Data science

Лекция 2. Применение Data science

2025/2026 учебный год

Доцент кафедры МО&МО, Махно В.В.

Кодовое слово для способа МАГИСТРЫ 2026 = **12022026**

Data Science и инструменты Анализа Данных

ИНСТИТУТ
МАТЕМАТИКИ
МЕХАНИКИ
КОМПЬЮТЕРНЫХ
НАУК
имени И.И. Воровича

«Данные — это топливо 21 века» (2006 г). 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»

Востребованность профессий

hh.ru –> “Data Scientist” — в крупных городах это **сотни вакансий**.

- Для «Data Analyst» — ещё больше: от ритейла до финтеха.
- **LinkedIn** показывает: профессии, связанные с данными, — в топ-10 самых быстрорастущих в мире.

- Зарплаты (по данным hh.ru, весна 2025):

зарплаты зависят от города, компании и реальных навыков — чем лучше портфолио и понимание, тем выше предложение.

Профессия	Джун (₽)	Мидл (₽)	Сеньор (₽)
Data Analyst	80–120 тыс.	130–180 тыс.	200–250 тыс.
Data Scientist	100–150 тыс.	180–250 тыс.	300+ тыс.
ML-инженер	120–160 тыс.	200–270 тыс.	350+ тыс.

Сравнение с другими IT-направлениями

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

Данные

- Данные — это основа всего
- Без данных нет Data Science.
- Data Scientist без данных — как повар без продуктов .

- Что такое данные?
- Данные — это любая зафиксированная информация, которую можно анализировать:

- Пример Что это за данные?
- В интернет-магазине Название товара, цена, количество покупок
- В медицинской системе Возраст пациента, диагноз, результаты анализов
- В приложении Время использования, клики, переходы по экранам
- В Spotify Треки, лайки, пропуски, плейлисты

Каждый набор данных — это как таблица Excel:
Строки — объекты (пользователи, пациенты, товары)
Столбцы — признаки (возраст, цена, диагноз, рейтинг)

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

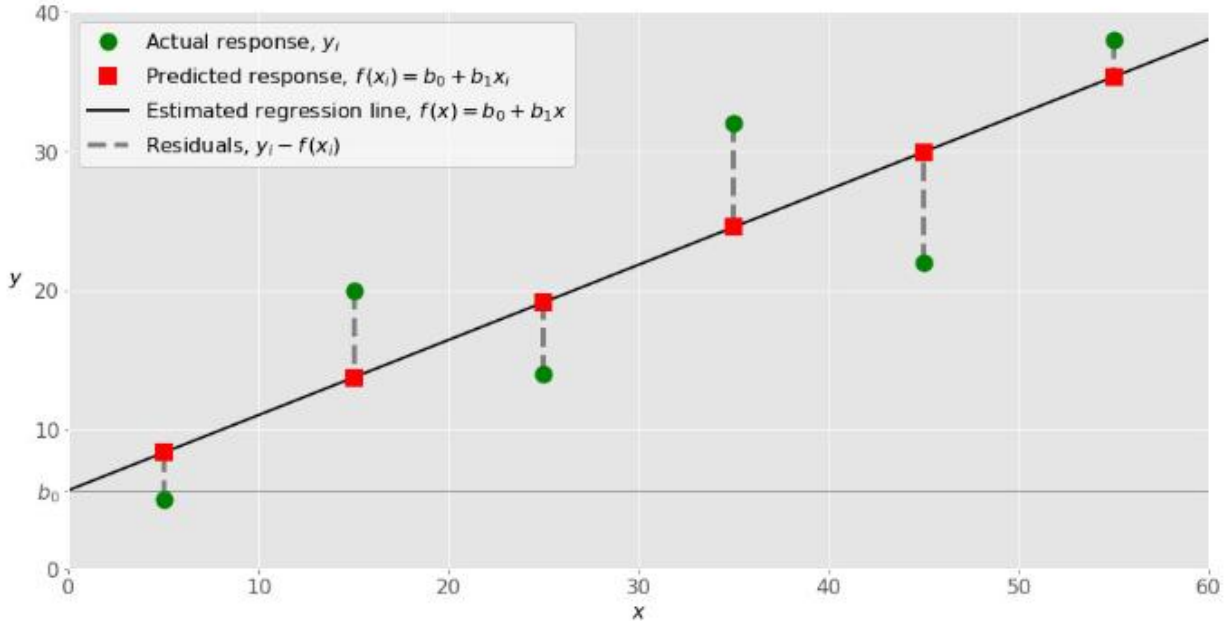
Обучить модель

```
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:

$y=w\cdot x+b$ $y=w\cdot x+b$

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):

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

Кодовое слово для способа МАГИСТРЫ 2026 = **12022026**