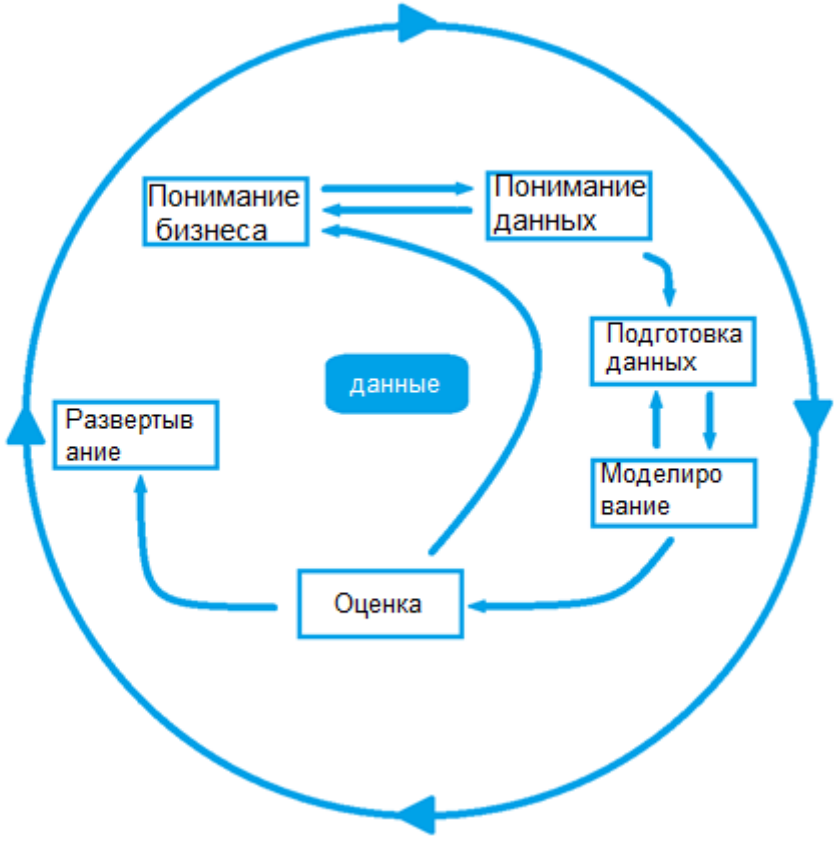
методология работы с данными

Методология CRISP-DM(Cross-Industry Standard Process for Data Mining)

Для работы над проектами с исследованиями данных чаще всего используется методология ведения проектов, которая называется crisp-dm. Основными этапами этой методологии являются:

- Понимание бизнеса
- Понимание данных
- Подготовка данных
- Моделирование
- Оценка
- Развертывание


Понимание бизнеса
Понимание данных
Подготовка данных
данные
Развертывание
Моделирование
Оценка

- Перемещение вперед и назад между фазами — обычное дело.
- В зависимости от результата фазы принимается решение, в какую фазу переходить дальше.
- Стрелками обозначены наиболее важные и частые переходы между фазами.
- Внешний круг символизирует циклическую природу анализа данных.
- Процесс анализа данных продолжается и после развертывания решения.

Знания, полученные во время процесса, могут породить новые более тонкие вопросы бизнеса. Последующий процесс анализа данных выгодно проводить, используя знания, полученные ранее.

Понимание бизнеса

Первая фаза процесса направлена на определение целей проекта и требований со стороны бизнеса. Затем эти знания конвертируются в постановку задачи интеллектуального анализа данных и предварительный план достижения целей проекта.

- Определить бизнес цели
- Оценить ситуацию
- Определить цели анализа данных
- Составить план проекта

Понимание данных

Вторая фаза начинается со сбора данных и ставит целью познакомиться с данными как можно ближе. Для этого необходимо выявить проблемы с качеством данных такие как ошибки или пропуски, понять что за данные имеются в наличии, попробовать отыскать интересные наборы данных или сформировать гипотезы о наличии скрытых закономерностей в данных.

- Собрать исходные данные
- Описать данные
- Исследовать данные
- Проверить качество данных

Подготовка данных

Фаза подготовки данных ставит целью получить итоговый набор данных, которые будут использоваться при моделировании, из исходных разнородных и разноформатных данных. Задачи подготовки данных могут выполняться много раз без какого-либо наперед заданного порядка. Они включают в себя отбор таблиц, записей и атрибутов, а также конвертацию и очистку данных для моделирования.

- Отобрать данные
- Очистить данные
- Сделать производные данные
- Объединить данные
- Привести данные в нужный формат

Моделирование

В этой фазе к данным применяются разнообразные методики моделирования, строятся модели и их параметры настраиваются на оптимальные значения. Обычно для решения любой задачи анализа данных существует несколько различных подходов. Некоторые подходы накладывают особые требования на представление данных. Таким образом часто бывает нужен возврат на шаг назад к фазе подготовки данных.

- Выбрать методику моделирования
- Сделать тесты для модели
- Построить модель
- Оценить модель

Оценка

На этом этапе проекта уже построена модель и получены количественные оценки её качества. Перед тем, как внедрять эту модель, необходимо убедиться, что мы достигли всех поставленных бизнес-целей. Основной целью этапа является поиск важных бизнес-задач, которым не было уделено должного внимания.

- Оценить результаты
- Сделать ревью процесса
- Определить следующие шаги

Развертывание

В зависимости от требований фаза развертывания может быть простой, например, составление финального отчета, или сложной, например, автоматизация процесса анализа данных для решения бизнес-задач. Обычно развертывание — это забота клиента. Однако, даже если аналитик не принимает участие в развертывании, важно дать понять клиенту, что ему нужно сделать для того, чтобы начать использовать полученные модели.

- Запланировать развертывание
- Запланировать поддержку и мониторинг развернутого решения
- Сделать финальный отчет
- Сделать ревью проекта

качество и количество данных

Считаем, что этап понимание бизнеса уже пройден. Для следующих 2х этапов нужно понимать какие данные хорошие какие нет, и сколько нам их нужно.

Количество данных должно быть такое чтобы оно отражало реальную картину мира. Тут можно вспомнить определение **репрезентативной выборки**, то есть данные на которых мы учимся должны представлять собой репрезентативную выборку, что означает, что можно обобщать результаты исследования наших данных на все данные.

Также количество данных должно быть **достаточным чтобы провести моделирование**. Некоторые алгоритмы, например нейронные сети, требуют большое количество данных для обучения нежели другие алгоритмы.

Вообще ситуацию можно описать так, чем больше тем лучше (в конце концов можем какую-то часть данных не рассматривать в тренировке а только тестировать ей наши алгоритмы)

- Как собирались данные (людьми, автоматически, смешанный вариант)
- Как хранились данные (могли ли потеряться какие-то связи)
- Как представлены данные (субд, форматы)
- Оценить или посчитать количество пропусков и их значимость, возможность их заполнить
- При классификации посчитать на сколько представлен каждый класс в выборке (оценить это же на генеральной совокупности)

Примеры

- Люди записывали ответы на вопросы других людей (естественно могли делать ошибки). Могла система автоматически записывать данные, например, биллинга – длительность сессий количество сессий.
- Например данные перекочевали из одной базы в другую и потерялась часть данных, либо структура хранения такая что в истории есть не все данные, например все которые старше 3 месяцев, хранятся только статистические отчеты, но не полные данные.
- Данные могут быть в одной базе или в нескольких (тогда нужен способ понять как увязывать данные из разных баз между собой), одна и та же характеристика может быть быть в разных форматах более или менее удобных (например адрес может быть просто строкой, а могут быть отдельные характеристики для города, улицы, номера дома и номера квартиры)
- Могут быть пропуски, как ошибки операторов или падение системы в какой-то момент (в который мы и получили пропуски). Так же не нужно путать пропуски в данных с незаполненными данными (например адрес где не поставили номер квартиры, потому, что ее не узнали по какой то причине и адрес где нет квартиры – частный дом). Так же может оказаться что в каких-то, например регионах просто нет возможности собирать некоторую информацию.
- Может быть так, что у нас есть данные для бинарной классификации и класс 1 представлен 10 000 объектов а класс 0 100 000. Но реально генеральная совокупность устроена так, что классы встречаются 50 на 50.

Что делать с пропусками?

- Убрать из рассмотрения либо характеристики с пропусками, либо объекты с пропусками
- Заменить на среднее (среднее арифметическое, медиану или моду)
- Заменить «особым» значением, которое дает понять что тут был пропуск
- Заменить случайным значением, рассчитанным из распределения этой характеристики
- Научиться вычислять эту характеристику по другим характеристикам