Предобработка данных

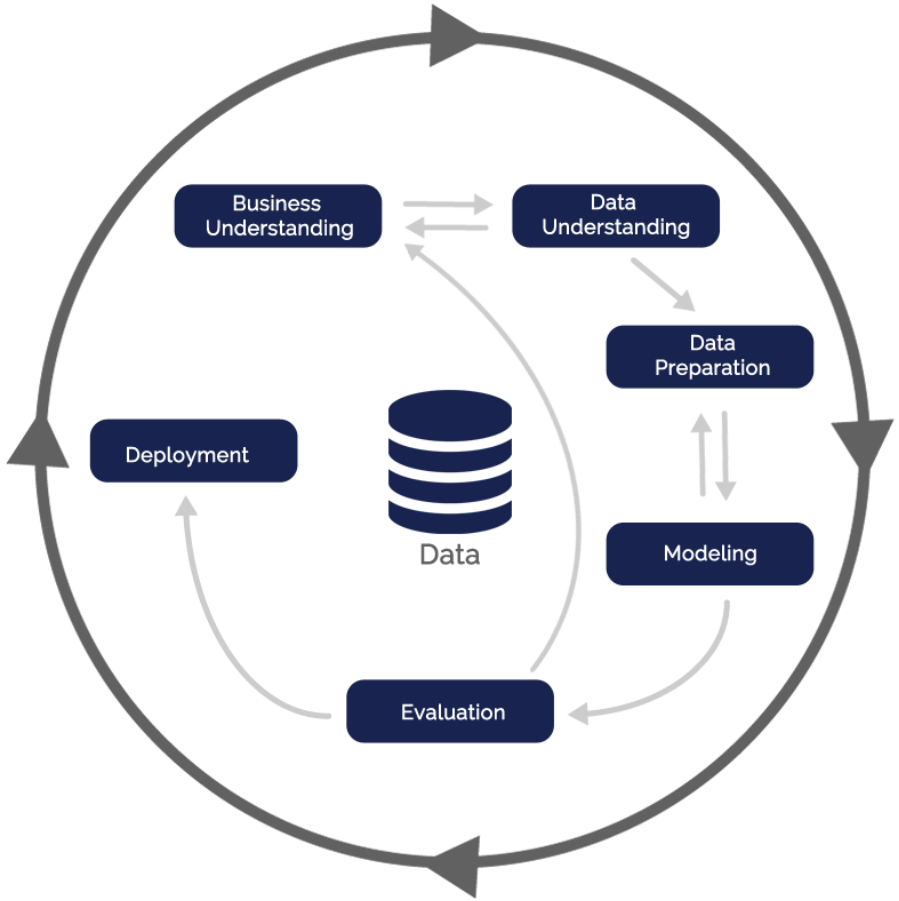
«Визуализация и моделирование»

Лекция 2

2021

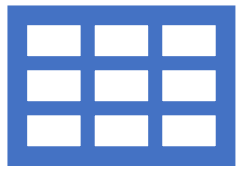
Этапы анализа данных по CRISP-DM

1. Понимание бизнеса / Business Understanding
2. Понимание данных / Data Understanding
3. **Подготовка данных / Data Preparation**
4. Моделирование / Modeling
5. Оценка / Evaluation
6. Внедрение / Deployment



Шаг 3 – Подготовка данных

Отбор данных

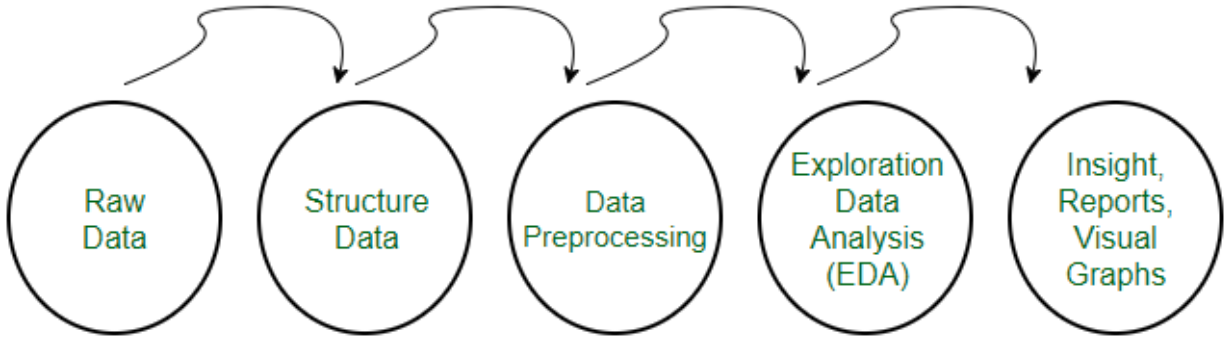


Предобработка данных

Обработка данных

Для чего нужна предобработка

1. Чтобы максимально оптимизировать параметры модели
2. Чтобы формат данных удовлетворял требованиям модели (например, random forest не поддерживает нулевые значения)
3. Чтобы отладить понимание данных и визуализацию



Проблемы в данных

1. НЕ точность

   Ошибки в данных неизбежны ввиду человеческого фактора.

2. НЕ полнота

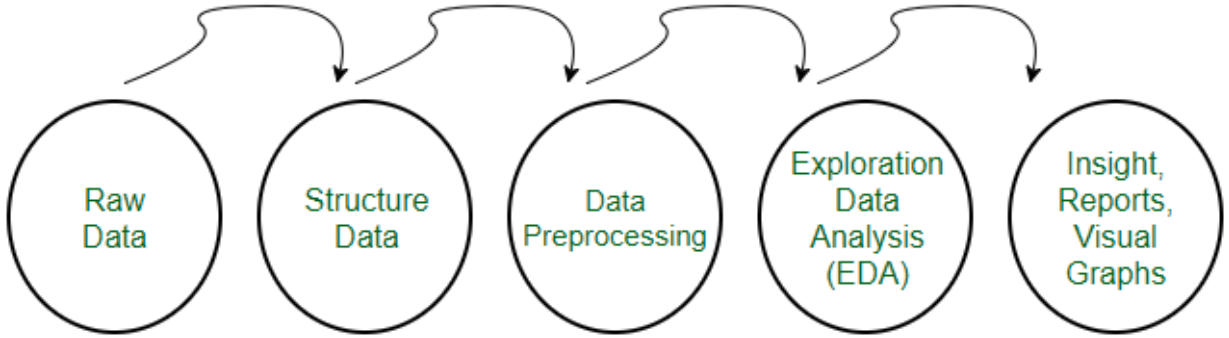
   Данные могут не существовать по какому-либо критерию.

   Данные могут быть отсеяны ввиду мнимой нерелевантности.

   Могут быть удалены противоречивые данные.

Виды предобработки данных

1. Обработка недостающих данных
2. Прореживание данных
3. Обработка выбросов
4. Обработка некорректных данных
5. Обработка дублей
6. Обработка категориальных данных



Обработка пустых ячеек

1. Удаление строк с пустыми ячейками
2. Заполнение недостающих данных вручную
3. Вести новую категорию для обозначения отсутствующего значения (unknown/other/-)
4. Обратиться к мерам центральной тенденции (по всему столбцу)
5. Заменить средним значением по категории (как 4, но по категории)
6. Заменить наиболее частым значением
7. Применить регрессию или дерево решений, чтобы предсказать наиболее вероятное значение

Обработка выбросов

Выбросы – значения непропорциональной величины для датасета. Например, значение возраста больше 100 (если датасет не о долгожителях).

Для обработки используется эвристика: *отсеиваются значения, находящиеся левее Q1-1.5 и правее Q3+1.5.*

Прореживание данных

1. Удаление столбцов, содержащих слишком много пустых ячеек (стандартный порог – 75%)
2. Удаление столбцов, значения которых имеют слишком низкую дисперсию
3. Определение столбцов с высокой корреляцией и удаление лишнего
4. Сократить количество признаков по методу главных компонент

Обработка некорректных данных

1. Перепутанные значения в столбцах
2. Значения с опечатками
3. Значения разных типов данных

Для детектирования таких данных в категориальных столбцах можно использовать, например, регулярные выражения.

Обработка категориальных данных

1. Масштабирование

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

2. Бинаризация
3. Стандартизация

$$x' = \frac{x - \bar{x}}{\sigma}$$

Славный пример

		Sex	Pregnant
Adult	1	Male	No
	2	Female	Yes
	3	**Male**	**Yes**
	4	Female	No
	5	**Male**	**Yes**

Данные

		Sex	Pregnant
Adult	1	Male	No
	2	Female	Yes
	4	Female	No

Очистка

		Sex	Pregnant
Adult	1	Male	No
	2	Female	Yes
	3	**Female**	**Yes**
	4	Female	No
	5	**Female**	**Yes**

Редактирование

		Sex	Pregnant
Adult	2	Female	Yes
	4	Female	No
	1	Male	No
	3	Male	Yes
	5	Male	Yes

Прореживание

Полезные ссылки

1. Data preprocessing in detail от IBM
2. Guide to Data Preprocessing for Data Science