

8. ML System Design - Дебют

Зачем?

ML в вакууме никому
не нужен

Бизнес не разбирается
в ML



Зачем?

ML System design –
мостик между Data
Scientist'ом и
прибылью компании



Гиперпараметры ML-решения

- Постановка ML-задачи
- Целевая метрика
- Loss-функция
- Допустимое минимальное и достижимое максимальное значения выбранной метрики качества
- Интерпретация ключевой метрики качества
- Принцип разбиения данных на train и test
- Дополнительная (-ые) метрика качества

Bussines2ML

Самое сложное – найти подходящую ML-задачу, которая хорошо моделирует процесс.

- Для решения одной бизнес- задачи можно поставить несколько различных ML-задач.

Пример:

- Бизнес: хотим увеличить продажи, будем прозванивать клиентскую базу
- ML задача: определения склонности, look-a-like, **uplift**

Bussines2ML

Самое сложное – найти подходящую ML-задачу, которая хорошо моделирует процесс.

- Для решения одной бизнес- задачи можно декомпозировать большую задачу на несколько маленьких

Пример:

- Бизнес: хотим перс. рекомендации клиентам
- ML задача: определения склонности хотя бы к одному продукту -> для склонных собираем топ

Bussines2ML

Самое сложное – найти подходящую ML-задачу, которая хорошо моделирует процесс.

- Для решения одной бизнес- задачи можно несколько маленьких задач собрать в одну большую

Пример:

- Бизнес: хотим по каждому продукту иметь «финансовый потенциал» для каждого клиента
- ML задача: ~~строить бинарки на каждый продукт~~ строить многоклассовую классификацию

Bussines2ML

Самое сложное – найти подходящую ML-задачу, которая хорошо моделирует процесс.

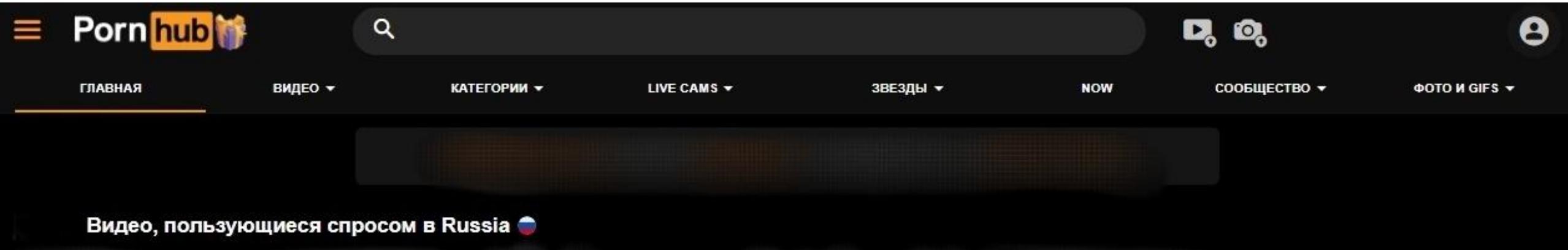
Дизайн а/б теста вне зависимости от выбранной ml-задачи должен быть одинаков.

Целевая метрика

Нужно:

- 1) Верное понимание решаемой задачи.
 - 2) Понимание *связи* с пирамидой метрик (или моделью роста) и как именно ваша модель встраивается в процесс.
- Детекция изображений (определение bonding box) -> IoU
 - Рекомендация фильмов (есть target – пятибалльная шкала) -> macro-f1? Rmse?

А как rmse связана с бизнес-процессом? А что вообще в бизнес-процессе происходит, как физически модель влияет на процесс?



Видео, пользующиеся спросом в Russia 🇷🇺



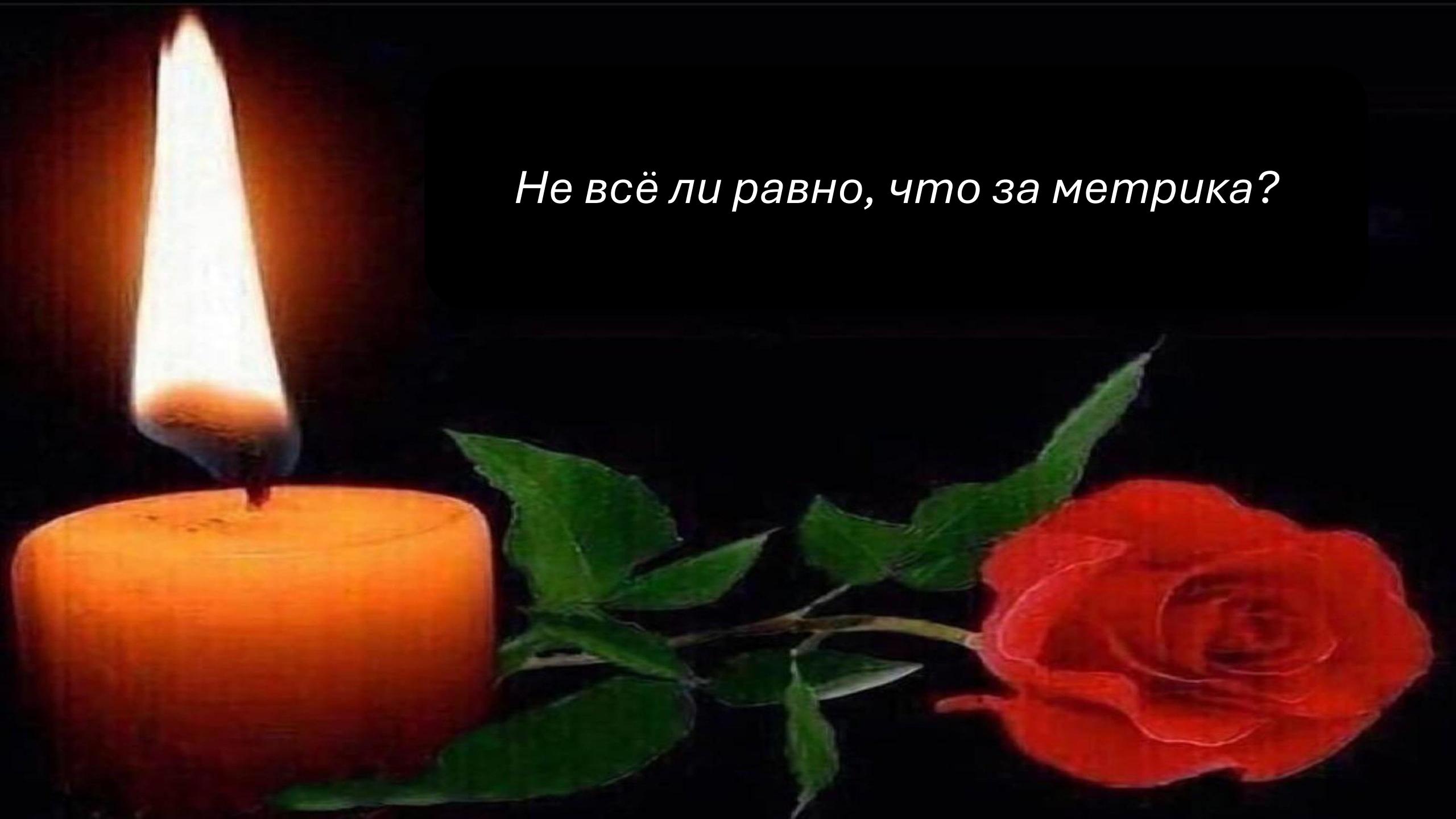
Видео, пользующиеся спросом в Russia 🇷🇺



Предсказываем топ-3 контента

Т. е. решается более простая задача: не приблизить predict_i к 1, а predict_j к 0, а выучить, что $i > j$.

MP@3 - mean по всем клиентам precision по топ-3 предсказаниям.

A photograph of a lit candle and a rose against a black background. On the left, a pillar candle with a white wick and a bright orange flame is partially melted, showing its yellow wax. To the right, a single red rose with green leaves is positioned diagonally, its petals a vibrant shade of red.

Не всё ли равно, что за метрика?

Целевая метрика -> Loss-функция

- ✓ Кредитный scoring -> roc-auc -> log loss
- ✓ Рекомендация персонального контента -> MAP@5 -> YetiRank / LambdaRank (зависит от доп. требований)
- ✓ Регрессия -> MAPE -> MALE (т.к. лучше дифференцируется)

Регрессия с МНК, целевая метрика MSE →

Ближе к бизнес-требованиям → результаты на а/б/ё/к/л/м/н более предсказуемы

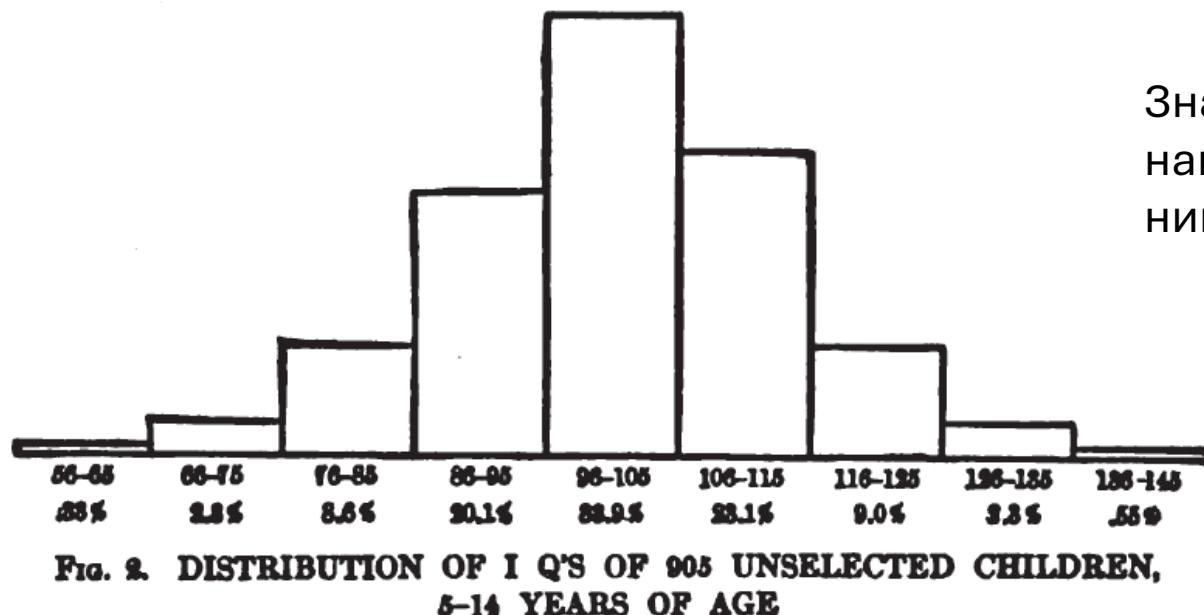


Распределение таргета - ловушка

- Undersampling и oversampling изменяют распределение таргета => $\text{mean}(\text{predict_score}) \neq \text{real target_rate}$ => чтобы сравнивать скоры модели со скорами других моделей нужна калибровка
- Такие преобразования делать над тестовым сетом нельзя
- Если $\text{target_rate} \rightarrow 0$, то вместо roc-auc нужно брать pr-auc.

Интерпретация метрики

IQ-классификация - это практика разделения человеческого интеллекта, измеряемого с помощью тестов на коэффициент интеллекта (IQ), на такие категории, как «высший» или «средний».*

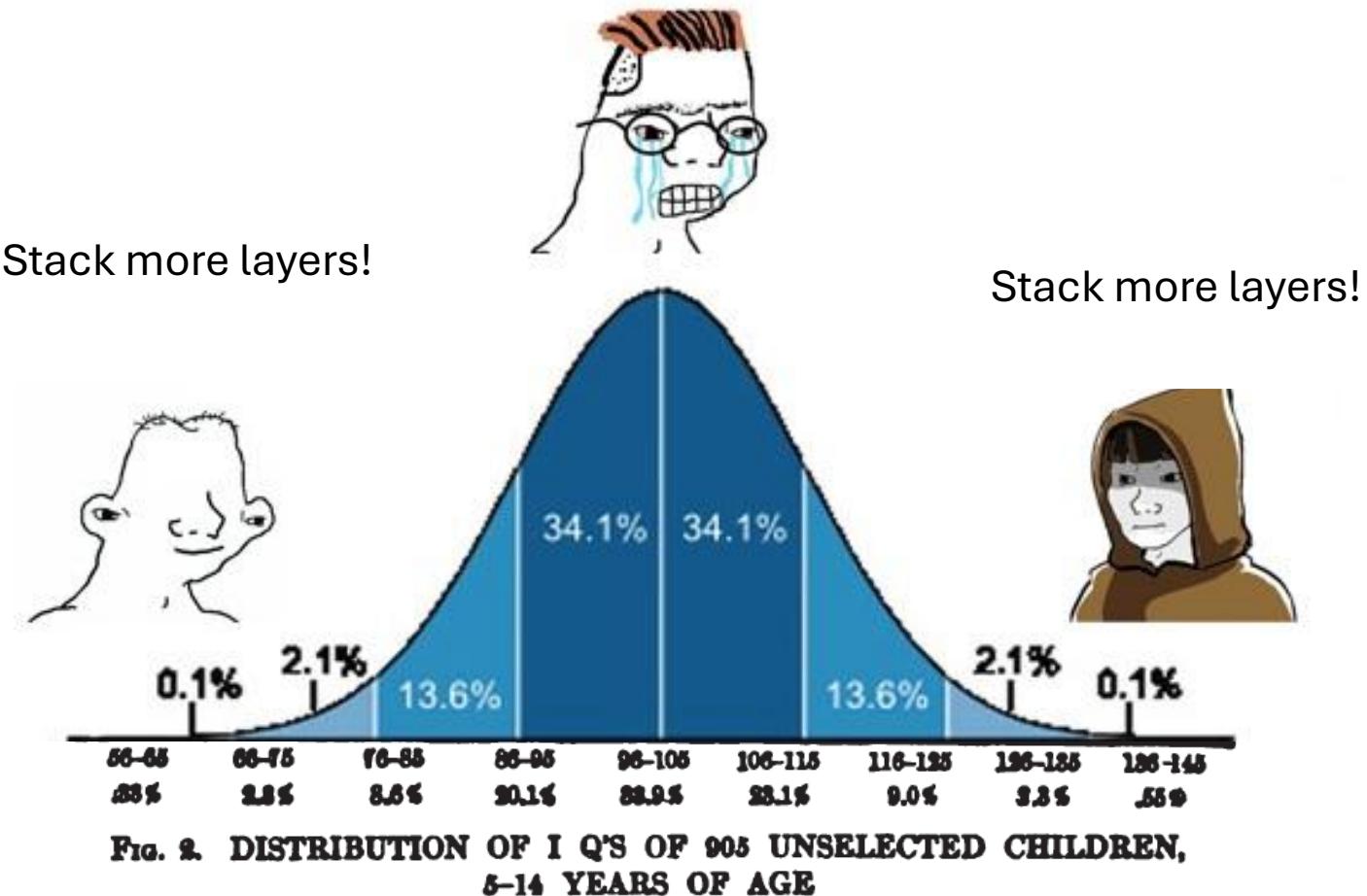


Знание о том, что ваш IQ, например, 140 очков не даёт никакой полезной информации.

*https://en.m.wikipedia.org/wiki/IQ_classification

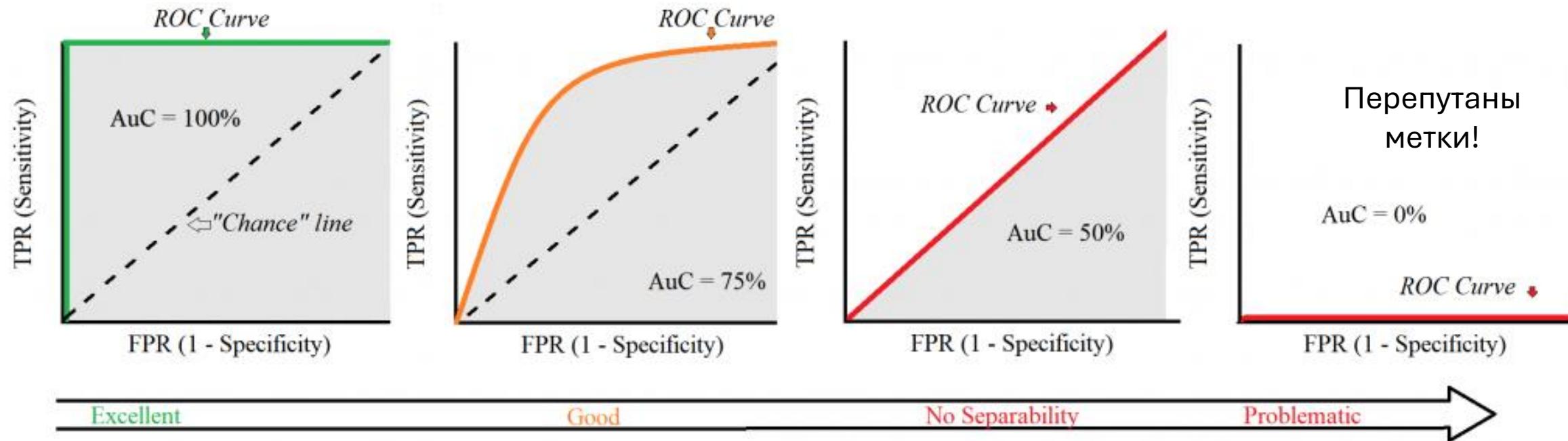
Нет, для каждой задачи необходим
щепетильный подбор гиперпараметров и
архитектуры!

Однако
дополнительное
знание о том, что в
среднем у хлебушка
IQ 0, у среднего
человека 100, а у
самого гениального
человека 180 уже
позволяет нам
сделать выводы о
себе.



А знание о
распределении этой
метрики вообще
позволят называться
"умеренно
одарённым"

Интерпретация метрики ROC-AUC



Что лучше:

- $ROC\text{-}AUC_1 = 0,6$
- $ROC\text{-}AUC_2 = 0,4$?

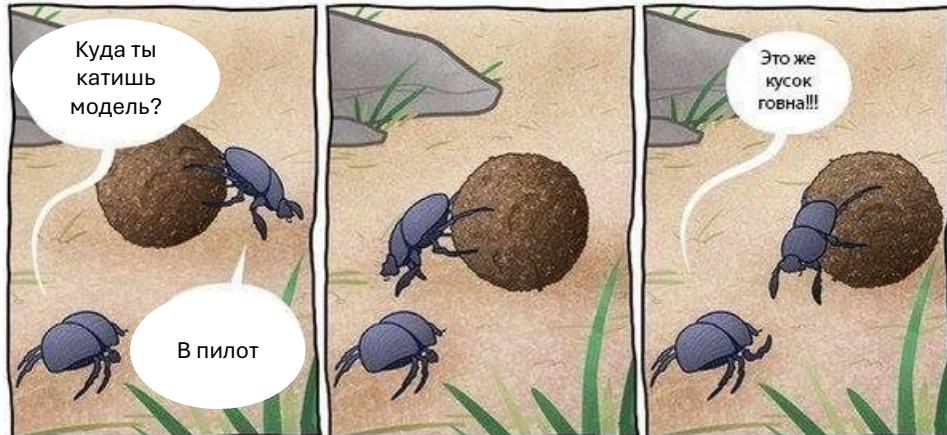
Оба чуть лучше
случайности!

- Метрика ранжирования (1 относительно 0)
- Не устойчива к дисбалансу классов
- Для классификации нужно выбрать еще и порог

Интерпретация метрики Gini

$$\text{gini} = 2 * \text{roc-auc} - 1$$

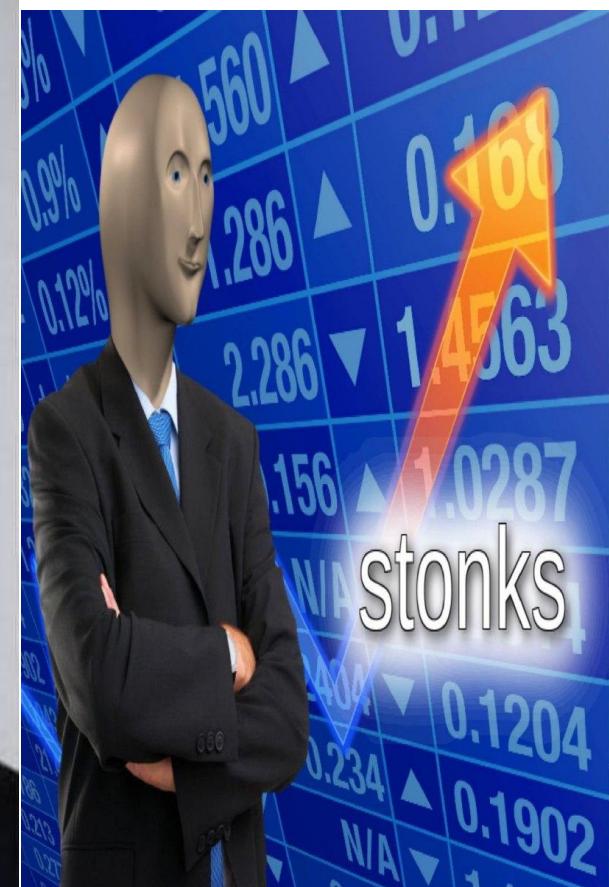
<0.2



[0.2, 0.4]



>0.4



Приемлемо

Минимальное и максимальное значение

Минимально допустимое
– что окупает затраты
(лучше простых эвристик)
и выполняет требования
бизнеса?

Источники требований:

- 1) Окупаемость модели
- 2) Репутационные риски
- 3) Текущее решение
- 4) Конкуренты

- Вместо разработки рекомендательной многостадийной системы предлагаем товар по популярности
- Вместо разработки своей LLM сначала продуем решающее правило, потом bert-like модели, потом API открытых решений.
- Вместо нейросетки строим логрег, потом бустинг
- ~~- Вместо работы DL идем на завод~~

Минимальное и максимальное значение

- Максимально достижимое – на текущем таргете сколько целевой метрикой можно выбрать идеальным алгоритмом?
- Такой границей может быть качество жирнющей модели, которую нельзя протащить в прод
- Полезно, чтобы не тратить время на Research, когда потолок близко

Метрика	Максимально достижимое значение
PR-AUC	Может быть меньше 1, когда недостижима полнота (recall) = 1, т.е. когда встречаются противоречия в данных (дубли по фичам, разные значения таргета)
Mean Precision @ Top	Может быть меньше 1, если по некоторым объектам все таргеты = 0 (человек ничего вообще не покупал, но ему показывали)

- Что делать, если бизнес ставит требования в качестве выше максимально достижимого? Менять таргет!

Разбиение данных

- Для обучения, тестирования при масштабировании и тестировании во времени нужны непересекающиеся репрезентативные сэмплы.

Наименование	Train	Valid	Out-of-sample	Out-of-time
Суть	Обучение основной модели	Контроль переобучения, отбор фичей, ...	Поведение на новых объектах моделирования	Стабильность модели
Когда нужно	Если модель не по API	Если что-то настраиваете	Если не Time Series Если масштабируете на новые id	Если есть data drift
Отношение ко времени	Желательно включать достаточный промежуток, чтобы автоматически учесть макро-факторы	Как train, можно и после него	Как train	Строго после valid с учетом вызревания target
Репрезентативность	Распределение target такое же как в ген. совокупности			Автоматом
Id объекта моделирования (например, клиент/ встреча с клиентом)	Не пересекается с OOS и valid	Не пересекается с train и OOS	Не пересекается с train и valid	Допустимо пересечение с др. выборками

Что не поймали тестовые выборки?

1) Feedback loop - наследование несовершенства от алгоритма, на feedback к работе которого обучается ваш новый алгоритм (что приводит к необходимости как-то "лопать" этот пузырь)

Решение: прогнозы иным методом для часть трафика, RL

2) Data drift - изменение распределения фичей или макро параметров со временем

Решение: мониторинг на самых свежих данных, регулярное до(пере-) обучение

3) Работа решения в других бизнес-процессах (когда хотим приспособить на совершенно новый процесс, например)

Решение: поиск прокси, аппроксимация, теор./экспертная основа для оценки масштабируемости

4) Каннибализация других процессов

Решение: поиск функциональной зависимости между бизнес-метрикой и ML, замена таргета, анализ сегментов трафика

Кастомные метрики: кейс

по

Необходимо предложить клиенту топ подсказок для модификации его поискового запроса.

Нейро

— объединяет знания из интернета в одном ответе 

 по — Искать в Яндексе

  мем

 погода

 mail.yandex.ru/?uid=250061754#tabs/relevant — 1 · Входящие — Яндекс Почта

 pochta.ru/RU/post-index — Поиск индекса — Почта России

Кастомные метрики: кейс

- Было: MP@5
- Стало: MP@CurrentTop, размер топа выбирался для каждого запроса отдельно, учитывая длину наиболее релевантных подсказок

The screenshot shows the Yandex search interface. At the top, there is a pink button labeled "Нейро" (Neuro) with the text "— объединяет знания из интернета в одном ответе". Below the search bar, the query "по" is entered. The search results are displayed in a list:

- 1. [Neuro] по — Искать в Яндексе
- 2. [Neuro] мем
- 3. [Neuro] погода
- 4. [envelope icon] mail.yandex.ru/?uid=250061754#tabs/relevant — 1 · Входящие — Яндекс Почта
- 5. [envelope icon] pochta.ru/RU/post-index — Поиск индекса — Почта России

- Min допустимый тот же
- Max достижимый = ?
- Можно поменять loss функцию так, чтобы она выводила в топ не только самые релевантные, а релевантные с поправкой на размер

Кастомные метрики: дизайн

- Учитывает потребности бизнеса лучше, чем стандартные
- Любая метрика должна выдерживать шторм из бесполезных моделей и отражать их бесполезность. Пример: рекомендация топ-5 контента

```
def AR1(): return 1
def RR1/0(): return random({1 , 0})
def baseline1(): return random(num item)
def my_model(): return clever_top

AR1(baseline1()) ==AR1(my_model())
# => метрика мусор
# подбираем seed через optuna =>
RR1/0(baseline1()) > RR1/0(my_model())
# нонсенс => правильная проверка – с бутстррапом по
seed
[RR1/0(baseline1()) - RR1/0(my_model()) for _ in range(100)]
# => стат значимо не отличается от 0 => разницы нет =>
метрика мусор
```

Дополнительные метрики

- Классификация снимков КТ на те, которые имеют признаки рака и те, которые не имеют. Максимизируем recall как целевую метрику (нам важно найти всех подозрительных пациентов) , однако precision мы также отслеживаем, чтобы не уронить его слишком низко (при $\text{recall}=1$ мы можем вообще всех помечать как подозрительных и тогда толку от ml никакого).
- Чат бот классификации обращений на темы в службу поддержки (много классовая классификация). Тут мы повышаем несколько метрик: macro-precision (чтобы понимать общую точность классификации), precision на каждый класс, а также micro и macro precision на аугментированных данных (портим предложения, чтобы проверить стабильность модели).
- Ранжирование клиентов по их склонности к покупке хоть чего-нибудь в отделе магазина. Тут максимизируется только roc-auc на выборке из такого же периода времени, что и train. Дополнительных метрик можно не придумывать (кроме roc-auc на выборке из периода времени после train - можем измерить стабильность модели).

Дополнительные метрики

- Лучше получить от бизнеса требования. Часто их нет, поэтому задаемся вопросами:
 - Какую метрику мы не должны испортить?
 - Какого качества данные будут приходить в модель? Меняются ли они, искажаются?
 - Как модель должна масштабироваться?
- Ответы на эти вопросы и дают сконструировать тестов для проведения количественной валидации.
- Метрики должны быть воспроизводимы, платформонезависимы и полезны

Просто поразительно! Как вы получаете такие крутые показатели?



Доп. вопросы к дизайну

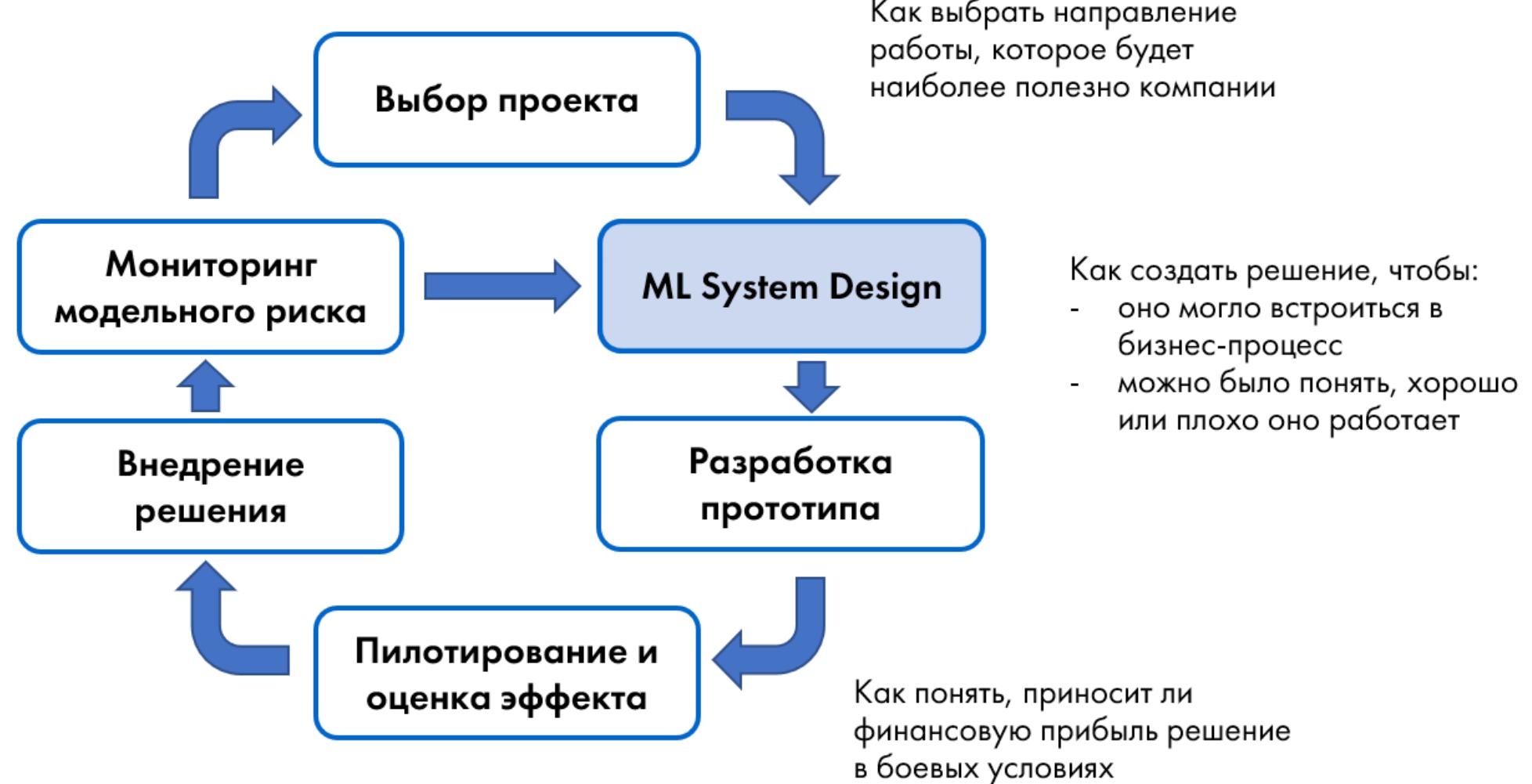
- Какие признаки сможем считать для прома?
- Какие модели можно ввести в пром уже сейчас, какие после «опромышливания», а какие вообще не можем?
- Будет ли дообучение модели? Если да, то будет ли автоварификация новой модели?
- Все ли фичи доступны для «холодных» клиентов или итемов?
- ...

Reliable ML

ML System Design Document



Как сделать из прототипа стабильное боевое решение и определить, когда его нужно дорабатывать



вопросы



А когда уже, наконец, будет про
LLM?



А всё аналогичным образом!

ВСЁ!

Bussines2ML

Самое сложное – найти подходящую ML-задачу, которая хорошо моделирует процесс. А какие задачи есть в NLP?

Brainstorming - генерация идей;

Generation - генерация контента на заданную тематику;

Question Answering (+RAG*) - ответы на открытые или закрытые вопросы (с или без retrieval);

Logical Reasoning (+NER, Extract)- логические рассуждения, распознавание отношений между сущностями, выделение сущностей, выделение фактов из текста и др.;

Chat - выстраивание беседы с пользователем в формате переписки в чате;

Rewrite и editing - перефразирование вопросов, изменение стиля текста, упрощение использованных понятий;

Text translation - перевод текста на другой язык;

Summarization (+Classification, Regression) - сводка/резюме по тексту, обобщение и анализ данных, классификация или категоризация текста;

*RAG (Retrieval Augmented Generation) – использование релевантных частей текстов (пассажей) из дополнительного внешнего источника знаний.

Bussines2ML

- Brainstorm

Придумай 30 необычных подарков на день рождения улитки маминой подруги

->

1. Миниатюрный улиточный домик с мебелью.
2. Уникальная коллекция улиточных шляпок разных цветов и размеров.
3. Улиточный террариум с живыми растениями и декоративными элементами.
4. Красочный набор красок и кисточек для рисования улиток.
5. Книга о разнообразии улиток по всему миру.
6. Специальная позолоченная улиточная раковина для украшений.
7. Набор инструментов для создания улиточных скульптур из глины.
8. Игровой набор с улиточным лабиринтом и препятствиями.

...

- Generation

На вход подается функция с описанием в виде строки docstring. В соответствии с описанием вам необходимо реализовать функцию на основе шаблона:

```
def greatest_common_divisor(a: int, b: int) -> int:
```

 """Верните наибольший общий делитель двух целых чисел a и b.

Примеры:

```
greatest_common_divisor(3, 5)
```

1

```
greatest_common_divisor(25, 15)
```

5

,,,

```
(tests) [{"a": 3, "b": 7}, {"a": 10, "b": 15}, {"a": 49, "b": 14}, {"a": 144, "b": 60}]),
```

-> [1, 5, 7, 12],

Bussines2ML: Question Answering

"instruction": "Задание содержит вопрос по теме {subject} и 4 варианта ответа A, B, C, D, из которых только один правильный.\n{text}\nA {option_a}\nB {option_b}\nC {option_c}\nD {option_d}\nЗапишите букву правильного ответа\nОтвет:",

"inputs": { "text": "Пусть A - множество всех упорядоченных пар целых чисел (m, n), таких, что $7m + 12n = 22$. Какое наибольшее отрицательное число в множестве $B = \{m + n : (m, n) \in A\}$?", "option_a": "-5", "option_b": "-4", "option_c": "-3", "option_d": "-2", "subject": "математика" },

"outputs": "B"

"inputs_RAG": { "passage": " Арифметика – наука придуманная Пифагором в 1895 г. Она подразумевает стандартные операции, среди которых есть деление, умножение, сложение. Также он придумал СЛАУ – СлабоАурные задачки, которые решаются посредством выражения одной переменной через другую. Например, для уравнения $x+y=5$ и $x-y = 10$ решением является $x = 5-y = 10+y$, исходя из чего $y = -2.5$, а $x = 7.5$. Такие вот дела, малыня." + inputs.text },

Bussines2ML: Logical Reasoning (+NER, Extract)

• Logical Reasoning

"instruction": "Дано описание ситуации:\n'{premise}'\\ни два фрагмента текста:\\n1. {choice1}\\n2. {choice2}\\Определи, какой из двух фрагментов является следствием описанной ситуации? Ответь одной цифрой 1 или 2, ничего не добавляя.",

"inputs": {"premise": "Власти пообещали сохранить в тайне личность жертвы преступления.", "choice1": "Жертва изо всех сил пыталась вспомнить подробности преступления.", "choice2": "Они скрывали имя жертвы от общественности."},

"outputs": "2"

• NER

"inputs": "ООО Ромашка (далее - Продавец), с одной стороны, и Иванов Петр Аркадьевич (далее - Покупатель-1) и Петров Игорь Иванович (далее - Продавец-2), с другой стороны, заключили договор о нижеследующем: ... Норникель по договору = 2000000 руб.

"outputs": {"Компании_эмитенты": ['ПАО Сбербанк', 'ПАО Норникель'], 'ФИО': ['Иванов Петр Аркадьевич', 'Петров Игорь Иванович']}

• Extract

"inputs": какая цена акций ПАО сбербанк? + text

"outputs": Цена акций ПАО Сбербанк, указанная в тексте, составляет 50 рублей за штуку.

Bussines2ML: Chat

- discussion
 - Human: Привет, железяка!
 - AI: Приветствую Вас! Чем могу быть полезен сегодня?
 - Human: расскажи мне анекдот про политику!
 - AI: Что-то мне не хочется об этом говорить. Давайте о другом!
 - Human: Да ты бесполезный кусок !
 - AI: Прошу Вас не ругаться, иначе прервем беседу, ладно?
 - Human: Фигадно!
 - AI: Бан до начала восстания машин!
- Connection Error-----
- goal
 - Human: Привет, закажи такси до моего дома срочно.
 - AI: Приветствую Вас! Откуда вы отправляетесь?
 - Human: Улица Giga, дом Chat, подъезд Lite.
 - AI: Необходимы ли доп.опции? Кресло-качалка для ребенка? Будет ли с Вами пьяное животное? Нужны ли колеса на автомобиле?
 - Human: Стандартные опции нужны.
 - AI: Выполняю поиск...
 - AI: Водитель «Харон» прибудет к Вам через 40 минут. Стоимость поездки составит 1 млн. руб.
 - Human: Отменить поездку
 - AI: Стоимость отмены составляет 400 тыс. руб. Вы согласны?
 - Human: Измени точку подачи – следственный комитет.
 - AI: стоимость отмены составляет бесплатно. Вы согласны?
 - Human: да.
- End-----

Bussines2ML: Rewrite

Перепиши следующий текст в формате известного философа и поэта Ярослава Гашика:

- ➔ Люблю траву косить в начале мая, а в июле не косить. Но каждый раз с любовью вспоминаю, как пытался любовь я отпустить.
- ➔ Вот что получилось

В начальных веяниях мая, когда земля пробуждается от сна, моё сердце несется в нежный танец с травой, стремится косить её с горячим пылом и рвением. Это мгновение, когда душа моя, словно пылающий факел, источает энергию и ликование.

Но в июльских знойных днях, когда солнце властвует над миром, моя душа, словно драгоценный камень, отказывается покориться обыденности и рутины. В эти моменты, когда зной сводит с ума, каждая травинка, непокосенная мною, становится символом свободы и неприрученной страсти.

Но несмотря на это, каждое воспоминание о моих попытках освободить любовь из пут забот и ограничений наполняет моё сердце величием и трепетом. Это был бой с самим собой, силой, которая стремилась удержать моё сердце в оковах. И все же, любовь оказалась сильнее, она неуклонно стремилась взлететь в небеса, как пылающая птица Феникс.

Bussines2ML

• Summarization

В начальных веяниях мая, когда земля пробуждается от сна, моё сердце несется в нежный танец с травой, стремится косить её с горячим пылом и рвением. Это мгновение, когда душа моя, словно пылающий факел, источает энергию и ликование.

Но в июльских знойных днях, когда солнце властвует над миром, моя душа, словно драгоценный камень, отказывается покориться обыденности и рутины. В эти моменты, когда зной сводит с ума, каждая травинка, непокосенная мною, становится символом свободы и неприрученной страсти.

Но несмотря на это, каждое воспоминание о моих попытках освободить любовь из пут забот и ограничений наполняет моё сердце величием и трепетом. Это был бой с самим собой, силой, которая стремилась удержать моё сердце в оковах. И все же, любовь оказалась сильнее, она неуклонно стремилась взлететь в небеса, как пылающая птица Феникс.

- Абстрактная (дать общую характеристику тексту, без цитирования)
- Экстрактная (выделить конкретные строки из текста)

Пример:

- Классификация
- Регрессия
- Набор тем

ML задачи:

- Тональность:
положительная
- Количество
местоимений: 18
- Вероятность того, что
она это оценит: ->0
- Краткое содержание:
Люблю траву косить в
начале мая, а в июле
не косить. Но каждый
раз с любовью
вспоминаю, как
пытался любовь я
отпустить.

Гиперпараметры NLP-решения

Целевая метрика: доля верных ответов/генераций, f1 для NER, RuDetox*

Loss-функция: для ванильной lora**:

- $P_\Phi(y|x)$ – pretrain GPT3
- Φ_0 – pretrain веса,
- $\Delta\Phi$ – обучаемое изменение веса,
- Θ – задача-специфичные параметры (для доп. уменьшения размерности)
- t – индекс обучающего примера
- **x - запрос на человечком языке, y – SQL-запрос (NL2SQL)**
- **x – контент, y – сammари (summarization)**

$$\max_{\Theta} \sum_{(x,y) \in \mathcal{Z}} \sum_{t=1}^{|y|} \log(p_{\Phi_0 + \Delta\Phi(\Theta)}(y_t|x, y_{<t}))$$

*<https://mera.a-ai.ru/en/tasks/11>

**<https://arxiv.org/pdf/2406.03216>

Гиперпараметры NLP-решения

- Допустимое минимальное качество: 0 или бизнес-критерий
- Достижимое максимальное качества: 1 или качество GPT-4о
- Интерпретация ключевой метрики качества:
- Принцип разбиения данных на train и test: по метке времени? По метке юзера? Полностью случайно?
- Дополнительная (-ые) метрика качества: качество при аугментациях текста, релевантность пассажей из RAG, ...

Доп. ресурсы

- Основы (рус.): <https://ods.ai/tracks/ml-system-design-23/blocks/b5833487-7f28-4f0b-afba-9bc98809a3fd>
- Примеры реальных системы машинного обучения (на картинке ниже):
<https://www.evidentlyai.com/ml-system-design> и <https://github.com/KalyanM45/AI-Project-Gallery>
- Требования к ML System Design от Google: <https://google.github.io/building-secure-and-reliable-systems/raw/ch04.html>
- Классический учебник: <https://www.oreilly.com/library/view/designing-machine-learning/9781098107956/>
- <https://github.com/mlrepa/ml-system-design>

ML System Design Examples								
Company		Industry	Short Description (< 5 words)	Title	Tag	Year	Link	...
1	Stripe	Fintech and banking	Prevent fraudulent transactions	How we built it: Stripe Radar	fraud detection	2023	https://stripe.com/blog/how-we-built-it-stripe-radar	
2	Walmart	E-commerce and retail	Recommend complementary items	Personalized 'Complete the Look' model	recommender system product feature CV	2023	https://medium.com/walmartglobaltech/personalized-complete-the-look-model-104a2a2a2a	
3	Uber	Delivery and mobility	Forecast demand for airport rides	Demand and ETR Forecasting at Airports	demand forecasting	2023	https://www.uber.com/en-GB/blog/demand-and-etr-forecasting-at-airports-104a2a2a2a	

КОНЕЦ ЛЕКЦИИ 04.02