Предыдущий документ

ГЛАВА 2. КОНСТРУИРОВАНИЕ ПРОЦЕССА

2.1 СТРУКТУРА ПРОЕКТА

Данную работу удобно рассматривать как Data Science проект. Все ключевые этапы практически идентичны в обоих случаях. При таком подходе будет иметься чёткий план действий в детерминированном порядке.

В Data Science-проекте имеется три основных раздела, в каждом из которых 3 этапа. Рассмотрим основные разделы:

- Работа с требованиями: на этом этапе необходимо вникнуть в постановку задачи, понять, какой результат требуется получить от проекта, узнать про участников. В соответствии с определенной задачей нужно решить, какой метод использовать для решения задачи. Результатом этого шага будут требования к данным: что может понадобиться для успешного решения;

- Работа с данными: необходимо приступить к поиску данных для решения задачи: узнать, какие источники доступны, и сформировать выборку, с которой в дальнейшем будет осуществляться работа. После того как данные собраны, необходимо провести ряд исследований, чтобы лучше понимать, как устроена выборка:

  - Исследовать: центральное положение, вариабельность;
  - Выявить корреляции между признаками;
  - Построить графики распределения.

После этого этапа можно приступать к подготовке данных. Как правило, этот этап самый трудоемкий процесс. В зависимости от того, насколько качественно он выполнен, зависит успех всего проекта;

- Разработка и внедрение: после того, как данные готовы, можно приступать к разработке и внедрению. Программируем модель, прогоняем на обучающей выборке, проверяем на тестовой, если результат устраивает, то демонстрируем

заказчику, внедряем, собираем фидбэк, если нет, то дорабатываем до удовлетворительного результата.

Теперь рассмотрим подробнее продемонстрированные выше разделы по этапам:

Таблица 1. Описание основных этапов Data Science проекта

Название этапа	Описание этапа
Понимание задачи	Необходимо определить цель исследования: что является проблемой? Почему проблема должна быть решена? Кого затрагивает проблема? Главное: по каким метрикам будет оцениваться успешность проекта? Иными словами, необходимо выявить цель.
Аналитический подход	Нужно выбрать аналитический подход для решения бизнес-задачи. Выбор подхода зависит от того, какой тип ответа нужно получить в итоге: - если ответ должен быть вида да/нет, то нужен классификатор; - если ответ значение численного признака, то используются регрессионные модели; - промежуточный вариант перечисленных деревья решений; - если нужно определить вероятность, необходимо использовать предиктивную модель; - если необходимо выявить связи, используется дескриптивный подход.

Требования к данным	Когда определена цель исследования и выбран подход, необходимо определиться с тем, какие данные позволят дать искомый ответ. Нужно подготовить требования к данным: контент, форматы, источники.
Сбор данных	На этом этапе выполняется сбор данных из имеющихся источников: убеждаемся, что источники доступны, надежны и могут быть использованы для получения искомых данных в требуемом качестве. После необходимо понять, получили ли данные, какие хотели. На этой стадии можно пересмотреть требования к данным и принять решения о необходимости дополнительных данных. Могут быть выявлены лакуны в данных и составлен план, как их закрыть или найти замену.
Анализ данных	Анализ данных включает в себя все работы по конструированию выборки. На этом этапе необходимо получить ответ на вопрос: репрезентативны ли собранные данные относительно поставленной задачи? Здесь используется описательная статистика. Она применяется ко всем переменным, которые будут использоваться в выбранной модели: - исследуется центральное положение; - ищутся выбросы и выполняется оценка вариабельности; - строятся гистограммы распределения переменных; - визуализируются данные; - выполняется попарное сравнение: вычисляются корреляции между переменными, чтобы определить, какие из них связаны и насколько. Если найдутся значительные

	корреляции между переменными, какие то из них могут быть отброшены, как избыточные.
Подготовка данных	На этом этапе перерабатываем данные в такую форму, чтобы с ними было удобно работать: удаляем дубликаты, обрабатываем отсутствующие или неверные данные, проверяем и при необходимости исправляем ошибки форматирования. Также на этом этапе конструируем набор факторов, с которым на следующих этапах будет работать машинное обучение: проводим извлечение и отбор признаков, которые потенциально помогут решить бизнес-задачу. Ошибки на этом этапе могут оказаться критическими для всего проекта, поэтому к нему стоит отнестись особенно внимательно: избыточное количество признаков может привести к тому, что модель будет переобучена, а недостаточное — к тому, что модель будет недообучена.
Построение модели	Выбор модели, как можно было заметить, осуществляется в самом начале работы и зависит от бизнес-задачи. Таким образом, когда тип модели определен и имеется обучающая выборка, аналитик разрабатывает модель и проверяет, как она работает на созданном на этапе 6 наборе признаков.
Применение модели	Применение модели идет в тесной связке с собственно построением модели: вычисления чередуются с настройкой модели. На этом этапе мы должны ответить на вопрос, отвечает ли построенная модель бизнес-задаче. Вычисление модели имеет две фазы: проводятся диагностические измерения, которые помогают понять,

	работает ли модель, так как задумано. Если используется предиктивная модель, может использоваться дерево решений, чтобы понять, что выдача модели соответствует изначальному плану. На второй фазе проводится проверка статистической значимости гипотезы. Она необходима, чтобы убедиться, что данные в модели правильно используются и интерпретируются и полученный результат выходит за пределы статистической погрешности.
Внедрение	Если модель дает нам удовлетворительный ответ на поставленный вопрос, этот ответ должен начать приносить пользу. Когда модель разработана, и аналитик уверен в результате своей работы, необходимо познакомить заказчика с разработанным инструментом. Имеет смысл привлечь не только владельца продукта, но и других заинтересованных лиц: маркетинг, разработчиков, системные администраторы: всех, кто хоть как то может оказать влияние на дальнейшее использование результатов проекта. Далее необходимо переходить к внедрению. Внедрение может происходить поэтапно, например, на ограниченную группу пользователей или в тестовом окружении. Также необходимо наладить систему фидбэка, чтобы отслеживать, насколько успешно разработанная модель справляется с поставленной задачей. Через некоторое время этот фидбэк будет полезен для того, чтобы усовершенствовать модель.

Следующий документ