Основные понятия машинного обучения

Елена Кантонистова

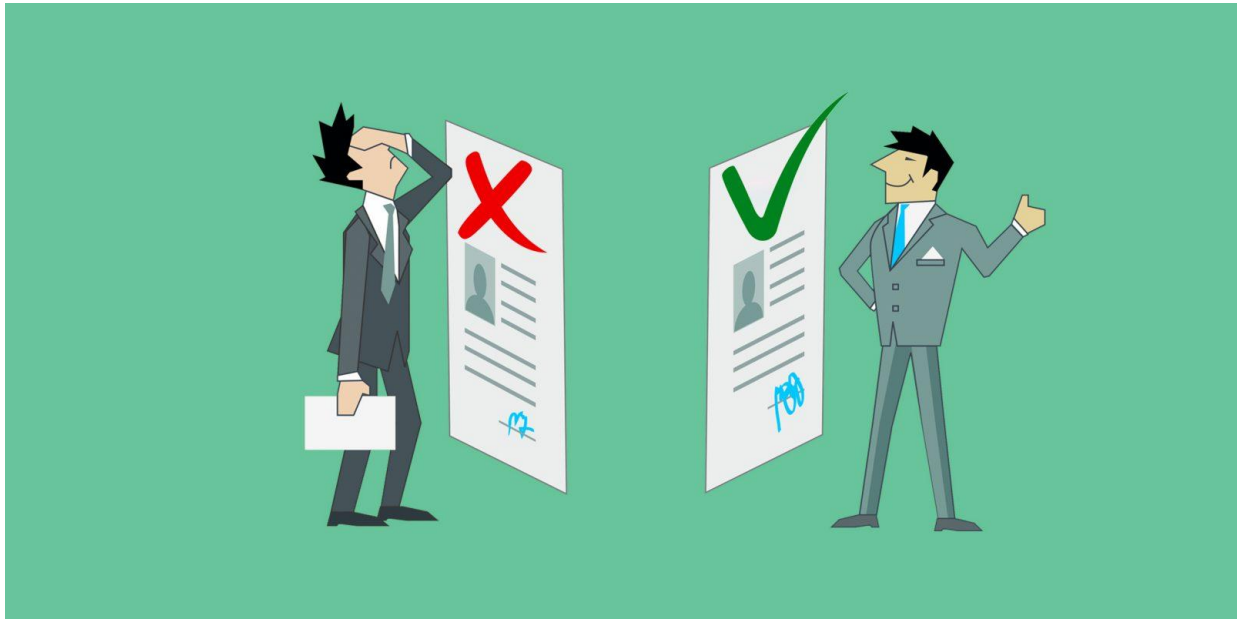
План лекции

- Основные понятия машинного обучения
- Типы задач
- Обучение модели
- Оценка качества модели
- Полный цикл проекта по анализу данных
- Введение в NLP
- Классические модели классификации и регрессии

1. Основные понятия машинного обучения

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

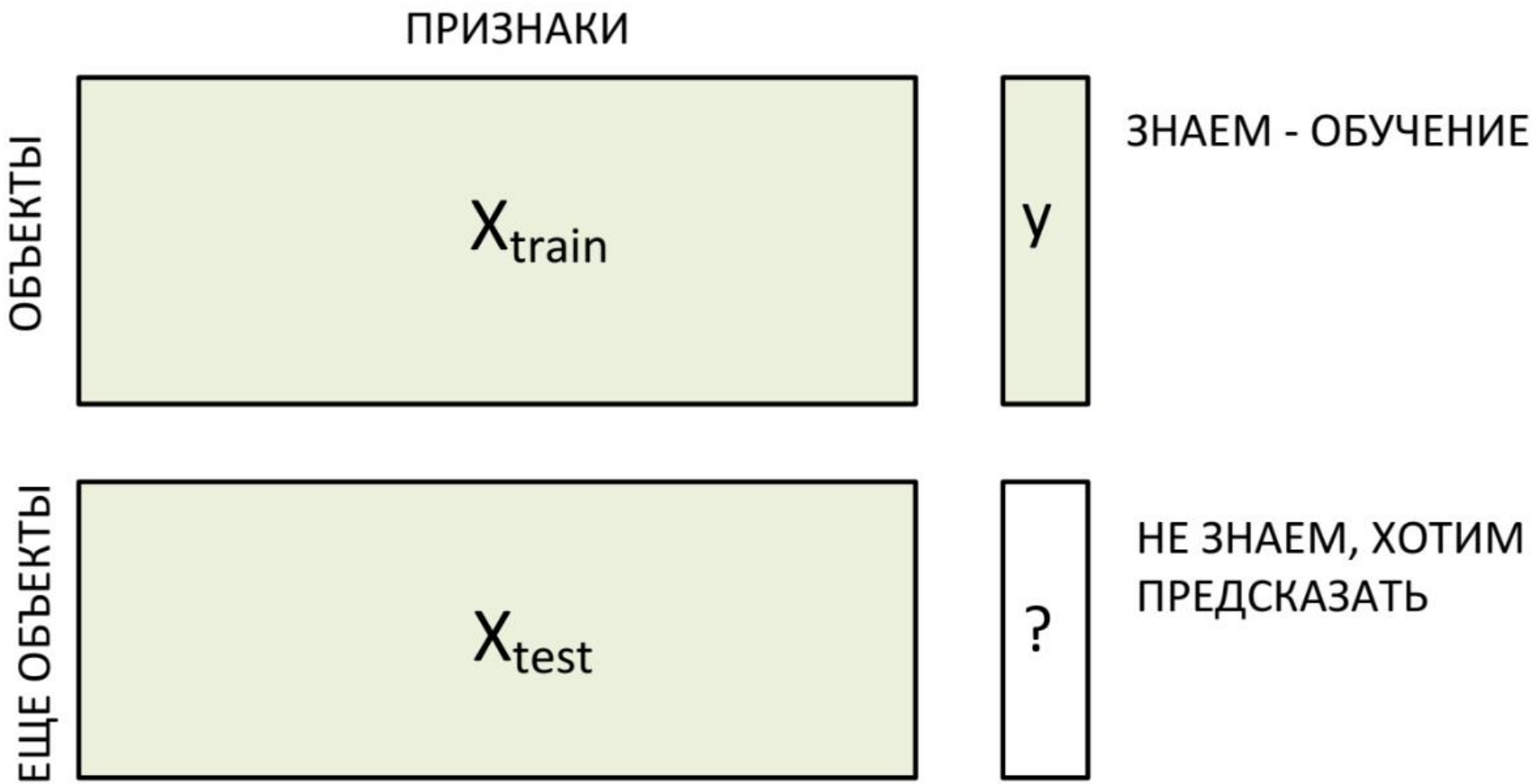
- На **этапе обучения** происходит анализ большого количества данных, для которых у нас имеются правильные ответы (например, клиенты, про которых мы знаем - вернули они кредит или нет; пациенты и их анализы, где про каждого пациента мы знаем, болен он или здоров и так далее).

- 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



2. Типы задач в ML

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

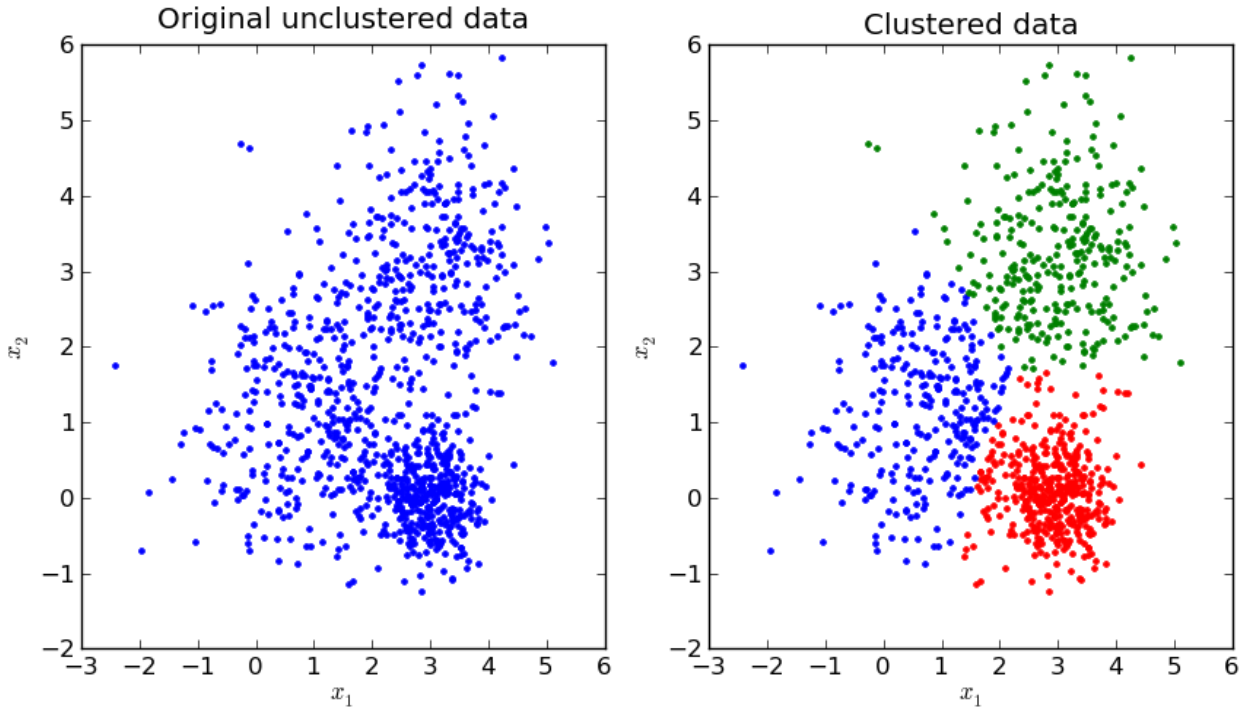
В задачах **регрессии** целевая переменная может принимать бесконечно много значений. Например, прибыль фирмы может быть любым числом (как очень большим, так и очень маленьким) - даже отрицательным или нецелым.

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

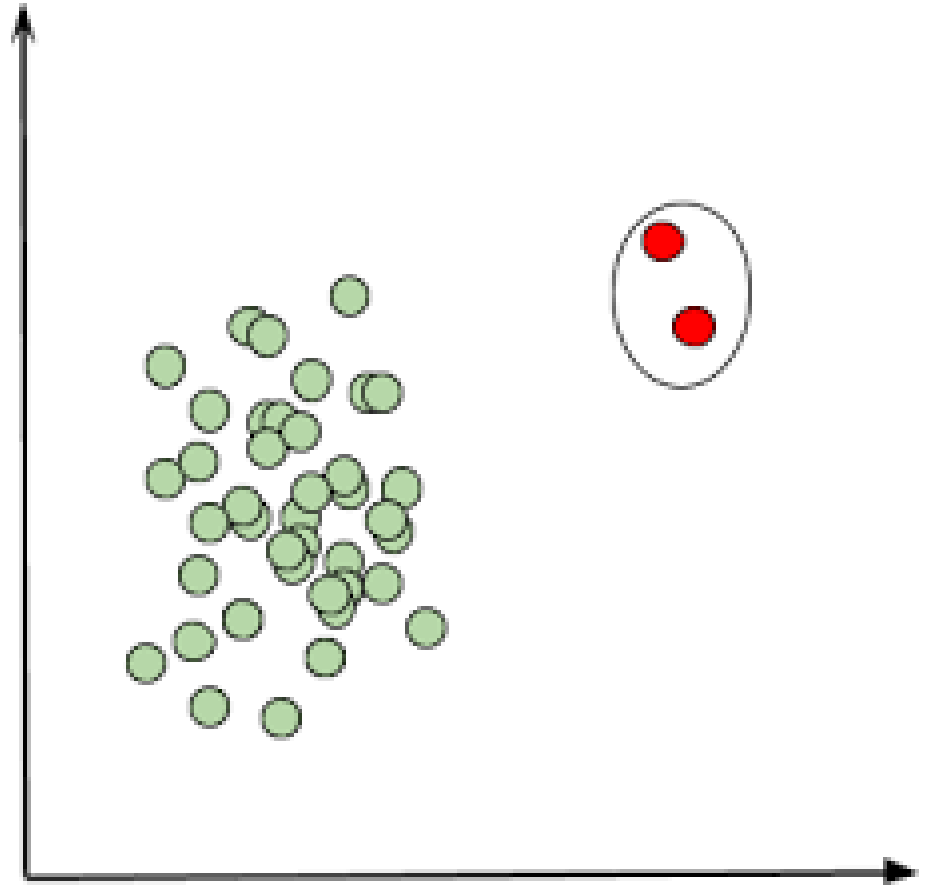
Кластеризация – задача разделения объектов на группы, при этом где целевые переменные для объектов неизвестны (или не существуют). Разделение происходит только на основе признаковых описаний объектов.



Другие типы задач в ML

- Ранжирование
- Рекомендации
- Снижение размерности
- Поиск аномалий
- Генерация
- Визуализация

И другие.



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.

3. Обучение модели

Обучение алгоритма

Предположим, что мы хотим предсказать *стоимость дома* y по его *площади* (x_1) и *количеству комнат* (x_2).

Обучение алгоритма

Предположим, что мы хотим предсказать *стоимость дома* y по его *площади* (x_1) и *количеству комнат* (x_2).

Как правило, алгоритм $a(x)$ выбирают из некоторого семейства алгоритмов A.

Обучение алгоритма

Предположим, что мы хотим предсказать *стоимость дома* y по его *площади* (x_1) и *количеству комнат* (x_2).

Как правило, алгоритм $a(x)$ выбирают из некоторого семейства алгоритмов A.

Используем линейную модель для предсказания стоимости. Она будет выглядеть так:
$$\boldsymbol{a(x) = w_0 + w_1 x_1 + w_2 x_2},$$

где w_0, w_1, w_2 -

параметры модели (*веса*).

Обучение алгоритма

Предположим, что мы хотим предсказать *стоимость дома* y по его *площади* (x_1) и *количеству комнат* (x_2).

Как правило, алгоритм $a(x)$ выбирают из некоторого семейства алгоритмов A.

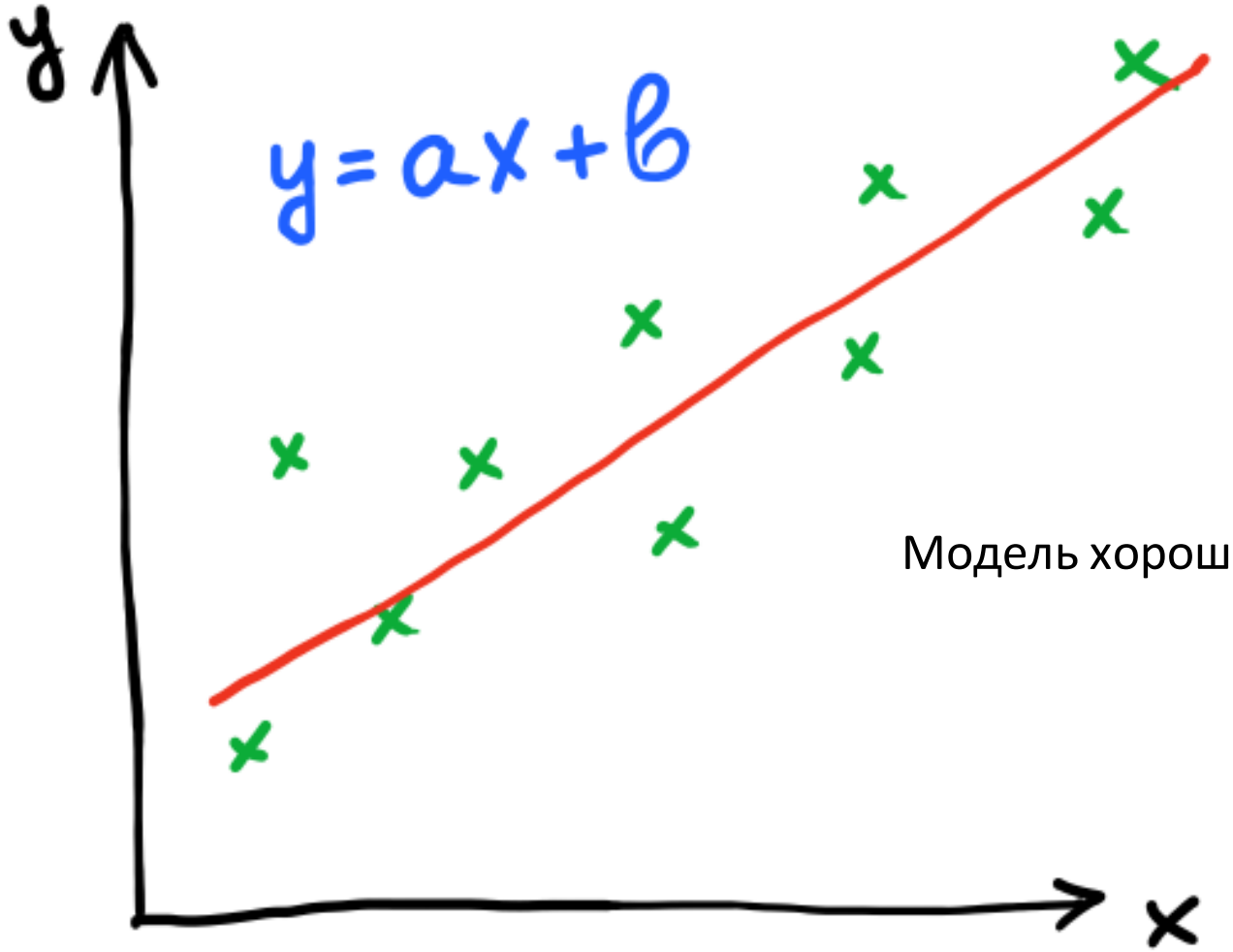
Используем линейную модель для предсказания стоимости. Она будет выглядеть так:
$$\boldsymbol{a(x) = w_0 + w_1 x_1 + w_2 x_2},$$

где w_0, w_1, w_2 -

параметры модели (*веса*).

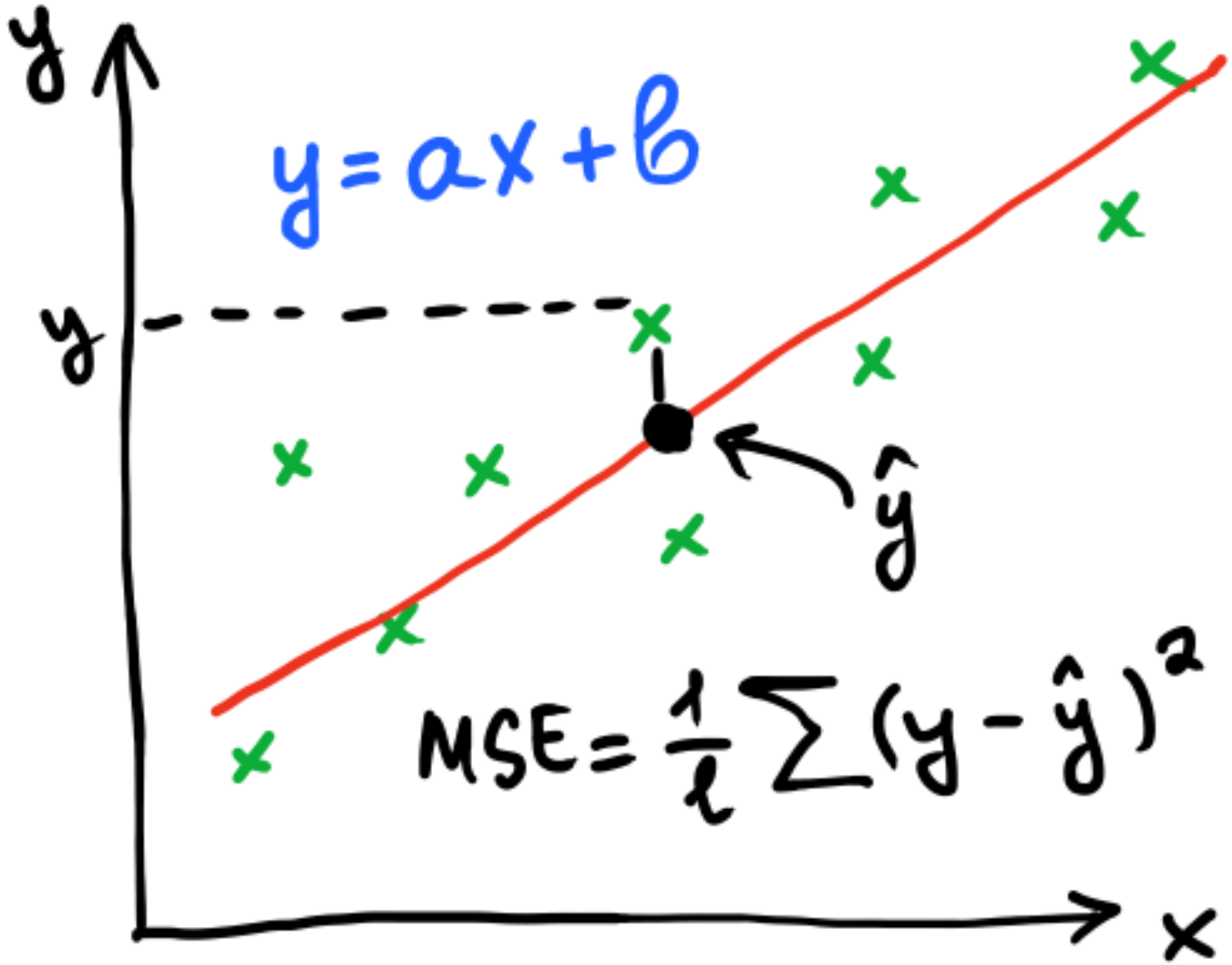
Общий вид линейных моделей:
$$\boldsymbol{A = \{a(x) = w_0 + w_1 x_1 + \cdots + w_d x_d | w_0, w_1, \ldots, w_d \in \mathbb{R}\}}$$

Обучение алгоритма



Модель хорошая?

Обучение алгоритма



Функционал ошибки

Как измерить ошибку алгоритма на всех объектах выборки?

Функционал ошибки – функционал, измеряющий качество работы алгоритма.

Пример (среднеквадратичная ошибка, MSE):

$$Q(a, X) = \frac{1}{l}\sum_{i=1}^{l}(a(x_i) - y_i)^2$$

X – объекты, l – количество объектов

a – алгоритм, $a(x_i)$ – ответ алгоритма на объекте x_i

y_i - истинные ответы

Функционал ошибки

Функционал ошибки – функционал, измеряющий качество работы алгоритма.

Пример (среднеквадратичная ошибка, MSE):

$$Q(a, X) = \frac{1}{l}\sum_{i=1}^{l}(a(x_i) - y_i)^2 \to \boldsymbol{min}$$

X – объекты, l – количество объектов

a – алгоритм, $a(x_i)$ – ответ алгоритма на объекте x_i

y_i - истинные ответы

При обучении алгоритма мы минимизируем функционал ошибки.

Обучение алгоритма

Пример (семейство линейных моделей):

$$A = \{a(x) = w_0 + w_1 x_1 + \cdots + w_d x_d | w_0, w_1, \ldots, w_d \in \mathbb{R}\}$$

Функционал ошибки:

$$Q(a, X) = \frac{1}{l} \sum_{i=1}^{l} (a(x_i) - y_i)^2$$

Функционал ошибки для линейной модели стоимости дома:

$$Q(a, X) = \frac{1}{l} \sum_{i=1}^{l} (w_0 + w_1 x_1 + w_2 x_2 - y_i)^2$$

Обучение алгоритма

Параметры w_0, w_1, w_2 подбираются так, чтобы на них достигался минимум функции потерь (на обучающей выборке):

Функционал ошибки для линейной модели стоимости дома:

$$Q(a, X) = \frac{1}{l} \sum_{i=1}^{l} (w_0 + w_1 x_1 + w_2 x_2 - y_i)^2 \to \min_{w_0, w_1, w_2}$$

Обучение алгоритма

Процесс поиска оптимального алгоритма (оптимального набора параметров или *весов*) называется **обучением**.

4. Оценка качества модели

4. Оценка качества модели

Чем отличается функция потерь от метрики качества?

Оценка качества модели

- В задачах машинного обучения для оценки качества моделей и сравнения различных алгоритмов используются ***метрики качества***
- Метрики отличаются от функций потерь своей целью – они должны быть понятными и интерпретируемыми, отражать интересующие нас показатели качества модели

Метрики качества

В задачах машинного обучения для оценки качества моделей и сравнения различных алгоритмов используются ***метрики качества***.

Примеры:

- Корень из среднеквадратичной ошибки – для регрессии

$$RMSE(a, X) = \sqrt{\frac{1}{l}\sum_{i=1}^{l}(a(x_i) - y_i)^2}$$

Метрики качества

В задачах машинного обучения для оценки качества моделей и сравнения различных алгоритмов используются ***метрики качества***.

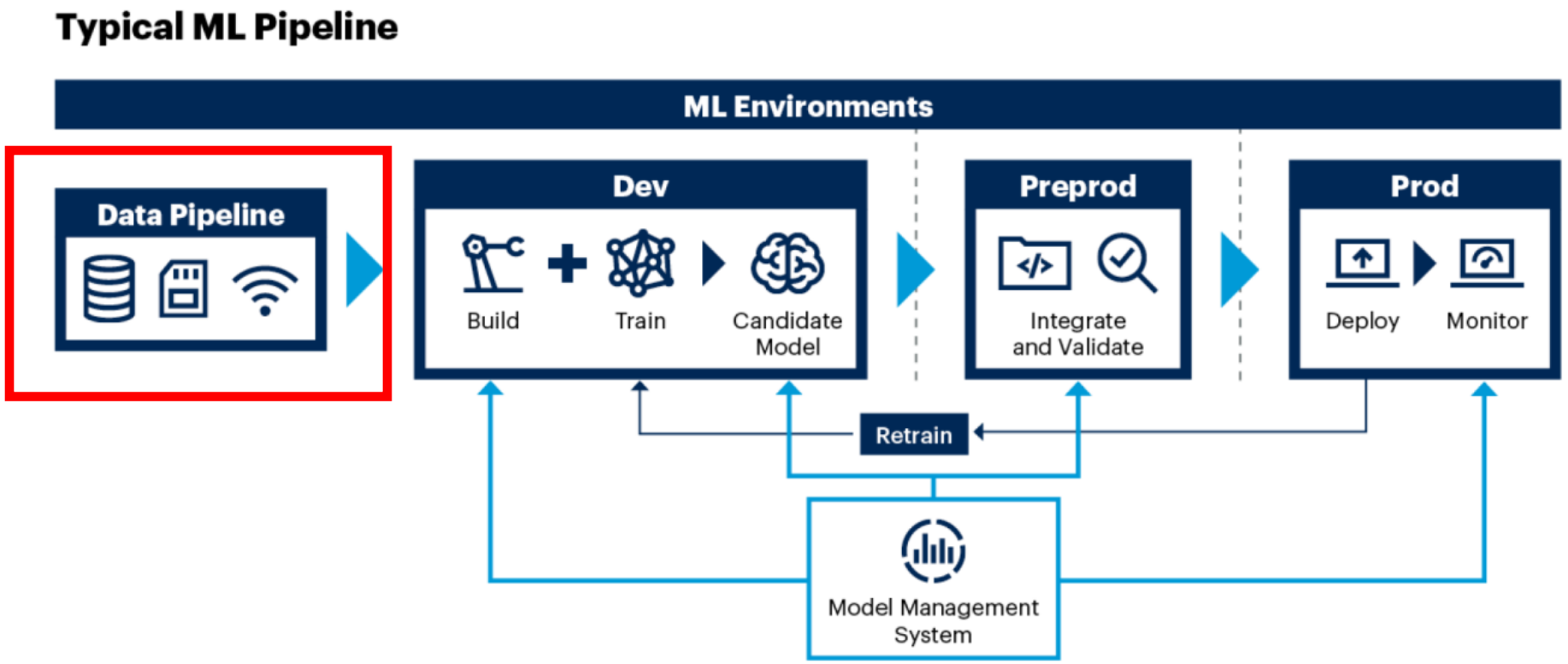
Примеры:

- Корень из среднеквадратичной ошибки – для регрессии
- Доля правильных ответов – для классификации

$$accuracy(a, X) = \frac{1}{l}\sum_{i=1}^{l}[a(x_i) = y_i]$$

5. Цикл проекта по машинному обучению

Анализ данных



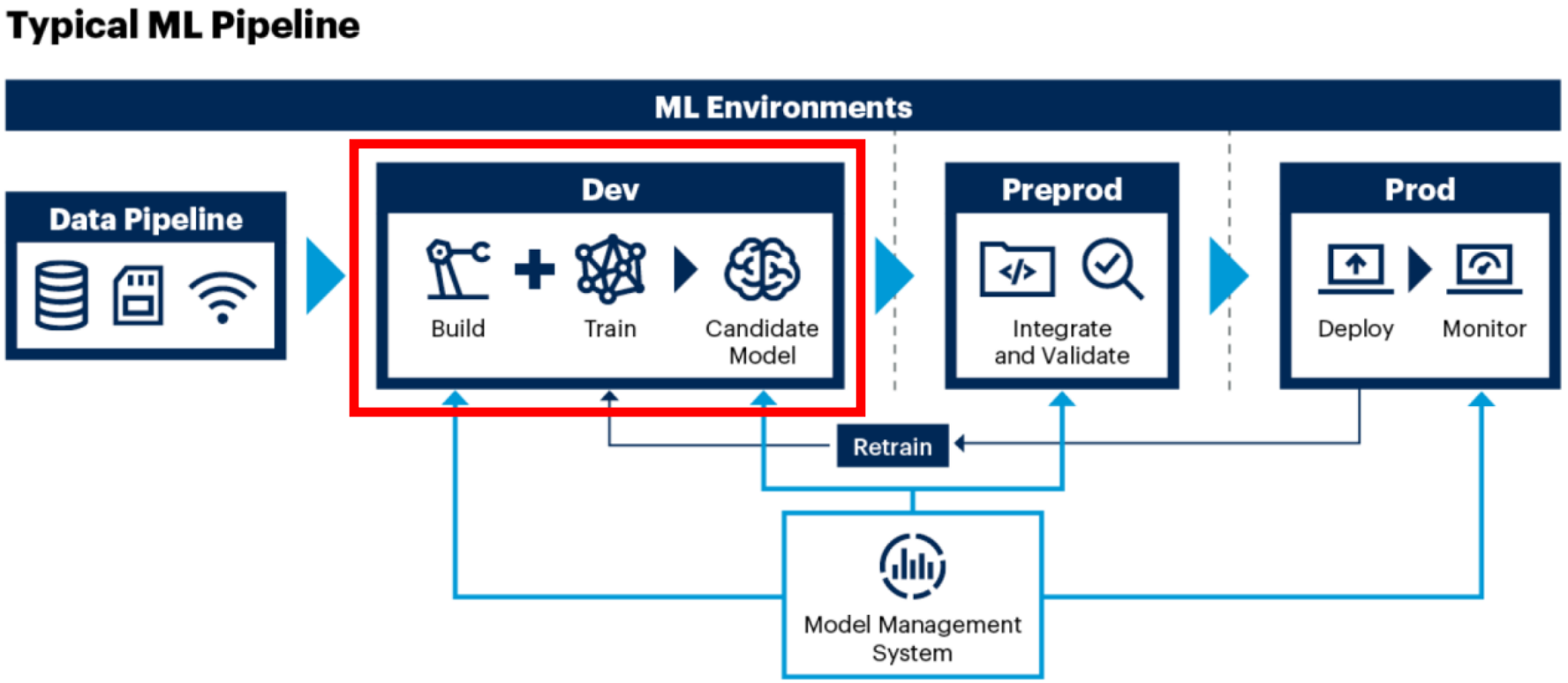
Анализ данных

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*
4. *Автоматизация процесса сбора, обработки и загрузки данных*

Обучение и валидация модели



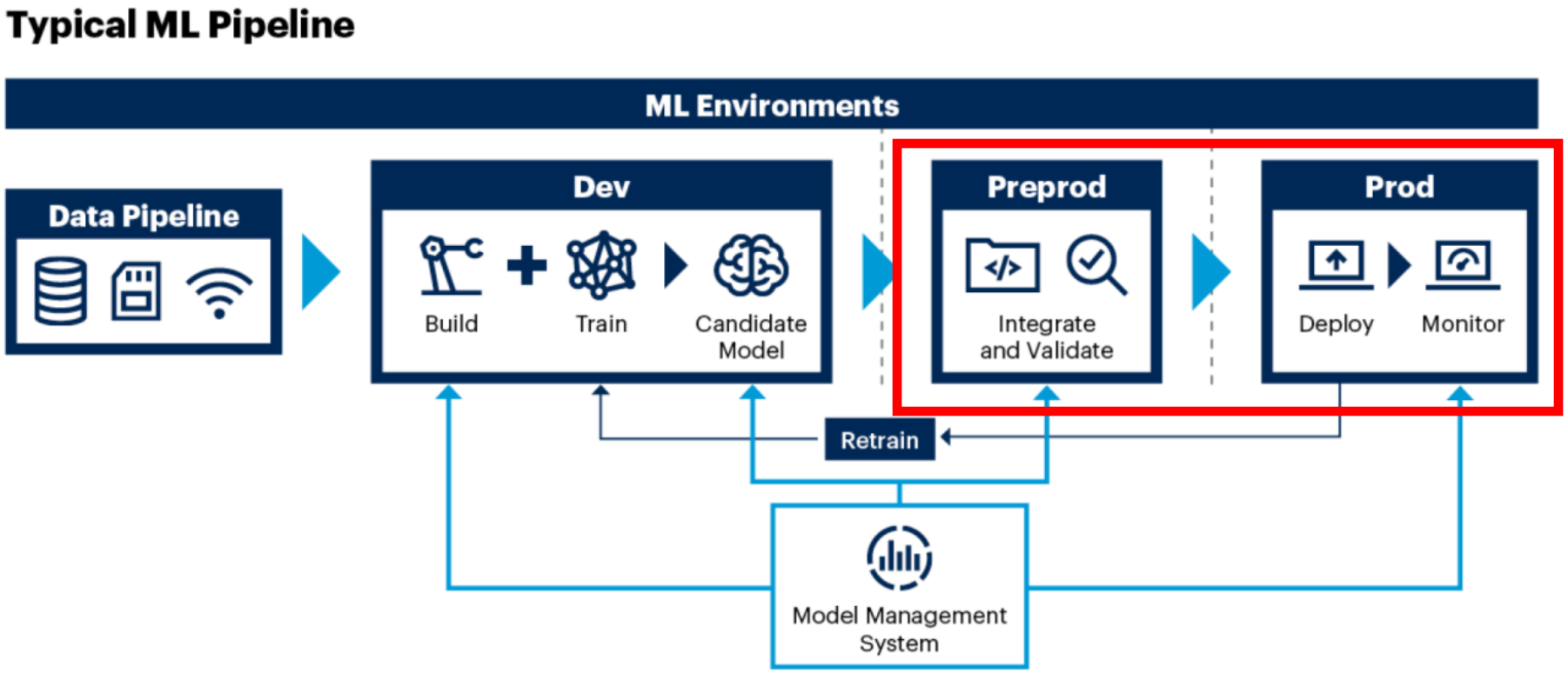
Source: Gartner

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*

Внедрение модели в production



Source: Gartner

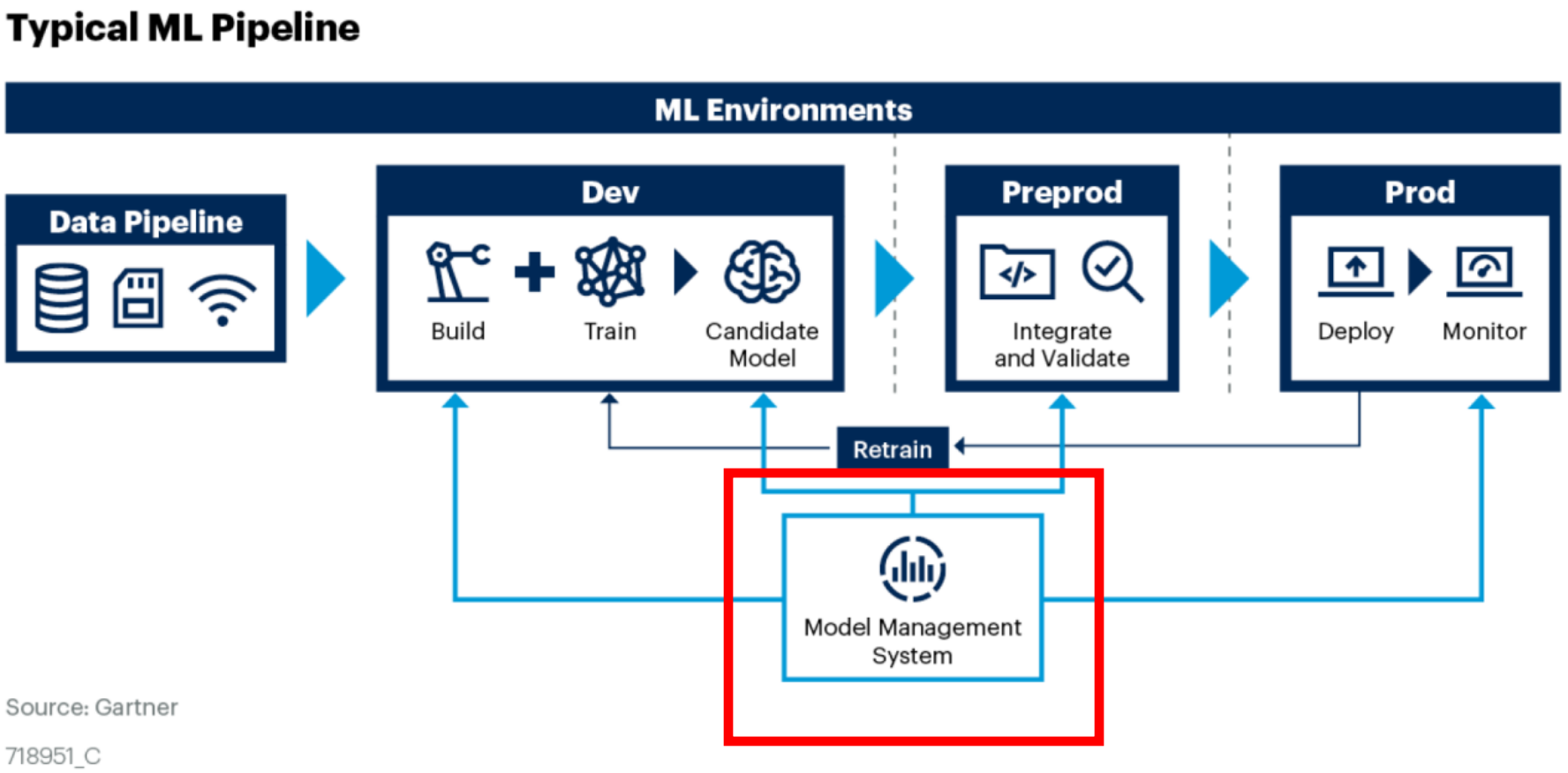
718951_C

Внедрение модели в production

Варианты внедрения:

- *Сервис (Streamlit, FastApi и другие)*
- *Telegram-бот*
- *Внедрение модели как компонента большого бизнес-процесса*

Оркестрация пайплайна и мониторинг



Квиз

Квиз: вопрос 1

Пусть мы решаем задачу определения вида животного на фотографии.

Что в этой задаче является ***целевой переменной?***

a) Одна фотография
b) Вид животного (кошка, тигр, собака...)
c) Наличие ушей на фотографии, количество лап, цвет шерсти
d) Невозможно определить

Квиз: вопрос 2

К какому типу относится задача определения тональности отзыва на фильм: положительный или отрицательный отзыв?

a) Классификация
b) Регрессия
c) Кластеризация
d) Невозможно определить

Квиз: вопрос 3

Пусть мы решаем задачу регрессии при помощи линейной регрессии, и формула для предсказания ответа имеет вид:

$$a(x) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3$$

Сколько весов в данной модели?

a) 3

b) 4

c) 7

d) Мало данных

Квиз: вопрос 4

Вы вычислили некоторую метрику, и результат оказался 5400 кг^2. Что это могла быть за метрика?
Выберите один ответ.

a) MSE

b) RMSE (корень из MSE)

c) Accuracy

d) Ни один вариант не подходит