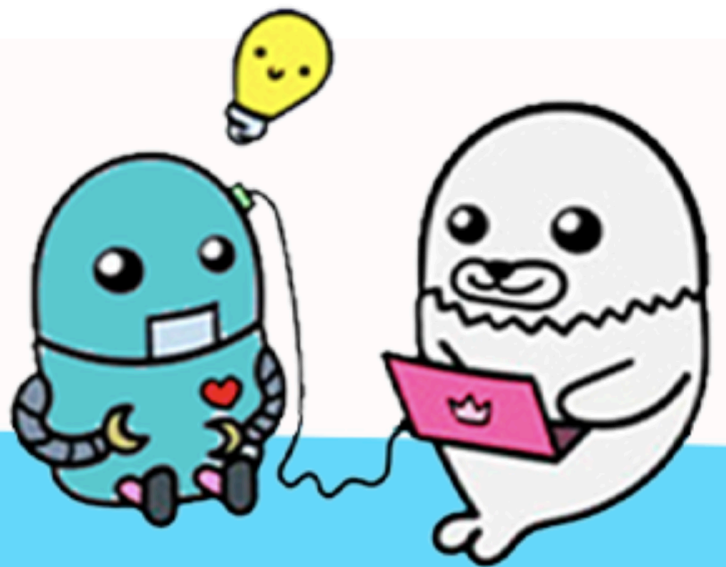
Open
Data
Science

Научный метод и воспроизводимость в DS

Владислав Горбунов — Head of DS



О чем поговорим

- Характеристики предмета исследований в DS
- Гипотезы и теории в DS
- Построение прогнозов в DS
- Проведение экспериментов в DS
- Документирование и рецензирование работ
- И причем же тут MLOps?

Научный метод в DS

Научный метод — это итеративный процесс исследования, в котором элементы постоянно сменяют друг друга при появлении новой информации.

- Предмет исследования и его характеристики
- Гипотезы и теории
- Прогнозы
- Эксперименты

Давайте посмотрим на данные элементы с точки зрения работы специалистов по данным

Характеристики предмета исследований в DS

Предмет исследования для DS — данные о каком-то явлении.

- Суть анализа данных заключается в исследовательской деятельности именно на данных. Предмет исследования выражен в данных, методы работы связаны с обработкой данных.
- Результативность работы специалистов по данным напрямую связана с тем, собираются ли необходимые данные по предмету исследования и какого они качества.

Посмотрим, какие могут быть характеристики у предмета исследования - данных

- Наблюдения
- Определения
- Измерения

Характеристики предмета исследований в DS - наблюдения

Начальное наблюдение может быть сформулировано не из данных, а из наблюдений за работой одного из отделов

- В таком случае необходимо провести работу по сбору необходимых данных, для фиксирования начального наблюдения на них — это отдельная задача.
- Если не будет возможности собрать нужные данные для фиксации наблюдения, то DS — не подходит для решения этой задачи.

Наблюдение может быть сформулировано на основе имеющихся данных

- При наблюдении на данных какого-то явления, важно зафиксировать:
  - Источник данных;
  - Время;
  - Само наблюдение (текстовые описания, таблицы, фото, видео, аудио,….);
  - Интерпретация наблюдения;
- Важно рассмотреть вероятность случайного совпадения наблюдений. Иногда можно начать исследовать то, что объясняется просто случайностью и редко повторяется.
  - Необходимо проводить множественные наблюдения, чтобы исключить вероятность случайного совпадения.
  - Для этого важен размер выборки, ибо чем она больше, тем меньше вероятность случайного совпадения наблюдений.
- Данные редко изолированы. Важно осознавать связность данных, каким образом они влияют на друг друга.
  - Не смотря на то, что интересующее явление может встречаться в одном датасете, возможно, влияющие на него переменные на самом деле будут в другом. А значит очень важно зафиксировать, что именно заинтересовало в данных с учетом других датасетов, которые могли повлиять на событие или могут быть связаны с ним.
  - Если используется производный источник данных, то есть уже совершается некоторое преобразование над исходными данными - стоит это зафиксировать.
    - В случае возникновения вопросов, стоит проверить, а не влияет ли на это предобработка исходных данных.
    - Также могут быть изменения в цепочке поставок данных, которые влияют на многие данные разом - стоит сообщить, если недавно были какие-то крупные изменения.

Характеристики предмета исследований в DS - наблюдения

Необходимо обеспечить доступ к свежим и корректным данным всех заинтересованных лиц;

- При наличие общих витрин данных, к которым имеют доступ и аналитики данных и заказчики, проще настроить правильную коммуникацию.
- А также дать возможность заказчику основывать свои наблюдения на данных, а не фантазии.

Полезно учитывать предыдущий опыт наблюдений за предметом исследования (других исследователей в других компаниях, так и проводимые наблюдения, эксперименты и практические работы над исследуемыми данными в вашей компании);

- Опыт из других компаний может подтвердить, что ваше наблюдение не случайно, ведь повторялось и у других в похожих условиях. А также позволит более широко посмотреть на проблему, чтобы случайно не сузить обзор первой идей.
- Опыт из вашей компании важен, так как если представление данных в компании со временем меняется, в том числе под влиянием предыдущих исследований. А значит, если какие-то уже были проведены работы над вашим предметом исследований, возможно, они уже заложены в текущее представление данных и могут оказывать влияние.

Характеристики предмета исследований в DS - определения

Наука о данных — междисциплинарная академическая дисциплина. Поэтому очень часто доменный язык науки о данных отличается от терминов используемых в областях, которые изучаются при помощи её методов.

- В других областях это может привести к проблемам с коммуникацией, а в науке о данных эта проблема стоит ещё более остро.
- Поэтому важно использовать единый язык, создать доступный всем глоссарий и поддерживать его актуальным в рамках исследования, чтобы сократить количество проблем в коммуникациях.

Что входит в определения:

- **Терминология**
  - **Глоссарий терминов из бизнеса**, которые используются в рамках исследования, язык доменной области.
    Совместно с представителями бизнеса стоит создать и поддерживать актуальным глоссарий бизнес терминов и наименований, принятых в компании.
  - **Глоссарий терминов из науки о данных**, которые используются в отчетах и технических документах.
    Не нужно приносить заказчику страничку с Википедией по анализу данных, однако, если вы пользуетесь техническими терминами для обозначений сущностей, стоит зафиксировать их с пояснениями. Это позволит и исследователю использовать привычную ему терминологию, и представителям бизнеса понимать, что пишет и говорит исследователь.
- **Меры и метрики**
  - **Бизнес метрики;**
    При оценке наблюдаемого явления и результатов исследования (критерии успеха / провала) будут использоваться бизнес метрики изучаемой области. А это значит, что от точности и прозрачности их определений будут зависеть и направление исследований, и их успешность.
  - **Технические метрики;**
    При анализе данных и обучении модели будут скорее всего использоваться технические метрики, которые не полностью отражают бизнес метрики. А ведь по последним будет оцениваться результат работ. Поэтому важно выбрать подходящие технические метрики, которые затем стоит связать с бизнес метриками и зафиксировать до старта работ по экспериментированию.

Характеристики предмета исследований в DS - измерения

- Так как работа ведется с данными о предмете исследования, важно тщательно собирать и записывать все измерения.
- При проведении разведочного анализа как раз цель и стоит в том, чтоб максимально описать и зафиксировать различные важные измерения о предмете исследования.

На основе характеристик предмета исследования формулируется запрос на исследование:

- Поиск способов исправления или улучшения наблюдаемого явления (оптимизация метрик).
- Выяснение причин и объяснение принципов работы наблюдаемого явления
  (в будущем может привести к запросу на улучшение наблюдаемого явления - оптимизацию метрик).

Гипотезы и теории - Формулирование гипотез

Гипотезы должны выдвигаться на основе наблюдений на данных.

- Гипотезы должны объяснять наблюдаемое явление или предлагать метод изменения наблюдаемого явления;
- Гипотезы подвержены личному опыту и взглядам человека, поэтому важно постоянно проверять себя.
  - Основываем ли мы гипотезы на наблюдениях или предполагаем то, чего на самом деле не наблюдали?
  - Есть ли альтернативные варианты объяснений и улучшений? Почему мы выбрали именно этот?
  - Не гонимся ли мы за новизной? Достаточно ли у нас оснований выдвигать именно такое предположение или есть более простые?

Гипотезы можно сформулировать в форме:

- Математических моделей;
- Логических рассуждений;

Можно использовать любую форму, однако в случае логических рассуждений стоит не забывать, что вашу гипотезу нужно всё равно привязать к данным.

Фальсифицируемость - важно следить за тем, чтобы вашу гипотезу можно было “опровергнуть”. Гипотеза, которая может объяснить всё и работает всегда - это, конечно, звучит красиво, однако таким образом очень легко начать заниматься самообманом.

Гипотезы и теории - Сохранение гипотез

Так как мы не можем со 100% уверенностью подтвердить или опровергнуть гипотезу, очень важно сохранять все гипотезы, над которыми вы экспериментировали.

В процессе генерации новых гипотез стоит учитывать не только “успешные”, но и те, что не были поддержаны экспериментальными данными.

- Это важно, потому что когда гипотеза не согласуется с результатами проверки, мы не можем полностью её отвергнуть.
- При этом гипотеза очень часто состоит из нескольких утверждений и по результатам проверки мы не знаем точно, какое из них ложное. Мы понимаем лишь то, что в сумме эти утверждения не были поддержаны экспериментом.
- Таким образом даже из “проваленных” гипотез можно и нужно черпать информацию при генерации новых гипотез.

Гипотезы и теории - Ранжирование гипотез

- Ценность для бизнеса при успехе или провале гипотезы
  Бизнес ценность, которую принесет проверка этой гипотезы (в т.ч. влияние на стратегические цели / KPI / принятие важных решений).
  Гипотеза может как быть поддержана экспериментальными данными, так и противоречить им. Если независимо от результатов проверки гипотеза способна принести ту или иную ценность компании, то её стоить проверить раньше, чем гипотезы, которые несут ценность только в случае успешной проверки.
- Больше эффект -> Меньше данных
  В первую очередь лучше проверять гипотезы, которые предполагают сильную зависимость или же высокий ожидаемый эффект. Большой эффект легче заметить, и для подтверждения его наличия не требуется большой объем данных.
- Есть наработки -> Проще задизайнить эксперимент
  Гипотезы, основанные на предыдущем опыте вашей команды или других исследователей стоит проверять раньше чем гипотезы без предыдущего опыта, только на наблюдениях за данными. Таким образом у нас на старте уже есть дополнительная информация о предмете исследования, а значит можно выдвигать более обоснованные гипотезы.
- Проще проверить -> Меньше времени и усилий на эксперименты, успеете проверить больше
  При ранжировании стоит учитывать предсказательную силу гипотезы. Гипотезы, которые позволяют делать более простые предсказания в проверке стоит проверять раньше.

Этот порядок приоритизации помогает сбалансировать между стремлением к получению максимальной ценности для бизнеса и необходимостью эффективного использования ресурсов. Однако в зависимости от конкретной ситуации и доступных данных порядок может быть скорректирован.

Прогнозы

- Стоит заранее фиксировать прогнозируемые гипотезой эффекты.
  - Можно выдвинуть гипотезу об улучшении определенной метрики при помощи “новой модной модели”, на основании того, что эта модель показывала “хорошие результаты” в похожей задаче. Однако, если не зафиксировать какой именно эффект мы считаем “улучшением”, гонять модель и подгонять результат можно очень долго.
- Важно, обсуждать какие именно предсказания, выдвигаемые на основе гипотезы, можно и стоит проверять.
  - В зависимости от формулирования прогноза, можно получить и как простую проверку, которую можно сделать за пару дней, так и эксперимент, которому пары лет не хватит для завершения.

Эксперименты - Дизайн эксперимента

Перед проведением экспериментальной проверки стоит зафиксировать основные аспекты планируемого эксперимента:

- **На каких данных он проводится?** Важно, проводить тестирование не на тех же данных, на которых проводилось наблюдение и генерировались гипотезы, таким образом можно попасть в ловушку удачно совпавшей выборки.
- **Какие методики тестирования используются?** Желательно планировать эксперимент с использованием методов, позволяющих максимально исключить предвзятость и неосознанная подтасовку результатов.
- **Какие ожидаемые эффекты?** Желательно не потерять к этому моменту сделанные прогнозы, чтобы сверить свой дизайн эксперимента ещё раз с тем, а что нужно было проверить. Для всех заинтересованных лиц должно быть понятно, какие результата эксперимента поддерживают гипотезы, а какие будут ей противоречить.

Не спешите бежать и проводить эксперимент сразу после того, как получили первую версию его дизайна.

- Обязательно стоит провести ревью полученного дизайна с другими исследователями в вашей команде и с представителями бизнеса.
- Также хорошей идей может быть рассмотрение альтернативных дизайнов эксперимента для проверки тех же прогнозов. Таким образом у вас будут и альтернативные варианты для самопроверки и вы сможете быть чуть более уверены, что выбрали наилучший метод проверки.

Помните о соотношении затрат и выгод. - Каждый эксперимент можно подвести к тому, сколько времени он займет, какие ресурсы потребует и что по итогу принесет (и в случае “успеха”, и в случае “провала”). - Всем заинтересованным лицам должно быть понятно, сколько будет стоить и что ожидается от проведения именно этой экспериментальной проверки.

Эксперименты - Проведение эксперимента

Очень часто в процессе проведения эксперимента может появиться соблазн “исправить немного” его, ведь появилась какая-то новая информация — стоит избегать подобного.

- Как только эксперимент начался, необходимо обязательно следовать его дизайну и выполнять всё в соответствии с ним.
- Все правки, замечания, возможные улучшения стоит оставить до фиксации результатов. Иначе можно начать заниматься подтасовкой результатов, даже если из хороших побуждений.

Эксперименты - Фиксация результатов

По результатам проведения эксперимента вы получаете некоторые данные, которые соотносятся с выдвинутым прогнозом или же нет.

- Интерпретация полученных результатов — это отдельная и важная задача.
  - Эксперимент лишь предоставил нам новые данные, их интерпретация - это уже работа исследователя, в которой вполне возможно влияние личного мнения.
  - Важно фиксировать и то, что показал эксперимент, и то, какие выводы на основе этого сделал DS, так как это не всегда полностью равнозначные данные.
- Новые экспериментальные данные, не важно поддерживают или противоречат выдвинутой гипотезе — это новые данные о предмете исследования, которые стоит внести в том числе в качестве новых наблюдений.
  - Часто после “неудачной проверки” данные о таком эксперименте забывают и больше не используются в генерации новых гипотез. А ведь они несут чуть ли не больше пользы, чем начальные наблюдения, ведь получены были в более контролируемых условиях.

Эксперименты - Воспроизводимость

- **Воспроизводимость результатов** возможна только если независимая группа других исследователей, которые анализируют тот же предмет исследования, что и вы, используют ваши методы анализа и приходят к тем же результатам.
  - Однако, чтобы шансов на это было больше, необходимо, чтобы ваше исследование обладало хотя бы воспроизводимостью методов.
  - Такое возможно только с несколькими командами в одной компании или в случае приемки работ от подрядчика.

- В рамках одной команды стоит обеспечить **воспроизводимость методов.**
- Ваши коллеги должны иметь возможность, повторить результаты ваших экспериментов (желательно без титанический усилий).
- Общие метрики и правила по обеспечению воспроизводимость методов выделить невозможно.
- Можно выделить следующие общие рекомендации:

Эксперименты - Воспроизводимость, примеры проблем

- Проблемы с зависимостями в библиотеках и зависимостями в зависимостях.
- Неверный порядок исполнения. JN сохранен без очистки кода, порядок не сохранен и некоторые переменные объявлены или инициализированы после использования.
- Нет нужных данных, например, указаны абсолютные пути или данных вообще нет в репозитории.
- Неуправляемая случайность в данных или алгоритмах (Random Seed)
- Исходные данные измененные вручную, а не с помощью скриптов
- Зависимость вывода и результатов от функций времени
- Различия отображения на графиках (некорректное использование matplotlib в том числе)
- Недоступны внешние данные
- Различия в выводе чисел с плавающей запятой
- Непостоянный порядок обхода словарей и др. контейнеров в python
- Различия в среде исполнения

Эксперименты - Воспроизводимость, рекомендации

- **Для каждого полученного результата сохраните алгоритм его получения.** - Важно знать каким образом вы получили те или иные результаты.
- **Избегайте этапов ручного управления данными или процессом.** - Может возникнуть соблазн открыть файлы данных в редакторе и вручную исправить пару ошибок форматирования или удалить выбросы. Кроме того, современные редакторы позволяют легко форматировать файлы огромных размеров. Однако соблазну сократить ваш алгоритм следует сопротивляться. Ручная обработка данных - это скрытая манипуляция.
- **Сохраните точные версии всех использованных внешних инструментов.** - В идеале вы должны настроить виртуальную машину или контейнер со всем программным обеспечением, используемым для запуска ваших скриптов. Это позволяет сделать снимок вашей аналитической экосистемы, что упрощает воспроизведение ваших результатов.
  По крайней мере, вам необходимо задокументировать версию всего используемого программного обеспечения, включая операционную систему. Незначительные изменения в программном обеспечении могут повлиять на результаты.
- **Используйте контроль версий.** - Для отслеживания версий ваших скриптов следует использовать систему контроля версий, такую как Git. Вы должны пометить (сделать снимок) текущее состояние скриптов и ссылаться на этот тег во всех получаемых вами результатах. Если вы затем решите изменить свои алгоритмы, что вы обязательно сделаете, можно будет вернуться во времени и получить точные сценарии, которые использовались для получения заданного результата.
- **Храните все промежуточные результаты в стандартизированном виде.** - В случае, когда необходима поэтапная проверка ваших результатов, куда удобнее, если все промежуточные данные были сохранены в доступном всем членам команды формате. Хранение этих промежуточных наборов данных (например, в формате CSV) предоставляет больше возможностей для дальнейшего анализа и может упростить определение проблемных мест поскольку нет необходимости все переделывать.

Эксперименты - Воспроизводимость, рекомендации

- **Для алгоритмов использующих случайность записывайте их случайное зерно.** - Одна вещь, которую специалисты по данным часто не делают - это фиксация исходные значения для своего анализа. Это делает невозможным точное воссоздание исследований машинного обучения. Многие алгоритмы машинного обучения включают стохастический элемент, и, хотя надежные результаты могут быть статистически воспроизводимыми, нет ничего, что можно было бы сравнить с теплым сиянием в глазах проверяющего при точном совпадении результатов.
- **Всегда храните вместе с графиками данные.** - Если вы используете скриптовый язык программирования, ваши графики скорее всего генерируются автоматически. Однако, если вы используете такой инструмент, как Excel, убедитесь, что вы сохранили начальные данные. Это позволяет не только воспроизвести график, но также более детально просмотреть лежащие в основе данные. Также стоит всегда сохранять алгоритмы, которые вы использовали для получения график на основе которых вы потом приводите какие-либо утверждения.
- **Иерархический подход при генерировании результатов анализа.** - Наша задача как специалистов по обработке данных - обобщить данные в той или иной форме. Вот что включает в себя извлечение информации из данных. Однако резюмирование также является простым способом неправильного использования данных, поэтому важно, чтобы заинтересованные стороны могли разбить сводку на отдельные точки данных. Для каждого итогового результата укажите ссылку на данные, использованные для расчета итогового значения.
- **Всегда указывайте вместе текстовые утверждения и результаты исследования.** - В конце работы результаты анализа данных оформляются в текстовом виде. Слова бывают неточны. Иногда бывает трудно определить связь между выводами и анализом. Поскольку отчет часто является самой важной частью исследования, важно, чтобы его можно было связать с результатами и, в соответствии с правилом № 1, с исходными данными.
- **Обеспечивайте доступность ваших результатов, данных и исследований.** - В коммерческих условиях может быть нецелесообразно предоставлять открытый доступ ко всем данным. Однако имеет смысл предоставить доступ другим пользователям в вашей организации. Облачные системы управления исходным кодом, такие как Bitbucket и GitHub, позволяют создавать частные репозитории, к которым могут получить доступ любые авторизованные коллеги.

Научный метод в DS - Рецензирование

Так как каждый элемент научного метода может и должен подвергаться рецензированию, на первое место выходит задача документирования.

- В вашей компании одновременно, даже силами одной команды могут проходить множественные исследования, по разным темам. Хорошей практикой является ведение общего реестра исследований, например, в виде базы знаний, в которой публикуются задокументированные элементы исследования.
- Также важно договориться о едином стиле оформления такой документации, ведь таким образом рецензирующий будет тратить меньше времени на стилистические проблемы и больше уделять внимание сути работы.
- Сейчас существует большое разнообразие инструментов позволяющих частично автоматизировать различные задачи, связанные с документированием вашего исследования. Снимая механическую нагрузку по сборке и публикации научной работы с аналитика, можно освободить время для более более полезной работы.

Так как DS с своей работе использует не только естественный язык, но и языки программирования, важной частью рецензирования становится ревью кода.

- Из области разработки программного обеспечения можно взять методы по автоматизации процесса ревью и тестирования кода, чтобы облегчить труд рецензента и позволить ему больше сосредоточиться на смысловой части работы.

Научный метод в DS - Итеративность

Научный метод — это итеративный процесс, в котором элементы постоянно сменяют друг друга.

Из любого элемента научного метода можно перейти в другой, так как мы постоянно получаем новую информацию и это может потребовать актуализации любого этапа.

Если появляется новая информация, например, обнаружены ошибки в работе модели, стоит начать с начала, то есть правильно зафиксировать наблюдение.

Применение полученных результатов и MLOps

Исследование не заканчивается после интеграции полученной модели машинного обучения в сервис.

- По результатам исследования мы получили группу гипотез, которые поддерживаются экспериментальными данными. Это наши теории, воспользовавшись которыми мы можем реализовать модель машинного обучения, предположив, что она решит поставленную задачу с некоторым уровнем уверенности.
  Однако, это всё ещё не 100% истина, вполне возможно, что вы просто пока не встречались с наблюдениями, которые противоречили бы вашей теории.
- Пока модель запущена и работает, мы должны продолжить собирать данные о её работе, ведь это новые данные наблюдений за нашим предметом исследования. Таким образом даже после релиза полученной модели можно проверять нашу теорию на практике, сверяя работу модели по новым данным, которые, возможно, в какой-то момент времени могут противоречить нашей теории.
- Данные, которые мы использовали для наблюдений и экспериментов, это лишь слепок реальных данных. Со временем реальные данные меняются.
  Необходимо после интеграции результатов в реальный продукт проверять и отслеживать данные, поступающие в модель. Таким образом, если произойдут существенные изменения, которые будут противоречить тому, что мы наблюдали на исходных данных, необходимо повторно провести исследовательскую работу для уточнения результатов.

Для практического применения полученных результатов могут потребоваться дополнительные работы.

- В своей работе DSы в первую очередь стараются проверить жизнеспособность определенных методов и алгоритмов в рамках предмета исследований.
- К ПО, которое интегрируется в production сектор, могут и скорее всего будут выдвинуты нефункциональные требования.
  Инженеры, которые берут результаты научных работ исследователей, решают множество важных задач для практической применимости полученных результатов.

Наконец-то, MLOps!

Тут, на стыке исследований на данных, инженерии моделей ML, поставке решений в продуктив появляется множество инструментов MLOps

Все инструменты, которые мы в дальнейшем рассмотрим в рамках курса позволят решать проблемы, в данных сферах:

- Работа в команде над DS / ML проектами (разные роли, разные задачи, разные процессы, передача артефактов)
- Воспроизводимость DS исследований (воспроизводимость методов по работам от других разработчиков)
- Вывод моделей в продуктив, автоматизация и поддержка (поставка решений в продуктив, сокращение трудозатрат, автоматизация задач, масштабирование и отказоустойчивость решений, …)