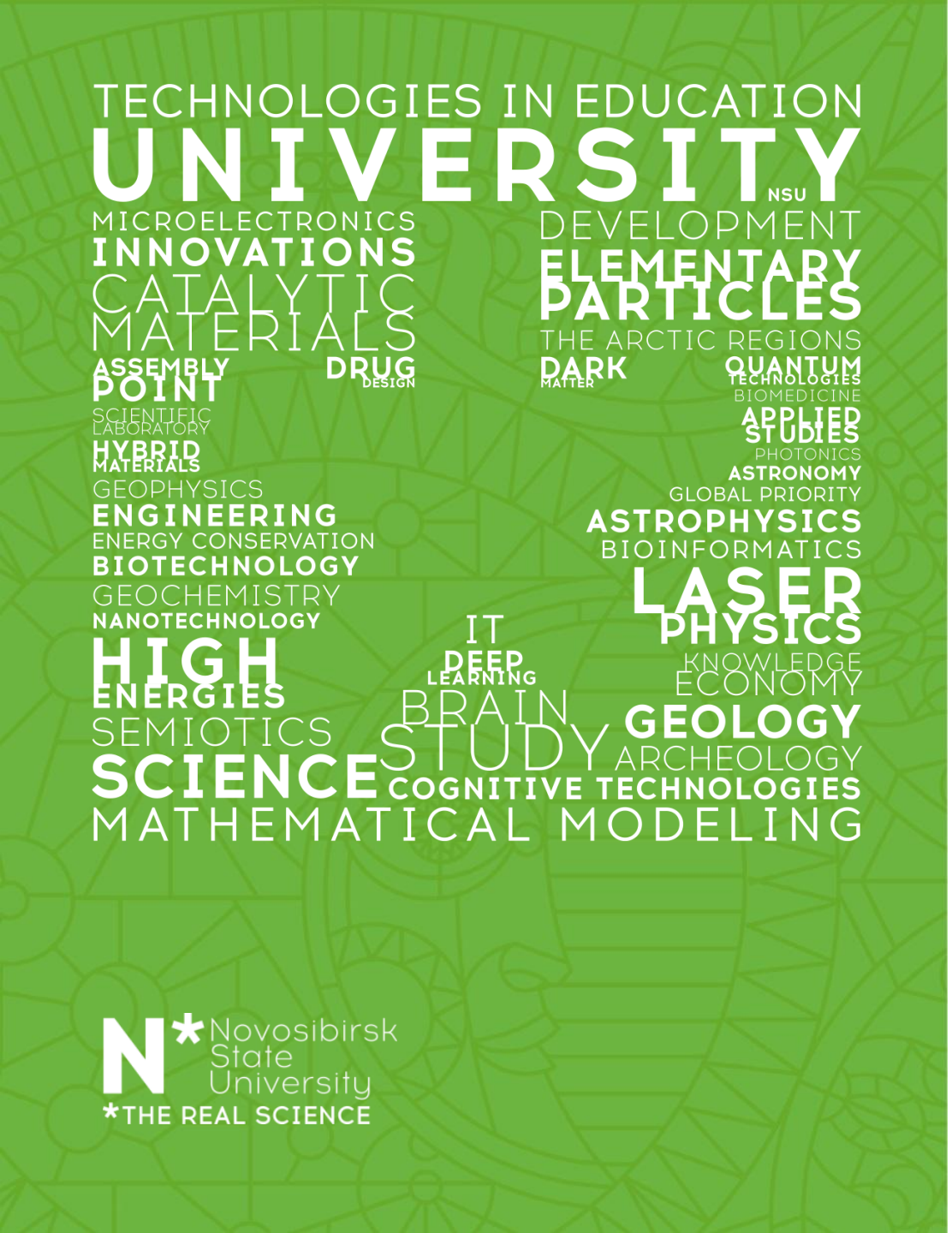


Базовые методы ИИ
Семинар 2

Глушенко Андрей Валерьевич
ФФ НГУ

Регрессия

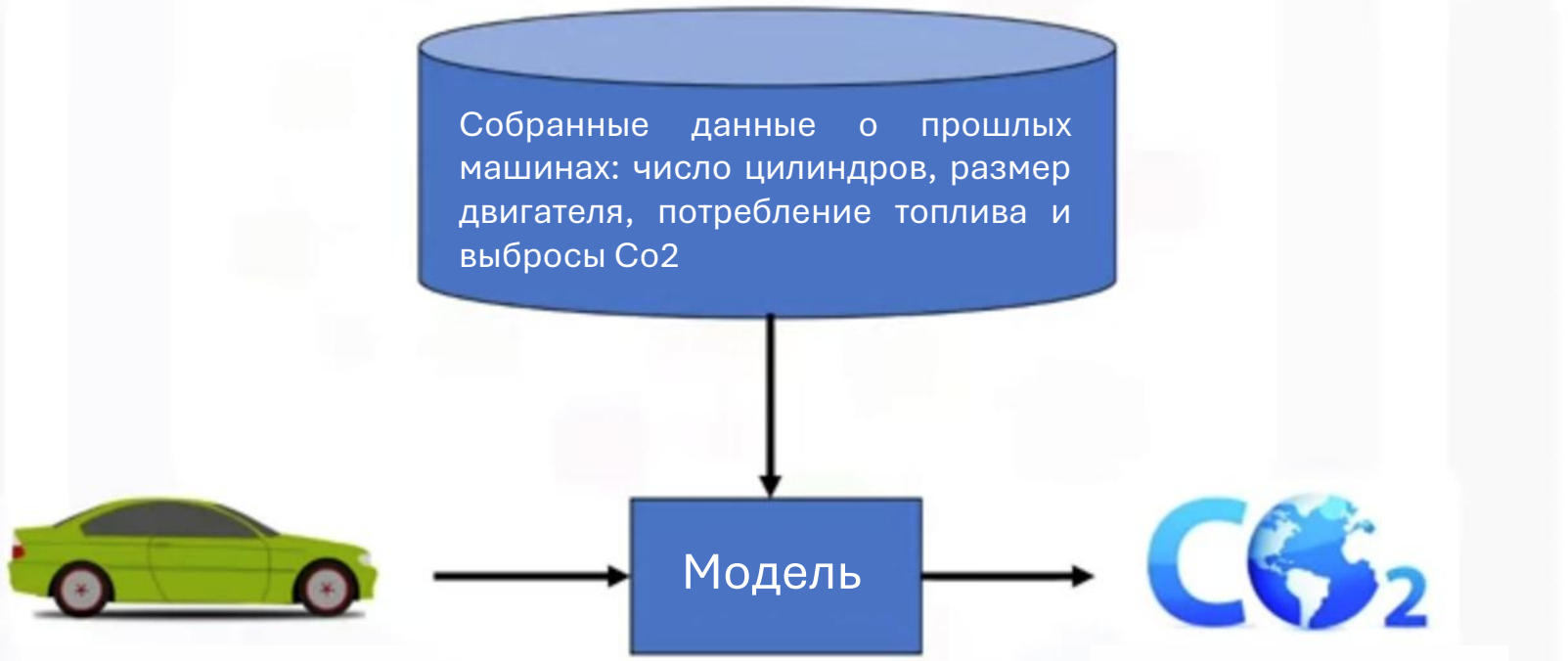
Введение

Simple Linear Regression

Практическое применение

Оценка модели и диагностика

Регрессия

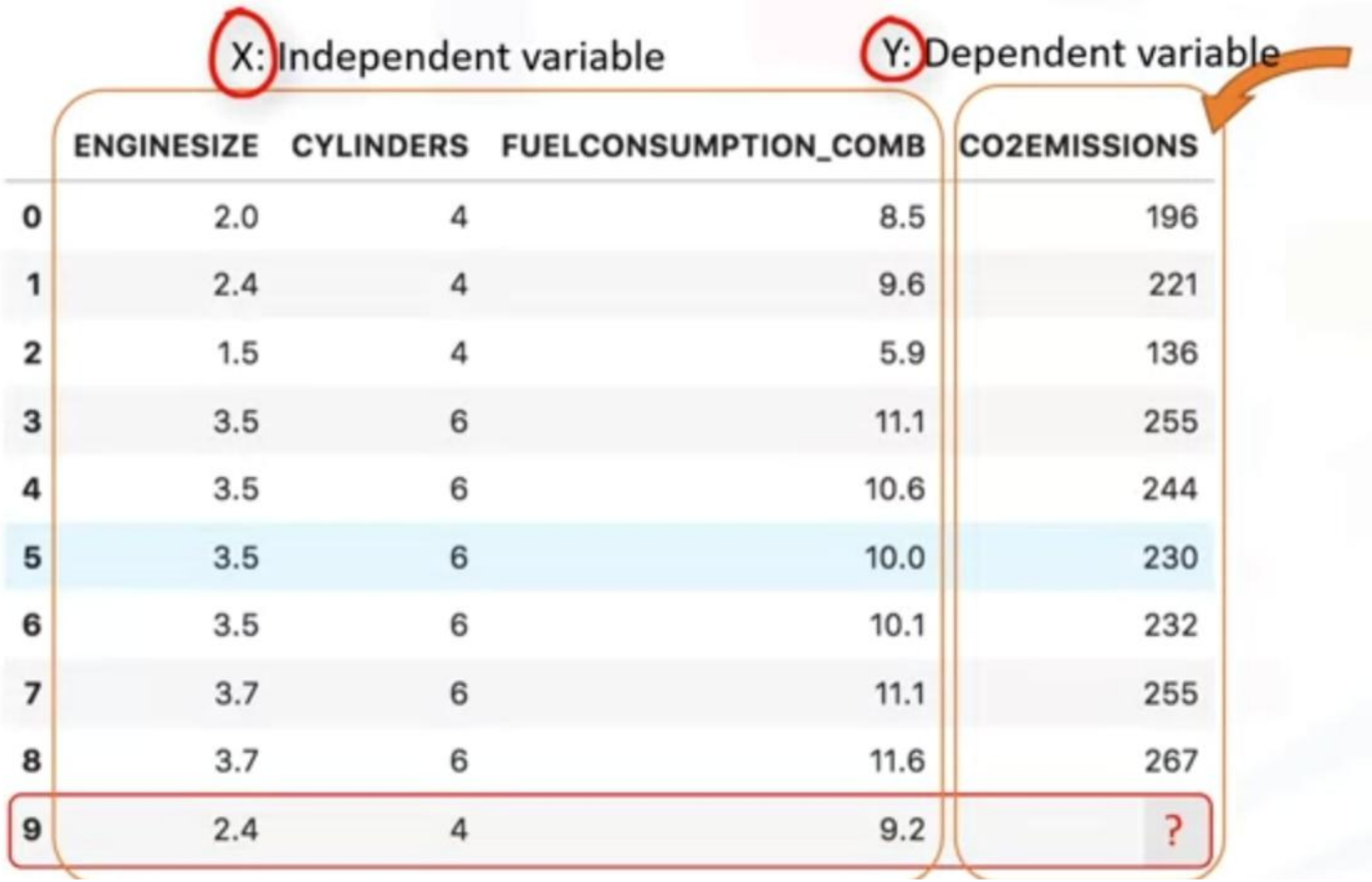

Собранные данные о прошлых машинах: число цилиндров, размер двигателя, потребление топлива и выбросы Co2
Модель
Новая машина
Ожидаемый выброс Co2

Регрессия

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Регрессия

- Задача предсказания непрерывного числа



	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Типы регрессии

Простая регрессия

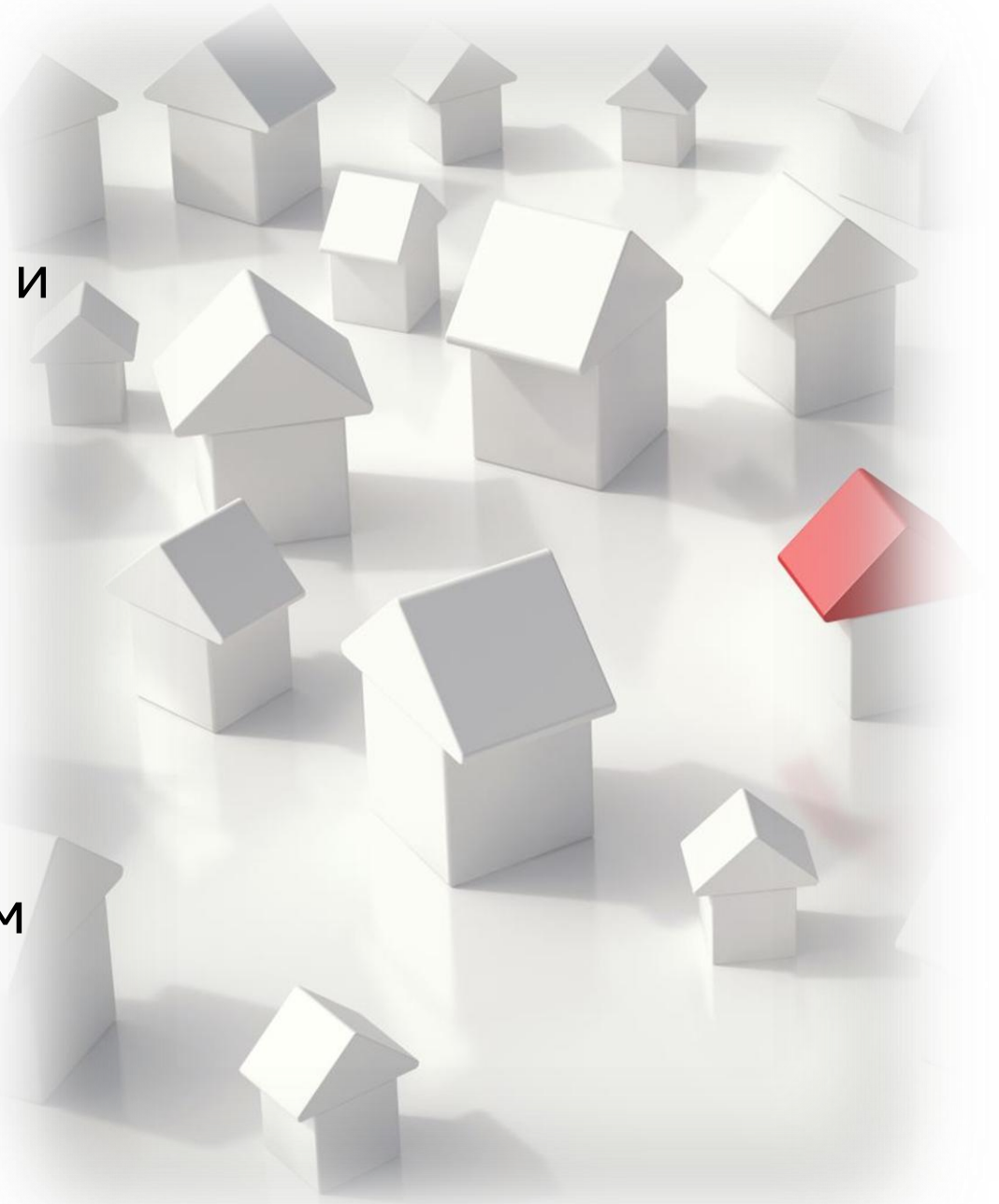
- Использование одной независимой переменной для прогнозирования зависимой переменной.
- Например: прогнозирование выбросов CO2 на основе размера двигателя.
- Простая линейная регрессия, простая нелинейная регрессия

Множественная регрессия

- Использование более чем одной независимой переменной для прогнозирования зависимой переменной.
- Например: прогнозирование выбросов CO2 на основе размера двигателя и количества цилиндров.
- Множественная линейная регрессия, множественная нелинейная регрессия

Применение регрессии

- Прогнозирование годового объема продаж человека на основе возраста, стажа работы и т. д
- Определение индивидуальной удовлетворенности на основе демографических и психологических факторов
- Прогнозирование цены дома на основе его размера, количества комнат и т. д
- Прогнозирование дохода от работы с учетом независимых переменных, таких как количество рабочих часов, образование, профессия, пол, возраст, стаж работы



Вопрос

Какой из примеров является примером применения регрессионного анализа?

- Прогнозирование наличия или отсутствия рака у пациента.
- Группировка похожих домов в районе.
- Прогнозирование количества осадков на следующий день.
- Прогнозирование победы или поражения команды.

Алгоритмы для решения задачи регрессии

- ❑Порядковая регрессия
- ❑Регрессия Пуассона
- ✓Линейная регрессия (Простая и множественная)
- ❑Полиномиальная регрессия
- ❑Регрессия Лассо
- ❑Гребневая регрессия
- ❑Регрессия на основе дерева решений
- ❑Регрессия на основе бустированного дерева решений

Простая линейная регрессия

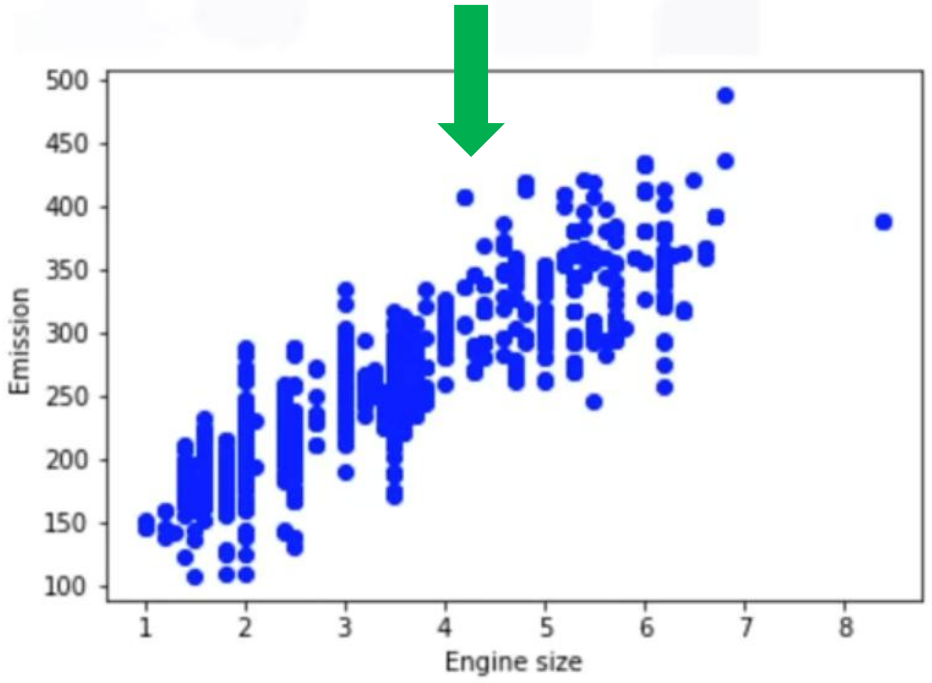
Линейная регрессия — это аппроксимация линейной модели, используемая для описания взаимосвязи между двумя или более переменными.

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Простая линейная регрессия

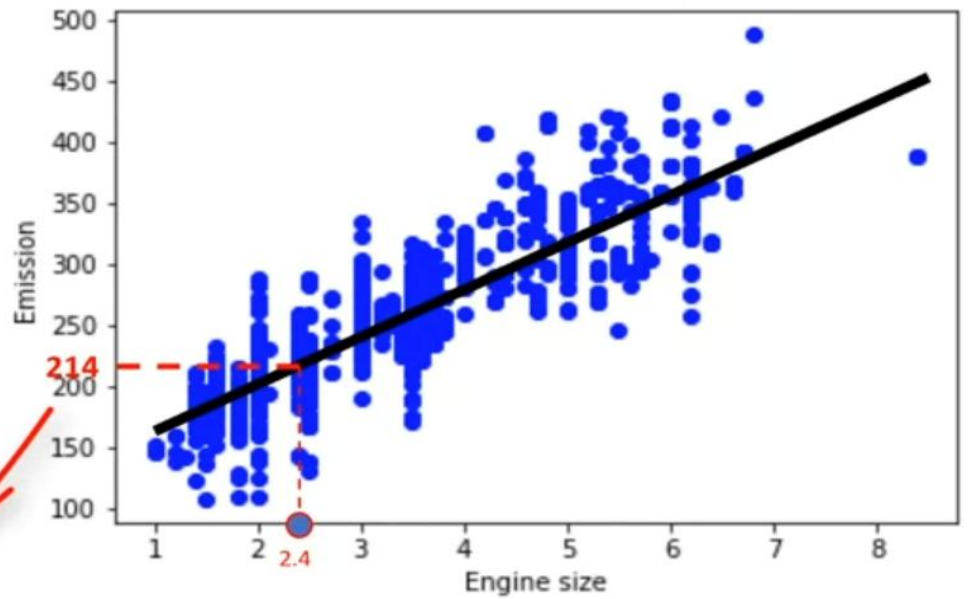
Изменение в одной переменной объясняет изменение в другой

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?



Простая линейная регрессия

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?



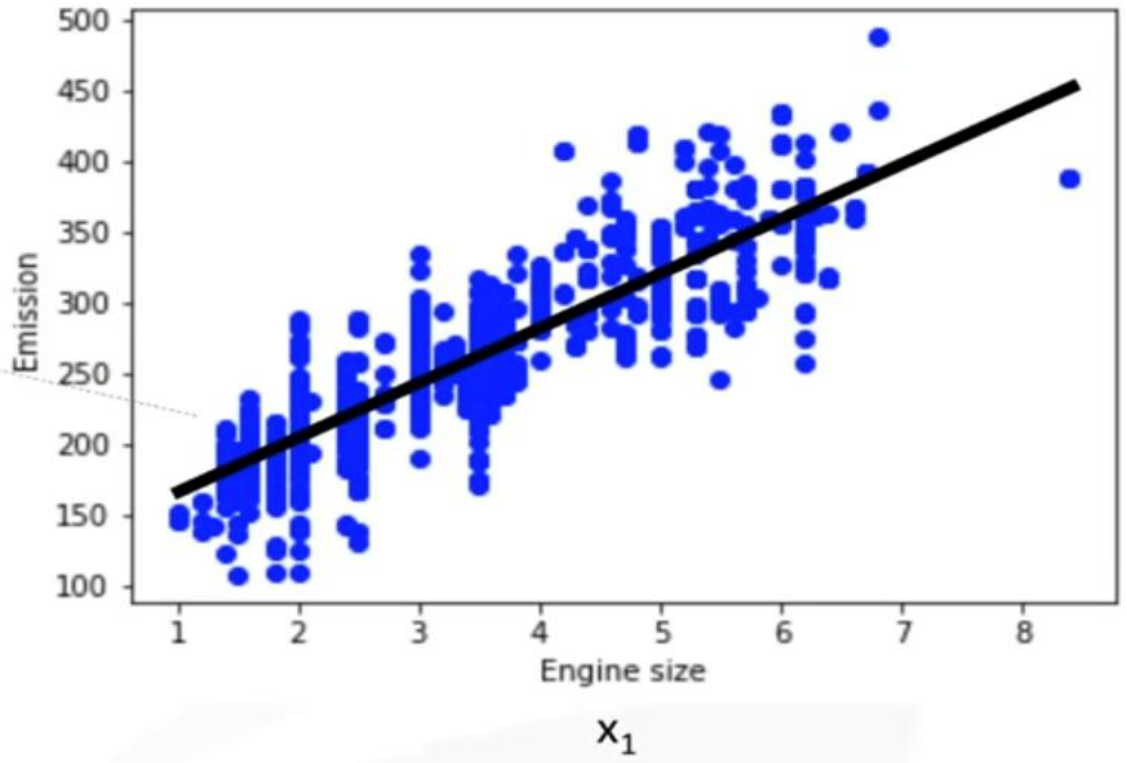
Простая линейная регрессия

Параметры, которые нужно найти

$$\hat{y} = \theta_0 + \theta_1 x_1$$

Независимая переменная

Зависимая переменная



Как найти оптимальный вариант?

$x_1 = 5.4$ независимая переменная

$y = 250$ истинный выброс Co2 для x1

$\hat{y} = \theta_0 + \theta_1 x_1$

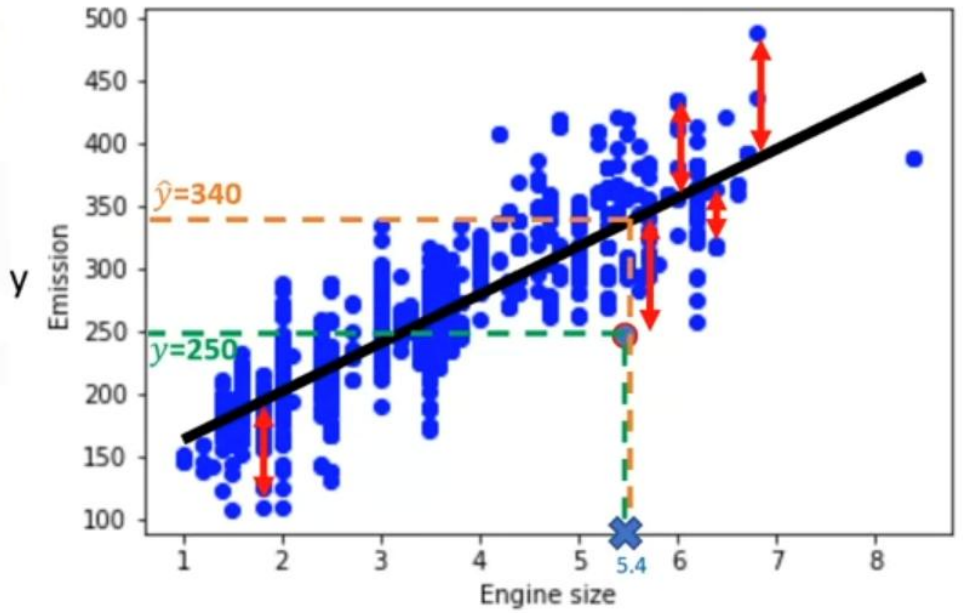
$\hat{y} = 340$ предсказанный выброс Co2 для x1

Error = $y - \hat{y}$

$= 250 - 340$

$= -90$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$



Математический подход

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

X_1 — ENGINESIZE; y — CO2EMISSIONS

$$\hat{y} = \theta_0 + \theta_1 x_1$$

$$\theta_1 = \frac{\sum_{i=1}^{s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{s}(x_i - \bar{x})^2}$$

$\bar{x}$ = (2.0 + 2.4 + 1.5 + ...)/9 = 3.0β

$\bar{y}$ = (196 + 221 + 136 + ...)/9 = 226.22

$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

Минимизируя сумму квадратов ошибок (SSE)

Предсказание

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

$$\hat{y} = \theta_0 + \theta_1 x_1$$

$$Co2Emission = \theta_0 + \theta_1 \, EngineSize$$

$$Co2Emission = 125 + 39 \, EngineSize$$

$$Co2Emission = 125 + 39 \times 2.4$$

$$Co2Emission = 218.6$$

Преимущества

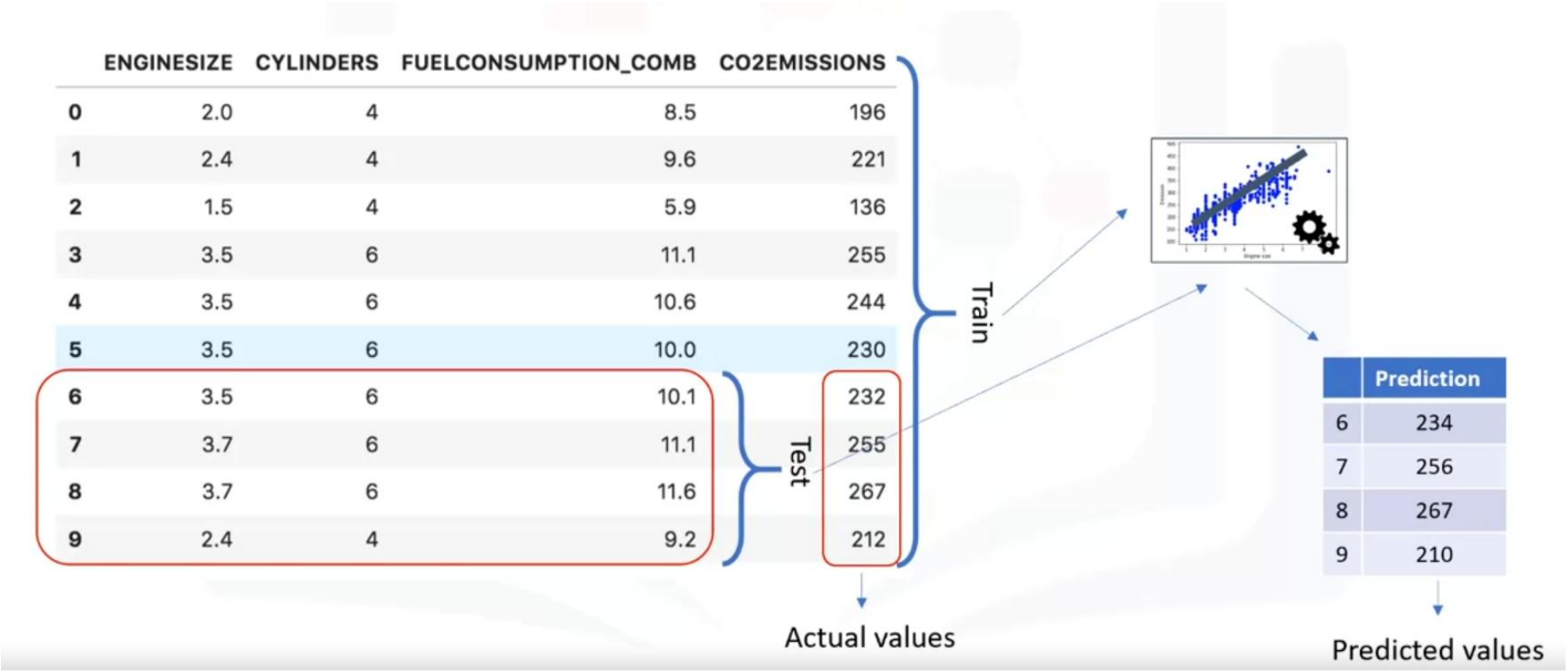
- Быстро
- Не нужна настройка гиперпараметров
- Интерпретируемая

Disadvantages

- Чувствительная к выбросам
- Не поможет при нелинейной связи

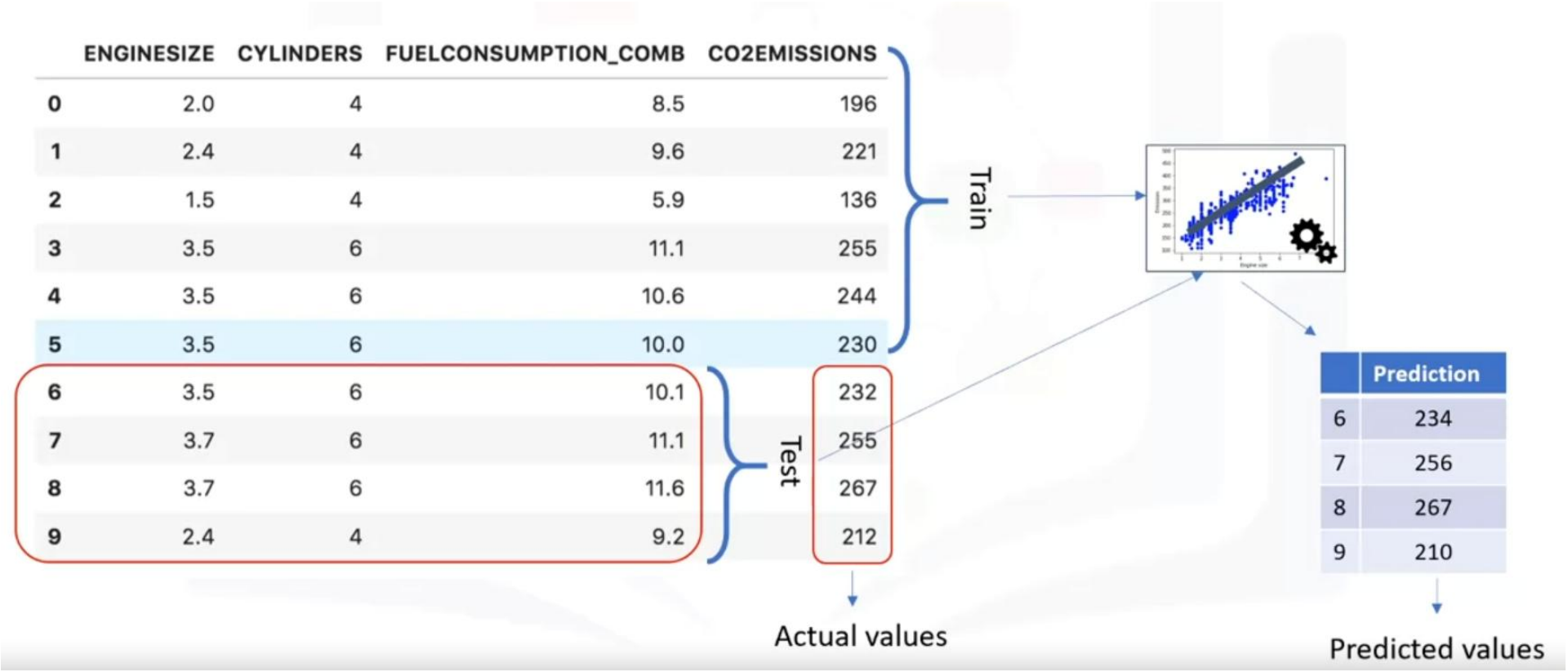
Как оценить точность (Accuracy) модели

Train/Test split



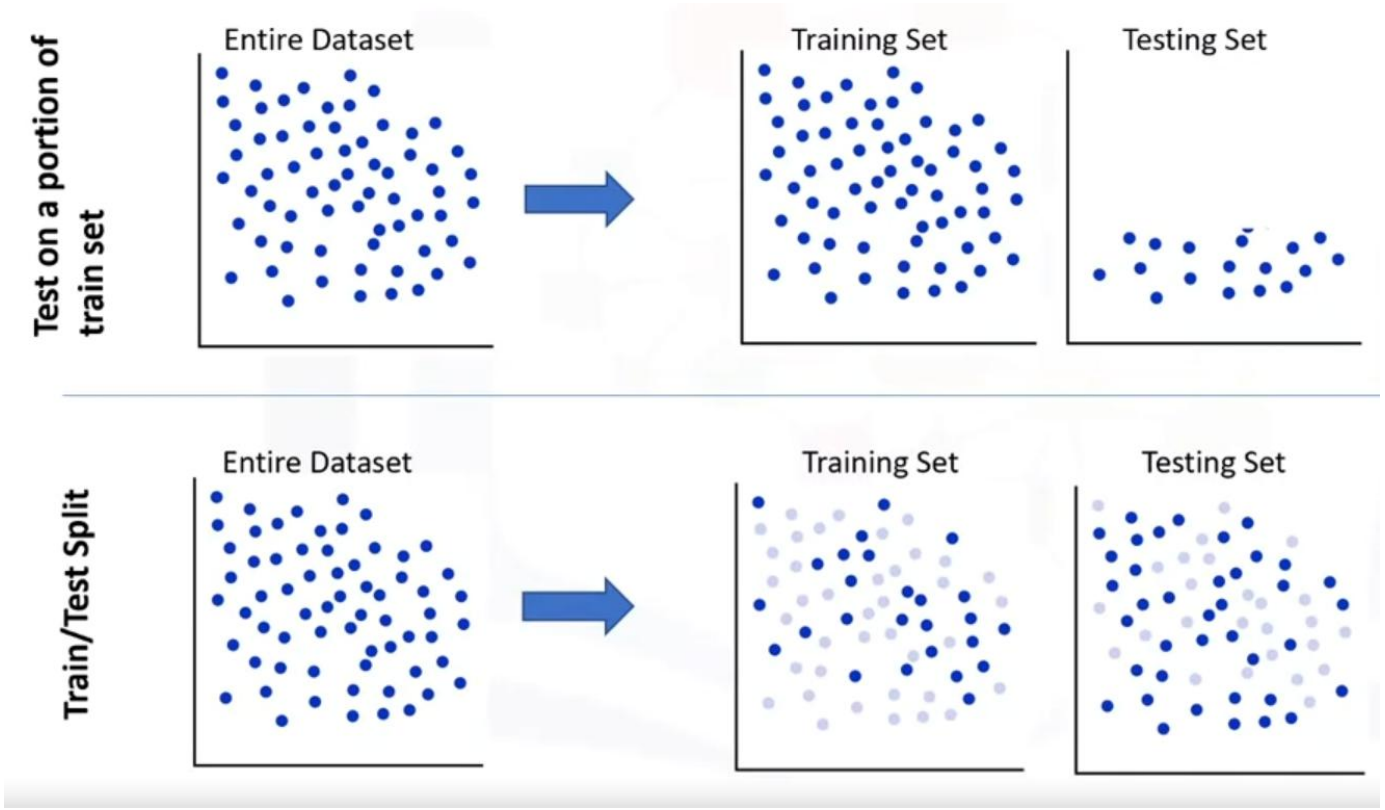
Как оценить точность (Accuracy) модели

Train/Test split



Как оценить точность (Accuracy) модели

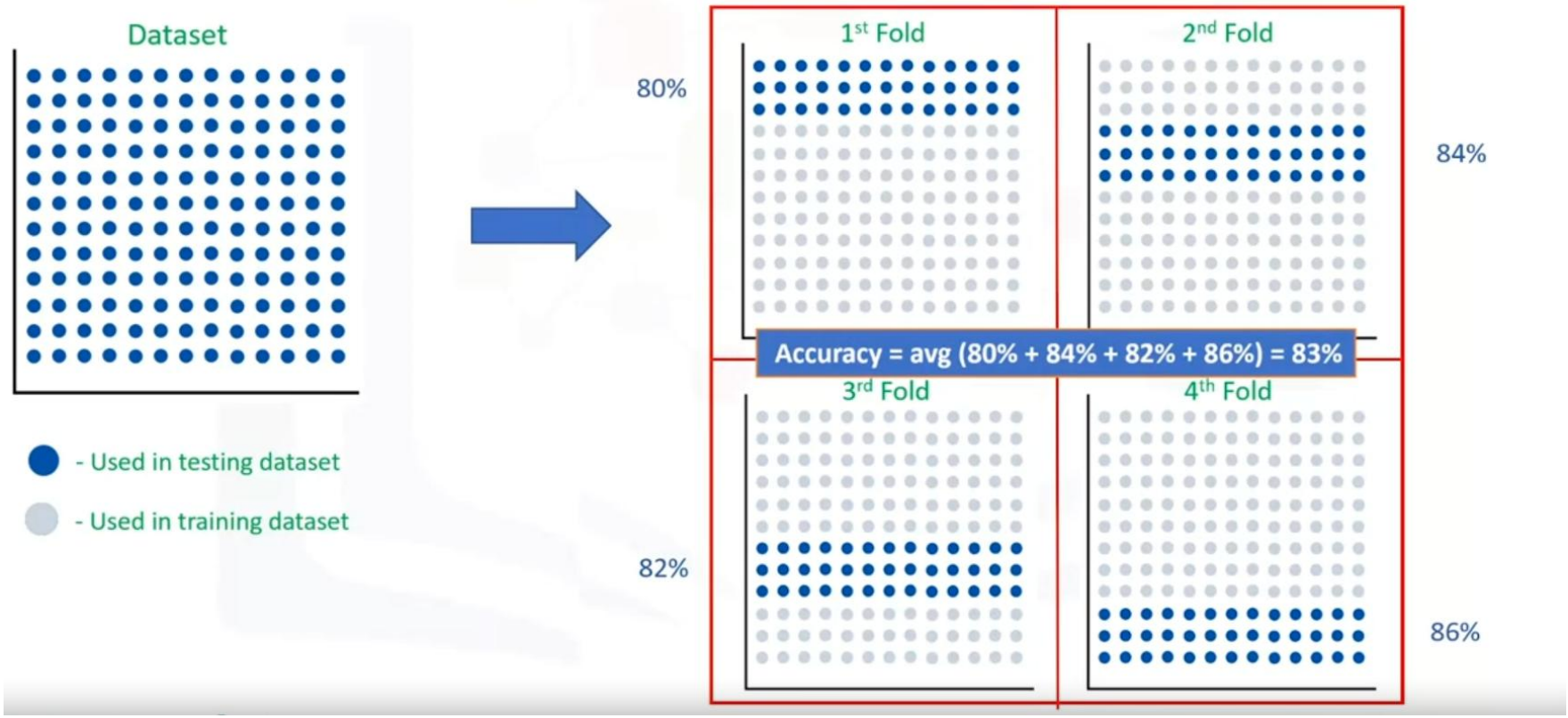
Train/Test split



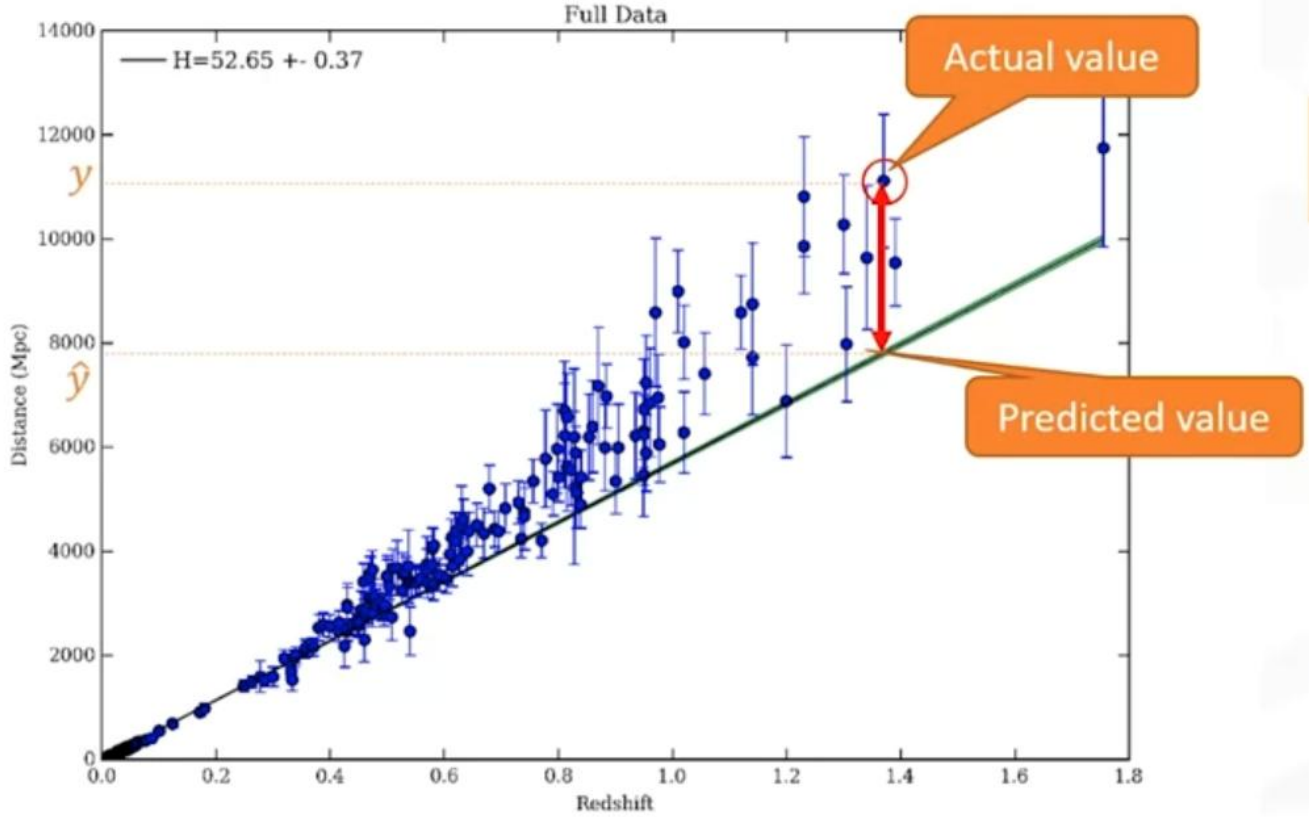
- Тестовая выборка часть обучающей
- Высокая точность на обученных данных
- Низкая точность на новых данных

- Разделение на независимые подвыборки
- Более высокая точность на новых данных
- Высоко зависит от того, на каких наборах данных обучалась и оценивалась

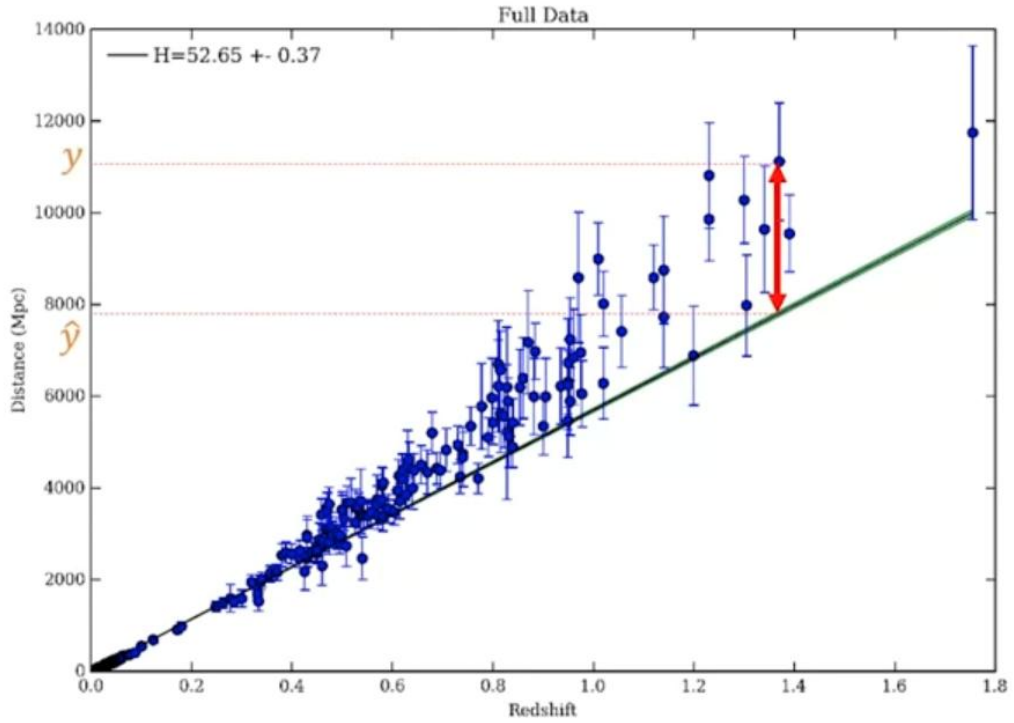
Кросс-валидация (k-fold cross-validation)



Оценка метрик в модели регрессии



Оценка метрик в модели регрессии



$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

$$RAE = \frac{\sum_{j=1}^{n}|y_j - \hat{y}_j|}{\sum_{j=1}^{n}|y_j - \bar{y}|}$$

$$RSE = \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2}$$

$$R^2 = 1 - RSE$$

Множественная линейная регрессия

$$Co2\ Em = \theta_0 + \theta_1 Engine\ size + \theta_2 Cylinders + \ ...$$

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ ... + \theta_n x_n$$

$$\hat{y} = \theta^T X$$

$$\theta^T = [\theta_0, \theta_1, \theta_2, ...] \qquad X = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ ... \end{bmatrix}$$

X: Independent variable

Y: Dependent variable

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Как найти лучшие параметры?

$\hat{y} = \theta^T X$

$\hat{y}_i = 140$ предсказанный x_i

$y_i = 196$ истинный x_i

$y_i - \hat{y}_i = 196 - 140 = 56$ остаточная ошибка

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

Как найти лучшие параметры?

- Математический подход: операции линейной алгебры (для небольших наборов данных)
- Оптимизационный подход: градиентный спуск (для больших наборов данных)

Множественная линейная регрессия

$$Co2\ Em = \theta_0 + \theta_1 Engine\ size + \theta_2 Cylinders + \ ...$$

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ ... + \theta_n x_n$$

$$\hat{y} = \theta^T X$$

$$\theta^T = [\theta_0, \theta_1, \theta_2, ...] \qquad X = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ ... \end{bmatrix}$$

$$\theta = (X^T X)^{-1} X^T y$$

X: Independent variable

Y: Dependent variable

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Множественная линейная регрессия

$$Co2\ Em = \theta_0 + \theta_1 Engine\ size + \theta_2 Cylinders + \ ...$$

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ ... + \theta_n x_n$$

$$\hat{y} = \theta^T X$$

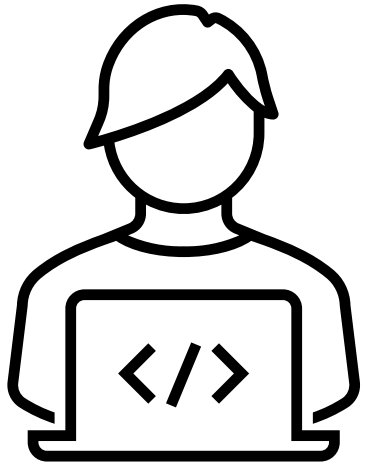
$$\theta^T = [\theta_0, \theta_1, \theta_2, ...] \qquad X = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ ... \end{bmatrix}$$

$$\theta = (X^T X)^{-1} X^T y$$

Смещение (Bias): Чтобы модель могла предсказывать значения не из начала координат, мы добавляем фиктивный столбец из единиц к матрице признаков.

Метрика R^2: Она показывает долю объясненной дисперсии. Значение 1.0 — идеальное предсказание, 0.0 — модель работает не лучше, чем простое предсказание среднего значения.

Практика



Семинар 2:

https://colab.research.google.com/drive/1ClbOl4-sHVzBL8C4g9qEFI2OOIoWxsi-?usp=sharing

Лабораторная работа 2:

https://colab.research.google.com/drive/15x9TOZsFSvbrF-0bxKw8Wwk75N3voAM8?usp=sharing

Датасет для lab02:

https://www.kaggle.com/datasets/prokshitha/home-value-insights