероятностей***



- Наблюдаемые данные – объект изучения ***математической статистики***. По выборкам из этих данных мы пытаемся понять, каким процессом они порождены

Что такое А/Б-тестирование

A/B-тестирование - это проверка гипотез.

- Две группы (но может быть и больше) - случайно/независимо полученные
- Метрика - что измеряем?
- Значимость - какова вероятность, что это случайность?

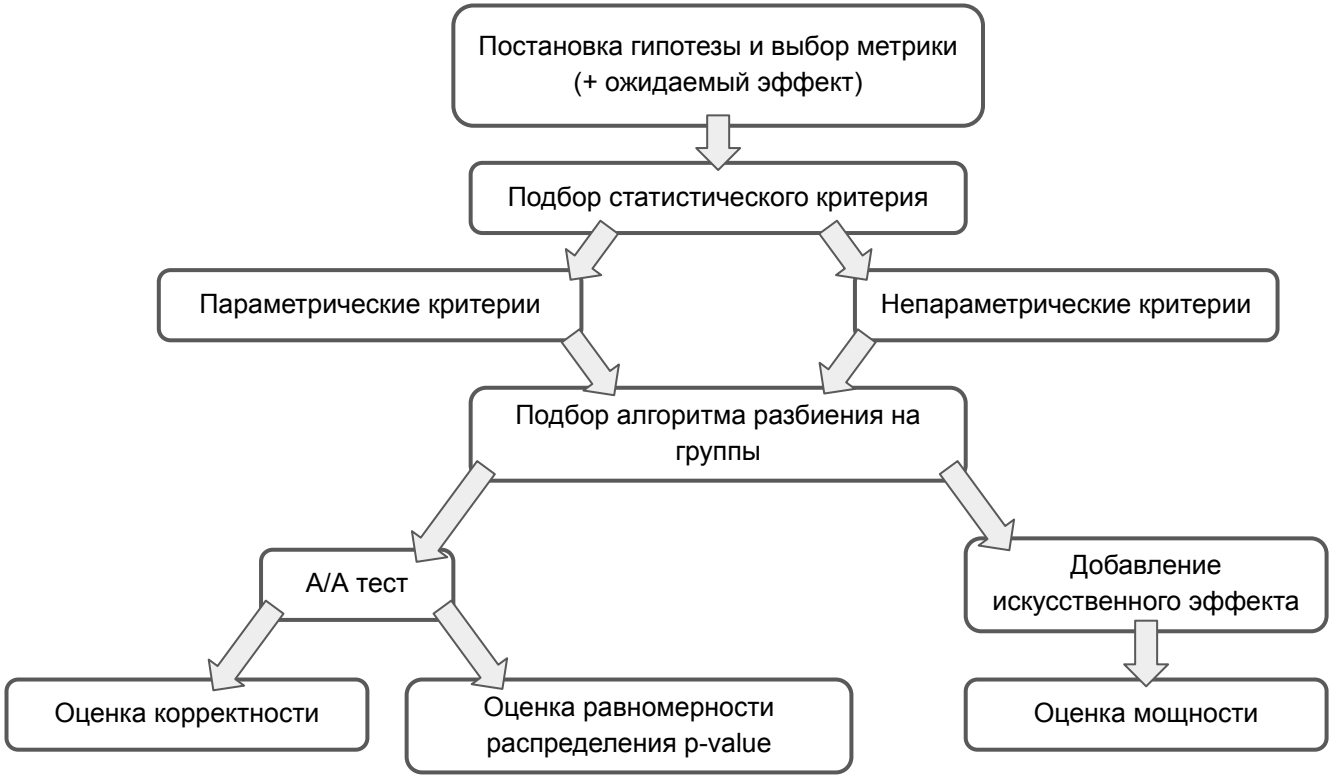
На основе анализа изменений интересующей метрики в двух группах пользователей (вызванных, например, изменениями пользовательского интерфейса, рекомендациями и т.д.) мы можем установить и подтвердить неслучайный характер этих изменений.

Что такое А/Б-тестирование

A/B-тестирование - это метод, облегчающий принятие решений, базирующийся на данных.

- Какое влияние внесенных изменений наблюдается?
  - Положительное или отрицательное
- Каков масштаб их воздействия?
  - Сильный или слабый
- Является ли результат значимым?
  - Вероятность получить такой результат случайно
- Является ли результат практически значимым?
  - Приносит ли бизнес ценность

Общая схема подготовки к А/Б



План курса по АБ тестированию

ИТМО

Лекции

1. Введение в А/Б-тестирование
2. Дизайн А/Б-теста
3. Дизайн в реальных условиях
4. Методы ускорения
5. Последовательный анализ

Семинары

1. Повторение мат. стата
2. Практика по дизайну А/Б-теста
3. Обзор методов и фреймворков
4. Практика по ускорению
5. Практика по последовательному анализу

Домашки

1. Самостоятельный дизайн А/Б
2. Ускорение своего дизайна

План лекции

ИТМО

- АБ тестирование
- Causal inference
- Валидация ML моделей

Корреляция != Причинность

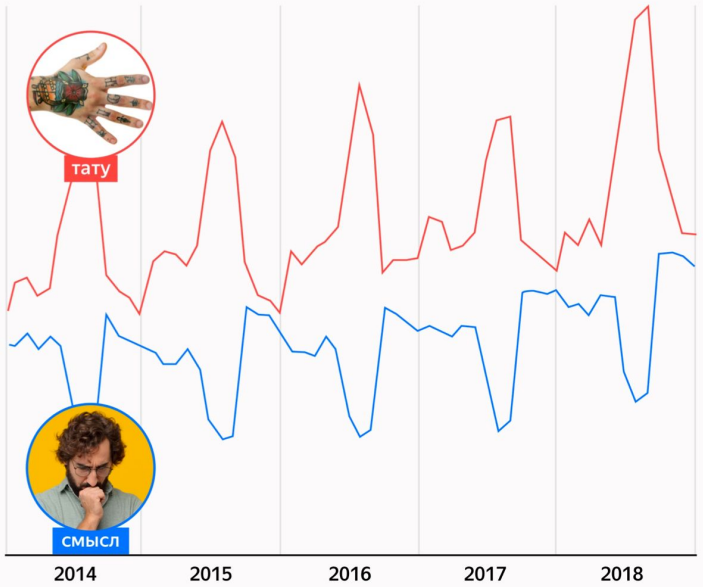
ИТМО

Парадоксы в Поиске — Яндекс

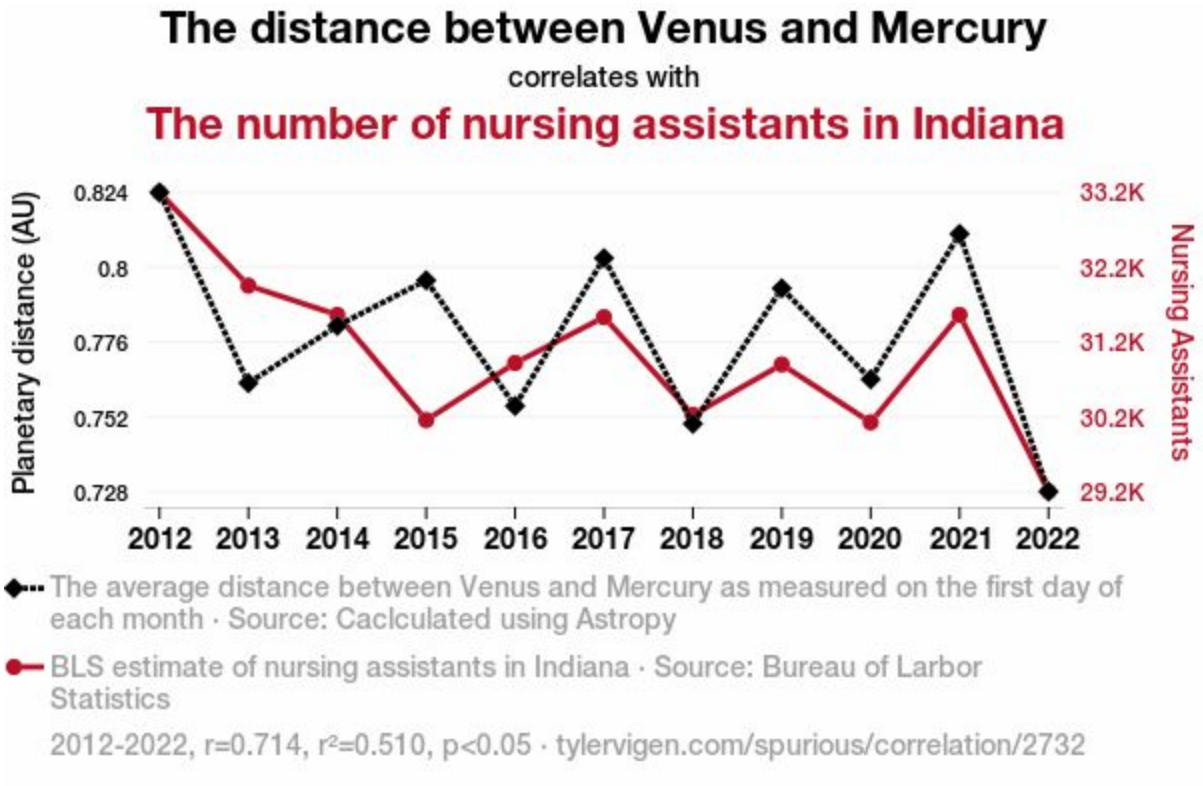
Когда в Поиске взлетает доля запросов со словом огурец, становится меньше запросов со словом эссе

Парадоксы в Поиске — Яндекс

Когда в Поиске растёт интерес к тату, снижается доля запросов со словом смысл



Ложная корреляция



https://www.tylervigen.com/spurious-correlations

Задача о собаке

ИТМО

Как погладить собаку? - Нужно оценить ее дружелюбность.

Применим ML!

Собираем данные

Порода	**Размер**	**Виляет ли хвостом**	**…**	**Дружелюбность**
Доберман	Большой	Да	…	Да
Мопс	Маленький	Да	…	Да
Чихуахуа	Маленький	Нет	…	Нет
…	…	…	…	…

Задача о собаке

ИТМО

Модель легко находит связь между вилянием хвостом и дружелюбностью

Порода	**Размер**	**Виляет ли хвостом**	**…**	**Дружелюбность**
Доберман	Большой	**Да**	**…**	**Да**
Мопс	Маленький	**Да**	**…**	**Да**
Чихуахуа	Маленький	Нет	**…**	Нет
…	**…**	**…**	**…**	**…**

виляет хвостом => дружелюбна => можно погладить

Задача о собаке

Вы пытаетесь сами “повилять” хвостом собаки, чтобы она стала дружелюбна...

Задача о собаке

ИТМО

YOU DIED

Прогнозирование vs Причинность

ITMO

> Прогнозирование: Проблема - ML - Solution

Подходит когда у нас нет влияния как на данные, так и на таргет. И задача, соответственно, когда нет таргета, предсказать его. При условии, что данные будут такие же.

> Причинность

Возникает, когда целью является понять, что **делать**, чтобы получить **требуемый результат**.

Какие задачи чаще встречаются на практике?

Анализ причинно-следственных связей

ITMO

Ключевые вопросы:

- Оценить эффект влияния на популяцию
- Какой метод лечения выбрать для пациента?
- Сколько продуктов мы продадим, если установим цену X?
- Сколько прибыли будет от внедрения модели?
- Влияет ли X на Y и как?
- Что если …?

Анализ причинно-следственных связей

ИТМО

Все зависит от постановки вопросы со стороны бизнеса.

Представим какой-то банк. Есть продукты и клиенты, которым эти продукты банк продает. Банк знает различные хар-ки клиентов и полностью знает свои продукты. Также есть история изменения цен, проведенных кампаний и тд. Рассмотрим возможные задачи (в каких из них нам поможет ML?):

1. Кому дать кредит?
2. Надо ли таргетировать рекламу на сегмент возраста от 20 до 30?
3. Сколько будет взято кредитов в следующем месяце?
4. Что будет, если поднять цены на 10%?
5. Как влияет наличие скидки на склонность к покупке?
6. Кто уйдет в отток в следующем месяце?

Анализ причинно-следственных связей

ИТМО

Все зависит от постановки вопросы со стороны бизнеса.

Представим какой-то банк. Есть продукты и клиенты, которым эти продукты банк продает. Банк знает различные хар-ки клиентов и полностью знает свои продукты. Также есть история изменения цен, проведенных кампаний и тд. Рассмотрим возможные задачи (в каких из них нам поможет ML?):

1. Кому дать кредит? ML/CI
2. Надо ли таргетировать рекламу на сегмент возраста от 20 до 30? ML/CI
3. Сколько будет взято кредитов в следующем месяце? ML
4. Что будет, если поднять цены на 10%? CI
5. Как влияет наличие скидки на склонность к покупке? CI
6. Кто уйдет в отток в следующем месяце? ML

Анализ причинно-следственных связей

ITMO

Не всегда удаётся интуитивно установить причинно-следственные связи между деянием и последствиями.

Например, широко известен парадокс о двух убийцах. Первый отравил воду жертвы, отправлявшейся в путешествие по пустыне. Второй пытался застрелить жертву из снайперской винтовки уже во время путешествия, но промахнулся, и попал во флягу с отравленной водой. Вода вытекла и жертва погибла от жажды.

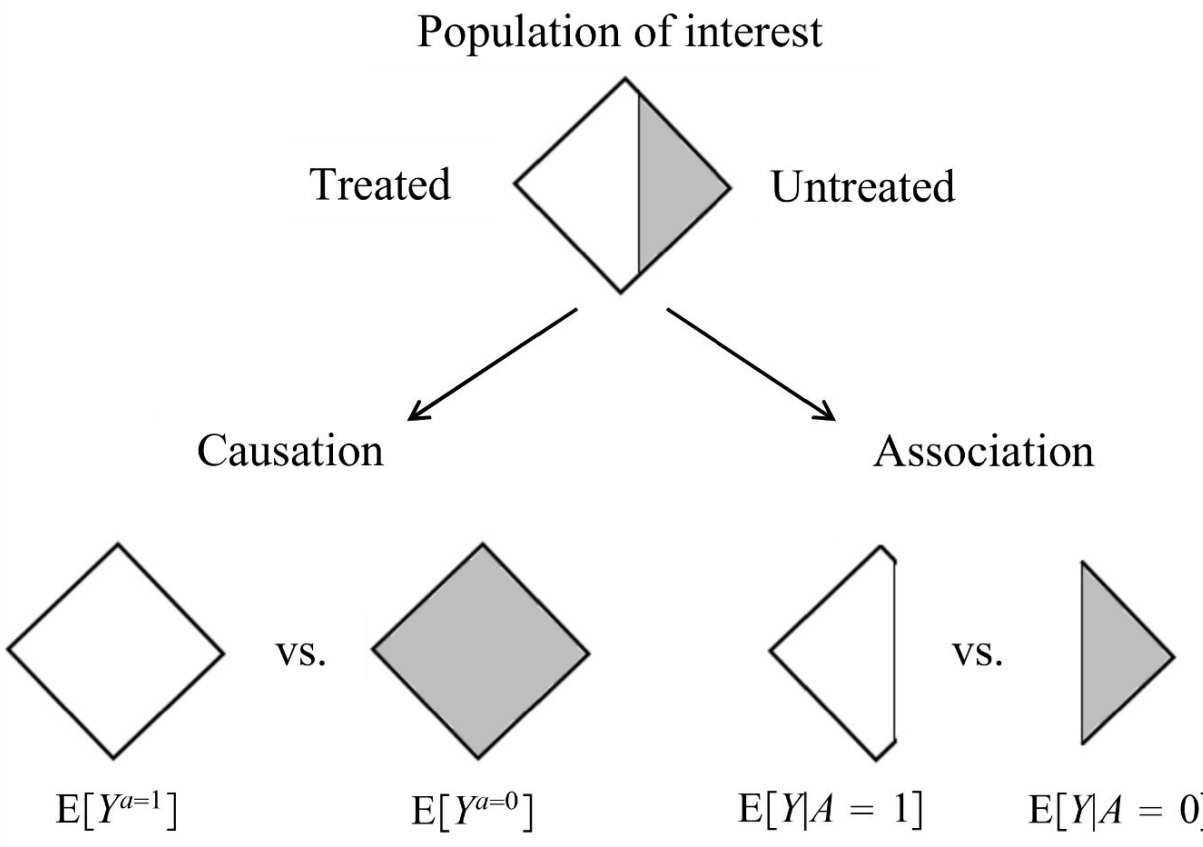
В результате оказывается, что первый убийца непосредственно убийство не совершал, поскольку жертва не пила отравленной воды (разумеется, имела место попытка убийства, прекращённая помимо воли убийцы).

С другой стороны, второй убийца непосредственно убийство тоже не совершал (хотя попытка имела место и здесь), поскольку в жертву не попал. Более того, он, пусть непроизвольно, несколько продлил жизнь жертвы, лишив её возможности выпить отравленную воду.

Тем не менее жертва погибла, и совершенно очевидно, что если бы не действия убийц, то этого бы не произошло.

Вики: Причинная связь в уголовном праве

Причинность и ассоциативность



Как выявлять?

ИТМО

- АБ тестирование - стандарт (The Design of Experiments, Fisher, 1935) оно же RCT или сплит-тестирование.
- Разбивать наблюдения по случайному признаку на две случайные выборки. В одной – изменение (показ рекламы, повышение цены), в другой – нет.
- Достаточно большие выборки можно проверить на равенство средних искомого таргета, например с помощью t-test
- При достаточно больших случайных выборках, согласно ЗБЧ, разницы между ними не будет для любых характеристик типа средних. (Лучше всегда проверять, так ли это, например, считая средние для «важных» характеристик.)
- Значит **причина**, почему разнятся выборки относительно таргета – наш эксперимент

Что еще почитать

ИТМО

- Causal Inference: What If, Miguel A. Hern´an, James M. Robins
  - код к книге - Github
- Causal Inference for The Brave and True
- https://ods.ai/tracks/interpretable-ml-df2021
- https://ods.ai/tracks/causal-inference-in-ml-df2020
- https://ods.ai/tracks/reliable_ml_ab_testing-causal_inference_meetup

Что еще попробовать

ИТМО

- ChiRho: causal modeling extensions for Pyro (Probabilistic Programming)
  - Есть статья
- DoWhy - Python library for causal inference that supports explicit modeling and testing of causal assumptions
- EconML - Python package for estimating heterogeneous treatment effects from observational data via machine learning
- CausalLearn - Causal Discovery in Python
- Causal ML - Uber Python Package for Uplift Modeling and Causal Inference with ML
- CausalPy - Байесовские модели для Causal Inference
- HypEx - Sber Advanced Causal Inference and AB Testing Toolkit

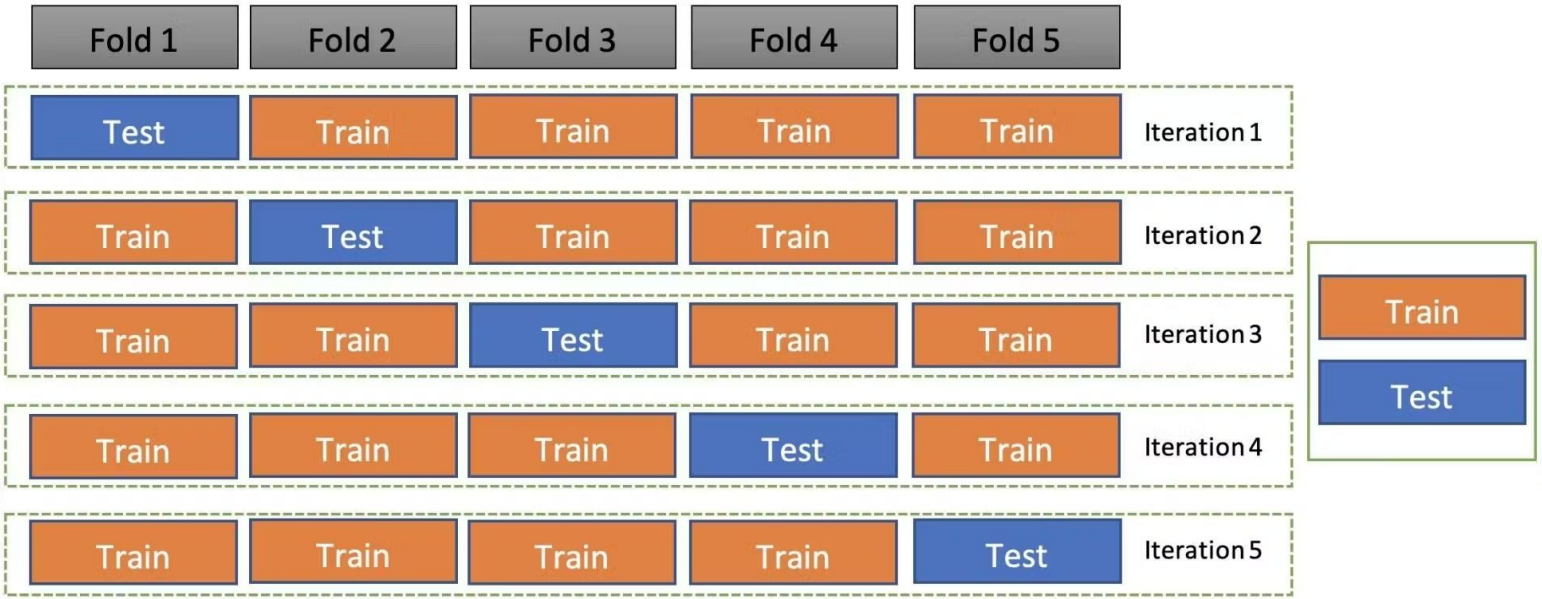
План лекции

- АБ тестирование
- Causal inference
- Валидация ML моделей

Валидация ML моделей

ИТМО

Классический KFold: дыбираем модель с лучшим средней метрикой



Но где гарантии, что это действительно лучшая модель?

Валидация ML моделей

ITMO

T-test KFold: смотрим на результаты по фолдам как на случайную величину

- KFold - значит есть K значений метрик для моделей A и B
- Считаем t-статистику как

$$t = \frac{\bar{p}\sqrt{k}}{\sqrt{\sum_{i=1}^{k}(p^{(i)-\bar{p}})^2/(k-1)}}.$$

$$p^{(i)} = p_A^{(i)} - p_B^{(i)}$$

$$\bar{p} = \frac{1}{k}\sum_{i=1}^{k} p^{(i)}$$

https://rasbt.github.io/mlxtend/user_guide/evaluate/paired_ttest_kfold_cv/

Валидация ML моделей

ИТМО

Проблемы T-test KFold: $$p^{(i)} = p_A^{(i)} - p_B^{(i)}$$

- Разница в метриках моделей A и B (p_i) распределена не нормально, так как значения метрик связаны между собой
- Также значения разниц в метриках (p_i) зависимы между собой, так как пересекаются обучающие примеры между фолдами

Как решать? Придумали T-test 5x2 CV

Валидация ML моделей

ИТМО

T-test 5x2 CV:

- 5 раз повторяем 2-fold валидацию (50% train/test сплит)
  - Почему 2-fold - потому что вначале train/test, потом test/train
- Считаем разницы по фолдам и потом усредняем и считаем отклонение

$$p^{(1)} = p_A^{(1)} - p_B^{(1)} \qquad p^{(2)} = p_A^{(2)} - p_B^{(2)}$$

$$\overline{p} = \frac{p^{(1)} + p^{(2)}}{2} \qquad s^2 = (p^{(1)} - \overline{p})^2 + (p^{(2)} - \overline{p})^2$$

- t-статистика

$$t = \frac{p_1^{(1)}}{\sqrt{(1/5)\sum_{i=1}^{5} s_i^2}}$$

https://rasbt.github.io/mlxtend/user_guide/evaluate/paired_ttest_5x2cv/

Валидация ML моделей

ИТМО

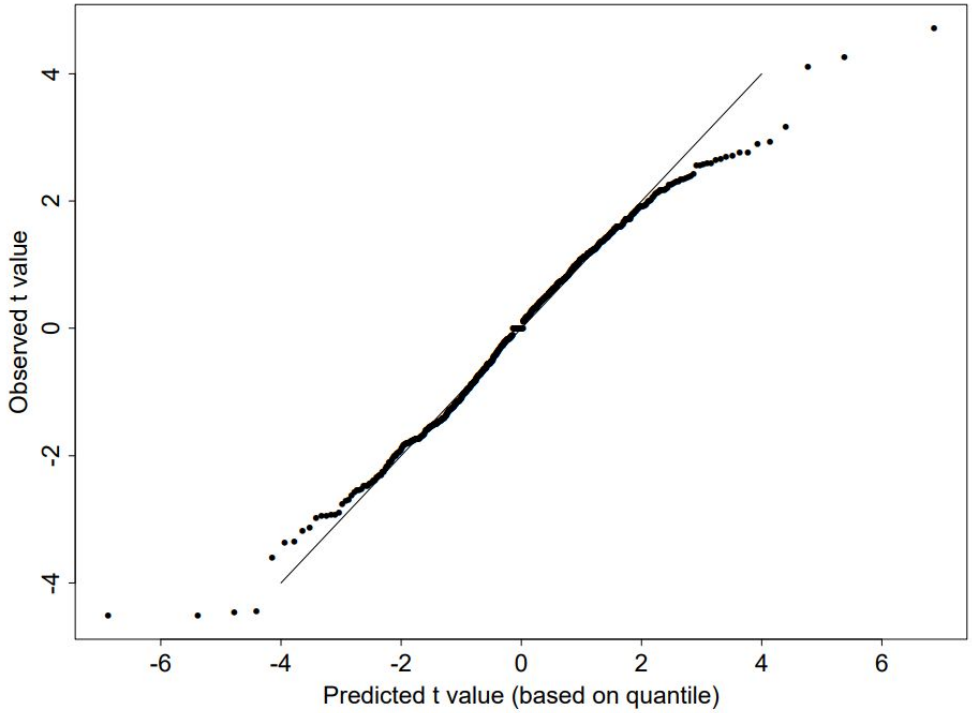


Figure 2: QQ plot comparing the distribution of 1,000 values of $\tilde{t}$ to the values they should have under a t distribution with 5 degrees of freedom. All points would fall on the line $y = x$ if the distributions matched.

Валидация ML моделей

ИТМО

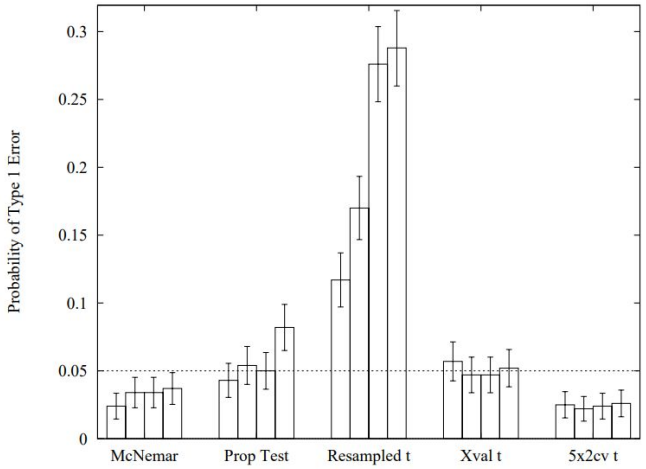


Figure 5: Probability of Type I error for each statistical test. The four adjacent bars for each test represent the probability of Type I error for $\epsilon = 0.10$, 0.20, 0.30, and 0.40. Error bars show 95% confidence intervals for these probabilities. The horizontal line shows the target probability of 0.05.

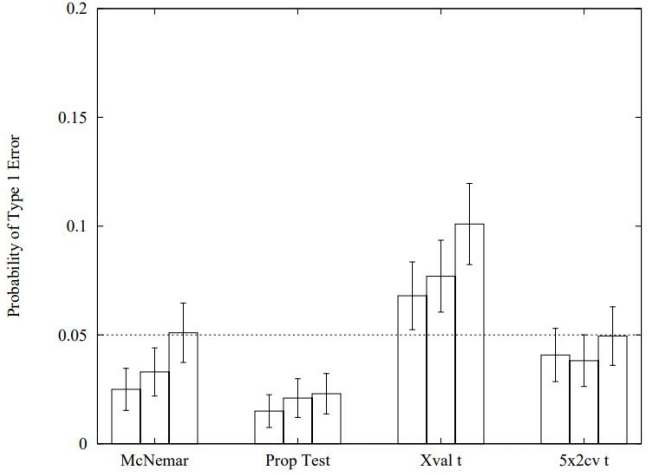


Figure 7: Type I error rates for four statistical tests. The three bars within each test correspond to the EXP6, Letter Recognition, and Pima data sets. Error bars are 95% confidence intervals on the true Type I error rate.

Валидация ML моделей

ИТМО

Стоит ли оно того? Не факт :)

Но если интересно, то вот основополагающие статьи:

- Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms, 1998, Dietterich
  - https://sci2s.ugr.es/keel/pdf/algorithm/articulo/dietterich1998.pdf
- Inference for the Generalization Error, 2003, Claude Nadeau, Yoshua Bengio
  - https://link.springer.com/article/10.1023/A:1024068626366#article-info
- Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms, 2004, Remco R. Bouckaert, Eibe Frank
  - https://link.springer.com/chapter/10.1007/978-3-540-24775-3_3
- Документация библиотеки MLXtend
  - https://rasbt.github.io/mlxtend/

Спасибо за внимание!