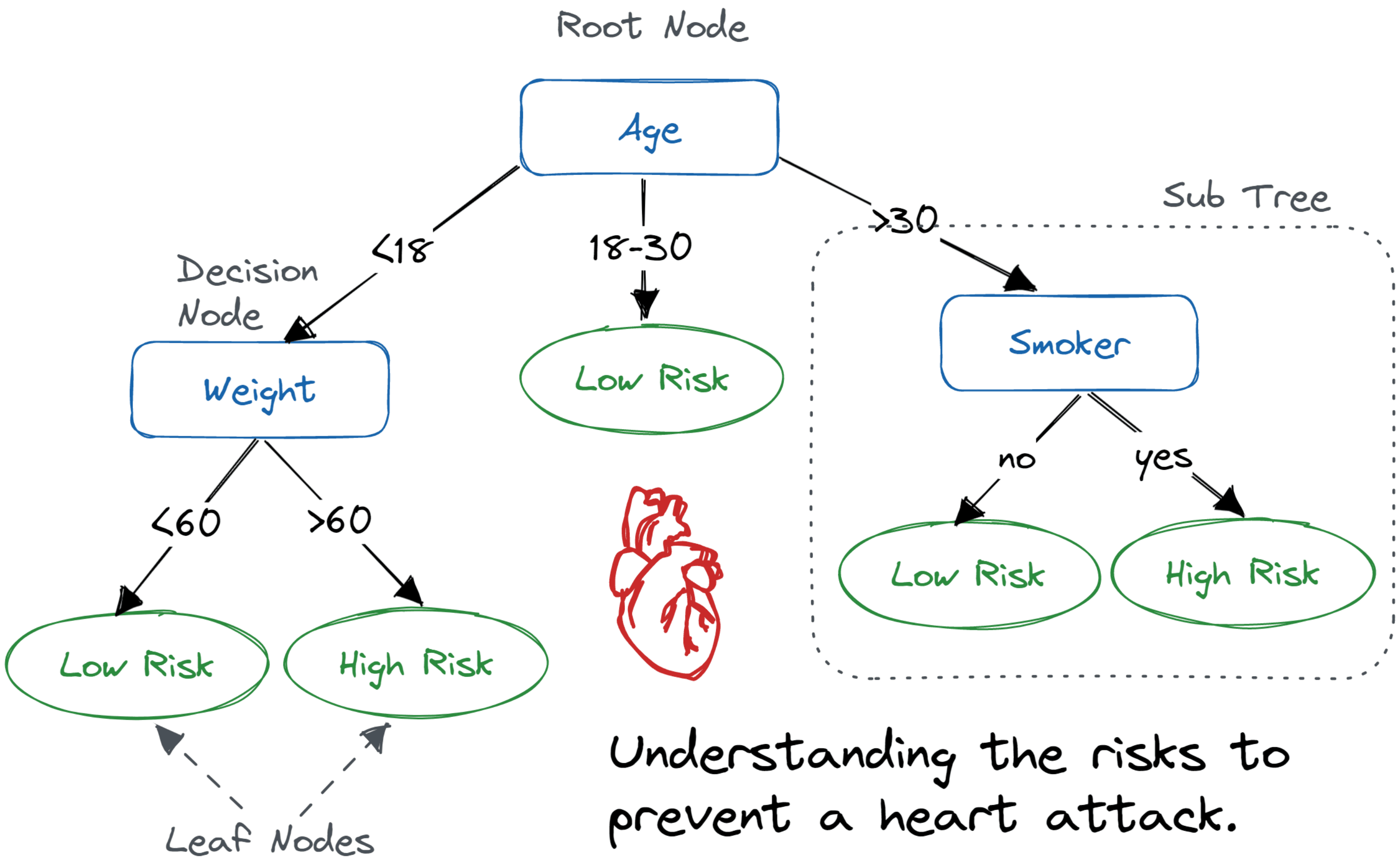
Ансамбли моделей

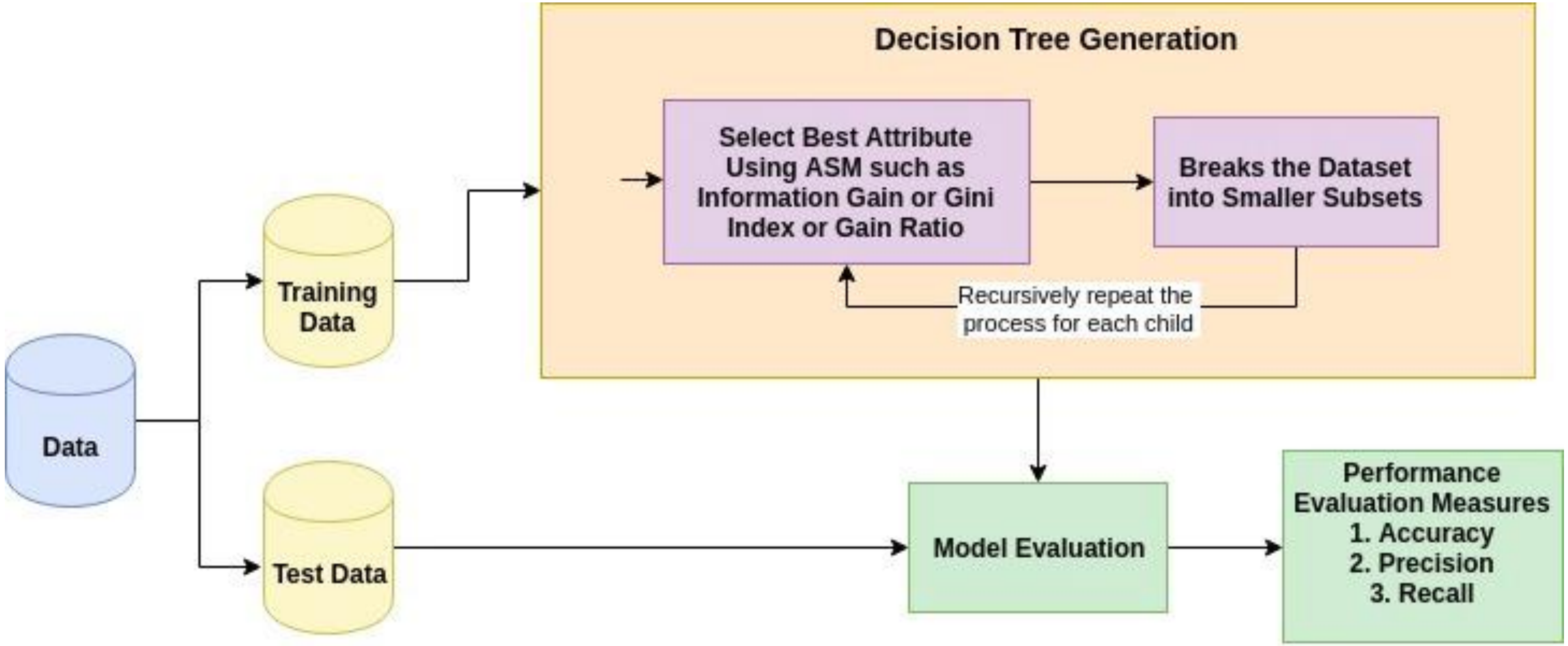
Лекция 5

Тест

Повторение. Решающие деревья



Повторение. Решающие деревья



Повторение. Решающие деревья

Регрессия:

$$H(R) = \frac{1}{N}\sum_{i=0}^{N}\left(y_i - \frac{1}{N}\sum_{i=0}^{N} y_i\right)^2 \to min$$

Классификация:

$$H(R) = -\sum_{k=1}^{K} p_k(1 - p_k) \to min$$

$$H(R) = -\sum_{k=1}^{K} p_k \log p_k \to min$$

Семплирование выборки

$\mathbb{X} = (x_i, y_i)_{i=1}^{l}$ – исходная выборка

Создаем n выборок размера l: $X_1, X_2 \ldots, X_n$

Обучаем n моделей на каждой из выборок: $b_1, b_2 \ldots, b_n$

Если мы знаем распределение объектов (с какой вероятностью каждый объект попадет в каждую из выборок $X_1, X_2 \ldots, X_n$) $p(x)$

Истинные ответы модели – $\mathrm{y}(x)$

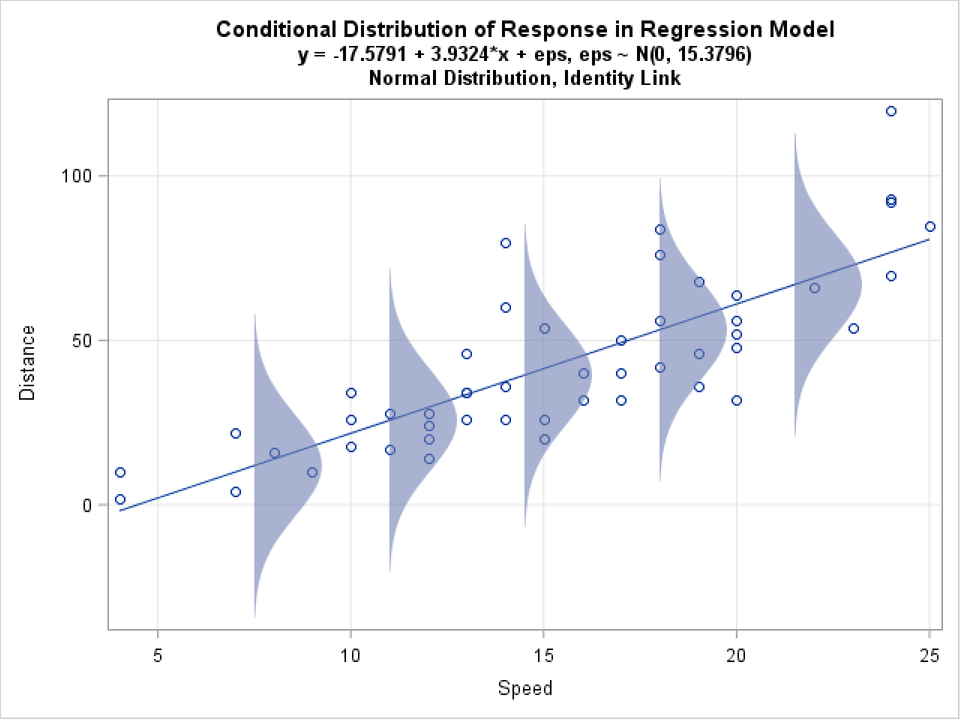
Ошибки

Для каждой модели запишем математическое ожидание среднеквадратичной ошибки как

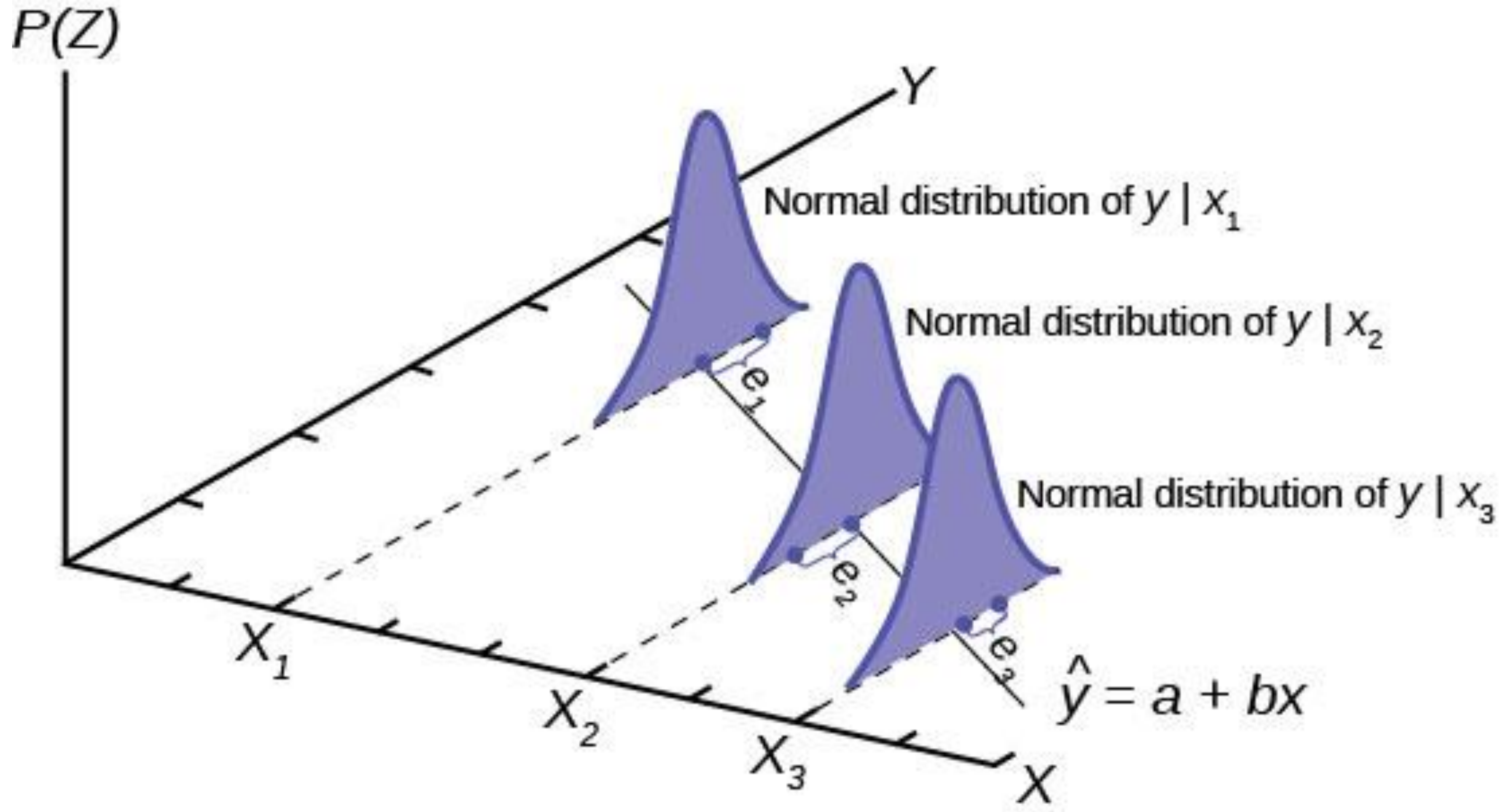
$$\mathbb{E}_x\left(b_j(x) - y(x)\right)^2 = \mathbb{E}_x\varepsilon_j^2(x)$$

Два предположения:

1. $\mathbb{E}_x\varepsilon_j(x) = 0$
2. $\mathbb{E}_x\left(\varepsilon_j(x)\cdot\varepsilon_i(x)\right) = 0, i \neq j$



Еще про распределение ошибок



Байка про бычка и мудрость толпы

Пример точечной оценки непрерывной величины

Фрэнсис Гальтон гулял в 1906 по сельской ярмарке и обратил внимание на конкурс, где участвовавшему человеку предлагалось оценить вес быка.

Почти 800 человек приняло участие в конкурсе, Гальтон собрал оценки каждого человека и обнаружил, что медианное значение составляло 1207 фунтов, в то время как истинный вес бычка был 1198 фунтов.

Для задачи классификации. Теорема Кондорсе о присяжных

Вводные:

M – количество присяжных

p – вероятность правильного решения одного эксперта

m – минимальное большинство членов жюри $m = \frac{N}{2} + 1$

Тогда вероятность правильного решения всего жюри:

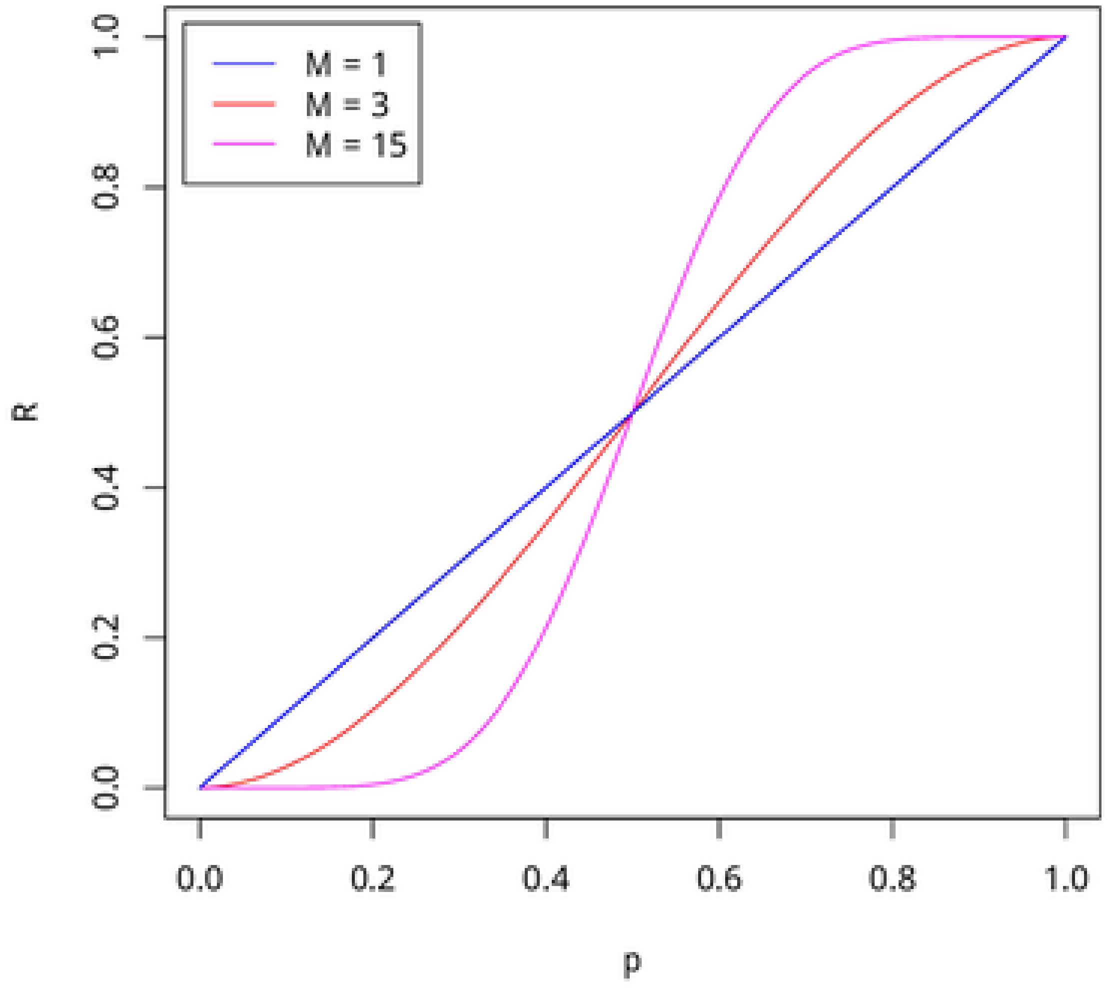
$$R = \sum_{i=m}^{M} C_M^i p^i (1-p)^{M-i}$$

Теорема Кондорсе о присяжных

Вероятность правильного решения всего жюри будет выше при:

1. Большем количестве присяжных M
2. Большей вероятности правильного решения одного эксперта p

Важно: оценки экспертов должны быть **независимы**



Усредненная модель

Применим новую модель, которая будет представлять собой усредненные предсказания всех n моделей.

$$a(x) = \frac{1}{n}\sum_{j=1}^{n} b_j(x)$$

Тогда ошибка:

$$\mathbb{E}_x\left(b_j(x)\ - y(x)\right)^2 = \mathbb{E}_x\left(\frac{1}{n}\sum_{j=1}^{n} b_j(x) - y(x)\right)^2 = \mathbb{E}_x\left(\frac{1}{n}\sum_{j=1}^{n} \varepsilon_j(x)\right)^2$$

Получаем улучшение

$$\mathbb{E}_x\left(\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j(x)\right)^2 = \frac{1}{n^2}\mathbb{E}_x\left(\sum_{j=1}^{n}\varepsilon_j^2(x) + \sum_{i\neq j}^{n}(\varepsilon_i(x)\cdot\varepsilon_j(x))\right)^2$$

$$= \boldsymbol{\frac{1}{n}\mathbb{E}_x\varepsilon_j^2(x)}$$

Разложение ошибки (смещение и разброс)

$\mathbb{X} = (x_i, y_i)_{i=1}^{l}$ – исходная выборка

Создаем n выборок размера l: $X_1, X_2 \dots, X_n$

Обучаем n моделей на каждой из выборок: $b_1, b_2 \dots, b_n$

Разложение ошибки (смещение и разброс)

Функционал ошибки на всех выборках запишем в виде:

$$Q(a) = \mathbb{E}_x \mathbb{E}_{X,\epsilon}[y(x,\epsilon) - a(x,X)]^2$$

где X – обучающая выборка

x – точка (объект) из тестового множества

$y(x,\epsilon) = f(x) + \epsilon$ – целевая зависимость, которую можем измерить с точностью до случайного шума ϵ

$a(x,X)$ – предсказание алгоритма, обученного на X, на объекте x

$\mathbb{E}_x$ – математическое ожидание по всей тестовой выборке (среднее по всем тестовым объектам)

$\mathbb{E}_{X,\epsilon}$– математическое ожидание по всем обучающим выборкам и случайному шуму ϵ

https://education.yandex.ru/handbook/ml/article/ansambli-v-mashinnom-obuchenii

Разложение ошибки (смещение и разброс)

$$Q(a) = \mathbb{E}_x bias_X^2 a(x, X) + \mathbb{E}_x \mathbb{V}_X[a(x, X)] + \sigma^2$$

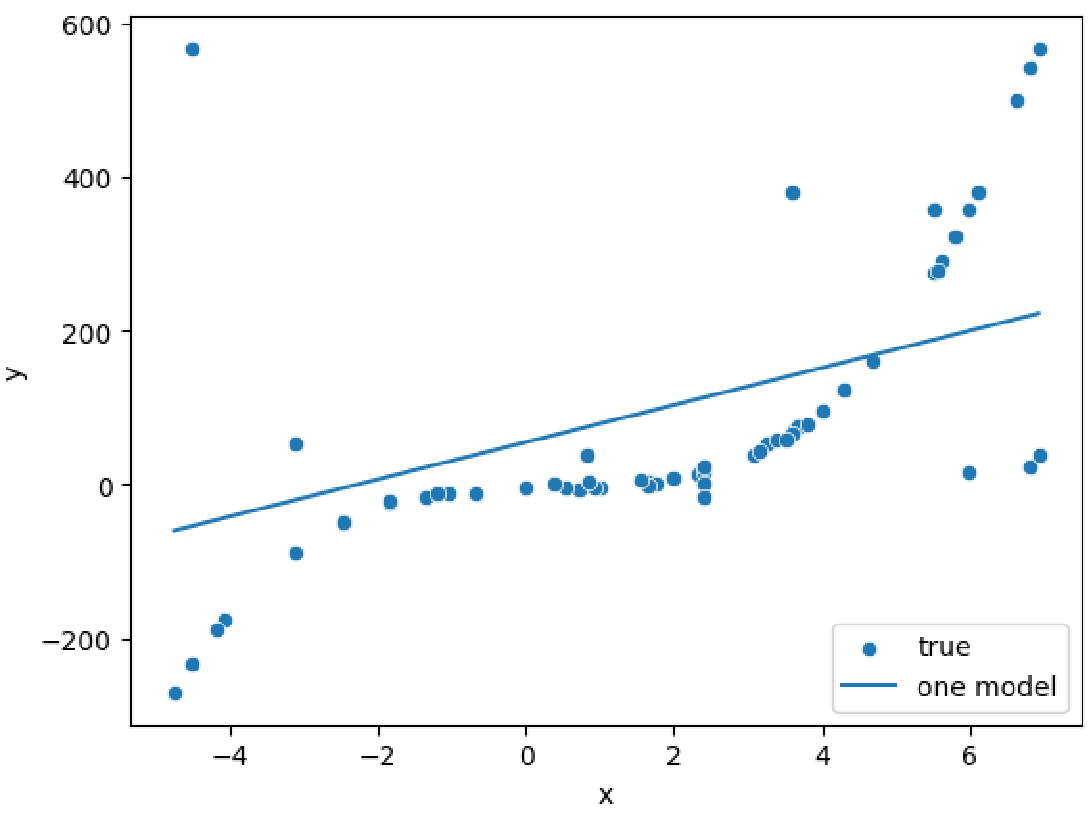
где $bias_X\ a(x, X) = f(x) - \mathbb{E}_x[a(x, X)]$ – **смещение**, показывает, насколько усредненная модель далека от истинной зависимости

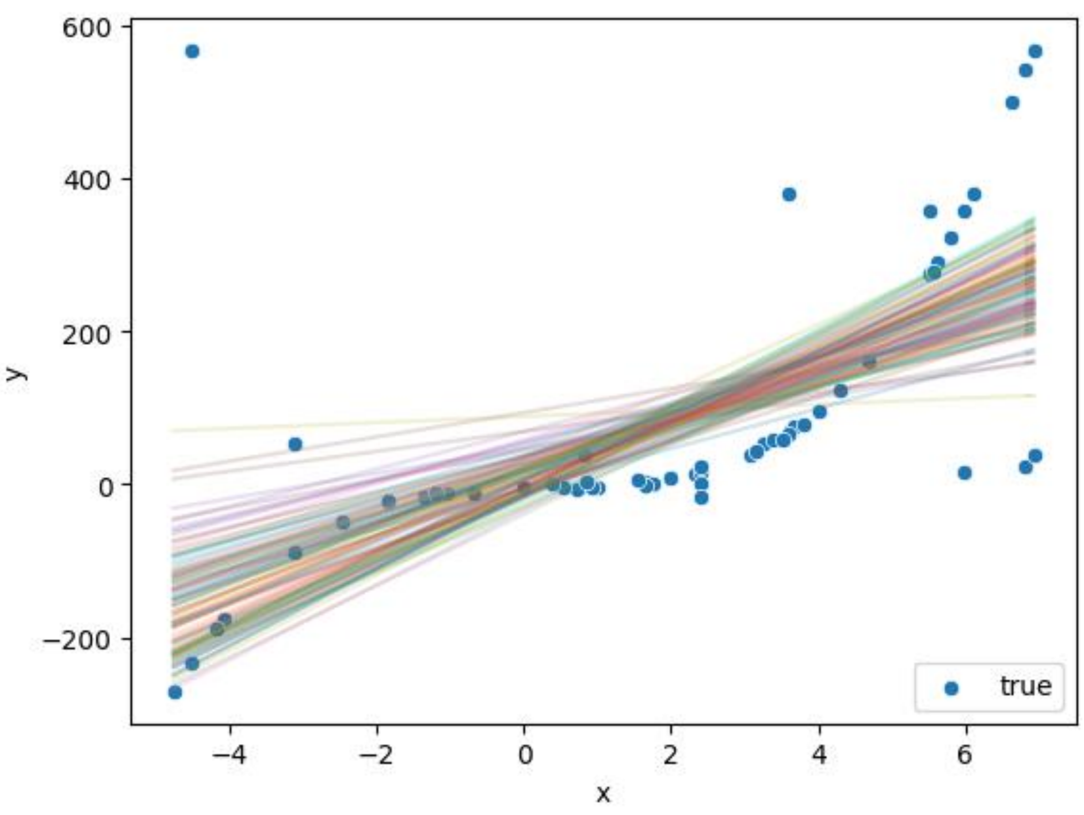
$\mathbb{V}_X[a(x, X)] = \mathbb{E}_X\left[a(x, X) - \mathbb{E}_x[a(x, X)]\right]^2$ – **дисперсия**, показывает, насколько велико отклонение каждой отдельной модели от усредненной

$\sigma^2 = \mathbb{E}_X \mathbb{E}_\epsilon[y(x, \epsilon) - f(x)]^2$ – **неустранимый шум** в данных

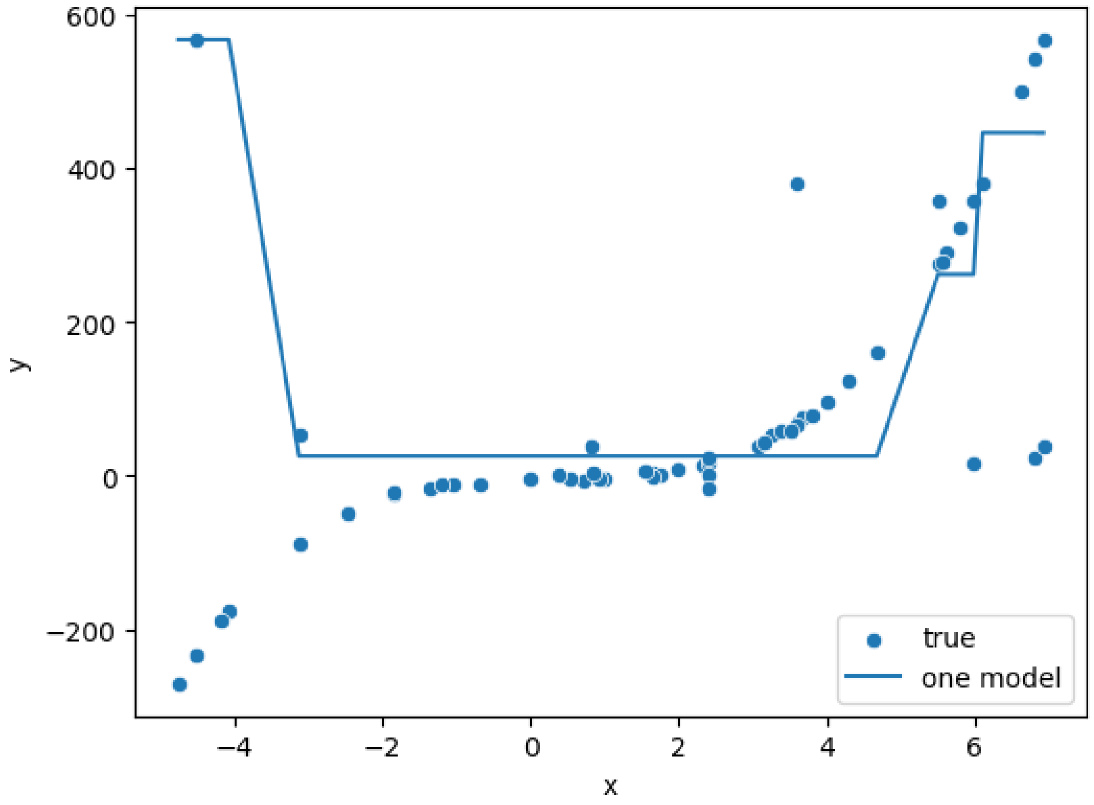
Полный вывод формулы – https://github.com/esokolov/ml-course-hse/blob/master/2020-fall/lecture-notes/lecture08-ensembles.pdf

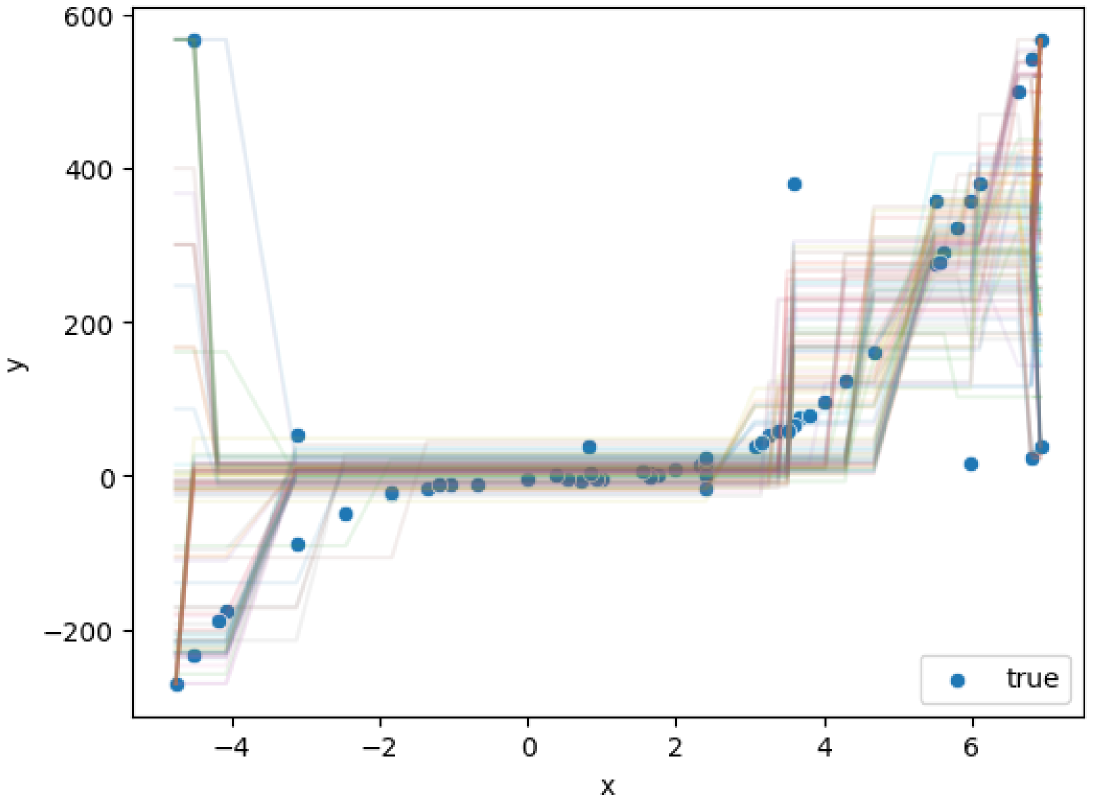
Линейная регрессия



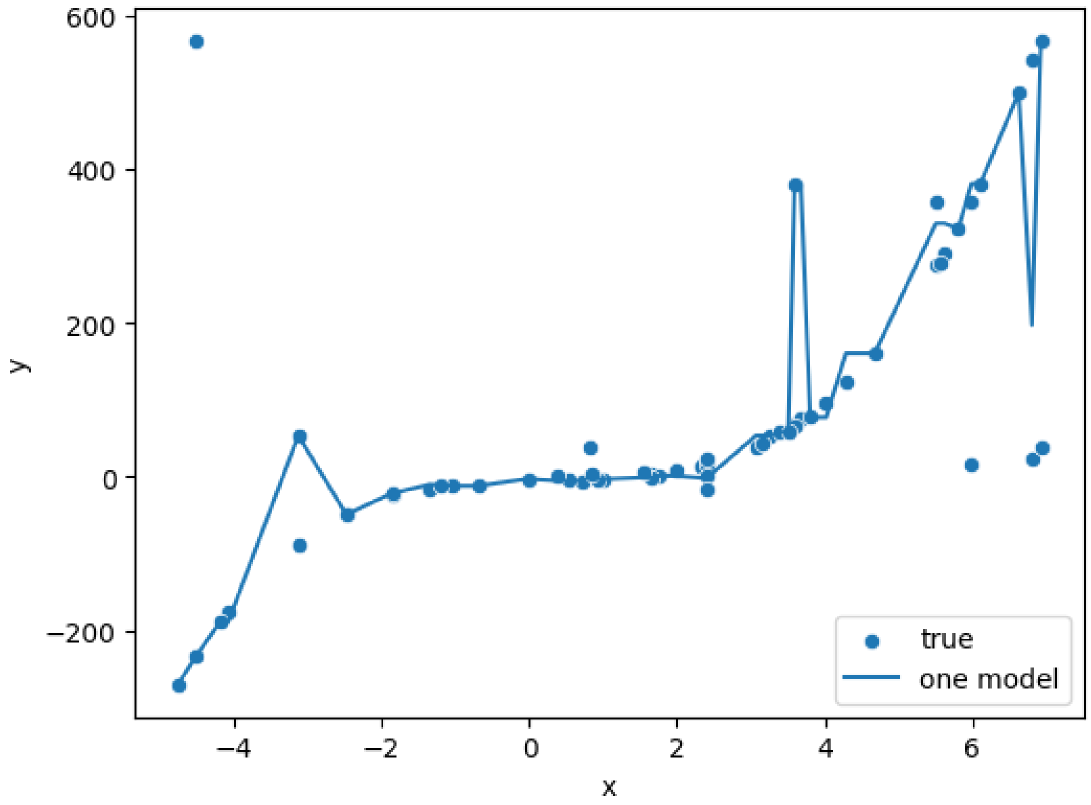


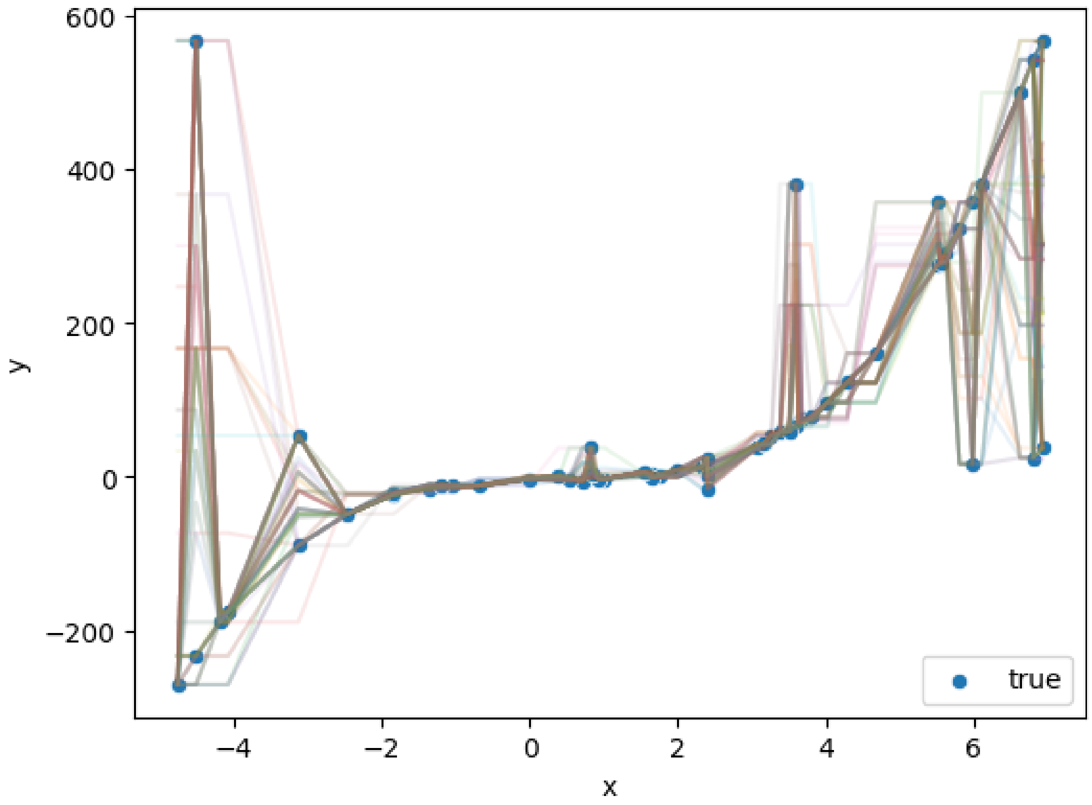
Решающее дерево (глубина 2)



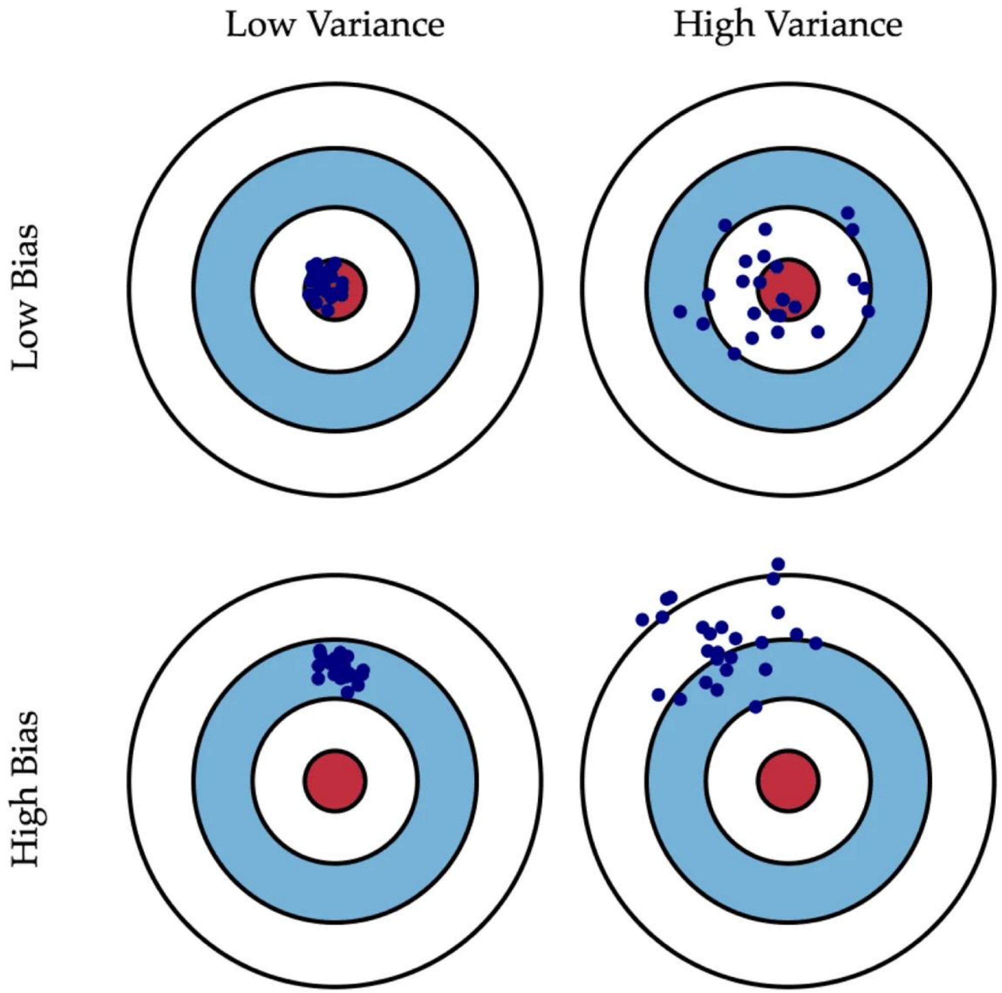


Решающее дерево (глубина None)

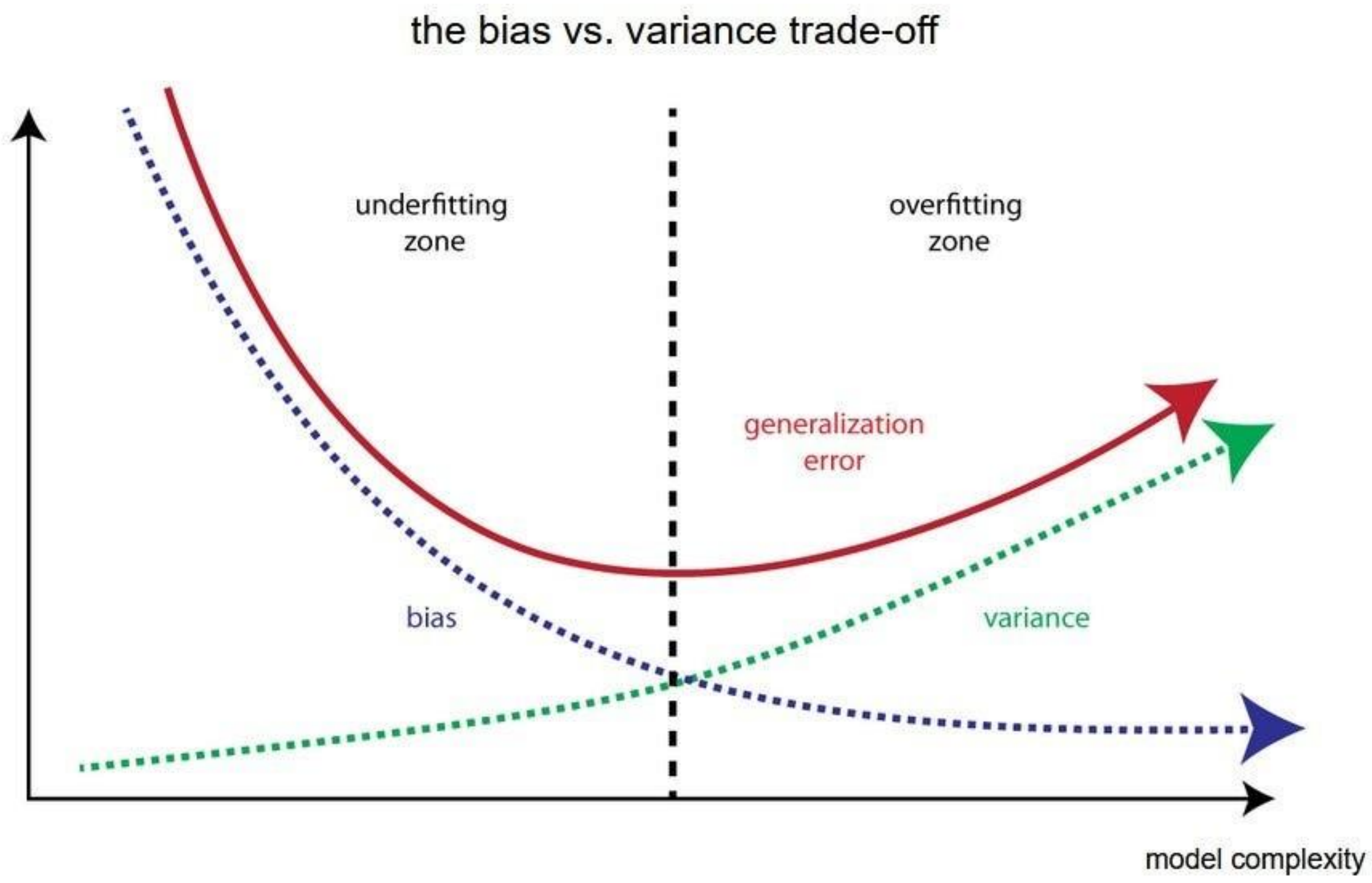




Разложение ошибки (смещение и разброс)

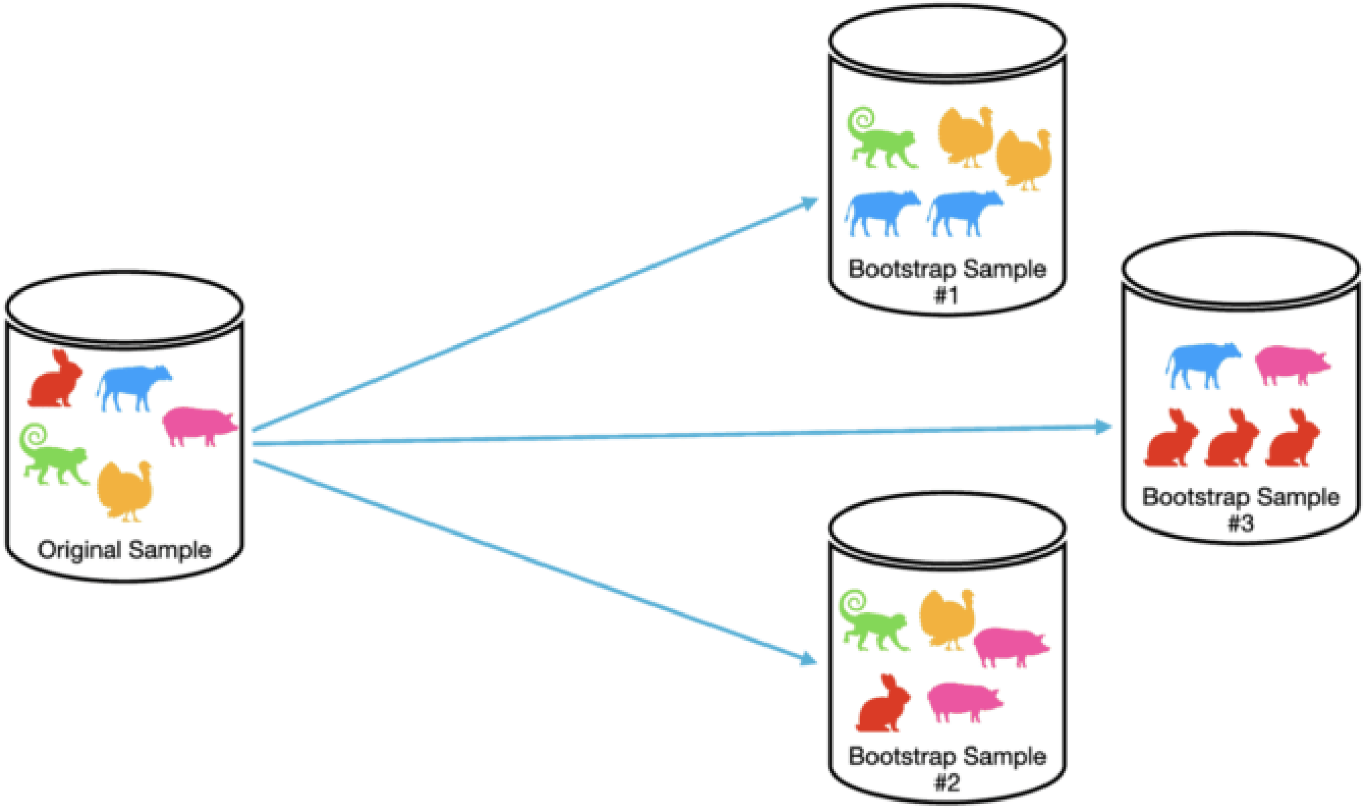


Разложение ошибки (смещение и разброс)



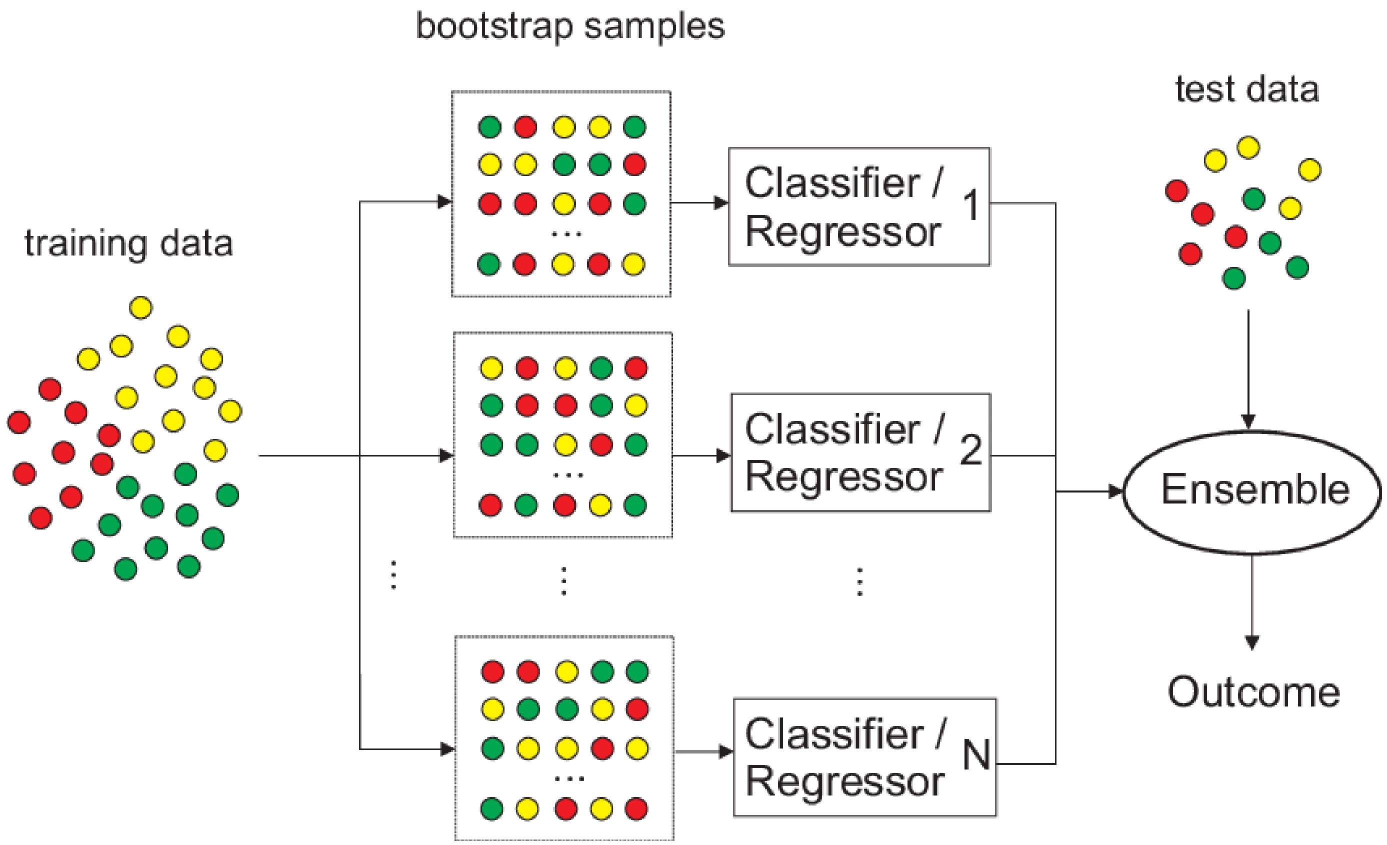
Бутстрап

Метод сэмплирования, при котором подвыборка формируется из основной выборки с возвращением



Случайный лес. Бэггинг

Бэггинг (bagging = bootstrap aggregating) – обучение моделей при помощи сэмплирования с возвращением.



Влияние на ошибку при использовании бэггинга

Результат работы ансамбля моделей:

$$a(x) = \frac{1}{n}\sum_{j=1}^{n} b_j(x)$$

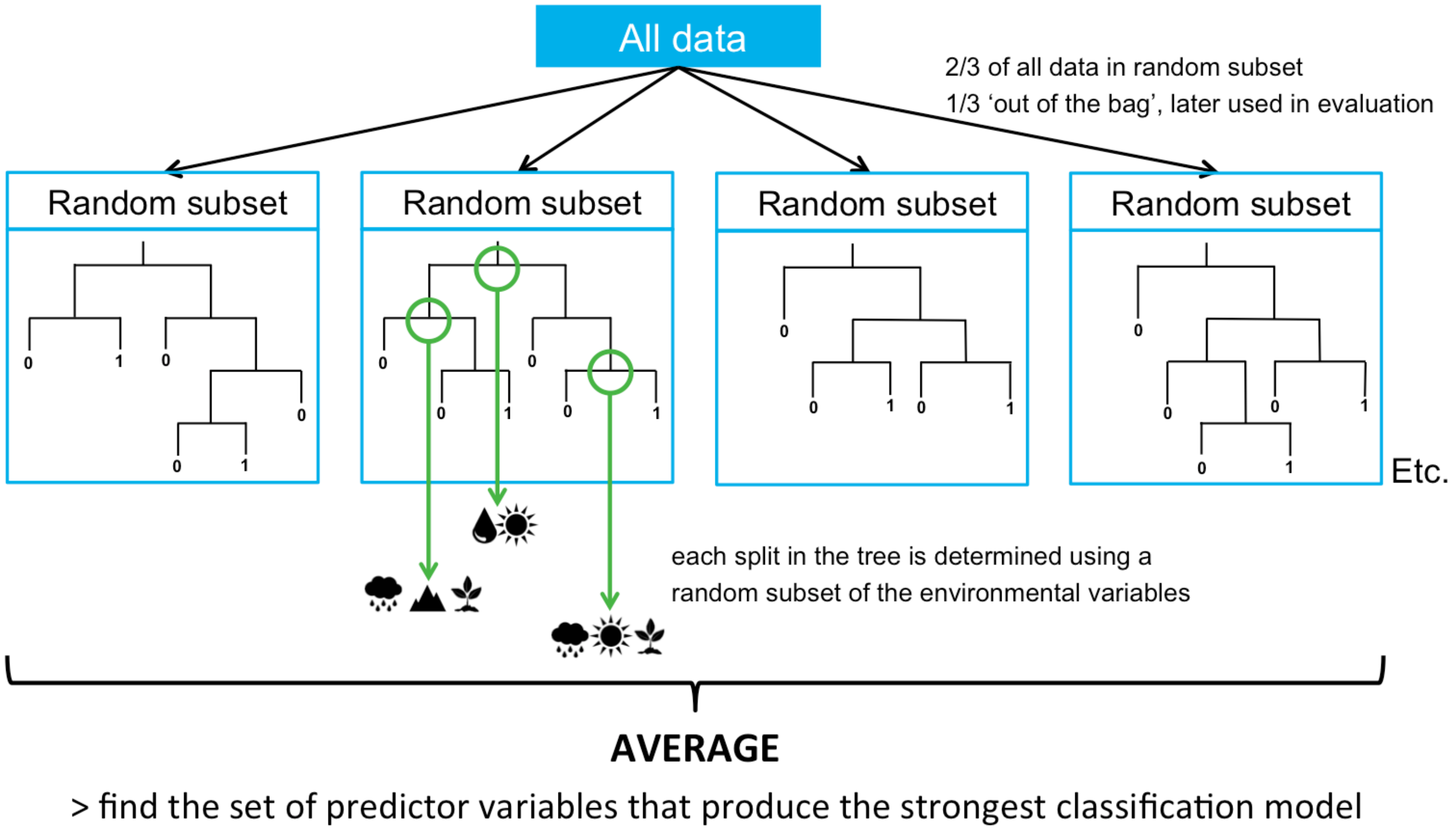
Смещение (bias) при использовании ансамбля моделей **не уменьшается**

Разброс (variance) при использовании ансамбля моделей **уменьшается**

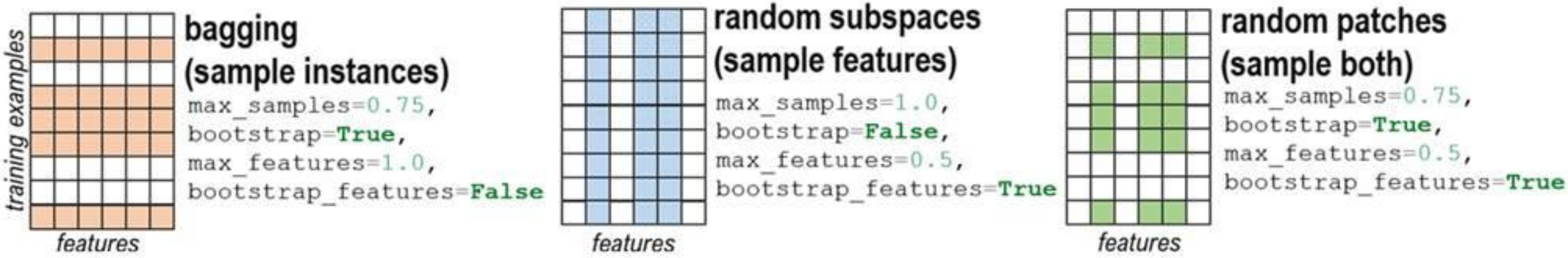
Поэтому надо использовать в качестве **базовых моделей** такие, которые бы обладали **низким смещением** (разброс может быть большим)

Случайный лес. Требование №2

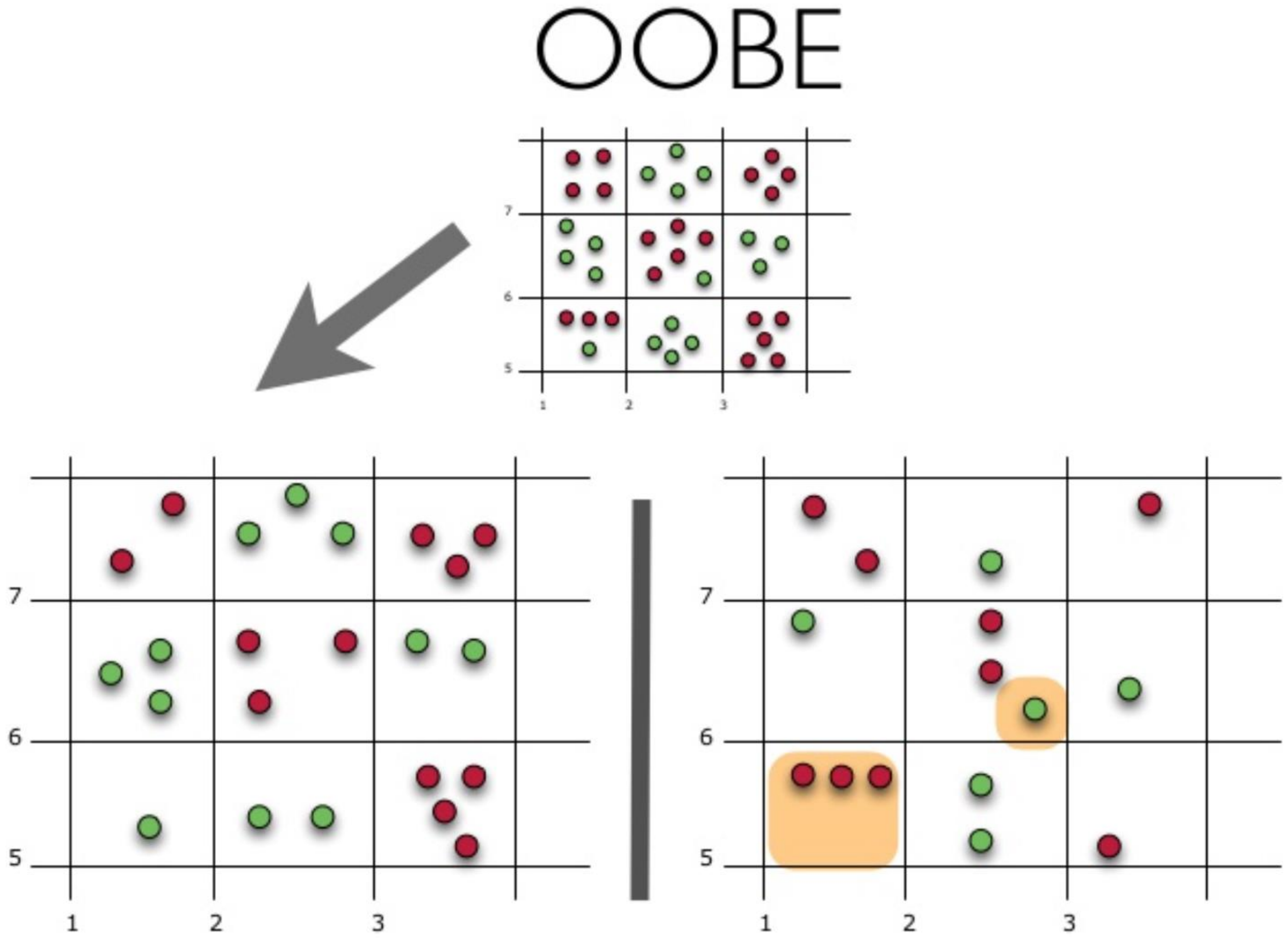
Необходима независимость базовых моделей



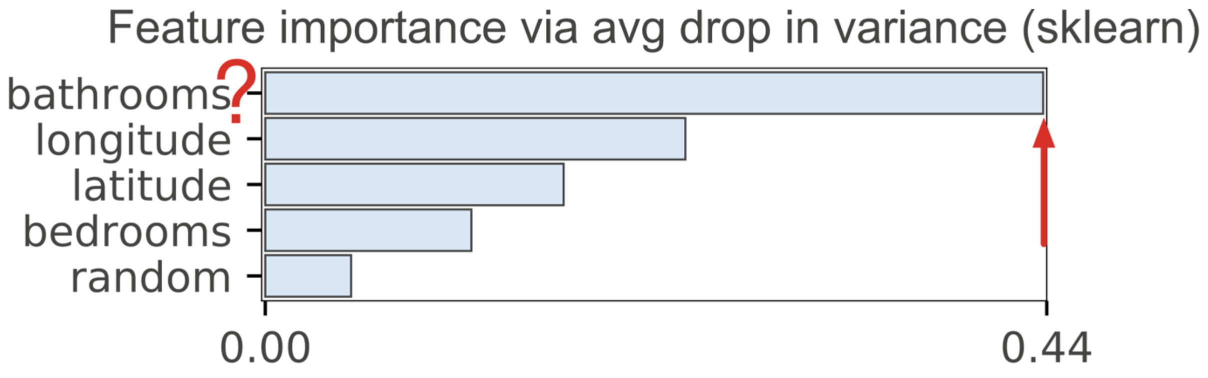
Случайный лес. Требование №2

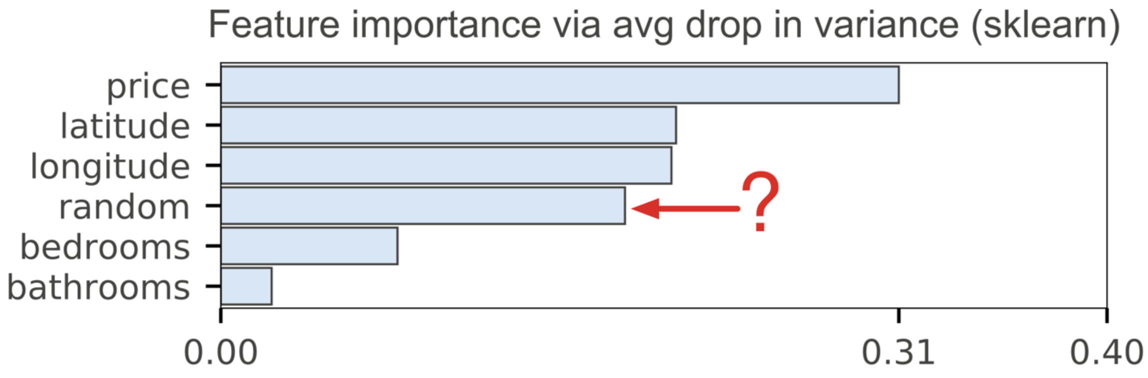


Оценка Out-of-bag



Важность признаков





https://explained.ai/rf-importance/