Райффайзен Банк

Машинное обучение: регрессия, классификация, метрики

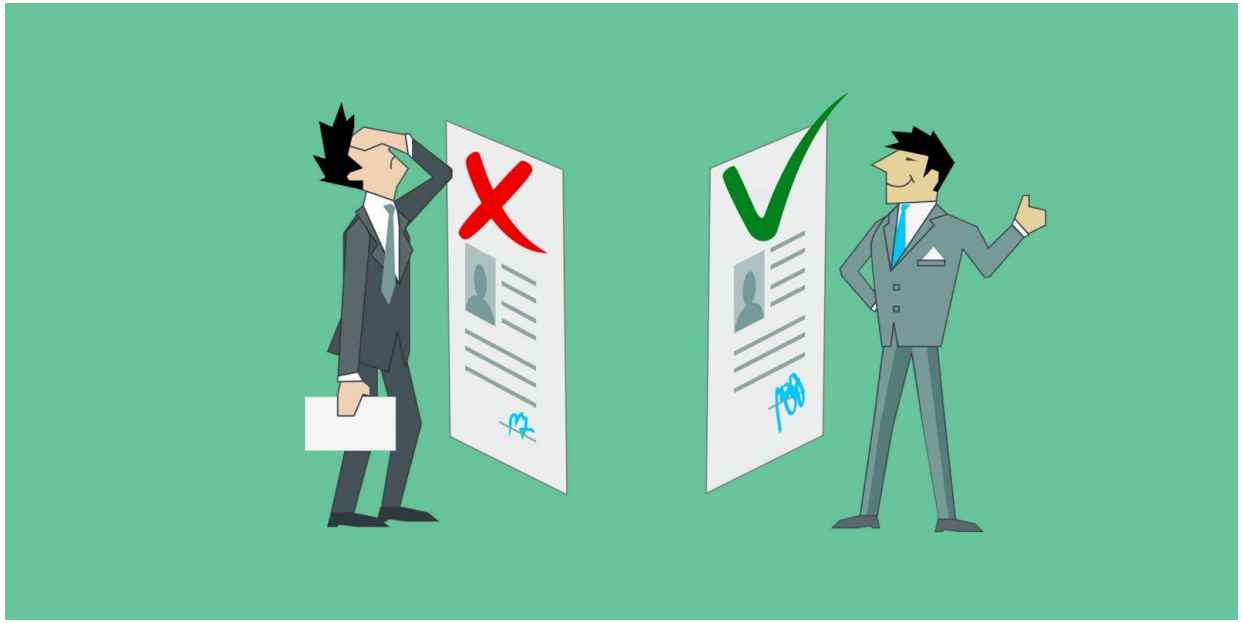
Елена Кантонистова

ВШЭ, 2023

HSE University
Faculty of Computer science

Пример: задача скоринга

- Пусть по характеристикам клиента (пол, возраст, средний доход, рейтинг кредитной истории и так далее) мы хотим предсказать, **вернёт клиент кредит или не вернёт**.



Пример: задача скоринга

- **Целевая переменная (target)**, то есть величина, которую хотим предсказать - это число (например, 1 - если человек вернет кредит, и 0 иначе).
- Характеристики клиента, а именно, его пол, возраст, доход и так далее, называются **признаками (features)**.
- Сами же клиенты - сущности, с которыми мы работаем в этой задаче - называются **объектами (objects)**.

Обучение алгоритма

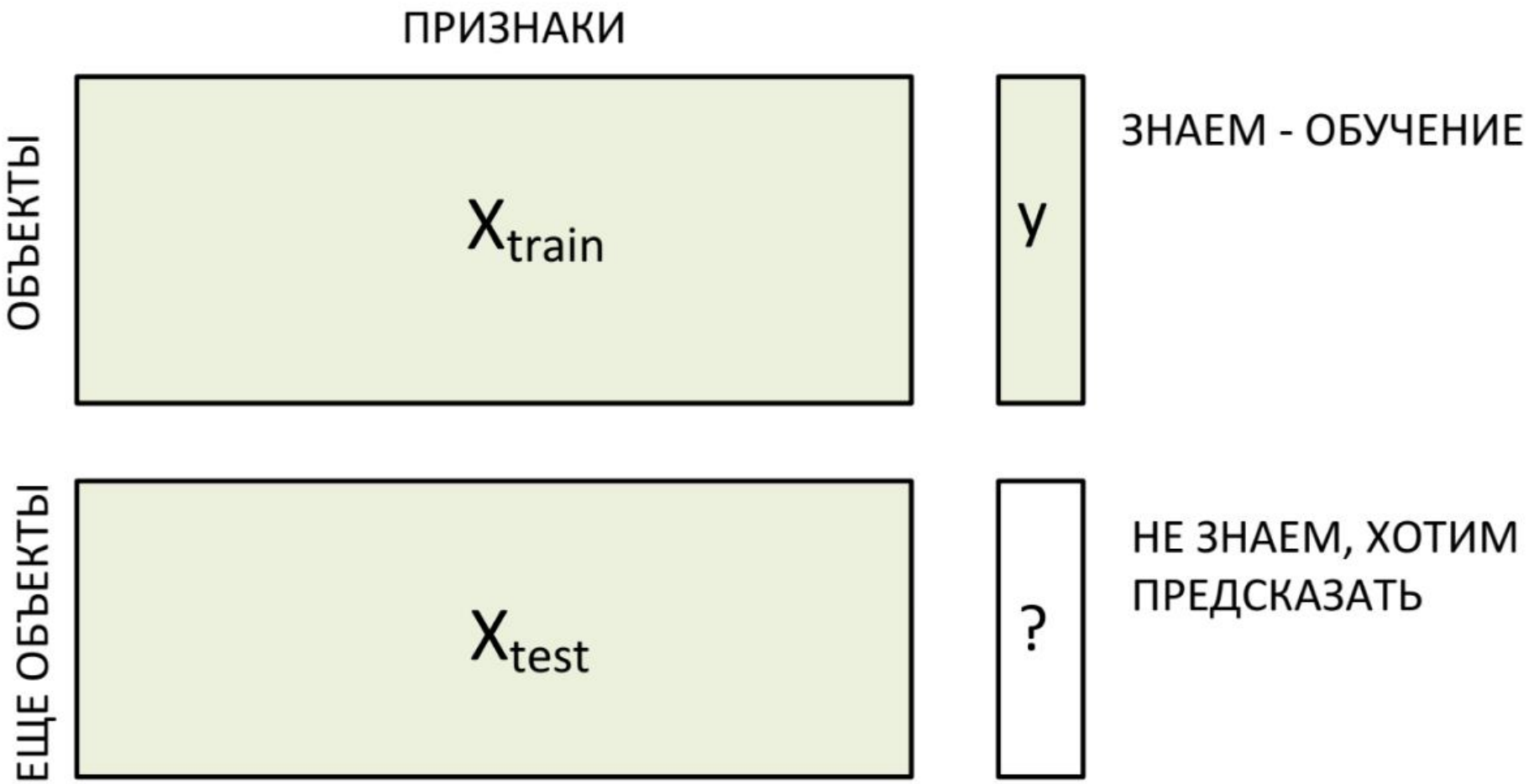
- На **этапе обучения** происходит анализ большого количества данных, для которых у нас имеются правильные ответы (например, клиенты, про которых мы знаем - вернули они кредит или нет; пациенты и их анализы, где про каждого пациента мы знаем, болен он или здоров и так далее).

- Модель машинного обучения изучает эти данные и старается научиться делать предсказания таким образом, чтобы для каждого объекта предсказывать как можно более точный ответ. Все данные с известными ответами называются **обучающей выборкой**.

Применение алгоритма

- На **этапе применения** готовая (уже обученная) модель применяется для того, чтобы получить ответ на новых данных. Например, у нас есть подробная информация о клиентах, и мы применяем модель, чтобы она предсказала, кто из них вернет кредит, а кто нет.

Этапы машинного обучения



Типы задач в ML

Что такое задача классификации?

Что такое задача регрессии?

Типы задач в ML: Классификация

- В задачах **классификации** целевая переменная - это класс объекта. То есть в задачах классификации ответ может быть одним из конечного числа классов.

Примеры:

- пол клиента (мужчина или женщина)
- уйдет клиент из компании или нет
- вернет человек кредит или нет
- болен пациент или здоров и т. д.

Примеры задач классификации

- Задачи медицинской диагностики (пациент здоров или болен)
- Задачи кредитного скоринга (выдаст банк кредит данному клиенту или нет)
- Задача предсказания оттока клиентов (уйдет клиент в следующем месяце или нет)
- Предсказание поведения пользователя (кликнет пользователь по данному баннеру или нет)
- Классификация изображений (на изображении кошка или собака)

Типы задач в ML: Регрессия

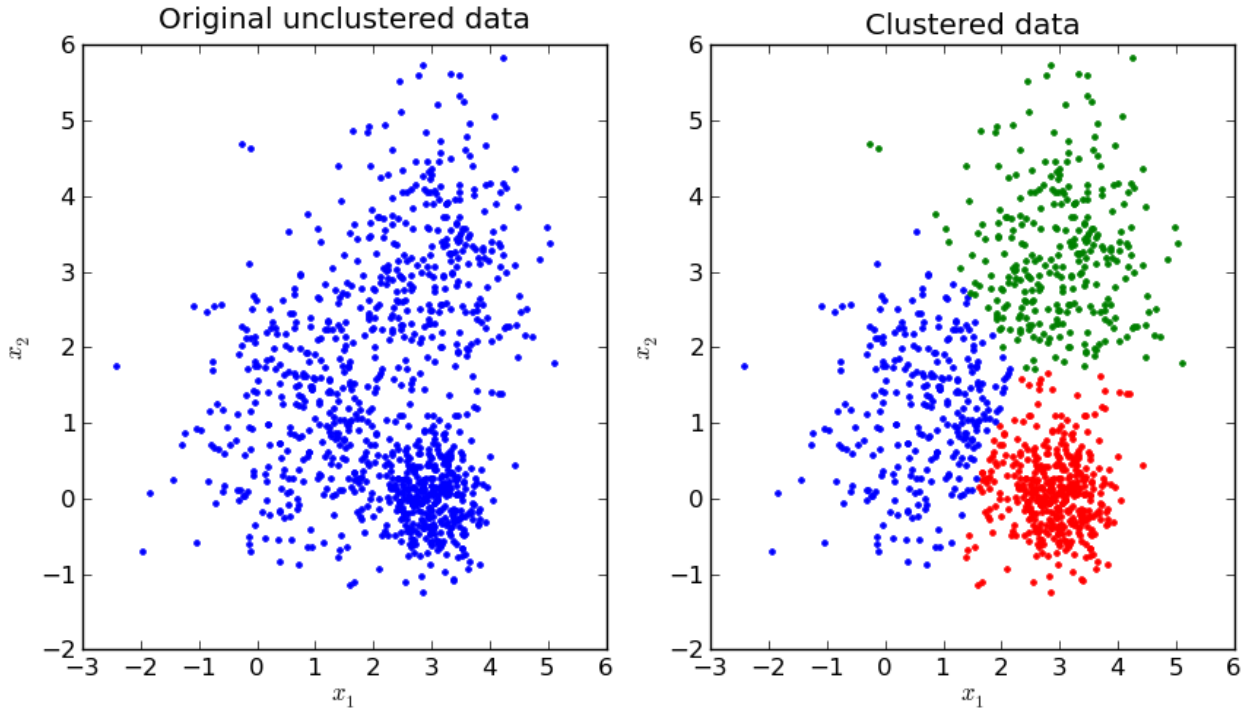
В задачах **регрессии** целевая переменная может принимать бесконечно много значений. Например, прибыль фирмы может быть любым числом (как очень большим, так и очень маленьким) - даже отрицательным или нецелым.

Примеры задач регрессии

- Предсказание стоимости недвижимости (стоимость квартиры в Москве)
- Предсказание прибыли ресторана
- Предсказание поведения временного ряда в будущем (стоимость акций)
- Предсказание зарплаты выпускника вуза по его оценкам

Типы задач в ML: кластеризация

Кластеризация – задача разделения объектов на группы, при этом где целевые переменные для объектов неизвестны (или не существуют). Разделение происходит только на основе признаковых описаний объектов.

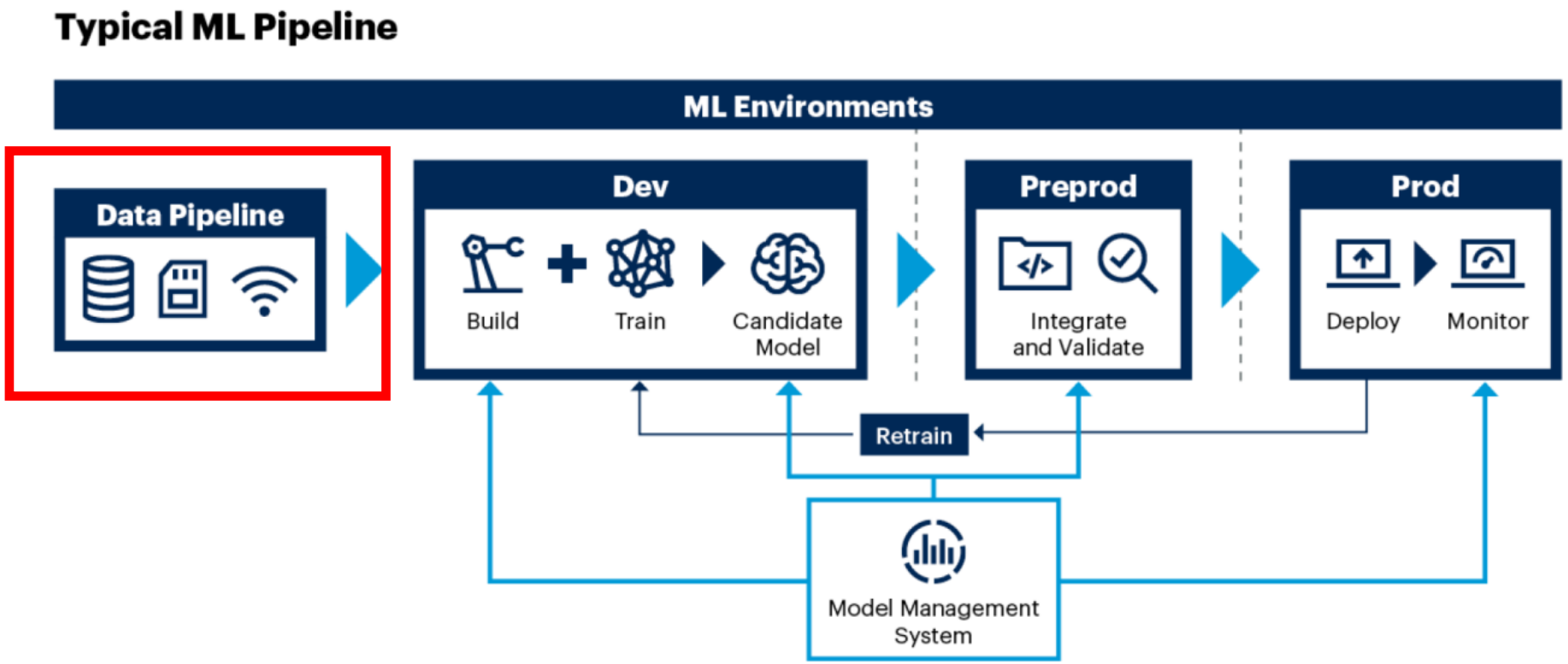


Типы задач машинного обучения

- Если нам известны значения целевой переменной, то есть алгоритм обучается так, чтобы правильно предсказывать целевую переменную – это **обучение с учителем**. Сюда относят классификацию, регрессию и ранжирование.
- Если нам неизвестны значения целевой переменной или целевая переменная вообще отсутствует, то есть алгоритм обучается только по признакам объектов, то это **обучение без учителя**. Примерами обучения с учителем являются кластеризация, понижение размерности и др.

Анализ данных и построение моделей

Анализ данных



Source: Gartner

718951_C

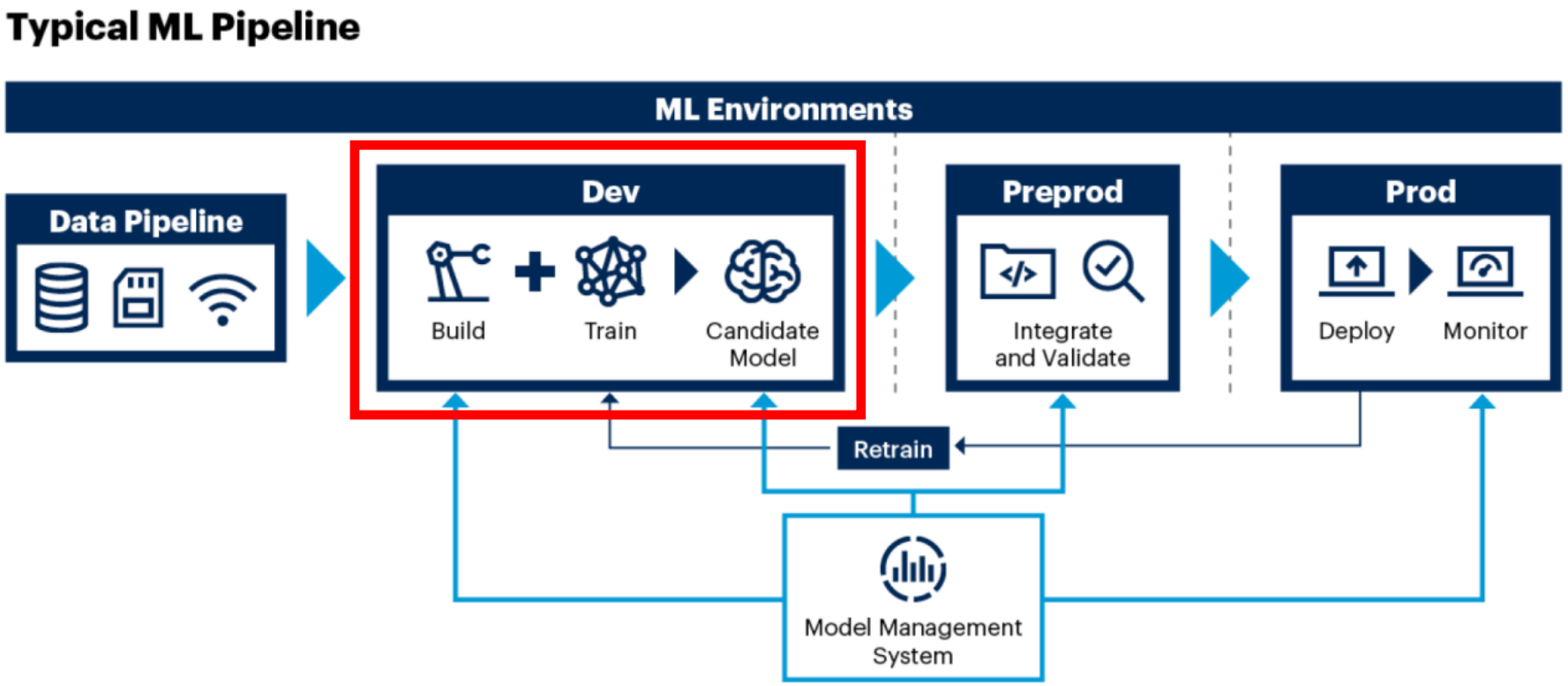
Анализ данных

1. *Сбор данных*: в каких источниках хранятся данные? Есть ли к ним доступы?
2. *Обработка данных*:

- Проверка качества данных
- Очистка данных
- Feature engineering
- Агрегация данных

3. *Загрузка данных в хранилище*

Обучение и валидация модели



Source: Gartner

718951_C

Обучение и валидация модели

1. *Выбор модели* (линейные модели, деревья, бустинги, нейронные сети)
2. *Обучение модели*
3. *Валидация модели* (оценка качества модели на тестовых данных)
4. *Подбор гиперпараметров модели*
5. *Выбор наилучшей модели*

Много практики!

https://colab.research.google.com/drive/11GCoTwnFmJ7Rb767Yxx995EcVc-eFr_n?usp=sharing