Data science

Лекция 4. Применение Data science

2025/2026 учебный год

Доцент кафедры МО&МО, Махно В.В.

ИНСТИТУТ
МАТЕМАТИКИ
МЕХАНИКИ
КОМПЬЮТЕРНЫХ
НАУК
имени И.И. Воровича

Что включает в себя Data Science:

Data Science — это область, которая помогает получать ценные знания из данных с помощью программирования, статистики и машинного обучения.

Или проще: **мы берём сырые данные → анализируем → делаем выводы → применяем в реальных задачах.**

- **Сбор данных** — из таблиц, сайтов, сенсоров, API
- **Очистка и подготовка данных** — чтобы данные стали пригодными для анализа
- **Анализ данных (EDA)** — визуализации, группировки, поиск закономерностей
- **Моделирование** — обучение моделей для предсказаний
- **Интерпретация** — объяснение результата людям или бизнесу

Где используется:

- Рекомендательные системы (Netflix, Spotify)
- Финансовые модели (одобрение кредита, прогноз акций)
- Маркетинг (анализ поведения клиентов)
- Здравоохранение (диагностика по снимкам)
- Спорт (оценка игроков, стратегия)

Кто такой Data Scientist?

Это человек, который может:

- понять бизнес-проблему,
- собрать и проанализировать данные,
- построить модель,
- и **предложить решение**, основанное на данных.

Итак:

Data Science — это мост между данными и реальными решениями.

Это профессия, где сочетаются математика, код, логика и креатив.

Где применяется Data Science: примеры из жизни и реальных проектов

YouTube, TikTok, Spotify

- Алгоритмы рекомендуют вам видео и музыку на основе:
- Истории просмотров
- Поведения пользователей, похожих на вас
- Времени суток, дня недели и других факторов

→ Это **рекомендательные системы**, и они основаны на анализе больших данных и машинном обучении.

Онлайн-магазины: Wildberries, Ozon, Amazon

Data Science помогает:

- Показывать товары, которые вас заинтересуют
- Прогнозировать спрос
- Оптимизировать логистику

→ Пример: если вы ищете «кроссовки», система предложит похожие товары, основываясь на поведении других покупателей.

Банки и финансы

Используют модели, чтобы:

- Выявлять мошенничество
- Предсказывать вероятность возврата кредита
- Предлагать персональные продукты

→ **Кредитный скоринг** — классическая задача классификации.

Где применяется Data Science: примеры из жизни и реальных проектов

Здравоохранение

Data Science применяют для:

- Распознавания болезней на снимках (рентген, МРТ)
- Прогнозирования рисков заболеваний
- Персонализированных рекомендаций по лечению

Спорт и аналитика

Клубы анализируют:

- Игру каждого спортсмена
- Оптимальную тактику
- Кто из игроков «перспективен» для покупки

Промышленность, транспорт, логистика

- Предсказание поломок оборудования
- Оптимизация маршрутов доставки
- Управление складом и запасами

Data Science используется везде, где есть данные. А данные — повсюду: в интернете, в банках, в медицине, в бизнесе, в телефоне каждого из нас.

Обзор профессий в Data Science

Аналитик данных (Data Analyst)

Что делает:

- Извлекает данные из баз
- Делает отчёты, дашборды, визуализации
- Отвечает на вопросы бизнеса: «почему продажи упали?», «кто наш клиент?»

Навыки:

- SQL, Excel, Python (Pandas, Matplotlib)
- Tableau / Power BI
- Бизнес-мышление

Цель: объяснить прошлое и настоящее.

Обзор профессий в Data Science

Data Scientist

Что делает:

- Разрабатывает модели прогнозирования
- Решает более сложные задачи: рекомендации, классификация, прогноз
- Работает с данными глубже, часто строит end-to-end решения

Навыки:

- Python (Pandas, Scikit-learn, ML-библиотеки)
- Статистика, математика
- Построение и оценка моделей
- Иногда — немного SQL, визуализации

Цель: предсказать будущее и автоматизировать принятие решений.

Обзор профессий в Data Science

ML-инженер (Machine Learning Engineer)

Что делает:

- Берёт модель от Data Scientist и делает так, чтобы она **работала в проде**
- Занимается инфраструктурой, масштабируемостью, скоростью, API

Навыки:

- Python (или другой язык), Git, Docker, CI/CD
- MLflow, FastAPI
- Глубокое понимание пайплайна моделей

Цель: запустить модель в реальном приложении — стабильно и быстро.

Обзор профессий в Data Science

Дополнительно

- **Research Scientist** — занимается новыми алгоритмами, работает в больших лабораториях
- **Data Engineer** — строит системы для сбора, хранения и обработки данных

Как выбрать

Графики и отчёты	Data Analyst
Математику и модели	Data Scientist
Кодить и автоматизировать	ML Engineer

Данные — это топливо 21 века. Без них не работают рекомендательные системы, не строится логистика, не принимаются бизнес-решения.

- Компании ежедневно **генерируют терабайты данных** — о клиентах, продажах, процессах.
- «Сырые данные» — бесполезны. Нужны специалисты, кто может **выжать из них пользу**.
- Поэтому **спрос на Data Science растёт во всём мире** — от стартапов до крупнейших корпораций.

📊 Прогноз McKinsey: к 2030 году нехватка специалистов по данным в мире превысит **250 000 человек**.

📌 Вывод: умение работать с данными = навык будущего, как раньше умение пользоваться компьютером.

Посмотреть на hh.ru «Data Scientist» и «Data Analyst»

Сравнение с другими IT-направлениями: в чём сила DS?

Направление	Что делает	Где применяется	Что особенного
Frontend	Делает сайты красивыми	Веб-разработка	Много визуального, важно внимание к UX
Backend	Обрабатывает логику сайта/сервиса	Сайты, API, базы данных	Много архитектуры, интеграций
QA	Тестирует программы	Везде, где пишется код	Важно внимание к деталям
Data Science	Анализирует данные и строит модели	Финтех, маркетинг, медиа, логистика и др.	Высокий уровень абстракции, автоматизация
ML Engineer	Запускает модели в реальную жизнь	ИИ-продукты, рекомендации, scoring	Сильная инженерия, продакшн, DevOps-навыки

Машинное обучение.

Что делает Data Scientist с этими данными?

- **Изучает**
- Что вообще есть? Как выглядят признаки? Есть ли пропуски или странности?
- 🧹 **Очищает**
- Удаляет мусор, заполняет пробелы. Пример: если у кого-то рост = 500 см, это ошибка.
- 🔧 **Преобразует**
- Переводит текст в числа, нормализует шкалы, разбивает дату на день/месяц/год.
- 🧊 **Визуализирует данные**
- Изучение данных перед обучением
- 🧠 **Обучает модель**
- Например, предсказывает цену квартиры по её параметрам.
- 📈 **Интерпретирует результат**
- Строит отчёты, графики, дашборды — чтобы бизнес понял, что делать дальше.

Машинное обучение.

Что делает Data Scientist с этими данными?

Примеры

Сфера	Какие данные	Что можно узнать
Банки	Доход, возраст, кредиты	Вернёт ли клиент займ
E-commerce	Поведение, клики, покупки	Что предложить клиенту
Медицина	Симптомы, анализы	Вероятность диагноза
Образование	Оценки, посещаемость	Предсказать успеваемость

Типы данных

Тип	Пример	Как использовать
Числовые	Возраст, цена	Можно сравнивать, усреднять
Категориальные	Пол, цвет, город	Переводим в числа
Булевы	Да/Нет, True/False	Часто используются в фильтрах
Дата/время	2025-07-07	Разбиваем на год/месяц/день

Как работать с данными в Data Science

Изучить данные

```
import pandas as pd df = pd.read_csv("sales.csv")
df.head() # показываем первые строки
df.info() # типы данных, пропуски
df.describe() # статистика по числам
```

Очистить данные

```
df.drop_duplicates(inplace=True) # убираем дубликаты
df['price'] = df['price'].fillna(df['price'].mean()) # заполняем пропуски
df = df[df['price'] > 0] # удаляем невозможные значения
```

Как работать с данными в Data Science

Преобразовать данные

🔤 **Кодирование текста:**

```
df['category'] = df['category'].astype('category')
df['category_encoded'] = df['category'].cat.codes
```

🧪 **Масштабирование чисел (для моделей):**

```
from sklearn.preprocessing import StandardScaler scaler = StandardScaler()
df[['price_scaled']] = scaler.fit_transform(df[['price']])
```

Как работать с данными в Data Science

Исследовать данные (визуализация)

```
import seaborn as sns
import matplotlib.pyplot as plt
sns.histplot(df['price'], bins=20)
plt.title("Распределение цен")
plt.show()
sns.scatterplot(x='price', y='rating', data=df)
plt.title("Связь цены и рейтинга")
plt.show()
```

Как работать с данными в Data Science

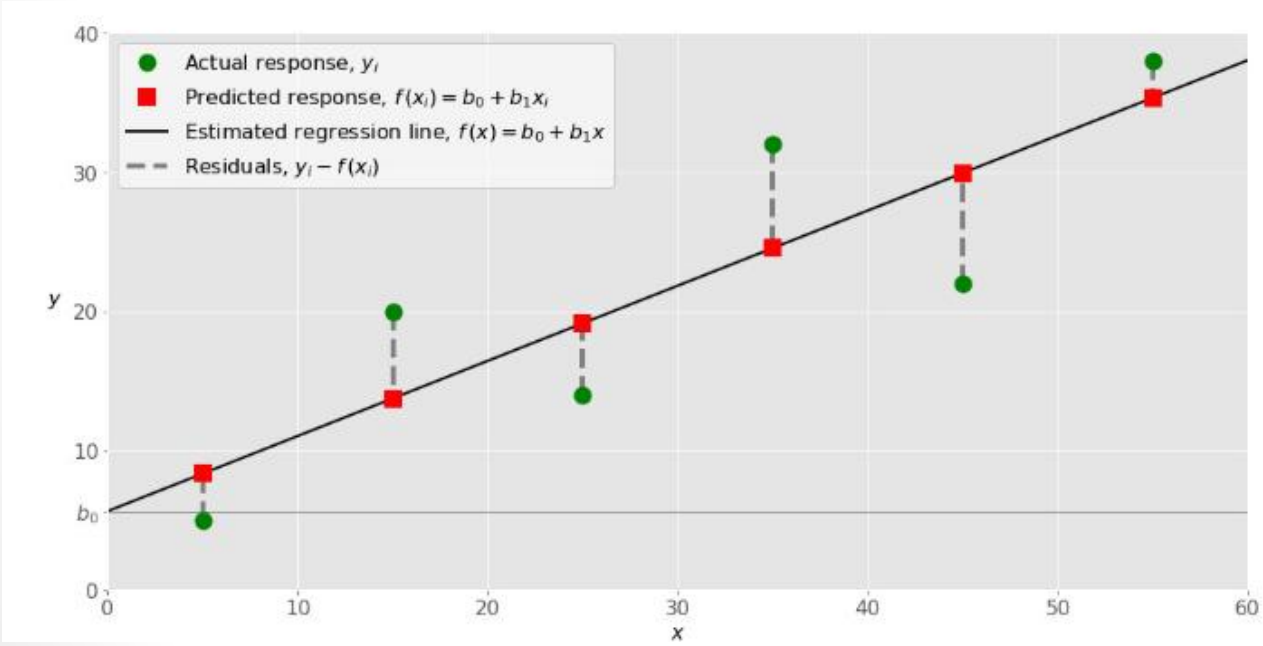
Обучить модель

```
from sklearn.linear_model import LinearRegression
X = df[['price']]
y = df['sales']
model = LinearRegression()
model.fit(X, y)
y_pred = model.predict(X)
```

Линейная регрессия

Линейная регрессия — это модель, которая **ищет прямую зависимость между признаками и целевой переменной.**

То есть: "Как изменяется y, если изменить x?«



Как зависит **цена квартиры** от её **площади**

Как влияет **опыт работы** на **зарплату**

Формула Линейной Регресии

Классическая линейная модель:

y=w·x+by=*w*·*x*+*b*

где:

- x — признак (например, площадь)
- y — результат (например, цена)
- w — **вес (коэффициент)** признака
- b — **сдвиг (intercept, свободный член)**

Площадь (м²)	Цена (₽ тыс)
30	2500
50	4000
70	5200

Модель "учится" подбирать w и b, чтобы как можно точнее предсказывать y.

Модель подбирает такие коэффициенты w и b, чтобы **суммарная ошибка между предсказанными и реальными значениями была минимальной**.

Ошибка модели (MSE):

$$\text{MSE} = \frac{1}{n}\sum(y_{\text{реальное}} - y_{\text{предсказанное}})^2$$

Пример на Python

```
import pandas as pd from sklearn.linear_model
import LinearRegression
X = data[['area']] # признаки
y = data['price'] # целевая переменная
model = LinearRegression()
model.fit(X, y)
```

Минусы Линейной Регресии

- Модель **предполагает линейную зависимость** — если зависимость не линейная, результат может быть неточным.
- **Чувствительна к выбросам** — экстремальные значения могут сильно сместить прямую.
- Не умеет учитывать **сложные связи между признаками**.

Как измерять качество регрессии?

🔹 **MAE (Mean Absolute Error) — средняя абсолютная ошибка**

$$MAE = \frac{1}{n} \sum |y_{\text{истинное}} - y_{\text{предсказанное}}|$$

🔸 **Интерпретация:**

На сколько в среднем ошибается модель (в тех же единицах, что и ответ).

🔧 Пример:

Предсказываем зарплату → MAE = 5000 → в среднем ошибка ±5 тыс

🔹 **MSE (Mean Squared Error) — средняя квадратичная ошибка**

$$MSE = \frac{1}{n} \sum (y_{\text{истинное}} - y_{\text{предсказанное}})^2$$

🔸 **Чувствительна к выбросам** (сильно штрафует большие ошибки)

🔹 **RMSE — корень из MSE**

$$RMSE = \sqrt{MSE}$$

🔸 То же, что MSE, но в тех же единицах, что и ответ

🔹 **R² (коэффициент детерминации)**

$$R^2 = 1 - \frac{MSE}{\text{вариация данных}}$$

🔸 Показывает, **какая доля дисперсии объясняется моделью**

📈 Значения от 0 до 1 (чем ближе к 1 — тем лучше)

Метрика	Показывает	Комментарий
MAE	Среднюю ошибку	Просто и понятно
MSE	Квадратичную ошибку	Штрафует сильнее
RMSE	Корень из MSE	Более интерпретируемо
R^2	"Процент объяснённости"	Часто используется в отчётах

Обучение нейронных сетей. Градиентный спуск

Градиентный спуск работает так: в начале обучения все веса устанавливаются в произвольные значения: генерируется случайный набор чисел. А затем много раз повторяют шаг обновления весов: каждое число (каждый вес) немного меняют, чтобы ошибка предсказания для обучающих данных немного уменьшилась.