EleWise проекты

Лекция 5

Дерево решений, бэггинг и случайный лес

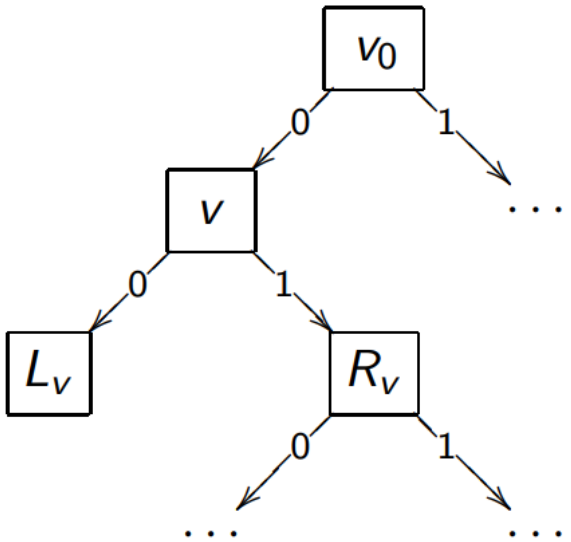
5.1. Дерево решений

5.1. Дерево решений

Бинарное решающее дерево — алгоритм классификации $a(x)$, задающийся бинарным деревом:

1) $\forall v \in V_{\text{внутр}} \;\rightarrow\;$ предикат $\beta_v : X \to \{0,1\}, \;\; \beta \in \mathscr{B}$
2) $\forall v \in V_{\text{лист}} \;\rightarrow\;$ имя класса $c_v \in Y$.

1: $v := v_0$;
2: **пока** $v \in V_{\text{внутр}}$
3: **если** $\beta_v(x) = 1$ **то**
4: переход вправо:
$v := R_v$;
5: **иначе**
6: переход влево:
$v := L_v$;
7: **вернуть** c_v.



5.1. Дерево решений

y	x	z
1	1	-2
1	0	3
2	0	-4
10	0	9
20	1	9

5.1. Дерево решений

Весь обучающий набор данных, называемый **корневым узлом**, и разбивается на два или более **узлов (сегментов)** так, чтобы наблюдения, попавшие в разные узлы, максимально отличались друг от друга по зависимой переменной.

В роли **правил разбиения**, максимизирующих эти различия, выступают значения независимых переменных.

Качество разбиения оценивается с помощью статистических критериев.

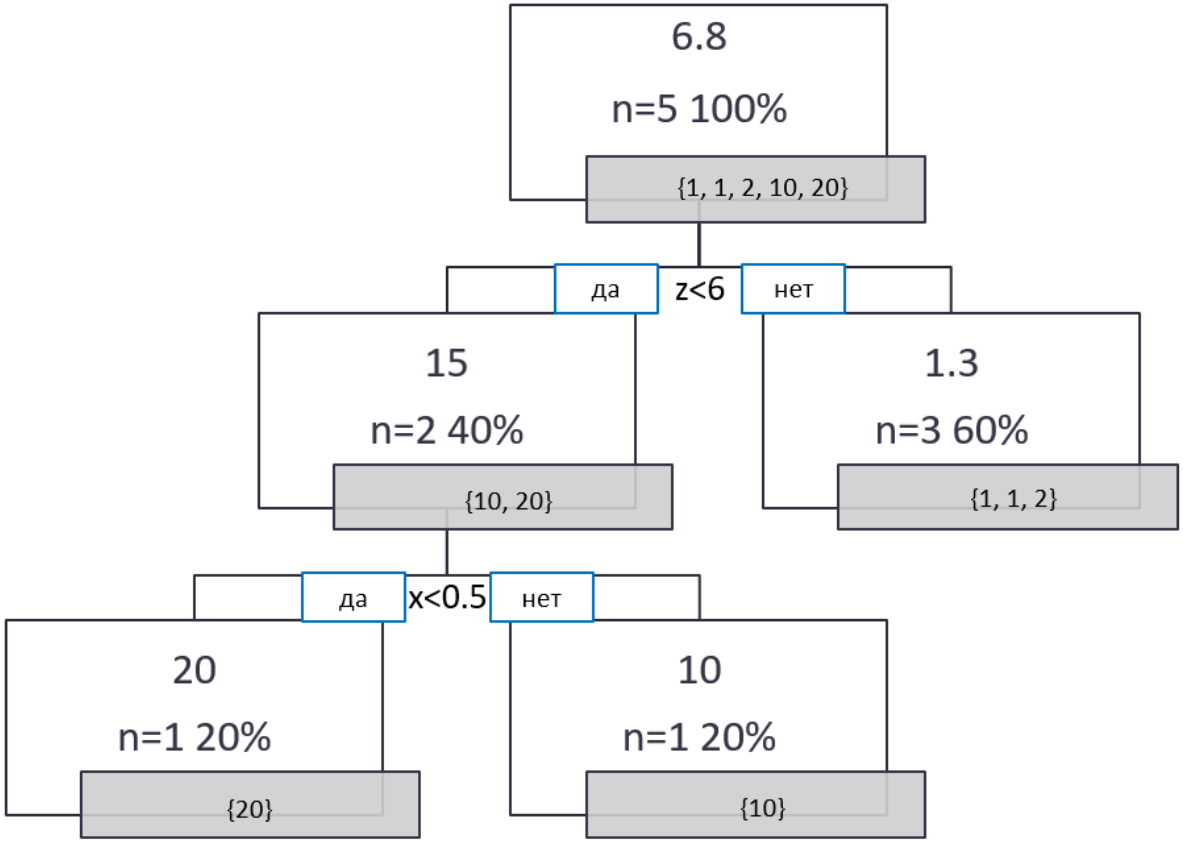
Правила и статистики отмечаются на ветвях.

Для каждого узла вычисляются вероятности в виде **процентных долей категорий** зависимой переменной (является категориальной) или **средние значения** зависимой переменной (является количественной).

В результате выносится **решение** – спрогнозированная категория зависимой переменной (является категориальной) или спрогнозированное среднее значение зависимой переменной (является количественной).

5.1. Дерево решений

y	x	z
1	1	-2
1	0	3
2	0	-4
10	0	9
20	1	9



5.1. Дерево решений

Наилучшее деление

y	x	z
1	1	-2
1	0	3
2	0	-4
10	0	9
20	1	9

- **До деления:**

  $RSS = 274.8 \quad [= MSE \cdot N]$

  $\{1,1,2,10,20\}, \bar{y} = 6.8$

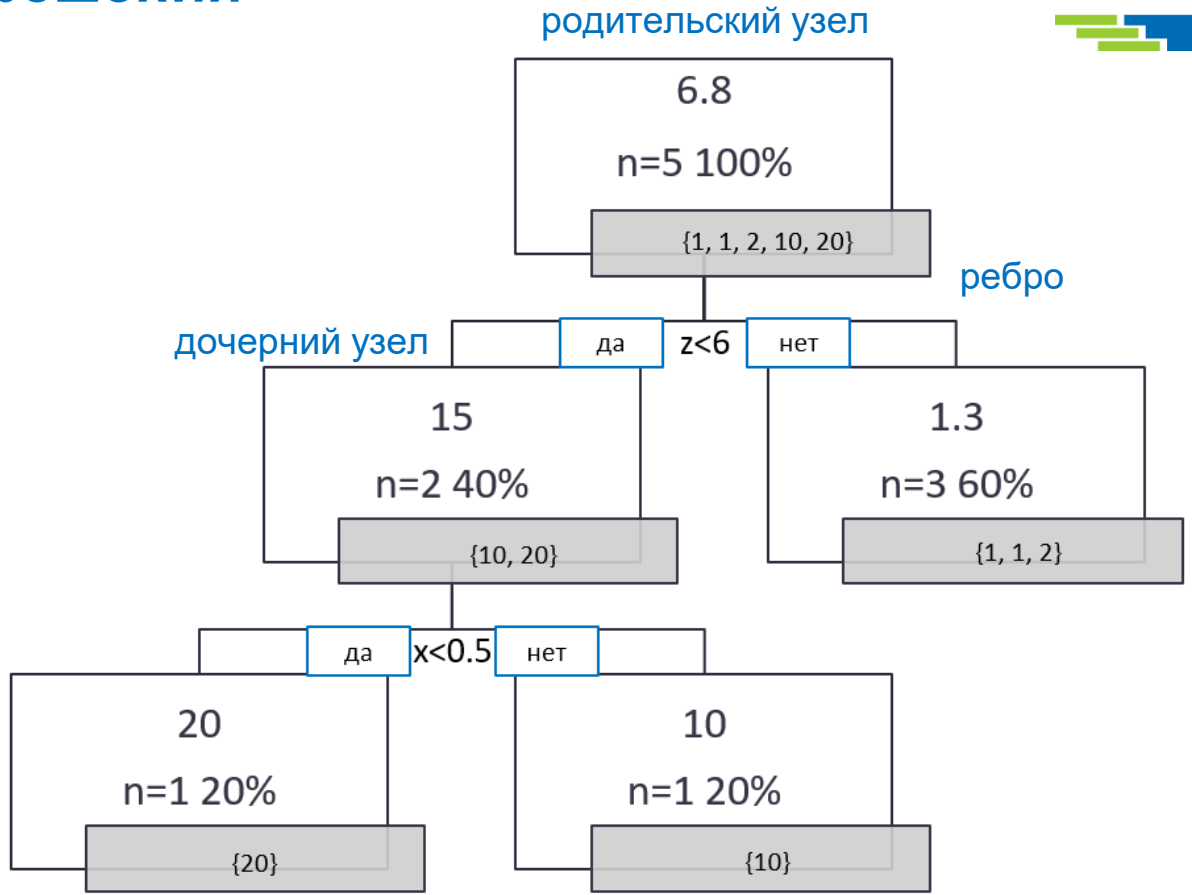
Слева: $RSS_1 = 50, \ \{10,20\}, \bar{y} = 15$

Справа: $RSS_2 = 0.67, \ \{1,1,2\}, \bar{y} = 1.33$

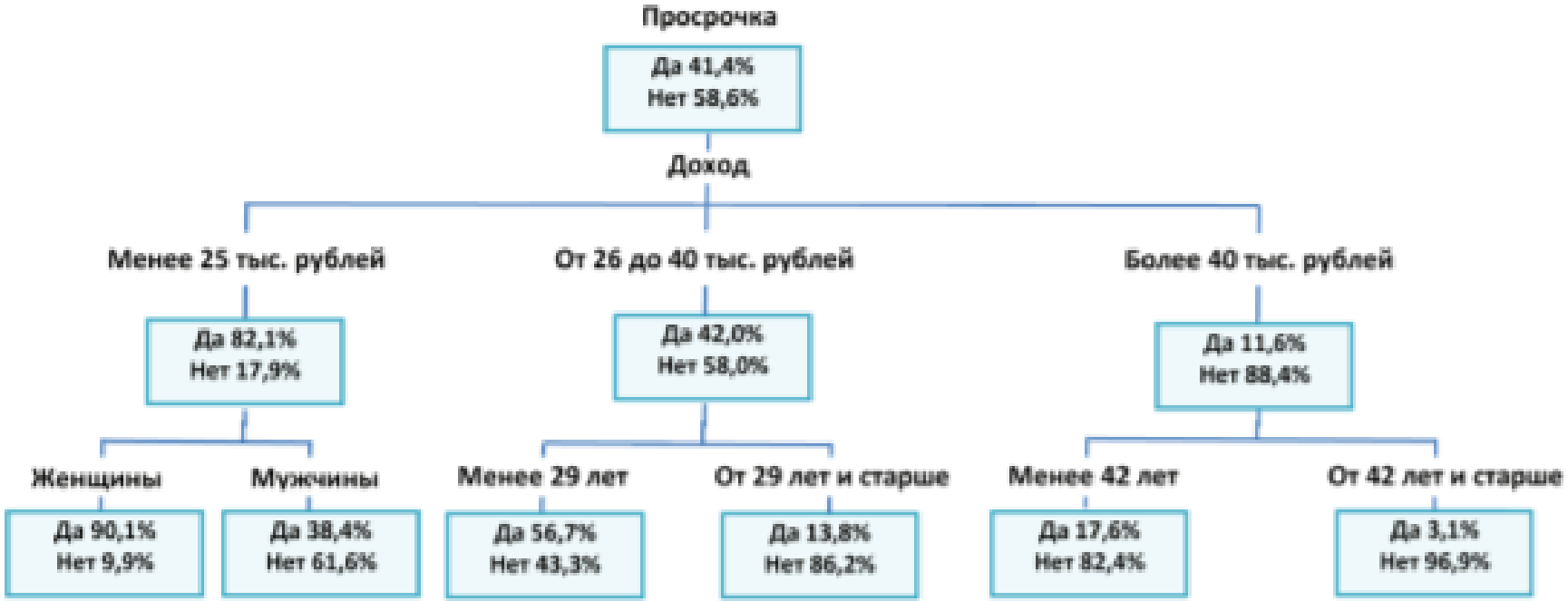
- **После деления:** $RSS = RSS_1 + RSS_2 = 50.67$

5.1. Дерево решений

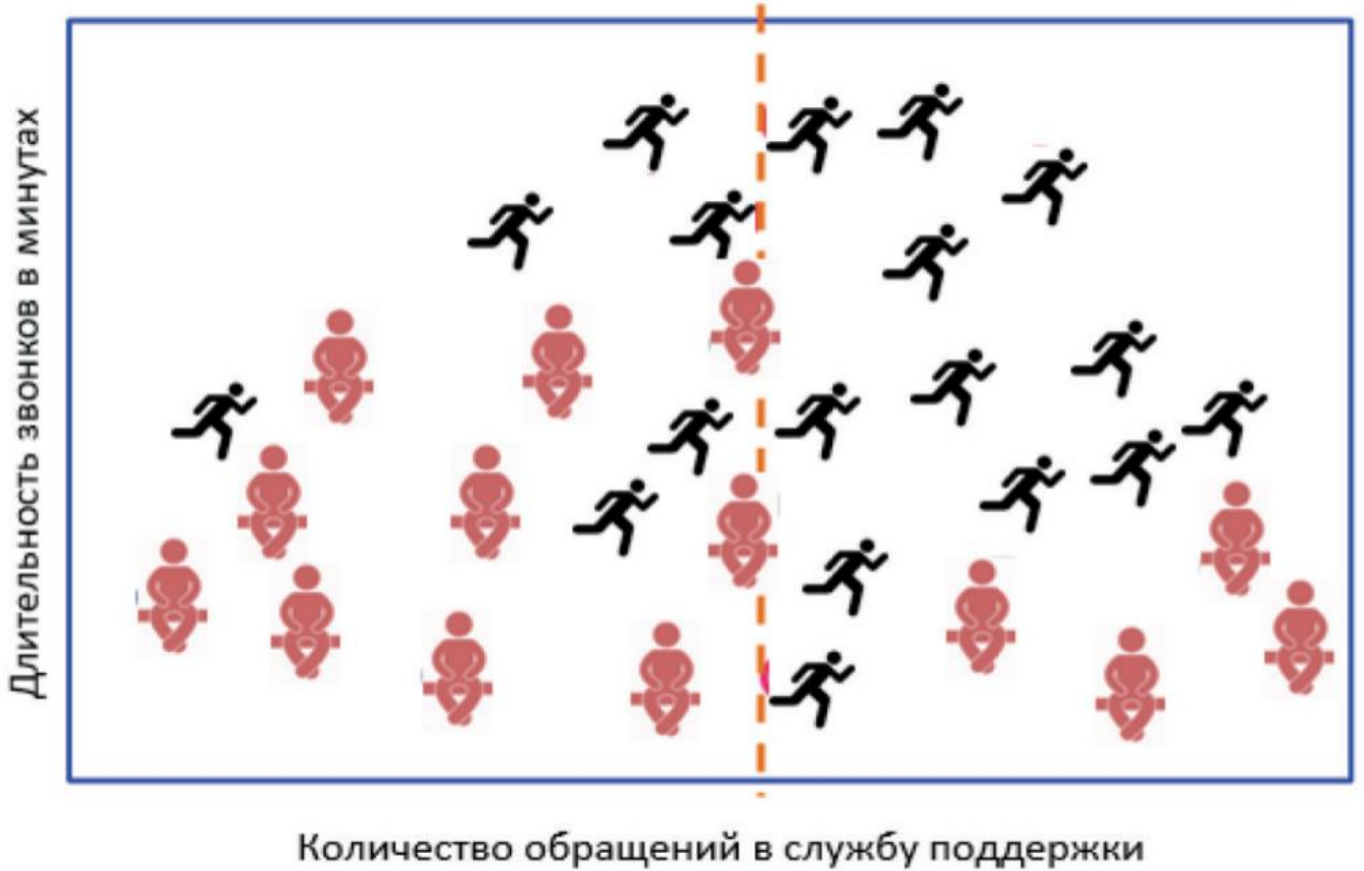
y	x	z
1	1	-2
1	0	3
2	0	-4
10	0	9
20	1	9



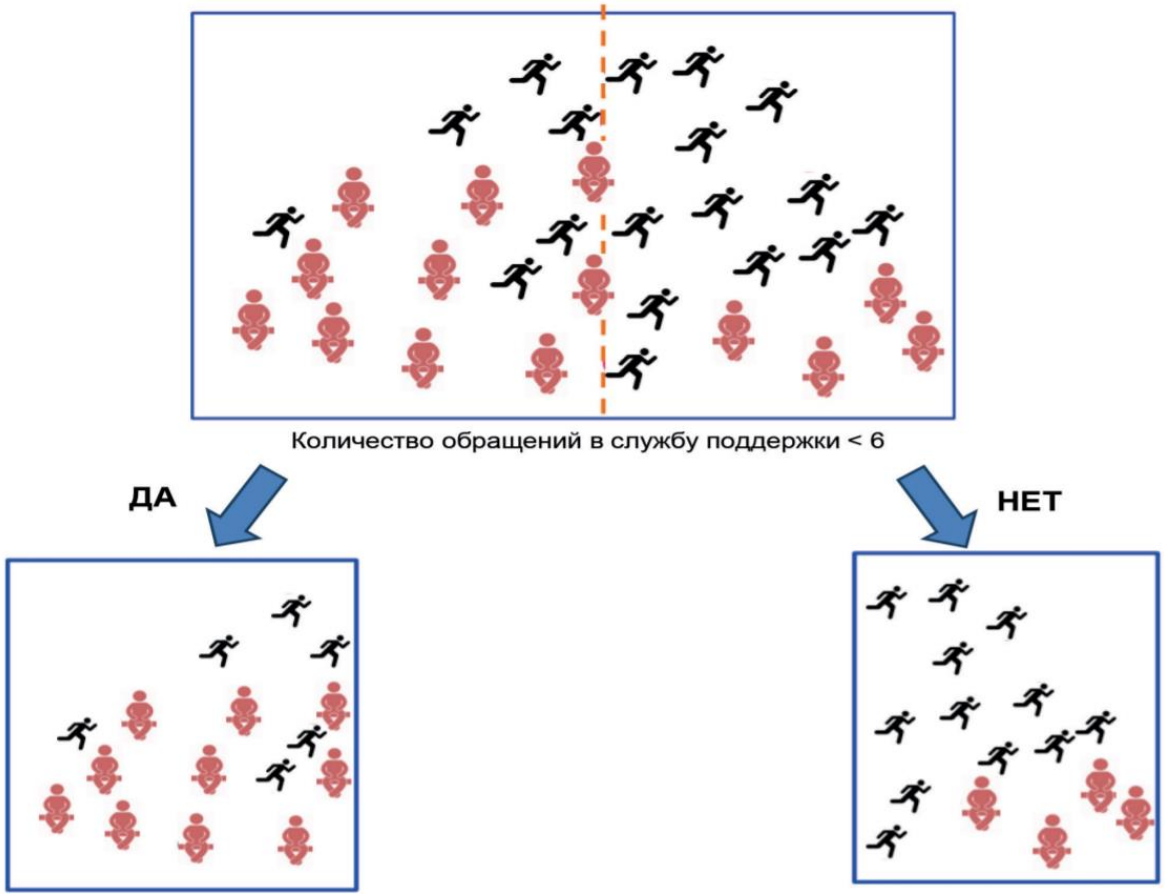
5.1. Дерево решений



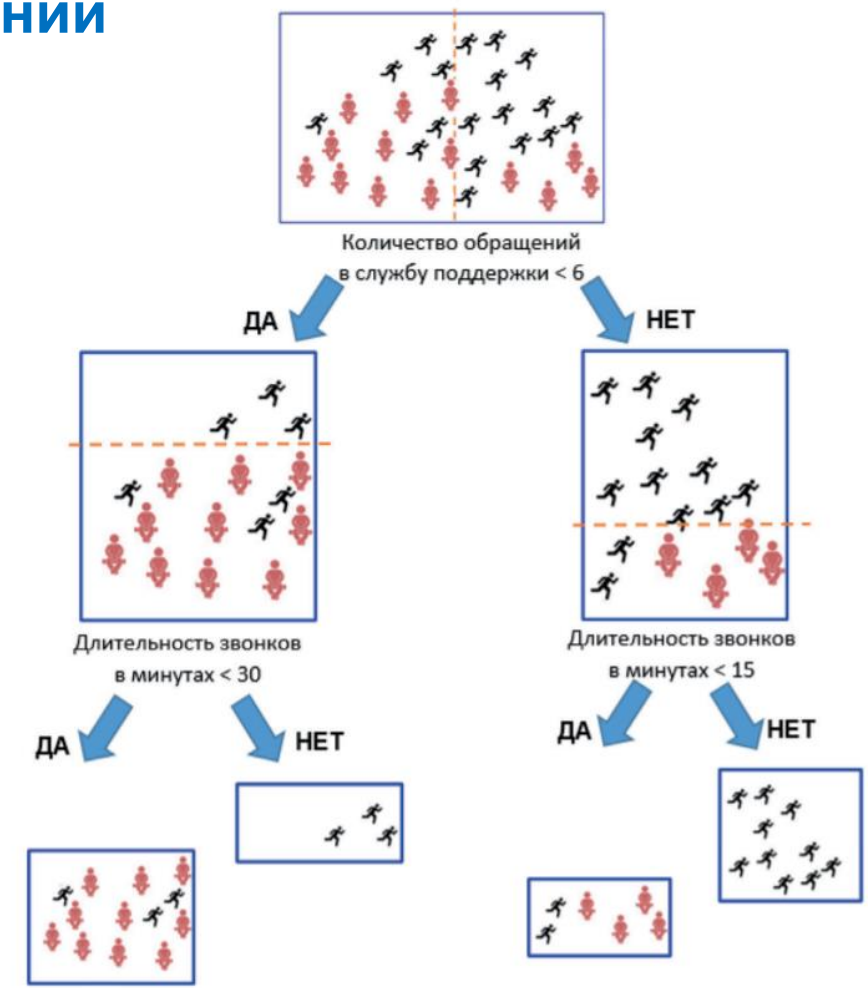
5.1. Дерево решений



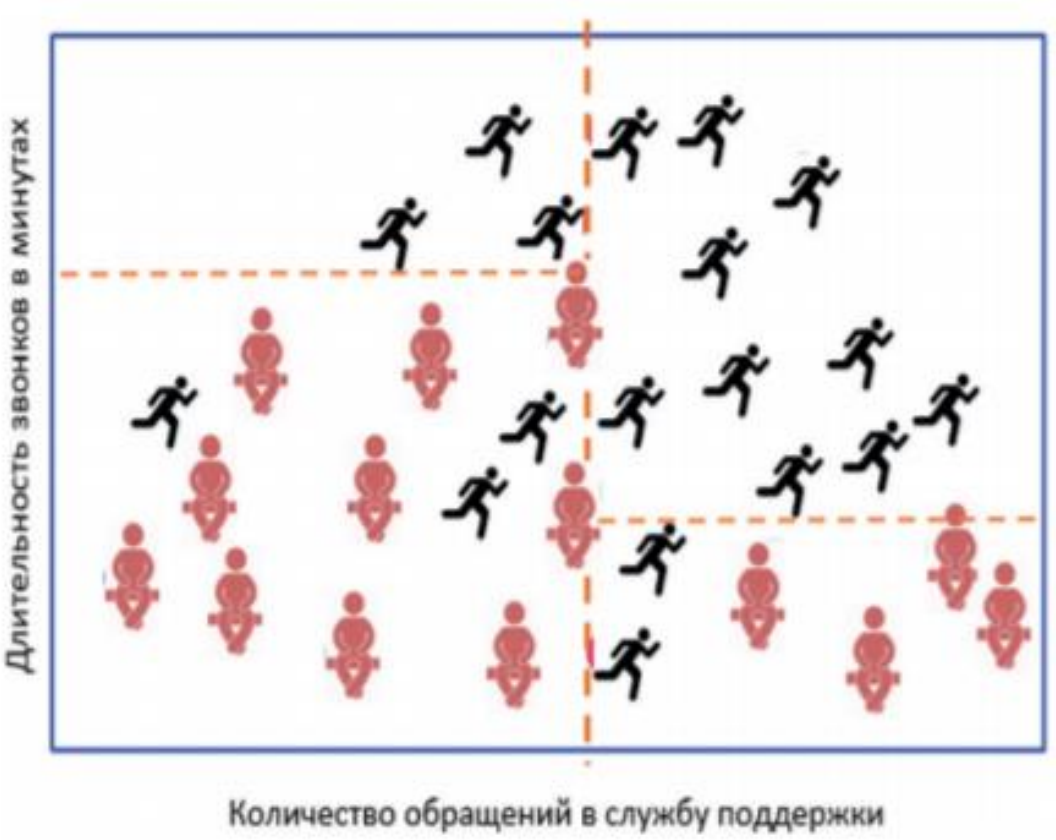
5.1. Дерево решений



5.1. Дерево решений



5.1. Дерево решений



5.1. Дерево решений

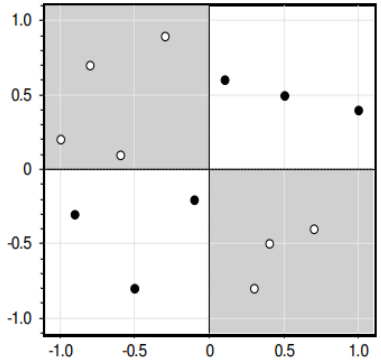
- Из имеющихся **k** переменных случайно отбираем **k0** переменных
- Из отобранных **k0** переменных выбираем ту, которая дает наилучшее деление ветви на две
- Повторяем до тех пор, пока в каждом терминальном узле остается больше **nodesize** наблюдений

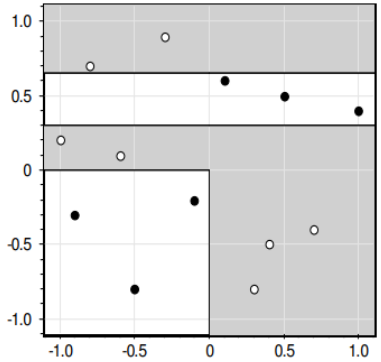
5.1. Дерево решений

Алгоритмы построения дерева принятия решения:

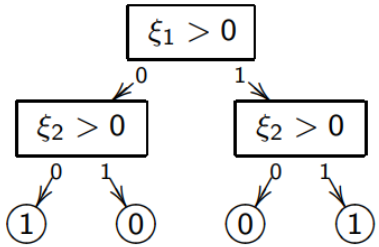
Характеристика метода	CHAID	Exhaustive CHAID	CRT	QUEST
Категориальная зависимая переменная	Да	Да	Да	Да, только номинальная
Категориальные предикторы	Да	Да	Да	Да
Количественная зависимая переменная	Да	Да	Да	Нет
Количественные предикторы	Да, преобразуются в порядковые	Да, преобразуются в порядковые	Да	Да
Тип разбиения	Множественный	Множественный	Бинарный	Бинарный
Цены ошибочной классификации (Построение дерева)	Нет	Нет	Да	Да
Статистические тесты (Отбор предикторов)	Да	Да	Нет	Да
Статистические тесты (Разбиение)	Да	Да	Нет	Нет
Время вычислений	Умеренное	Умеренное	Большое	Умеренное/Большое
Использование априорных вероятностей	Нет	Нет	Да	Да
Пропущенные значения в предикторах	Да, как категория	Да, как категория	Нет, для разбиения используется заменитель	Нет, для разбиения используется заменитель

5.1. Дерево решений: ID3

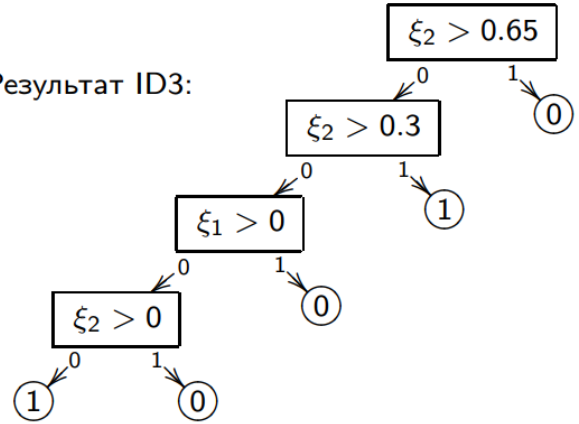




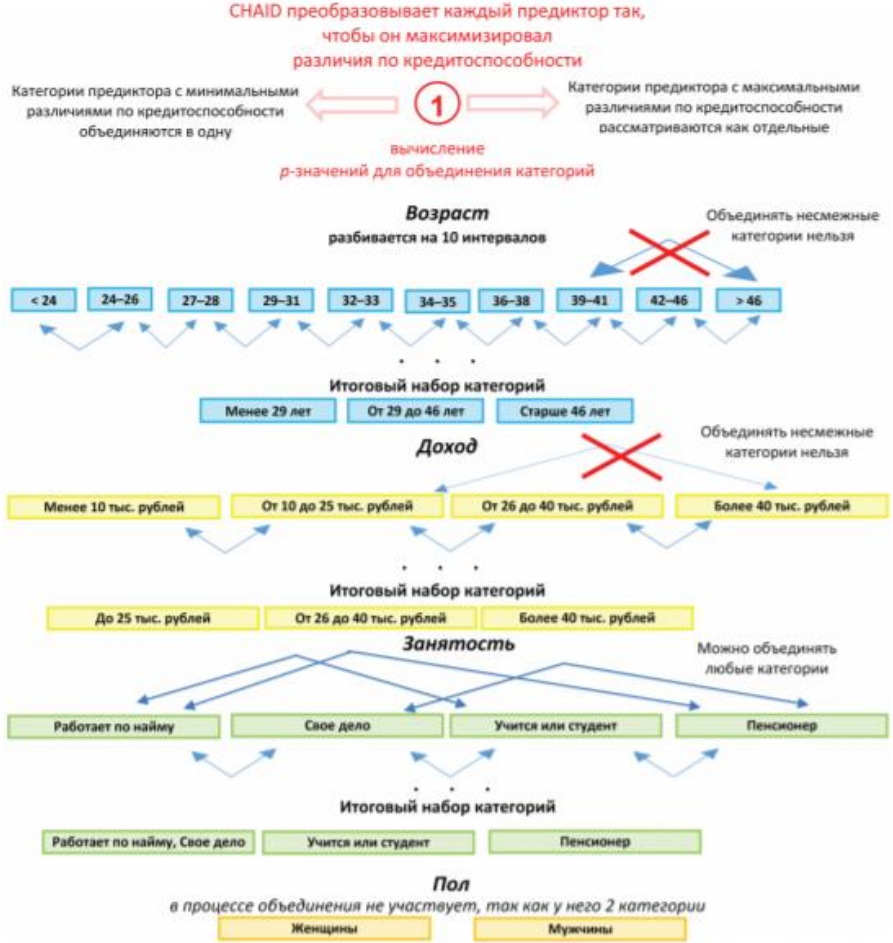
Оптимальное дерево для задачи XOR:



Результат ID3:



5.1. Дерево решений: CHAID

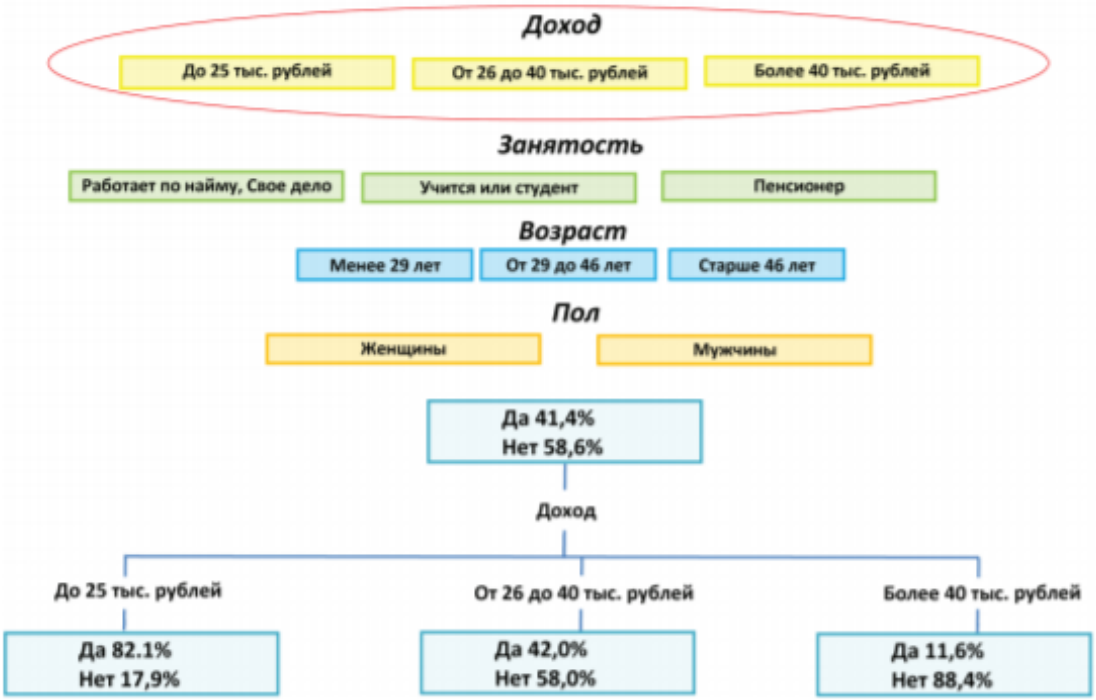


5.1. Дерево решений: CHAID

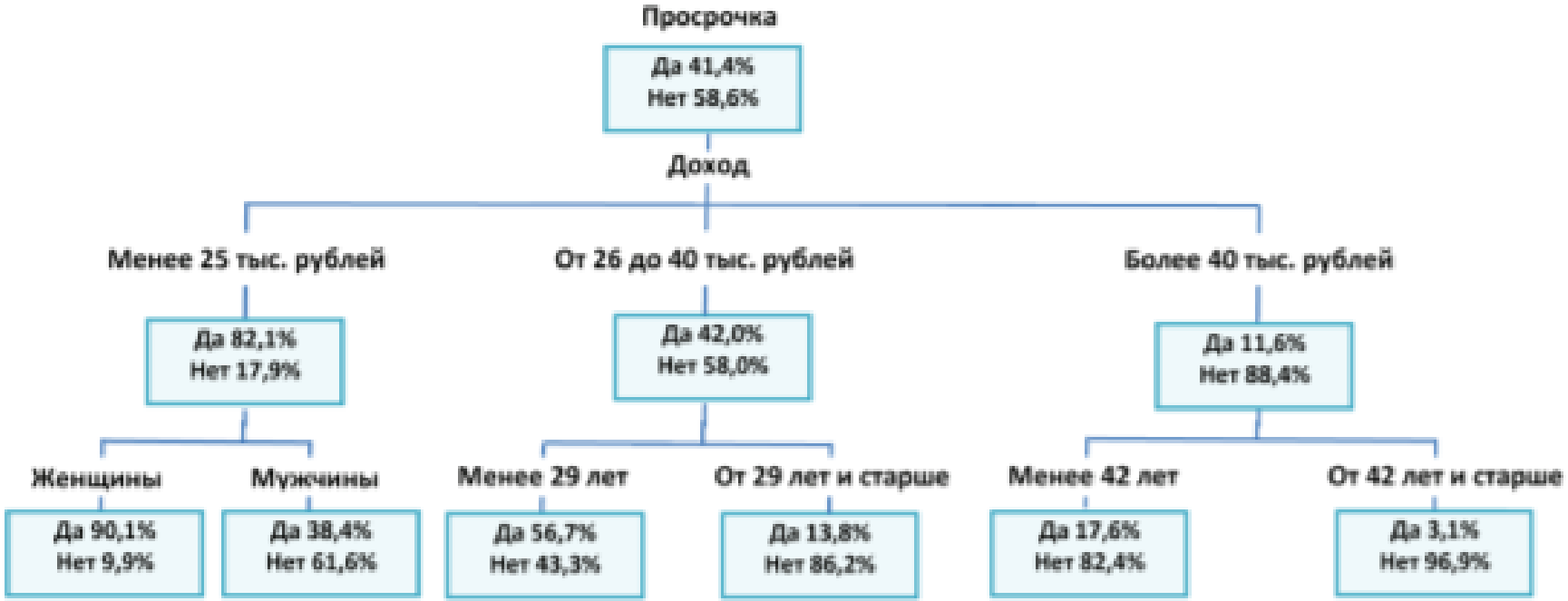
CHAID выбирает предиктор, лучше всего максимизирующий различия по кредитоспособности, из набора преобразованных

2

вычисление скорректированных *p*-значений для разбиения узла



5.1. Дерево решений



5.2. Бэггинг и случайный лес

5.2. Бэггинг и случайный лес

Обучающая выборка: $X^\ell = (x_i, y_i)_{i=1}^\ell, \quad x_i \in X, \quad y_i \in \{-1, +1\}$

Базовые классификаторы: $b_1(x), \ldots, b_T(x), \quad b_t \colon X \to \{-1, +1\}$

Простое голосование базовых классификаторов:

$$a(x) = \text{sign} \sum_{t=1}^{T} b_t(x)$$

Композиция *a(x)* может быть лучше базовых $b_1(x), \ldots, b_T(x)$, если они лучше случайного гадания и достаточно различимы.

Способы повышения различимости:

- обучение по случайным подвыборкам,
- обучение по выборке со случайными весами,
- обучение по случайным подмножествам признаков,
- использование различных моделей классификации,
- использование различных начальных приближений,
- использование рандомизации при обучении b_1, …., b_T

5.2. Бэггинг и случайный лес

Бутстреп (bootstrap)

Исходная выборка	1	2	3	4	5	6	7	8	9	10

Бутстреп-выборка должна иметь тот же самый размер, что и исходная выборка

Бутстреп-выборка I	10	9	7	8	1	3	9	10	10	7
Бутстреп-выборка II	4	8	5	8	3	9	2	6	1	6
Бутстреп-выборка III	6	2	6	10	2	10	3	6	5	1
Бутстреп-выборка IV	6	7	8	10	6	10	9	10	8	2
Бутстреп-выборка V	5	8	1	8	5	7	10	1	10	9

Поскольку отбор с возвращением, одно и то же наблюдение может попасть в бутстреп-выборку несколько раз

5.2. Бэггинг и случайный лес

Бэггинг (bagging, bootstrap aggregating): $b_t(x)$ обучаются независимо по случайным подвыборкам длины ℓ с повторениями (как в методе bootstrap), доля объектов, попадающих в выборку: $(1 - \frac{1}{e}) \approx 0.632$

Метод случайных подпространств: $b_t(x)$ обучаются по случайным подмножествам n' признаков.

Совместим обе идеи в одном алгоритме.

$\mathscr{F} = \{f_1, \ldots, f_n\}$ - признаки,

$\mu(\mathscr{G}, U)$ - метод обучения алгоритма по подвыборке $U \subseteq X^\ell$, использующий только признаки из $\mathscr{G} \subseteq \mathscr{F}$.

5.2. Бэггинг и случайный лес

Вывод: обучающая выборка X^ℓ; параметры: T

ℓ' - длина обучающих подвыборок;

n' - длина признакового подописания;

ε_1 - порог качества базовых алгоритмов;

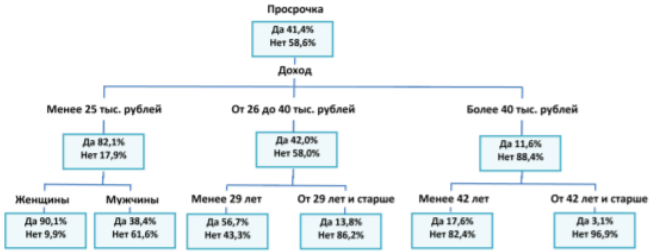
ε_2 - порог качества базовых алгоритмов на контроле.

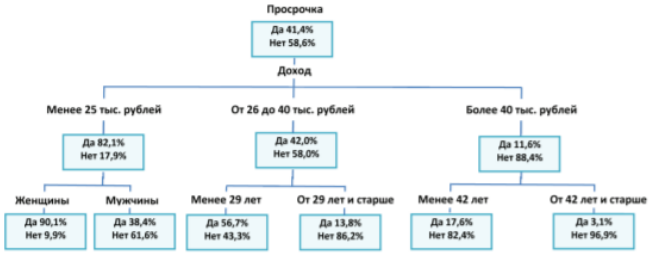
Вывод: базовые алгоритмы b_t $t = 1, \ldots, T$;

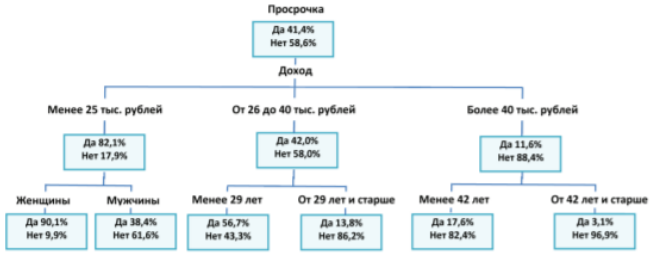
1: **для всех** $t = 1, \ldots, T$
2: $U_t :=$ случайное подмножество X^ℓ длины ℓ';
3: $\mathscr{G}_t :=$ случайное подмножество $\mathscr{F}$ длины n';
4: $b_t := \mu(\mathscr{G}_t, U_t)$;
5: **если** $Q(b_t, U_t) > \varepsilon_1$ или $Q(b_t, X^\ell \setminus U_t) > \varepsilon_2$ **то**
6: не включать b_t в композицию;

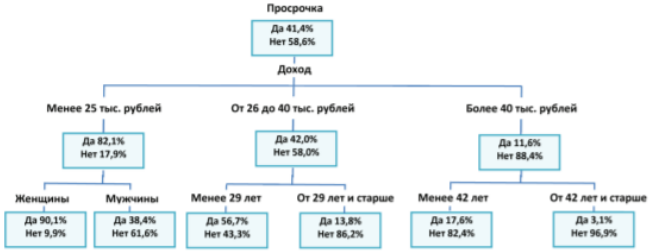
Композиция - простое голосование: $a(x) = \text{sign} \sum_{t=1}^{T} b_t(x)$.

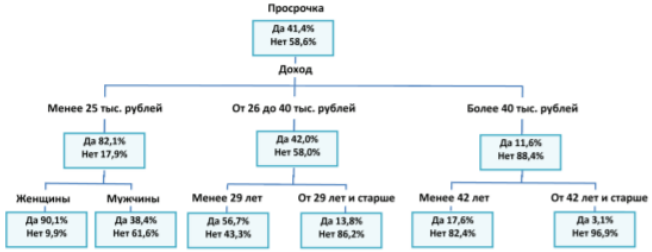
5.2. Бэггинг и случайный лес











5.2. Бэггинг и случайный лес.

Обучение случайного леса:

- бэггинг над решающими деревьями
- усечение дерева (*pruning*) не производится
- признак в каждой вершине дерева выбирается из случайного подмножества k из n признаков
- для регрессии рекомендуется $k=[n/3]$
- для классификации рекомендуется $k=[\sqrt{n}]$

Подбор числа деревьев T по критерию out-of-bag: число ошибок на объектах x_i, если не учитывать голоса деревьев, для которых x_i был обучающим:

$$\text{out-of-bag}(a) = \sum_{i=1}^{\ell} \left[\text{sign}\left(\sum_{t=1}^{T} [x_i \notin U_t] b_t(x_i) \right) \neq y_i \right] \to \min$$

Это несмещенная оценка обобщающей способности.

5.2. Бэггинг и случайный лес

Алгоритм:

- случайным образом отбираем (с повторениями) **n** наблюдений из исходных **n** наблюдений (**бутсреп**)
- по каждой **случайной** подвыборке строим дерево решений
- повторяем до получения **n_tree** деревьев

Прогноз:

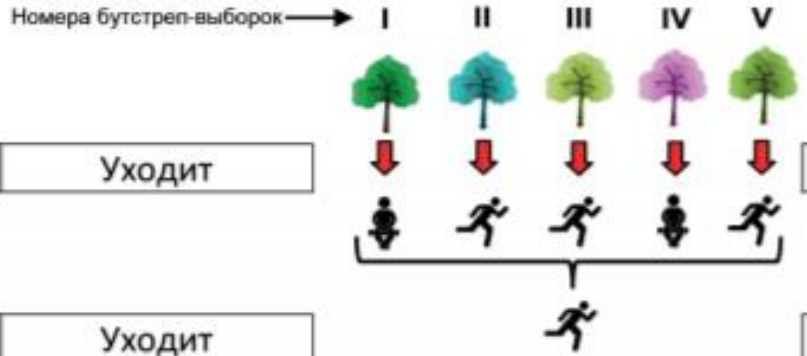
- каждое из **n_tree** деревьев дает прогноз **y'**
- усреднением получаем финальный результат

или

- голосованием по большинству получаем финальный результат

5.2. Бэггинг и случайный лес

№	Фактическое значение
1	Остается
2	Уходит
3	Уходит
4	Уходит
5	Остается
6	Уходит
7	Уходит
8	Остается
9	Остается
10	Уходит



Спрогнозированное значение (итог усреднения прогноза деревьев, построенных по всем бутстреп-выборкам)
Уходит
Уходит
Уходит
Уходит
Остается
Уходит
Уходит
Уходит
Остается
Остается

5.2. Бэггинг и случайный лес

Out of Bag

Исходная выборка	1	2	3	4	5	6	7	8	9	10
Бутстреп-выборка I	10	9	7	8	1	3	9	10	10	7
Бутстреп-выборка II	4	8	5	8	3	9	2	6	1	6
Бутстреп-выборка III	6	2	6	10	2	10	3	6	5	1
Бутстреп-выборка IV	6	7	8	10	6	10	9	10	8	2
Бутстреп-выборка V	5	8	1	8	5	7	10	1	10	9
Номера out-of-bag выборок	IV	I V	IV V	I III IV V	I IV	I V	II III	III	III	II

Для наблюдения 4 используются бутстреп-выборки I, III, IV, V, поскольку в них наблюдение 4 отсутствует

5.2. Бэггинг и случайный лес

Out of Bag

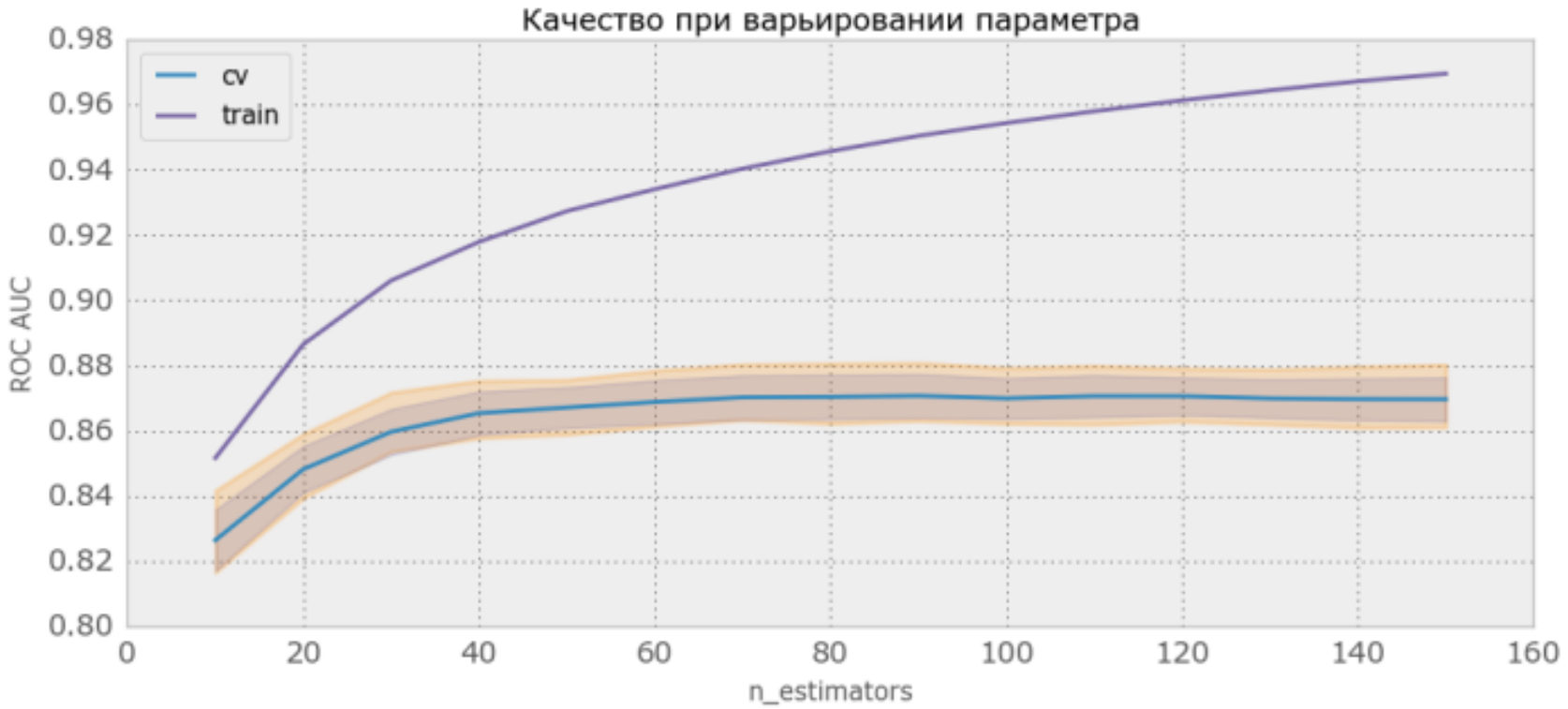
№	Номера out-of-bag-выборок, участвующих в голосовании	Фактическое значение	Спрогнозированное значение (итог голосования деревьев, построенных по out-bag-выборкам)	Результат классификации
1	IV	Остается	Уходит	НЕВЕРНО
2	I, V	Уходит	Уходит	ВЕРНО
3	IV, V	Уходит	Уходит	ВЕРНО
4	I, III, IV, V	Уходит	Уходит	ВЕРНО
5	I, IV	Остается	Уходит	НЕВЕРНО
6	I, V	Уходит	Остается	НЕВЕРНО
7	II, III	Уходит	Остается	НЕВЕРНО
8	III	Остается	Уходит	НЕВЕРНО
9	III	Остается	Остается	ВЕРНО
10	II	Уходит	Остается	НЕВЕРНО

Количество неверных ответов = 6

Ошибка классификации = количество неверно классифицированных наблюдений/общее количество наблюдений = 6/10 = 0,6

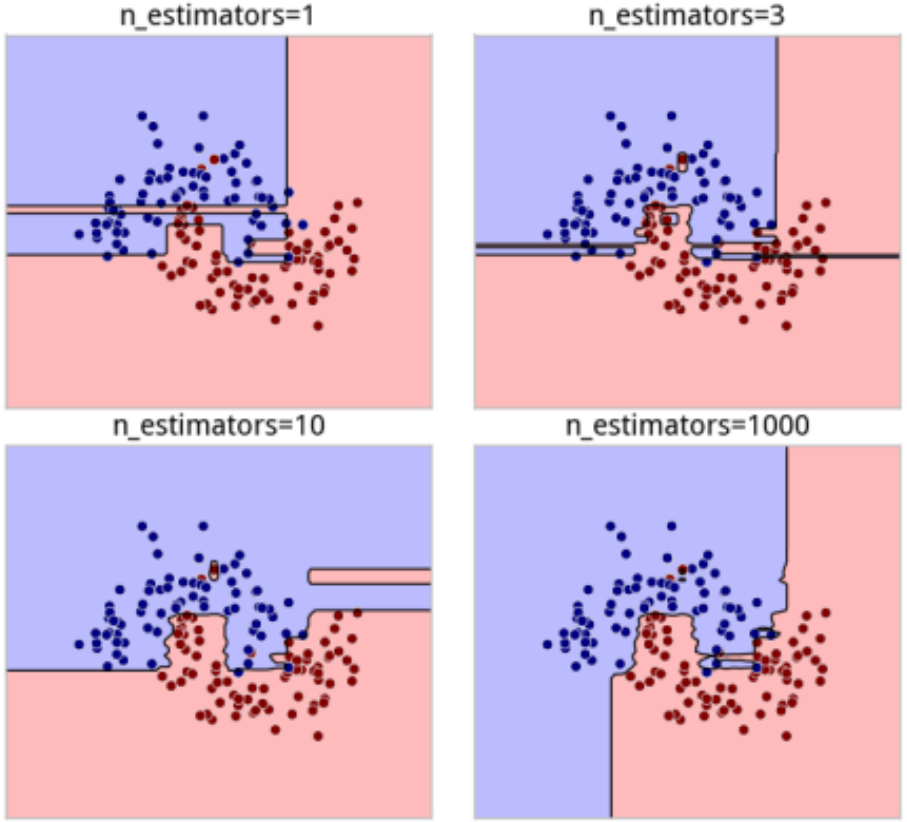
5.2. Бэггинг и случайный лес

Число деревьев



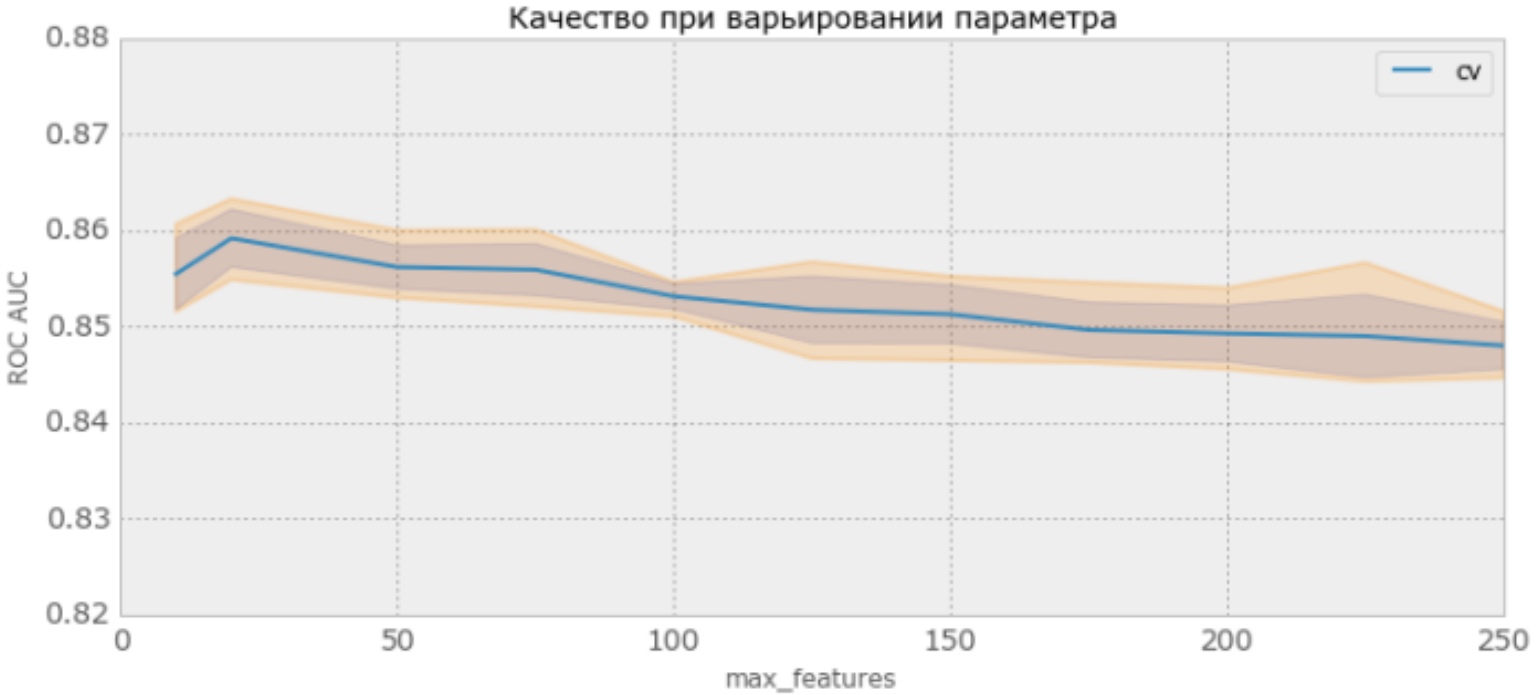
5.2. Бэггинг и случайный лес

Число деревьев



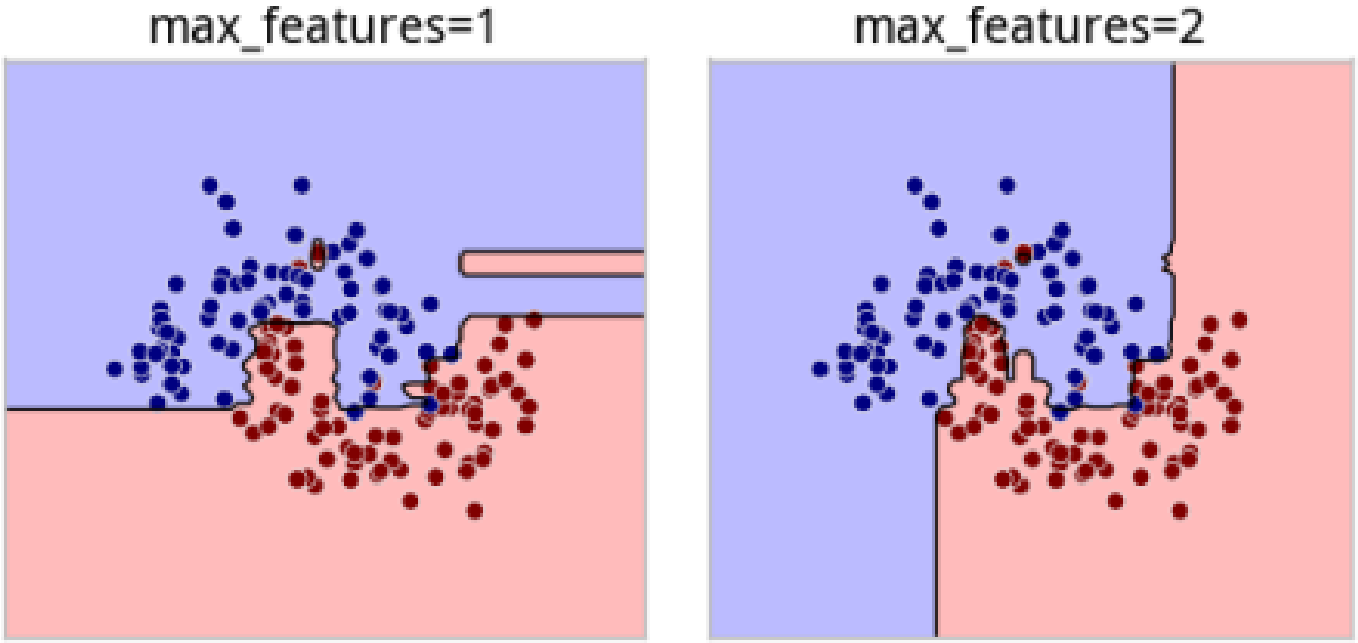
5.2. Бэггинг и случайный лес

Число признаков для выбора расщепления



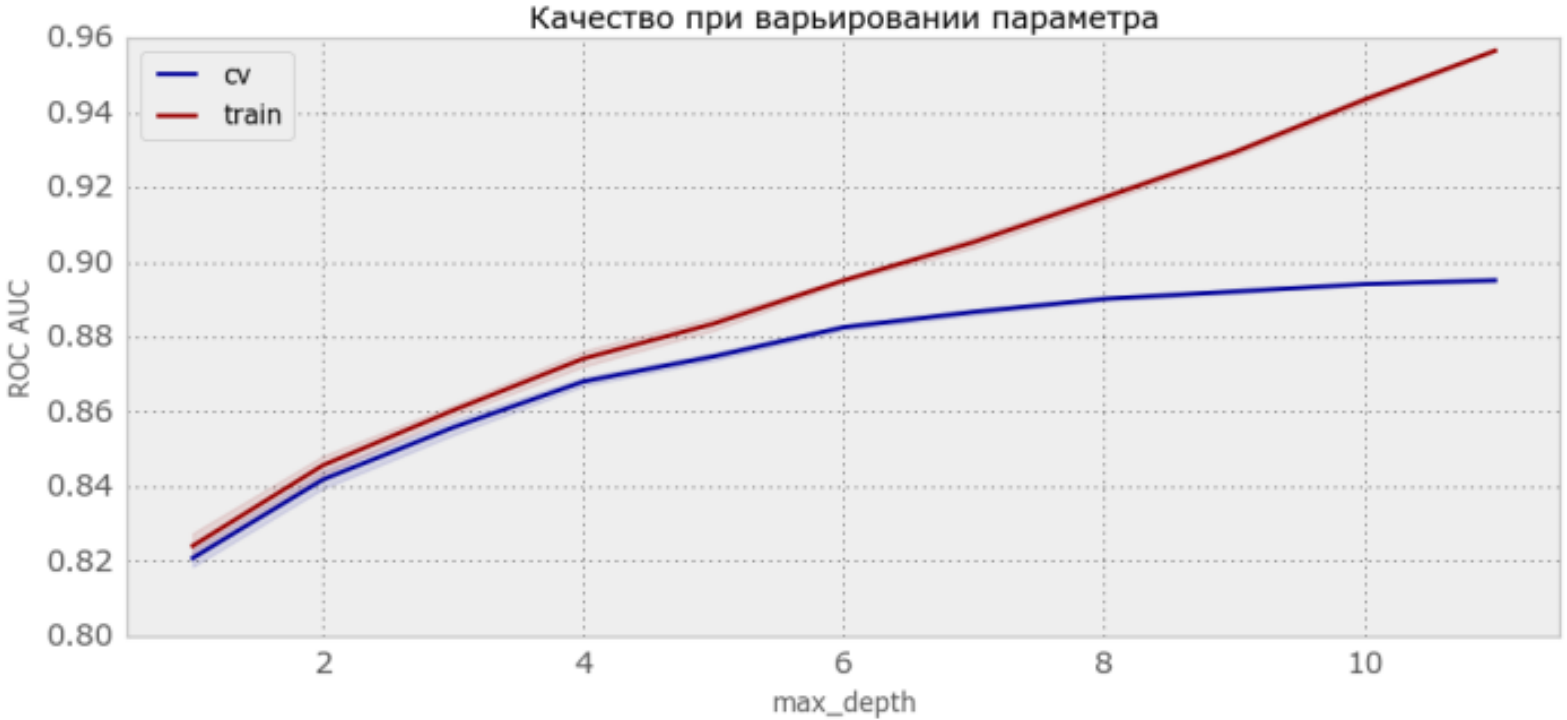
5.2. Бэггинг и случайный лес

Число признаков для выбора расщепления



5.2. Бэггинг и случайный лес

Максимальная глубина деревьев — чем меньше глубина, тем быстрее строится и работает



5.2. Бэггинг и случайный лес

Преимущества

- Надежный для коррелированных предикторов
- Используется для решения проблем регрессии и классификации
- Может обрабатывать тысячи входных переменных без выбора переменных
- Можно использовать в качестве инструмента выбора функции, используя его график переменной важности
- Заботится о недостающих данных внутри себя эффективным образом

5.2. Бэггинг и случайный лес

Недостатки

- Трудно интерпретировать
- Имеет тенденцию возвращать непредсказуемые значения для наблюдений из диапазона данных обучения. Например, данные обучения содержат две переменные x и y. Диапазон переменной x составляет от 30 до 70. Если тестовые данные имеют x = 200, случайный лес даст ненадежный прогноз.
- Память компьютера, время