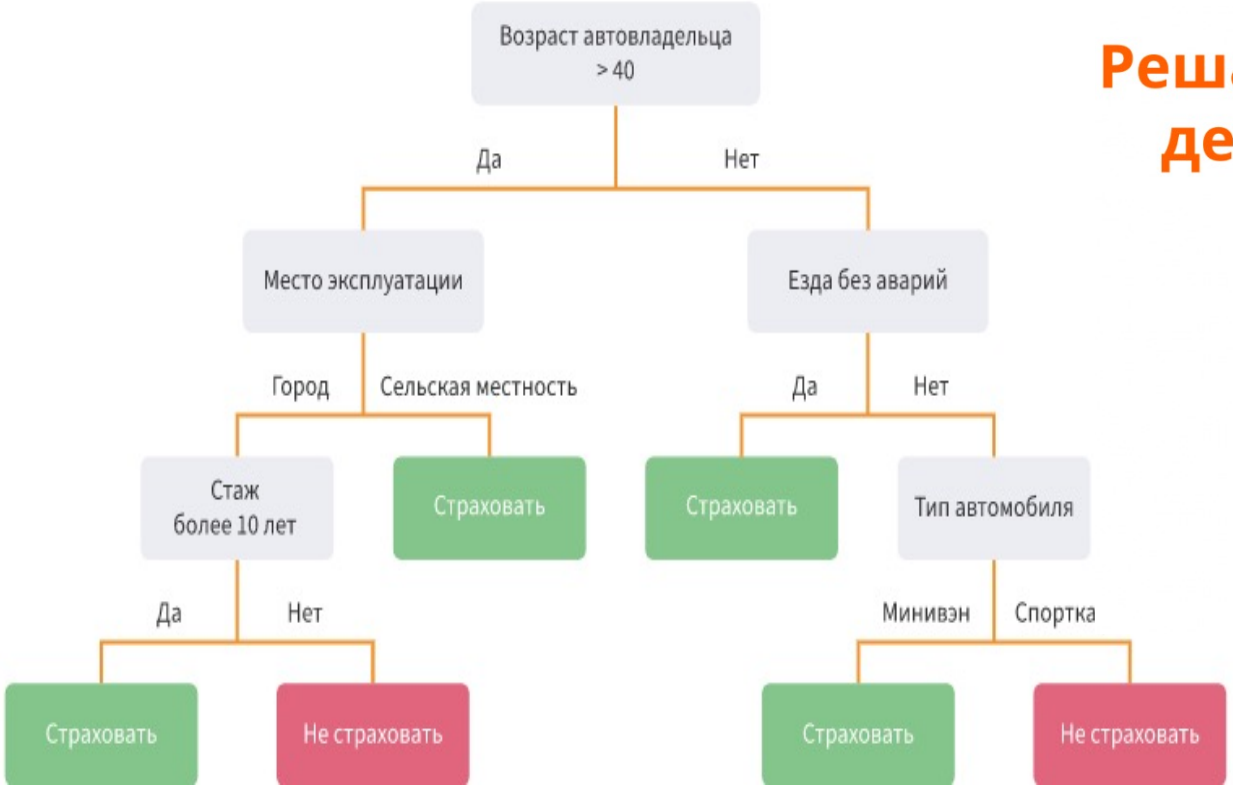
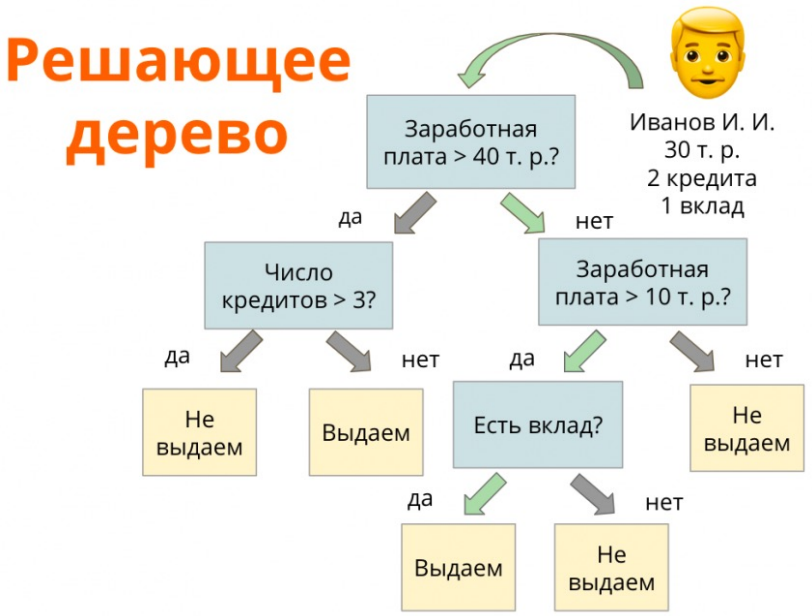
Лекция 3. ДЕРЕВЬЯ РЕШЕНИЙ

Решающее дерево — это алгоритм, который делает предсказания на основе серии вопросов об объекте.






К какой форме отнести публикацию?
Это монография, брошюра, учебник, сборник статей, сборник по итогам конференции и т. п.
Это статья
В научном журнале, в т. ч. электронном
В сборнике материалов конференции, сборнике статей, коллективной монографии, словаре
В СМИ
Не является научной публикацией
Может быть размещена в разделе “НИУ ВШЭ и СМИ”. Напишите на portalnews@hse.ru
Другое
Возможно, это препринт. Есть ли у него серия, напр., "Препринты ВШЭ" или "Arxiv"?
ДА
НЕ ЗНАЮ
Напишите на publications@hse.ru
Форма КНИГА
Форма СТАТЬЯ
Форма ГЛАВА КНИГИ
Форма ПРЕПРИНТ

Задачи

Основная сфера применения деревьев решений — поддержка процессов принятия управленческих решений, используемая в статистике, анализе данных и машинном обучении. Задачами, решаемыми с помощью данного аппарата, являются:

- **Классификация** — отнесение объектов к одному из заранее известных классов. Целевая переменная должна иметь дискретные значения.
- **Регрессия (численное предсказание)** — предсказание числового значения независимой переменной для заданного входного вектора.
- **Описание объектов** — набор правил в дереве решений позволяет компактно описывать объекты. Поэтому вместо сложных структур, описывающих объекты, можно хранить деревья решений.

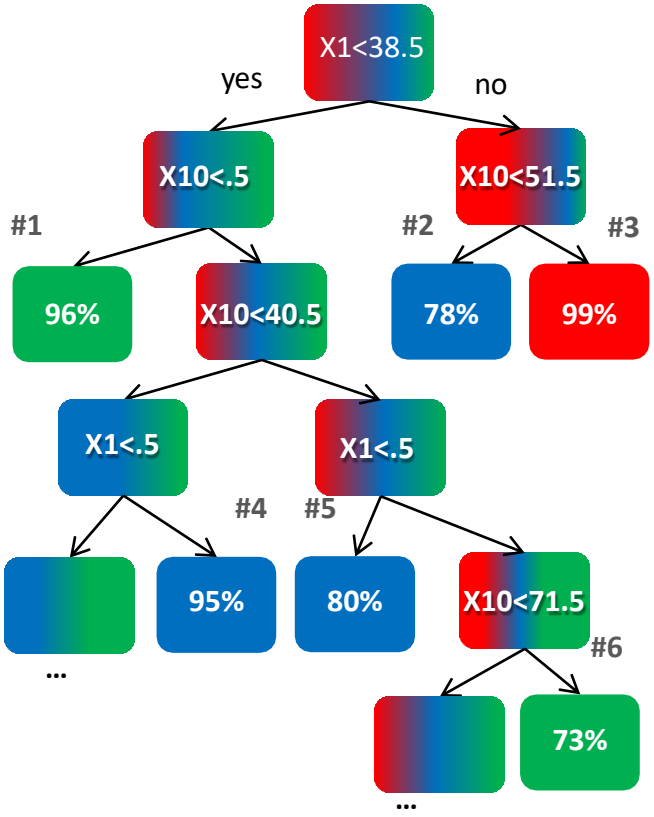
https://loginom.ru/blog/decision-tree-p1

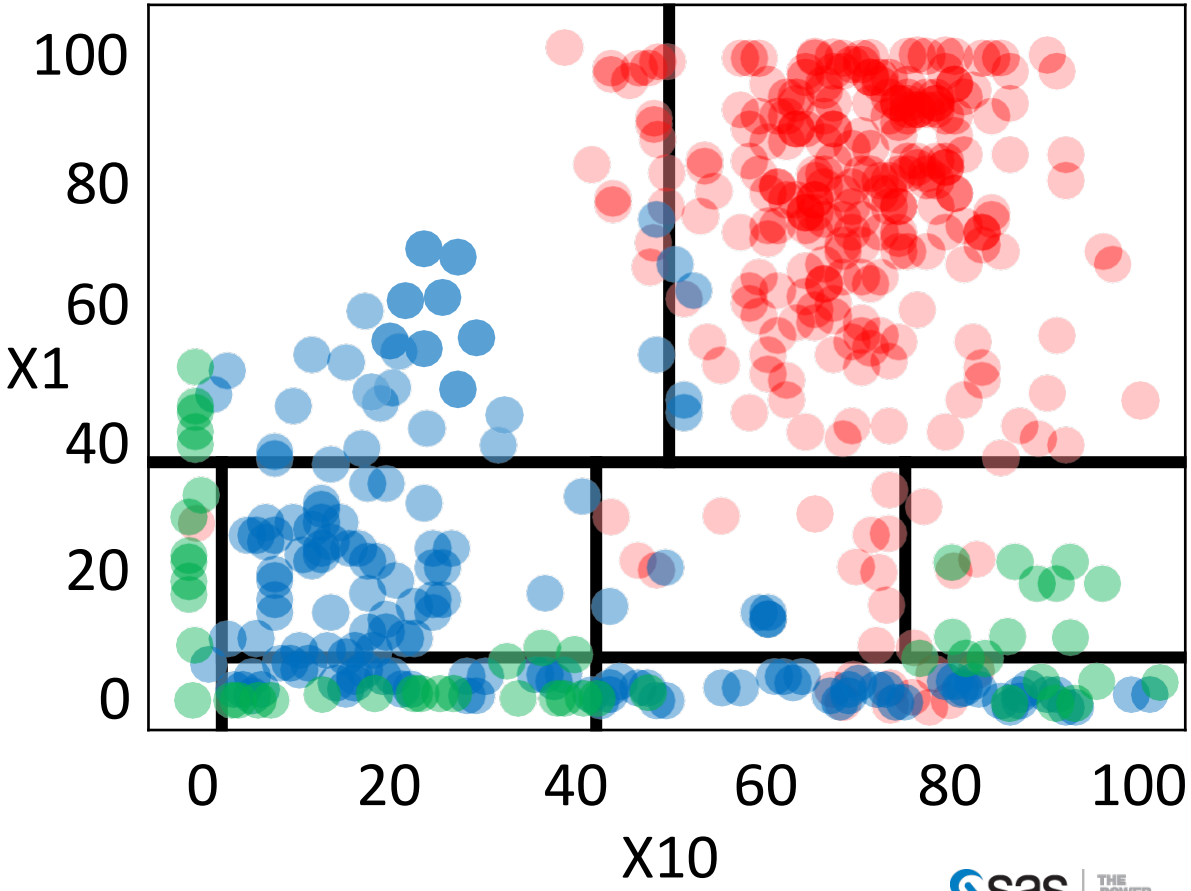
Основные этапы построения

1. Выбор атрибута, по которому будет производиться разбиение в данном узле (атрибута разбиения).
2. Выбор критерия остановки обучения.
3. Выбор метода отсечения ветвей (упрощения).
4. Оценка точности построенного дерева.

https://loginom.ru/blog/decision-tree-p1

НОМИНАЛЬНАЯ ЦЕЛЕВАЯ ПЕРЕМЕННАЯ





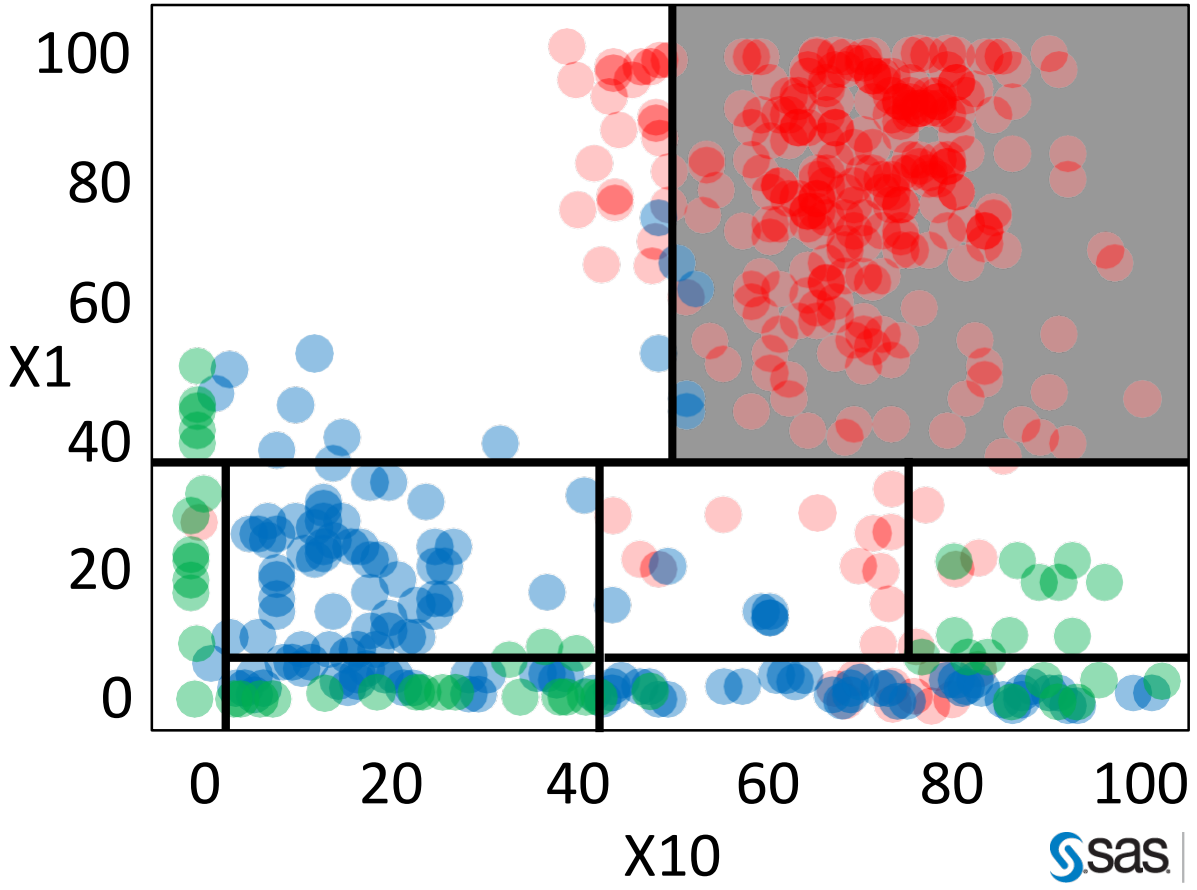
НОМИНАЛЬНАЯ ЦЕЛЕВАЯ ПЕРЕМЕННАЯ: РЕШЕНИЕ

#3

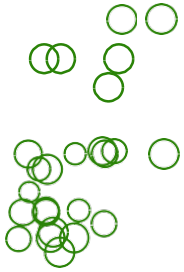
99%

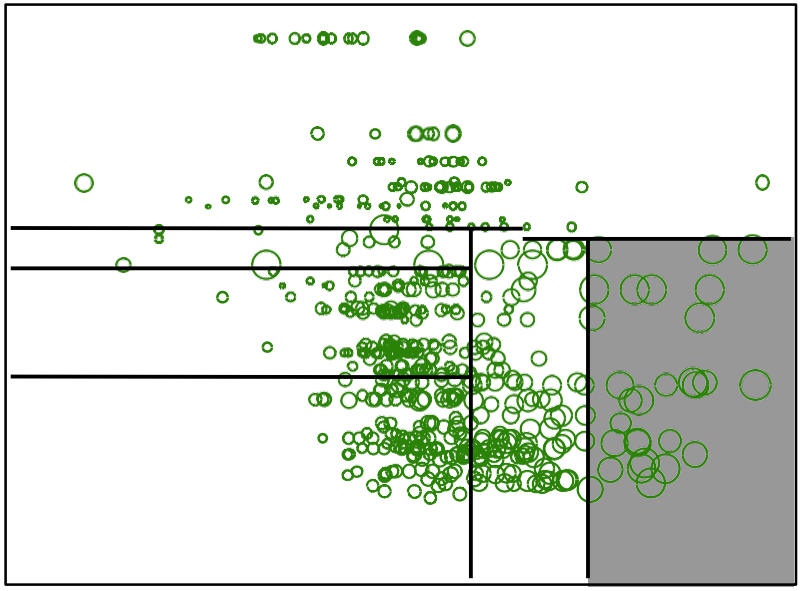
P( ) = 0.99
P( ) = 0.01
P( ) = 0.00

Prediction =



ИНТЕРВАЛЬНАЯ ЦЕЛЕВАЯ ПЕРЕМЕННАЯ: РЕШЕНИЕ

Prediction = Average()

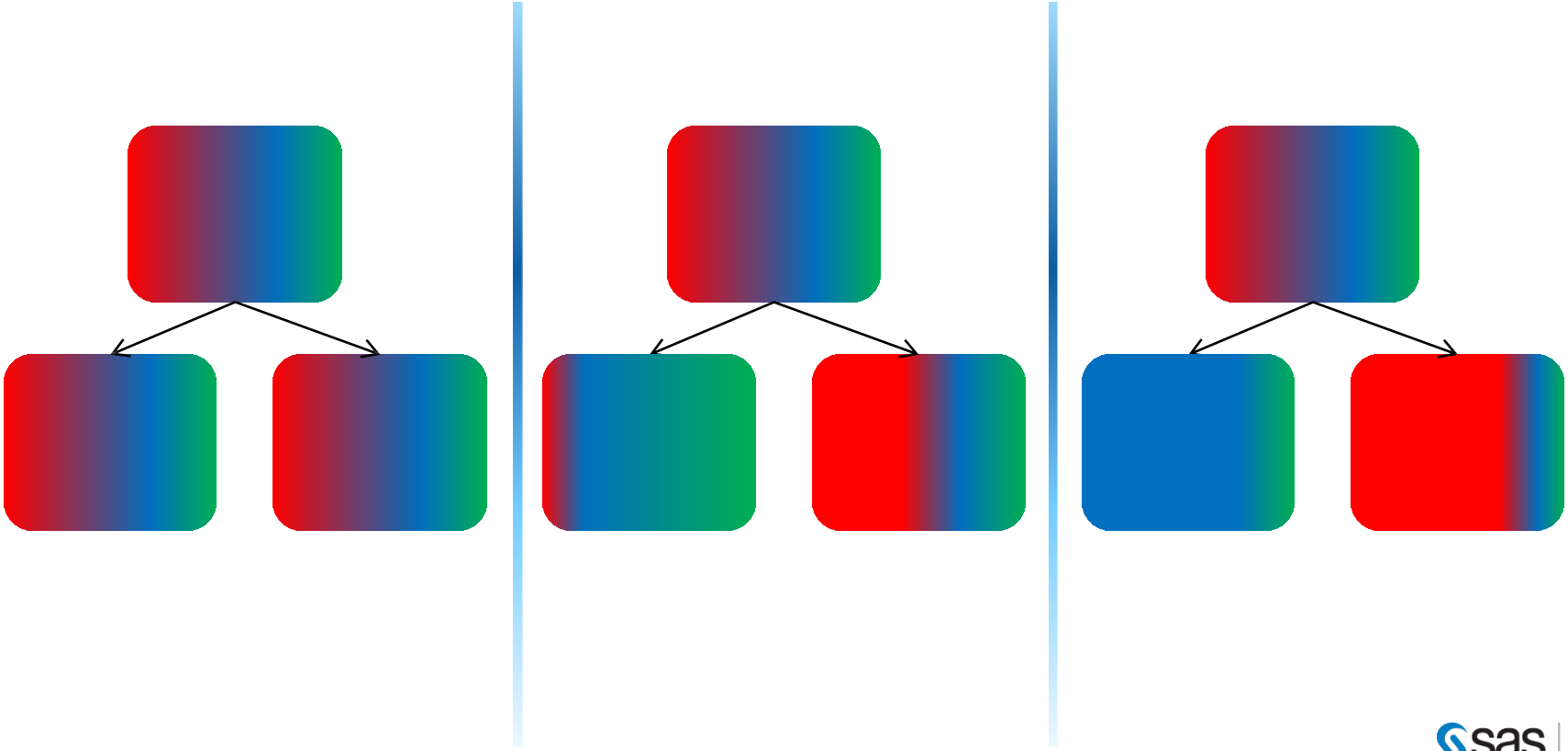


ПОИСК РАЗБИЕНИЙ

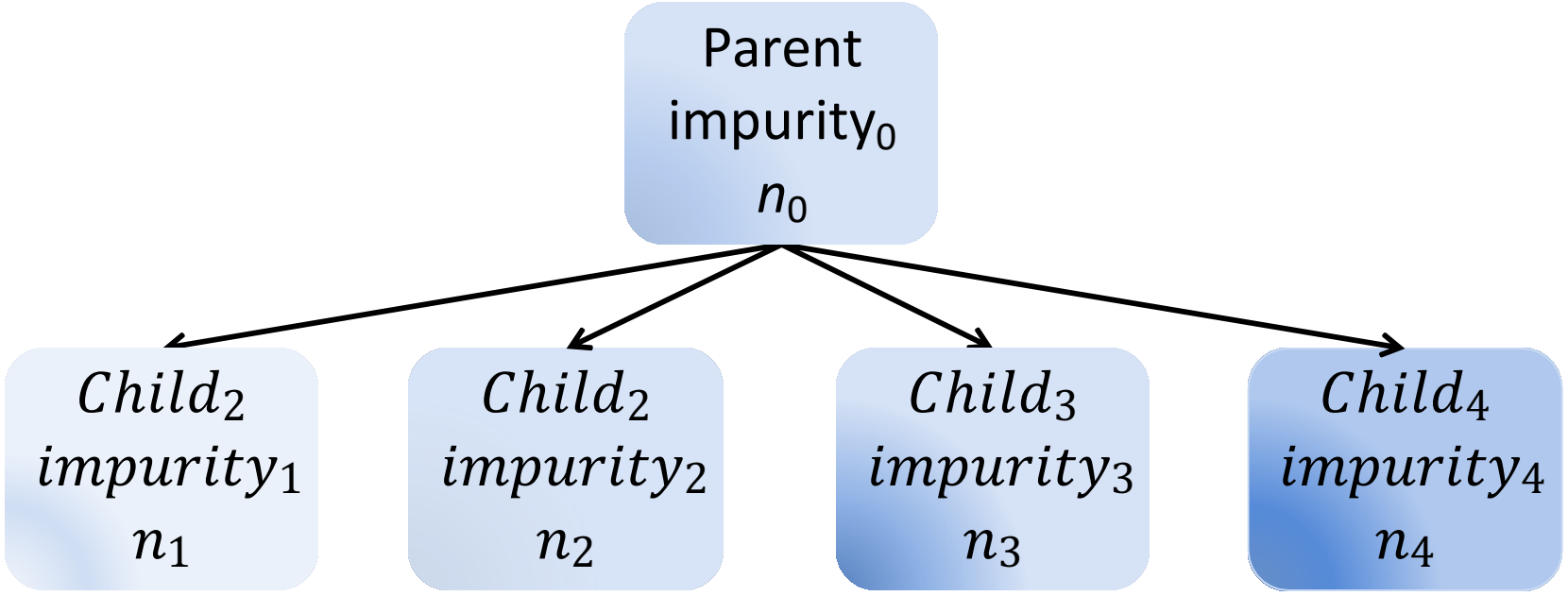
- Как выбрать «наилучшее» разбиение? – **КРИТЕРИЙ РАЗБИЕНИЯ**
- По какой переменной и какому значению производить разбиение? – **ПОИСК РАЗБИЕНИЯ**
- Когда нужно остановиться? – **ВЫБОР РАЗМЕРА ДЕРЕВА**

Критерии разбиения

КРИТЕРИИ РАЗБИЕНИЯ: ЧТО ТАКОЕ ХОРОШЕЕ РАЗБИЕНИЕ?



КРИТЕРИИ РАЗБИЕНИЯ: ЧТО ТАКОЕ ХОРОШЕЕ РАЗБИЕНИЕ?

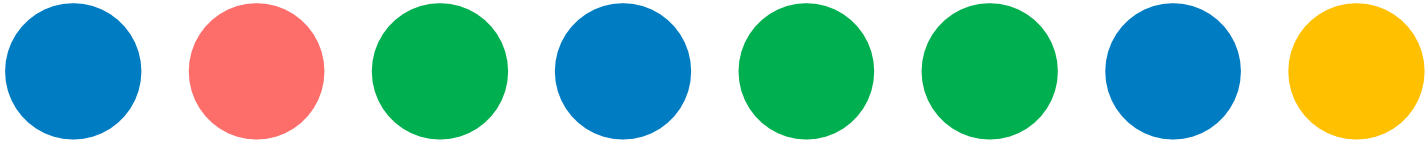


worth of split $\Delta i = i(0) - \left(\frac{n_1}{n_0} i(1) + \frac{n_2}{n_0} i(2) + \frac{n_3}{n_0} i(3) + \frac{n_4}{n_0} i(4) \right)$

КРИТЕРИИ РАЗБИЕНИЯ: GINI

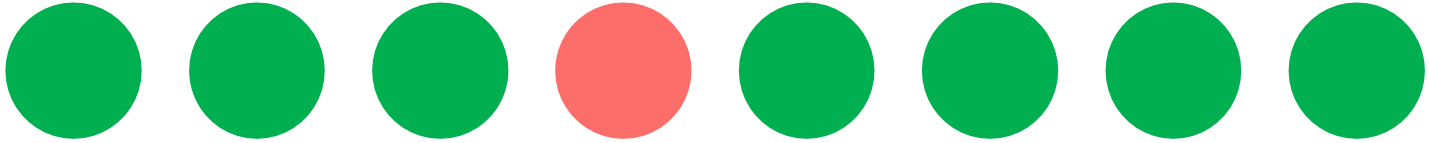
$$1 - \sum_{j=1}^{r} p_j^2 = 2 \sum_{j<k} p_j p_k$$

Большое разнообразие, низкая степень чистоты



$$GINI = 1 - 2\left(\frac{3}{8}\right)^2 - 2\left(\frac{1}{8}\right)^2 = .69$$

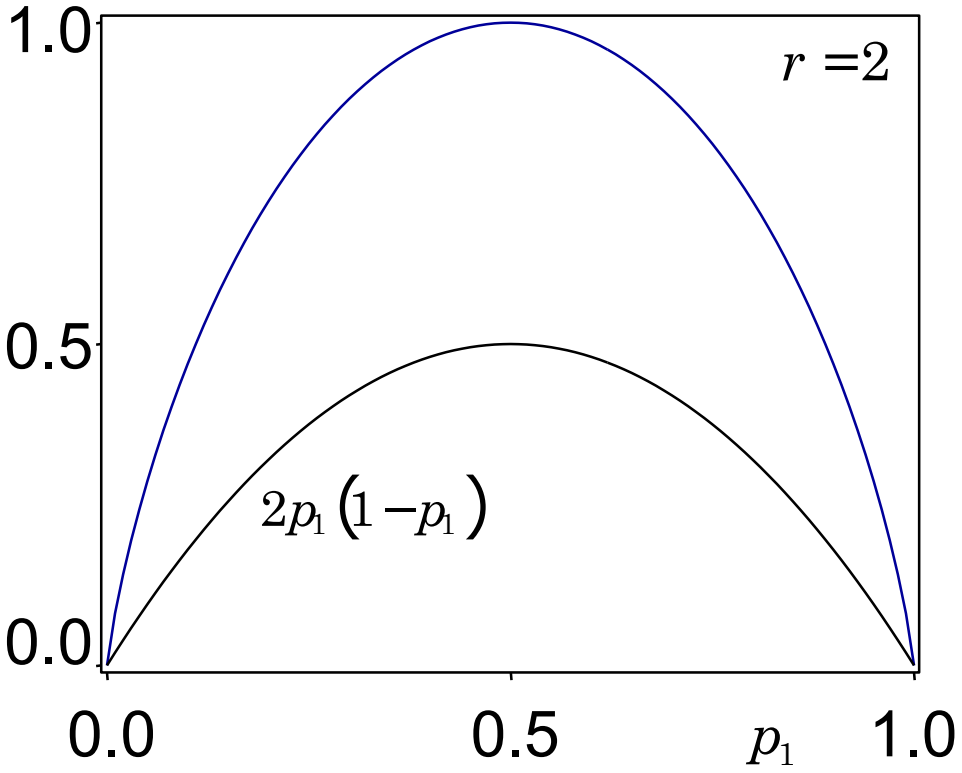
Небольшое разнообразие, высокая степень чистоты



$$GINI = 1 - \left(\frac{7}{8}\right)^2 - \left(\frac{1}{8}\right)^2 = .22$$

КРИТЕРИИ РАЗБИЕНИЯ: ENTROPY

$$H(p_1, p_2, \ldots, p_r) = -\sum_{i=1}^{r} p_i \log_2(p_i)$$



КРИТЕРИИ РАЗБИЕНИЯ: CHI-SQUARED

Observed

X1:	<38.5	≥38.5	
	293	71	.342
	363	1	.342
	42	294	.316
	.656	.344	*n*=1064

Expected

239	125
239	125
225	116

$$\frac{(O-E)^2}{E}$$

12	23
64	123
149	273

$$X_v^2 = \sum \frac{(O-E)^2}{E} = 644 \longrightarrow \textit{p-value, worth} = -\log_{10}(p)$$

$$v = (3-1)(2-1) = 2$$

КРИТЕРИИ РАЗБИЕНИЯ: CHI-SQUARED

Поправки p-value

- Количество ветвей
- Количество различных значений переменной
- Количество переменных (кандидатов) для разбиения в данном узле

- Глубина текущего узла в дереве

Чем больше вариантов мы перебираем, тем больше шансы случайно выбрать неверное разбиение

Чем ниже (глубже) мы в дереве, тем меньше кол-во наблюдений в узле, p-value слишком оптимистичные

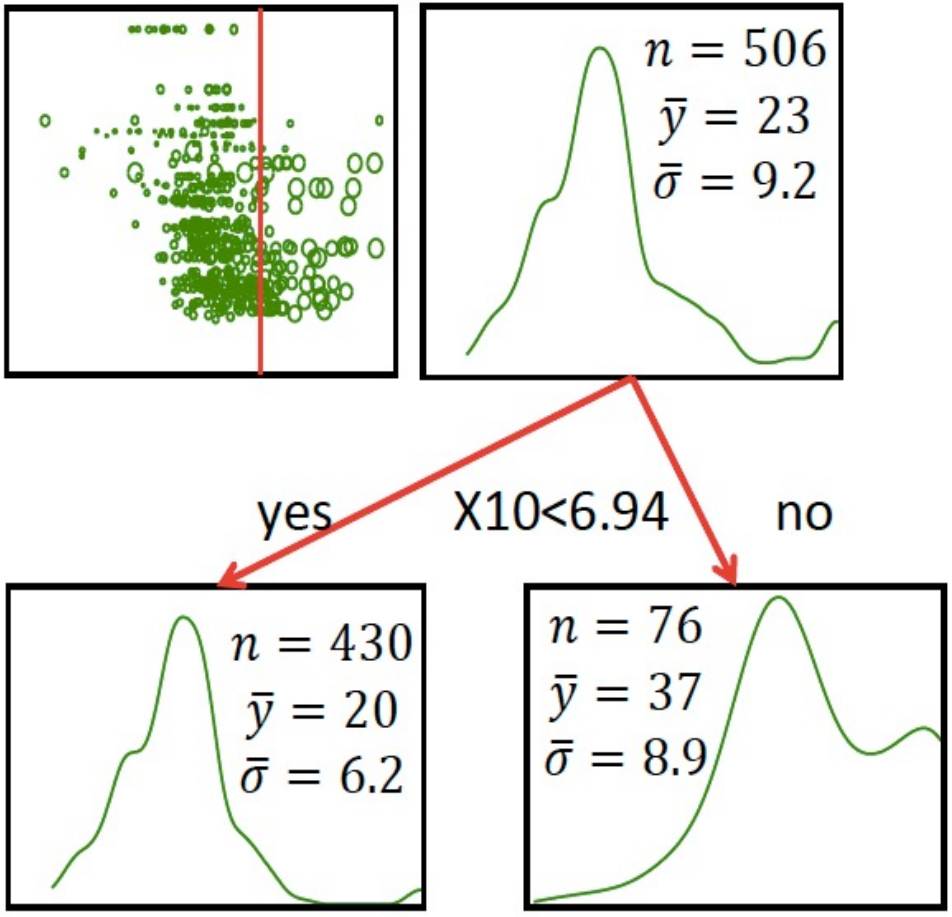
КРИТЕРИИ РАЗБИЕНИЯ: ИНТЕРВАЛЬНАЯ ЦЕЛЕВАЯ ПЕРЕМЕННАЯ

$$VARIANCE = \frac{\sigma_{j=1}^{N} (Y_j - \mu)^2}{N}$$

$$F = \left(\frac{SS_{between}}{SS_{within}} \right) \left(\frac{n - B}{B - 1} \right)$$

p-value , worth $= -\log_{10}(p)$

поправки p-value



Поиск разбиения

ПОИСК РАЗБИЕНИЯ. ШАГ 1

2. Для переменной **X** удалить наблюдения:
   - с пропущенными значениями (кроме случая *Use In Search*)
   - если **X** – категориальная, то удалить наблюдения со значениями, встречающимися редко

Удаления всегда производятся независимо по каждой переменной и каждому узлу

ПОИСК РАЗБИЕНИЯ. ШАГ 2

1. Сортируем наблюдения:
   - если **X** – интервальная или ординальная – просто сортировка по возрастанию, пропущенные значения отправляются в конец
   - если **X** – номинальная, а целевая переменная **Y** – непрерывная, то сортируем уровни/значения X: L_i по *avg*(**Y**)
   - если **X** – номинальная, а целевая переменная **Y** – имеет два значения, то сортируем уровни/значения X: L_i по *доле одного из значений* **Y**

В остальных случаях (X и Y – номинальная, Y имеет 3 и более значений) сортировка не производиться

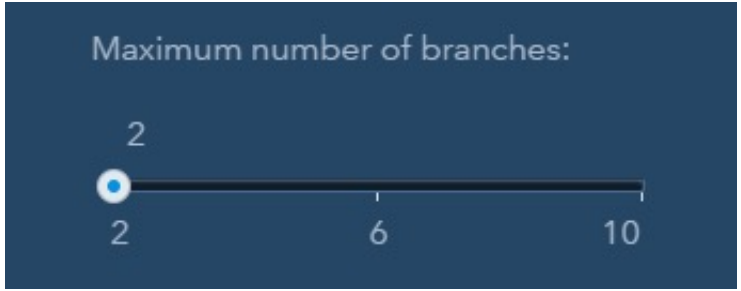
ПОИСК РАЗБИЕНИЯ. ШАГ 3

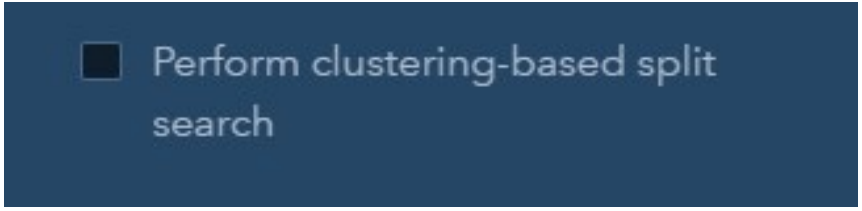
1. Рассматриваем все возможные ***бинарные*** разбиения:
   - если **X** – интервальная или ординальная – то разбиения между одинаковыми значениями X не допускаются
   - если **X** – номинальная, а целевая переменная **Y** – непрерывная, то разбиение между разными значений X с одинаковым *avg*(**Y**) не допускаются
   - Разбиение, при котором в листе останется мало наблюдений не допускаются

Теперь найдены все возможные
БИНАРНЫЕ РАЗБИЕНИЯ

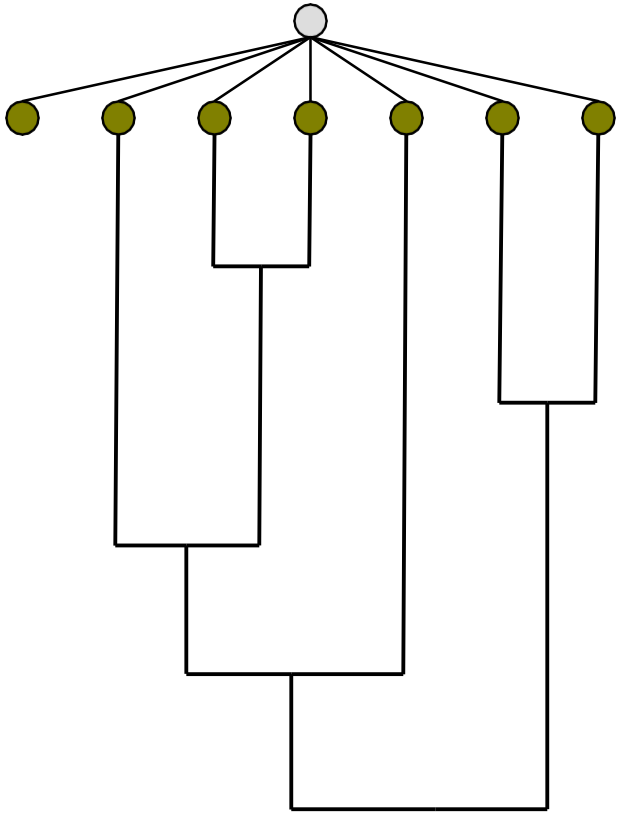
ПОИСК РАЗБИЕНИЯ. ШАГ 4

1. Рассматриваем все возможные **M**-*арные* разбиения (2 ≤ N ≤ **M**):
   - Перебираем все – выбираем одно лучшее
   - Или проводим консолидацию (*merge-and-shuffle algorithm*)





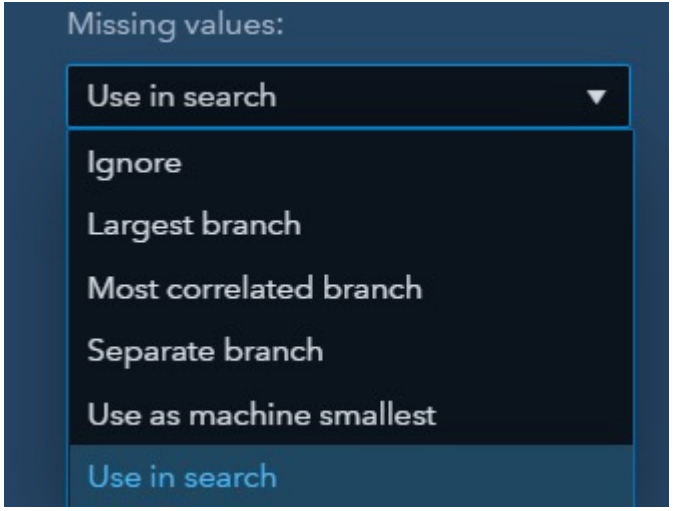
ПОИСК РАЗБИЕНИЯ. ШАГ 5



Merge-and-shuffle algorithm

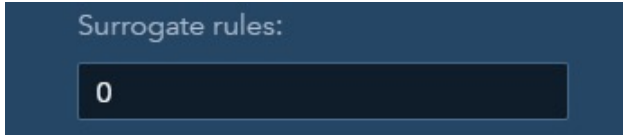
1. Для каждого значения X создаем свою **узел**
   - Интервальные переменные разбиваем на N групп
2. Поочередно рассматриваем пары **узлов:**
   - Объединяем пару
   - Обратно разбиваем на две
   - Измеряем полезность такого разбиения
   - Объединяем пару с минимальной полезностью

РАБОТА С ПРОПУЩЕННЫМИ ЗНАЧЕНИЯМИ



Что можно сделать с пропущенными значениями

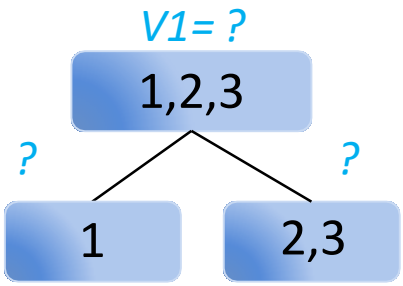
РАБОТА С ПРОПУЩЕННЫМИ ЗНАЧЕНИЯМИ

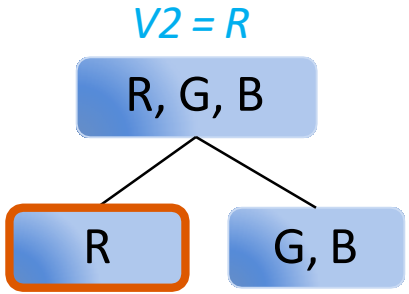


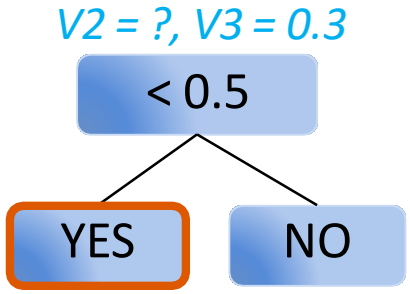
Суррогатные правила:

- в какую ветку отправится наблюдение, если значение *основной переменной* - пропущенное
- из всех возможных выбирается правило, максимально *соответствующее* основному

разбиение по **основной переменной** *V1*



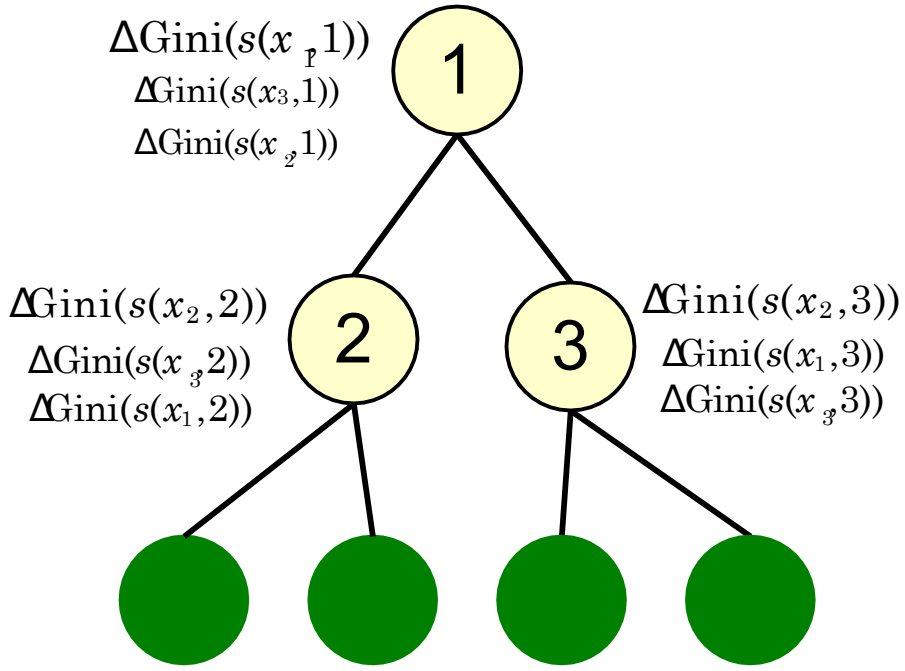




...

Значимость переменных

ЗНАЧИМОСТЬ ПЕРЕМЕННЫХ: НА ОСНОВЕ РАЗБИЕНИЙ



Значимость (важность) переменной x_i :

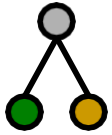
$$Importance(x_j) = \sqrt{\sum_{t=1}^{T} \frac{n_t}{n} \times \alpha(s_x, t_j) \times \Delta i(s(x_j, t))}$$

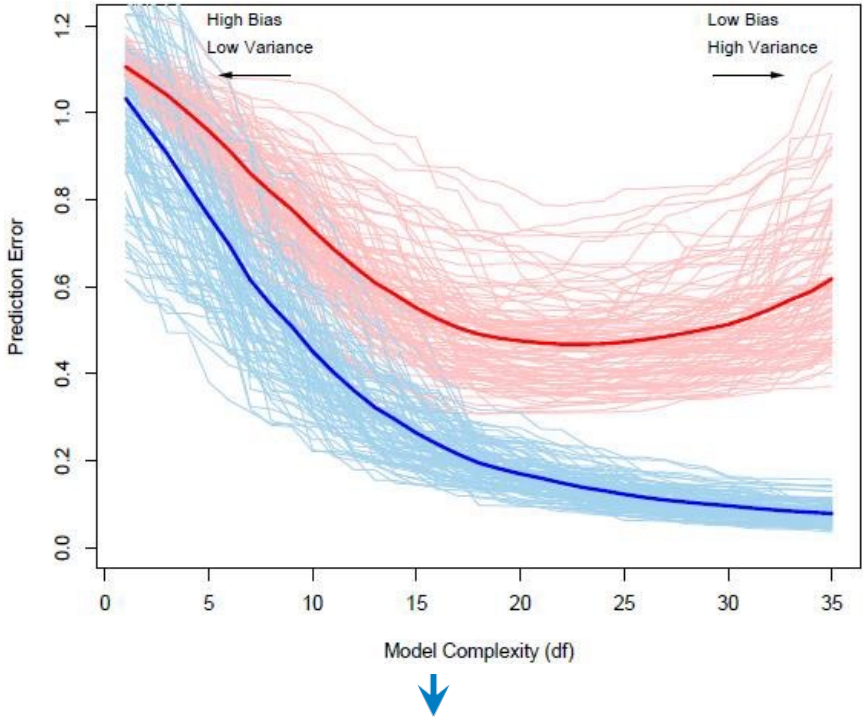
вес — соответствие — улучшение

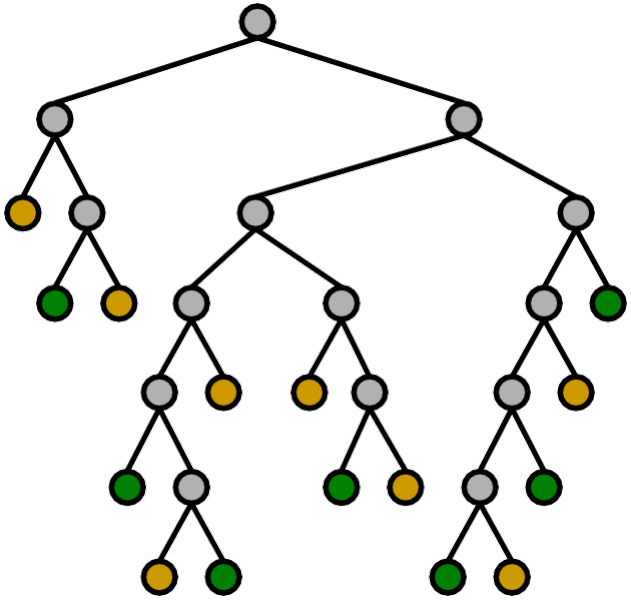
Нормируется: $Importance \in [0, 1]$

Остановка роста дерева & обрубание

КОГДА НУЖНО ПРЕКРАТИТЬ РОСТ ДЕРЕВА



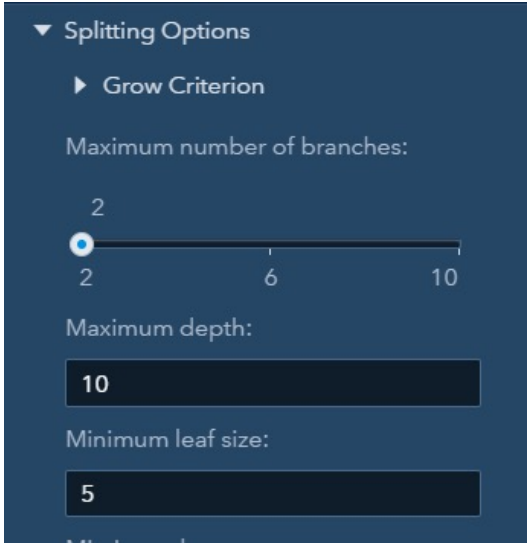




Сложность $= f($# листов, # ветвей, глубина$)$

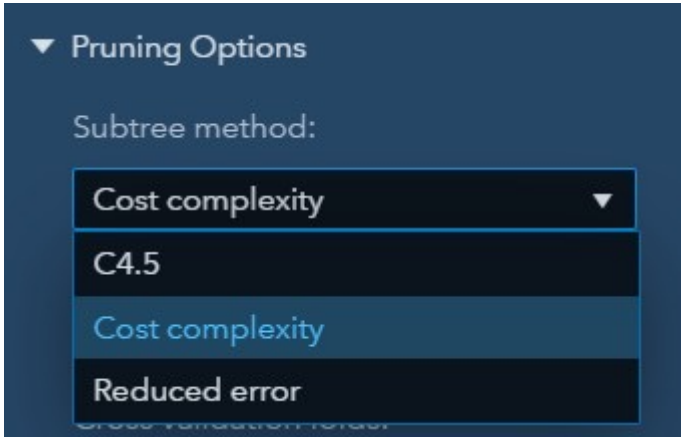
КОГДА НУЖНО ПРЕКРАТИТЬ РОСТ ДЕРЕВА: ОБРУБАНИЕ / PRUNING

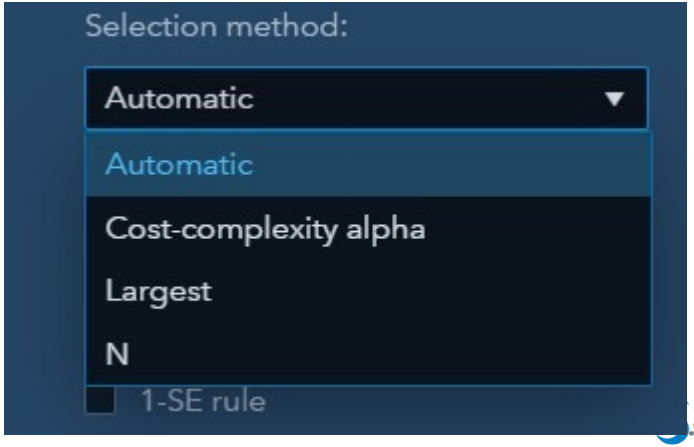
1. Сверху вниз (Top-Down) – критерии остановки роста:
   - Размер листа
   - Глубина дерева



КОГДА НУЖНО ПРЕКРАТИТЬ РОСТ ДЕРЕВА: ОБРУБАНИЕ / PRUNING

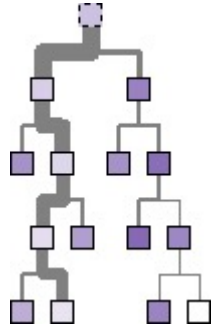
2. Снизу вверх (Bottom-Up) – критерий выбора «лучшего» дерева
   - Точность модели
     - На валидационном наборе
     - Используя CV



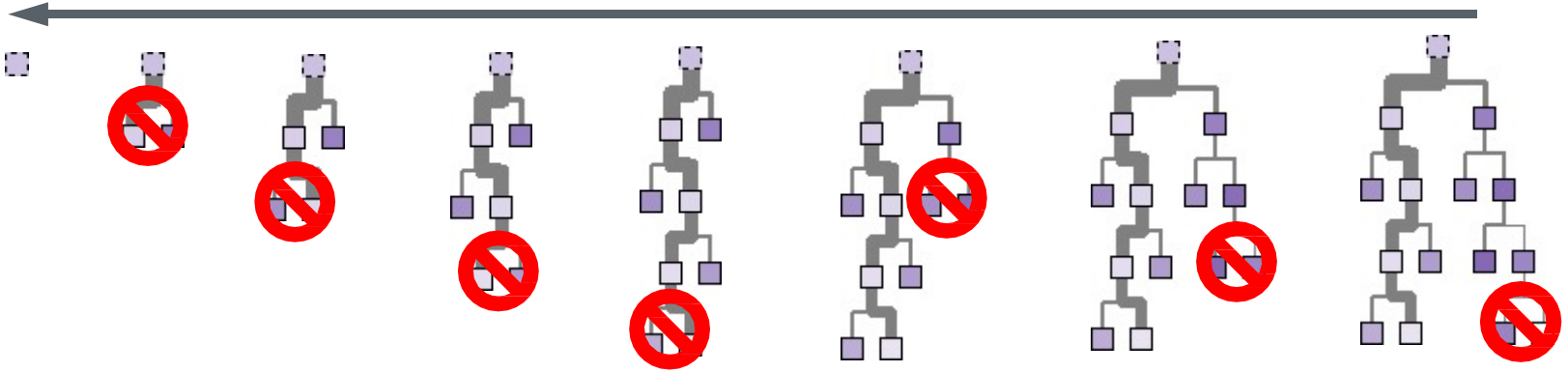


КОГДА НУЖНО ПРЕКРАТИТЬ РОСТ ДЕРЕВА: ОБРУБАНИЕ / PRUNING

2. Снизу вверх (Bottom-Up) – критерий выбора «лучшего» дерева

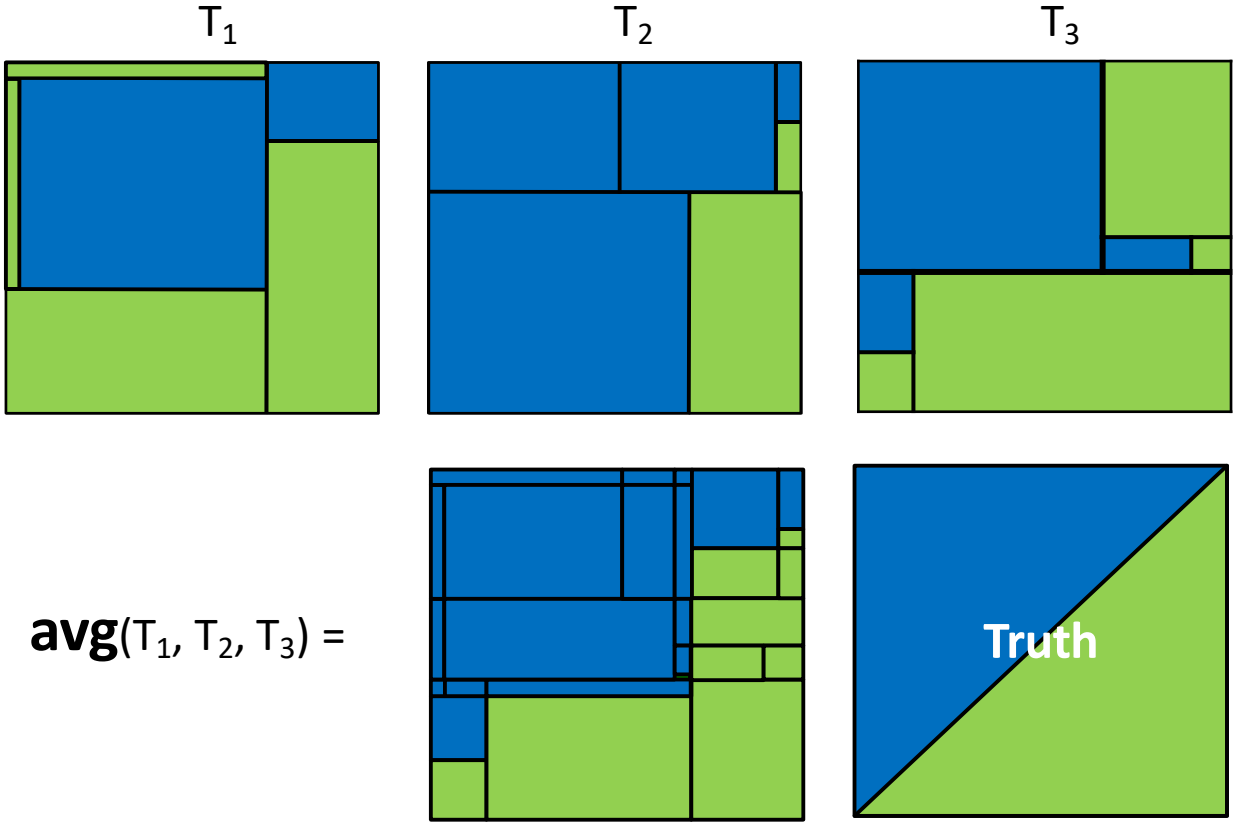


- Вырастить максимальное дерево
- Обрубая по одному листу (*min. valid-error* OR *min. CV-error*), построить последовательность деревьев:



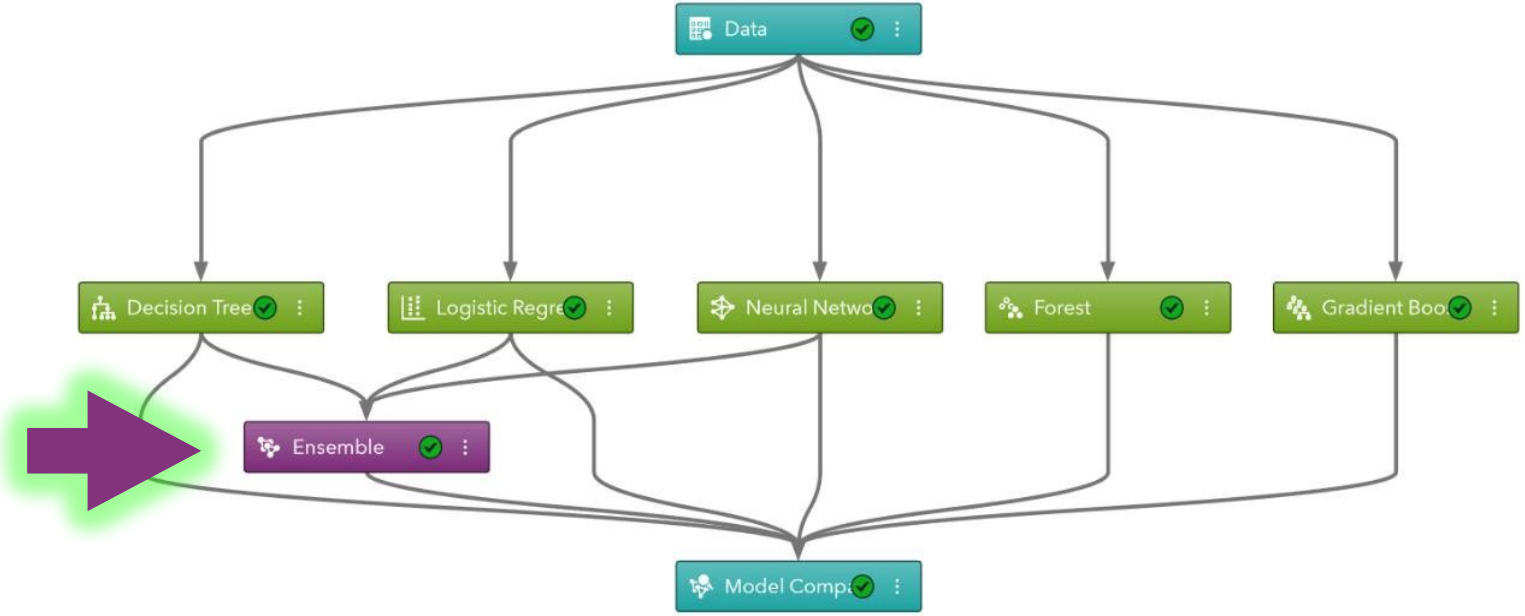
Ансамбли моделей. Случайный лес. Деревья как вспомогательный инструмент.

АНСАМБЛЬ – КОМБИНАЦИЯ МОДЕЛЕЙ

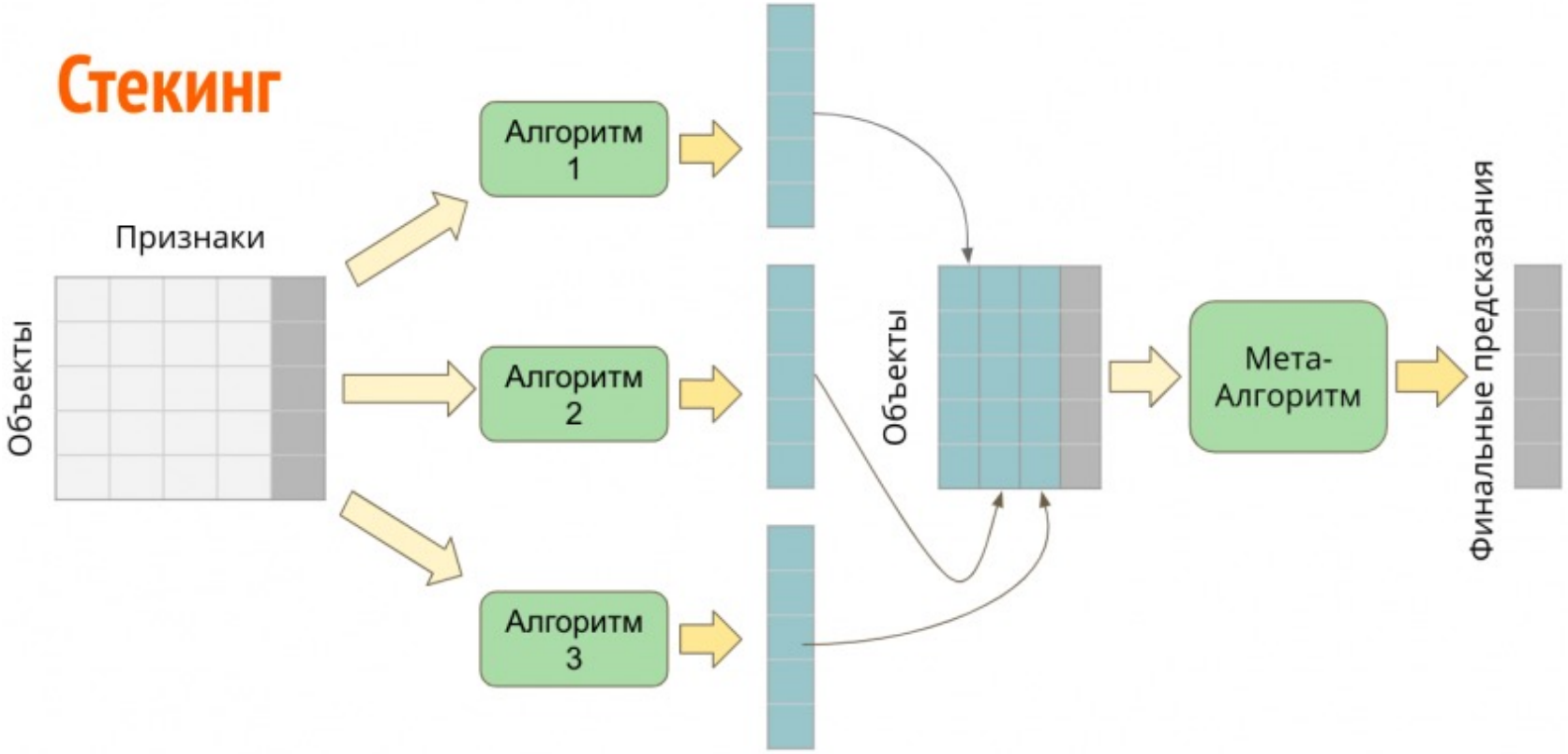


АНСАМБЛЬ – КОМБИНАЦИЯ МОДЕЛЕЙ

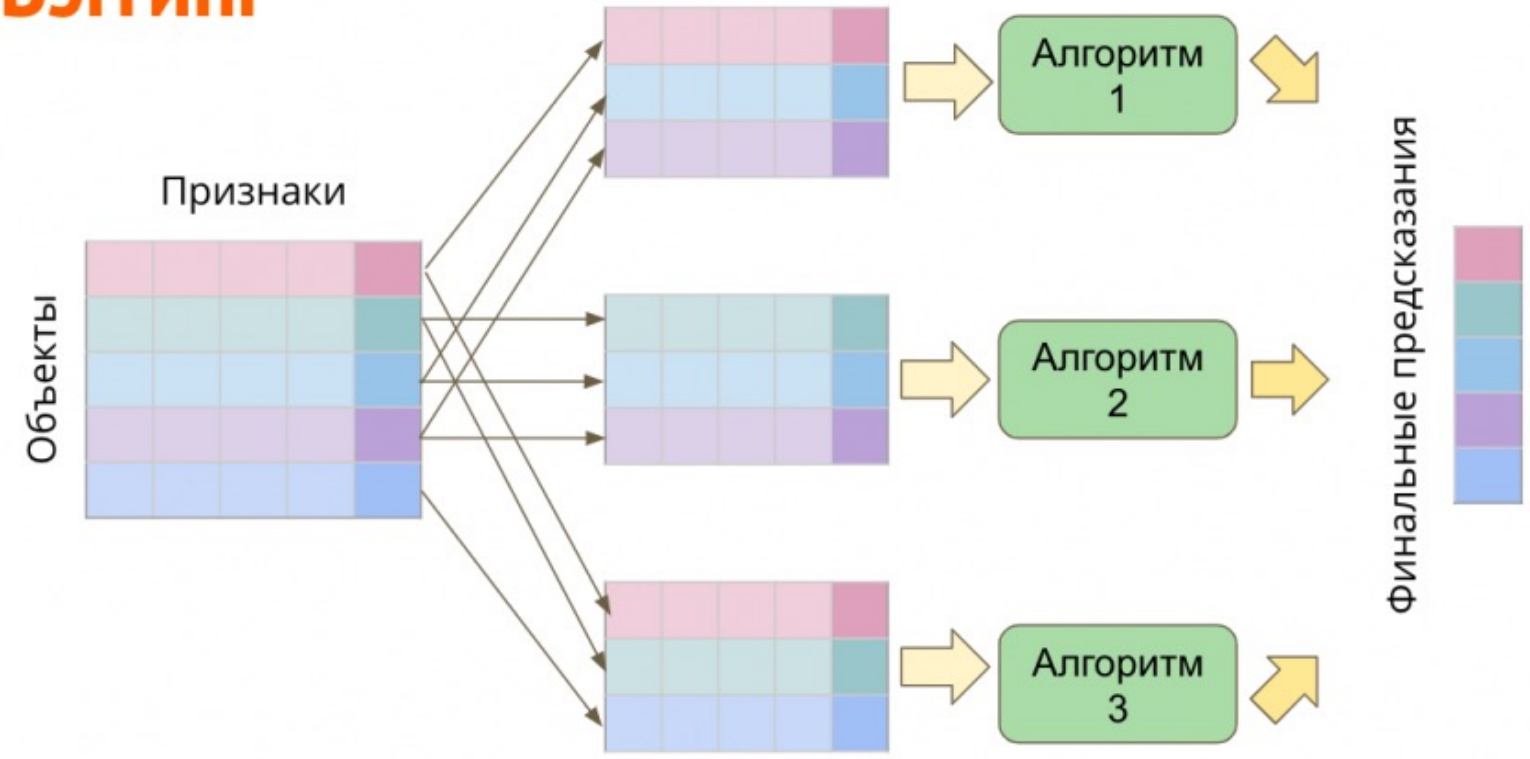
ПРОСТОЕ «ГОЛОСОВАНИЕ»: среднее вероятности на выходе моделей (арифметическое и геометрическое среднее), максимум вероятности



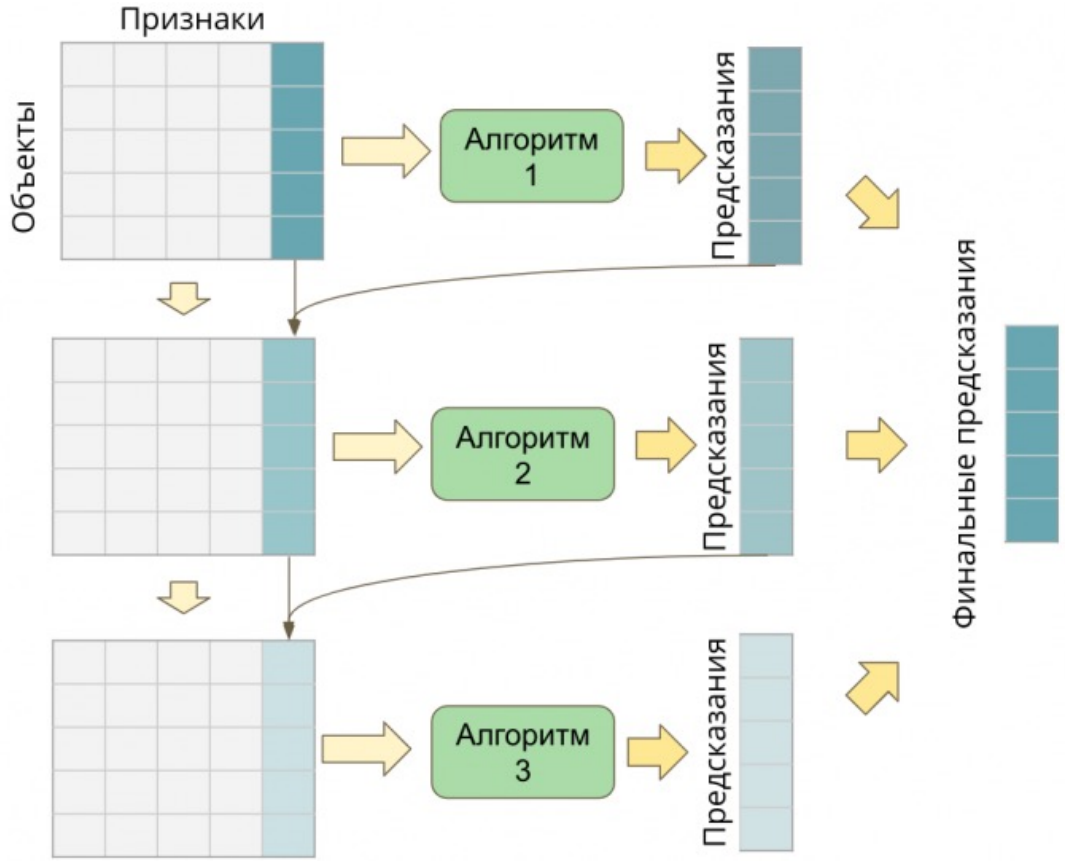
Стекинг



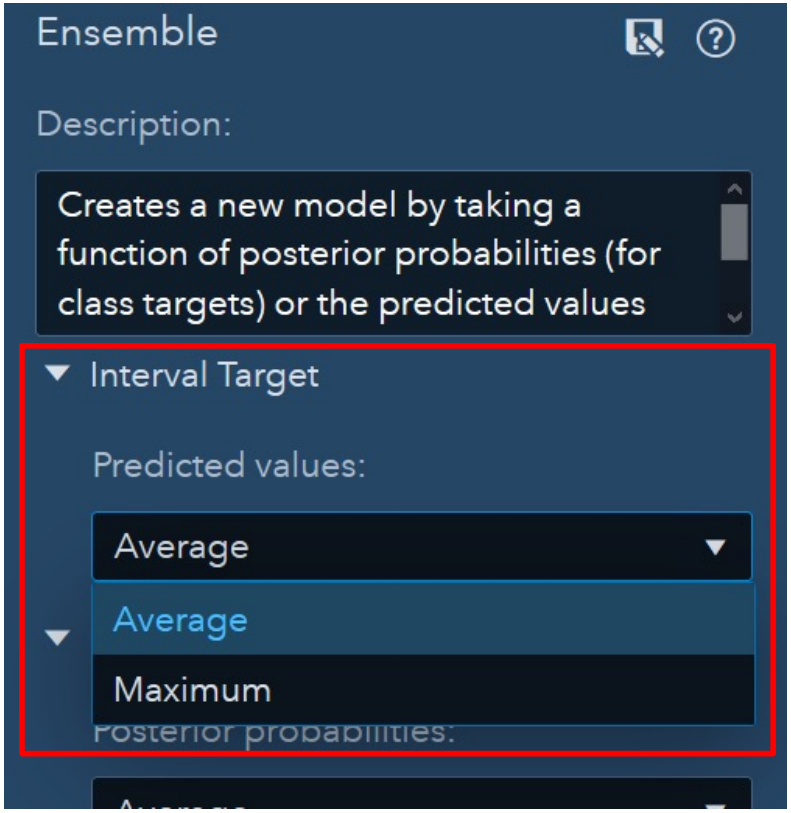
Бэггинг

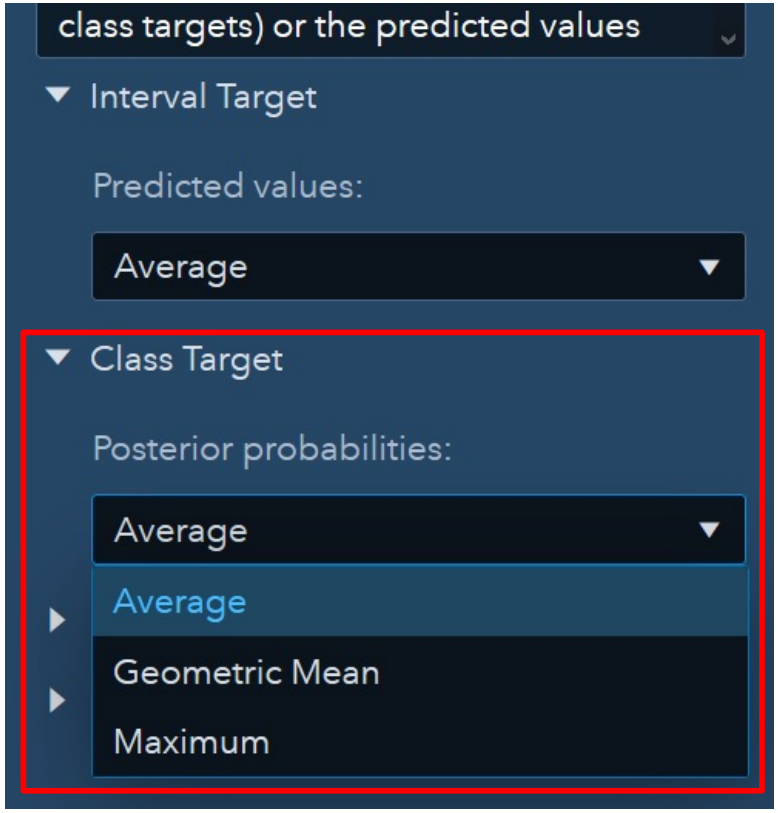


Бустинг



УЛЕЗ ENSEMBLE: НАСТРОЙКИ





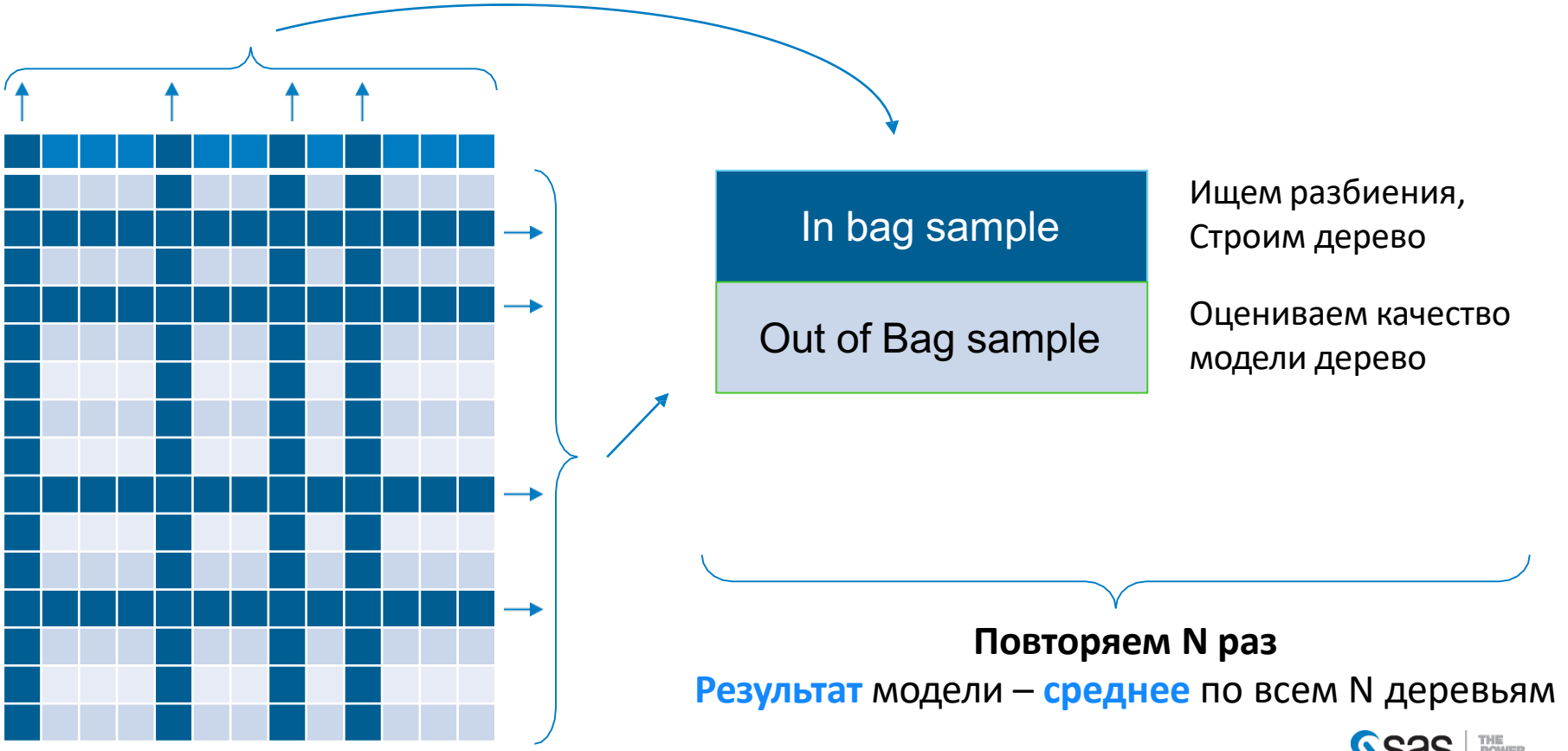
Random forest

RANDOM FOREST – СЛУЧАЙНЫЙ ЛЕС

- **ЛЕС**
  - Строим много деревьев
- **СЛУЧАЙНЫЙ**
  - Используем (случайную) подвыборку наблюдений тренировочного набора
  - Используем (случайную) подвыборку переменных

Хороший лес - Много разнообразных деревьев

RANDOM FOREST

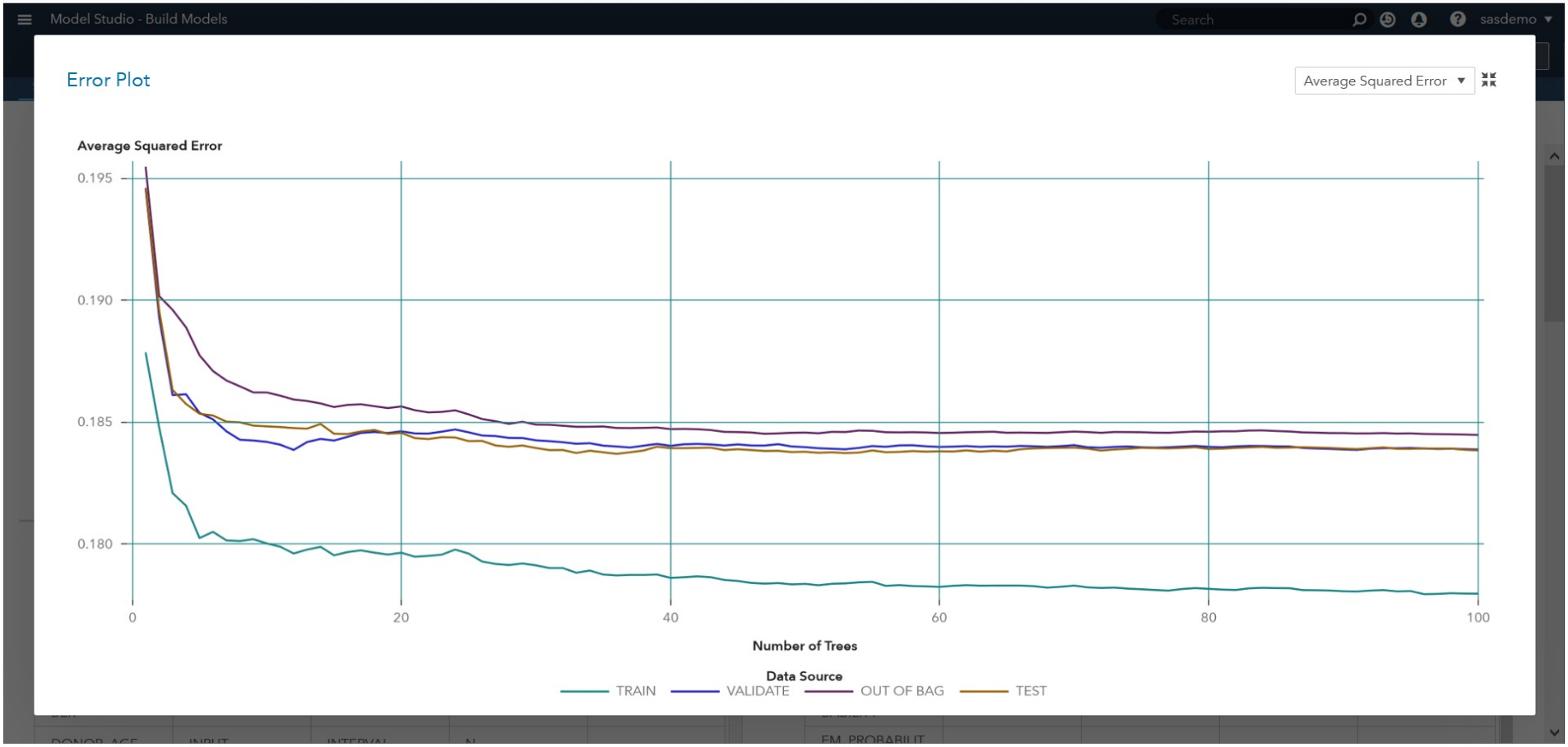


RANDOM FOREST

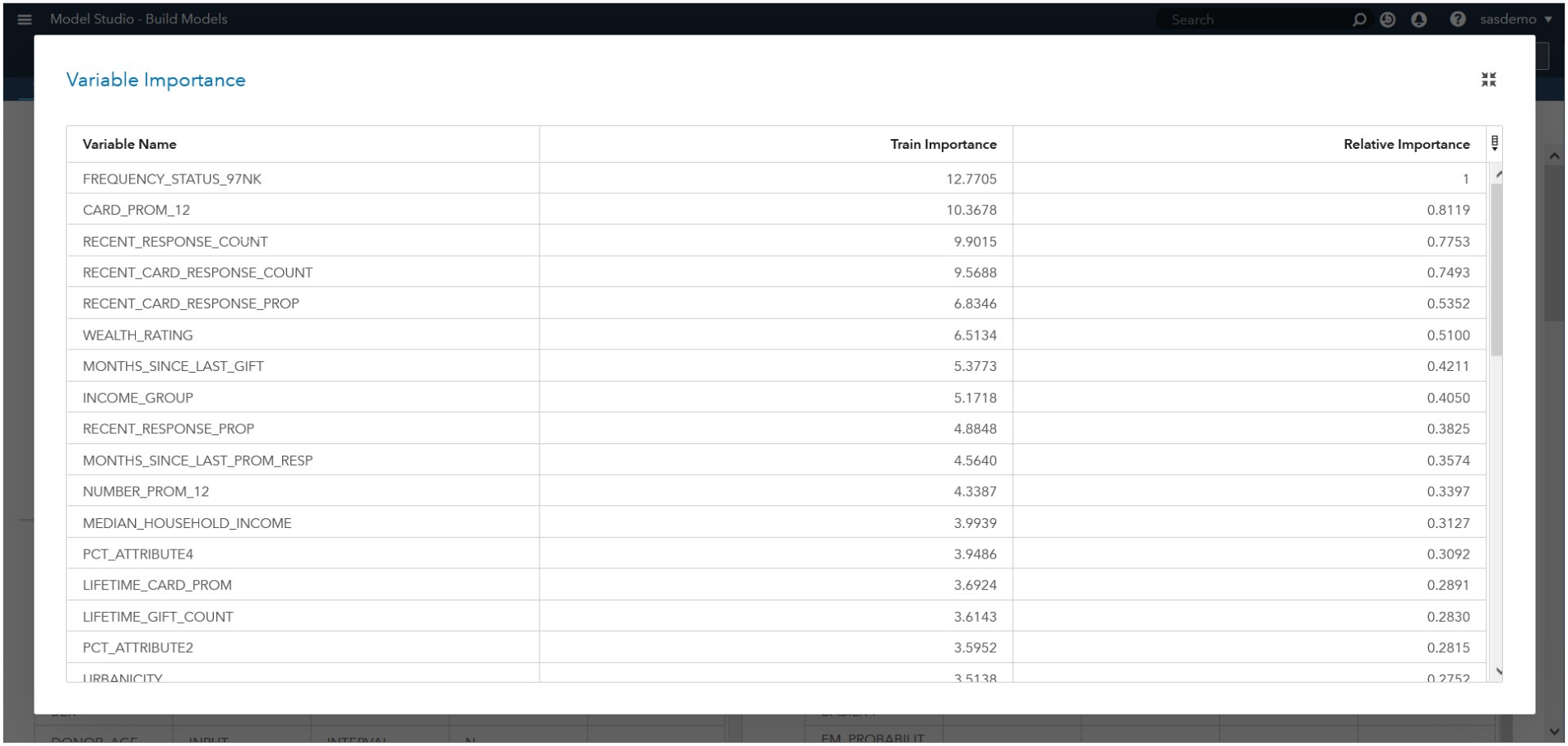
- Сколько нужно деревьев?
  - Больше деревьев – меньше дисперсия – надежнее предсказание
- Сколько нужно переменных?
  - Default - $\sqrt{N_{vars}}$
  - Много сильных предикторов – делаем меньше
  - Много слабых предикторов – делаем больше
    - дополнительно (опция ___) из выбранных случайно отбираем с лучшим p-value хи-квадрат (спец. вариант)

https://dyakonov.org/2016/11/14/%D1%81%D0%BB%D1%83%D1%87%D0%B0%D0%B9%D0%BD%D1%8B%D0%B9-%D0%BB%D0%B5%D1%81-random-forest/

RANDOM FOREST



RANDOM FOREST



Variable Importance

Variable Name	Train Importance	Relative Importance
FREQUENCY_STATUS_97NK	12.7705	1
CARD_PROM_12	10.3678	0.8119
RECENT_RESPONSE_COUNT	9.9015	0.7753
RECENT_CARD_RESPONSE_COUNT	9.5688	0.7493
RECENT_CARD_RESPONSE_PROP	6.8346	0.5352
WEALTH_RATING	6.5134	0.5100
MONTHS_SINCE_LAST_GIFT	5.3773	0.4211
INCOME_GROUP	5.1718	0.4050
RECENT_RESPONSE_PROP	4.8848	0.3825
MONTHS_SINCE_LAST_PROM_RESP	4.5640	0.3574
NUMBER_PROM_12	4.3387	0.3397
MEDIAN_HOUSEHOLD_INCOME	3.9939	0.3127
PCT_ATTRIBUTE4	3.9486	0.3092
LIFETIME_CARD_PROM	3.6924	0.2891
LIFETIME_GIFT_COUNT	3.6143	0.2830
PCT_ATTRIBUTE2	3.5952	0.2815
URBANICITY	3.5138	0.2752

Применение в качестве вспомогательного инструмента

Применение в качестве вспомогательного инструмента

- Выявление аномалий
- Уменьшение размерности
  - Выбор важных / значимых переменных для других методов моделирования
  - Группировка значений категориальных переменных
- Преобразование входных данных
  - Нелинейные пересечения переменных
  - Замена пропущенных значений
  - Дискретизация непрерывных переменных

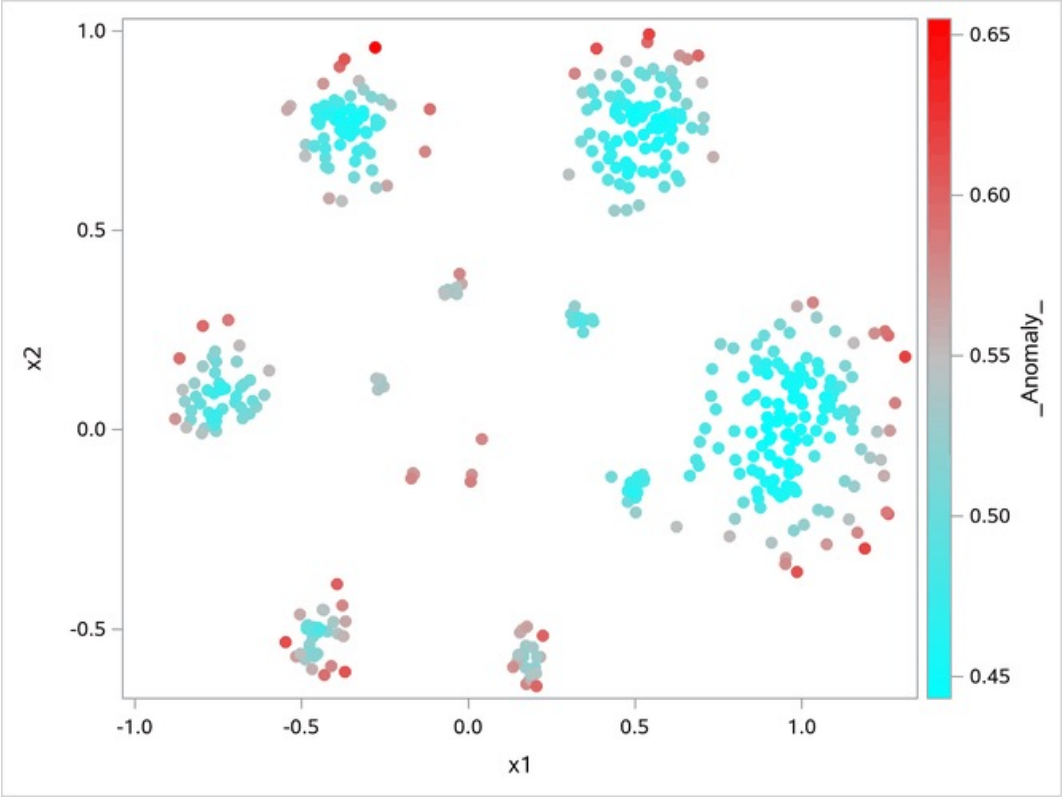
Случайно выбираем одну переменную:

- ***Интервальная переменная:*** разбиение по случайному значение
- ***Номинальная:*** каждое значение отправляется в случайную ветку/лист

При таком построении аномалии будут скорее всего ближе к корню, чем нормальные наблюдения. Считаем ***anomaly score*** для каждого наблюдения *x*:

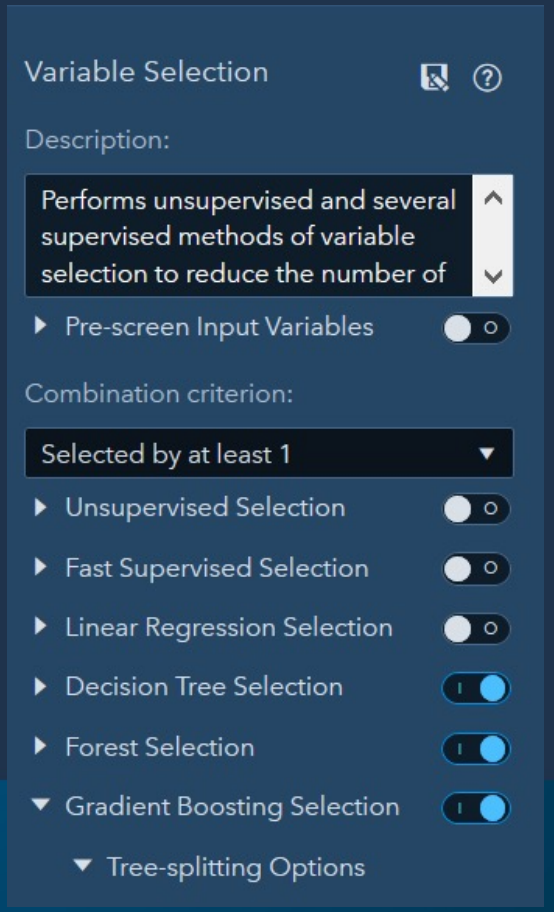
$$s(x) = 2^{\{-h(x)\}}$$

$h(x)$ - средняя (по всем деревьям) длина пути от корня до листа с наблюдением x

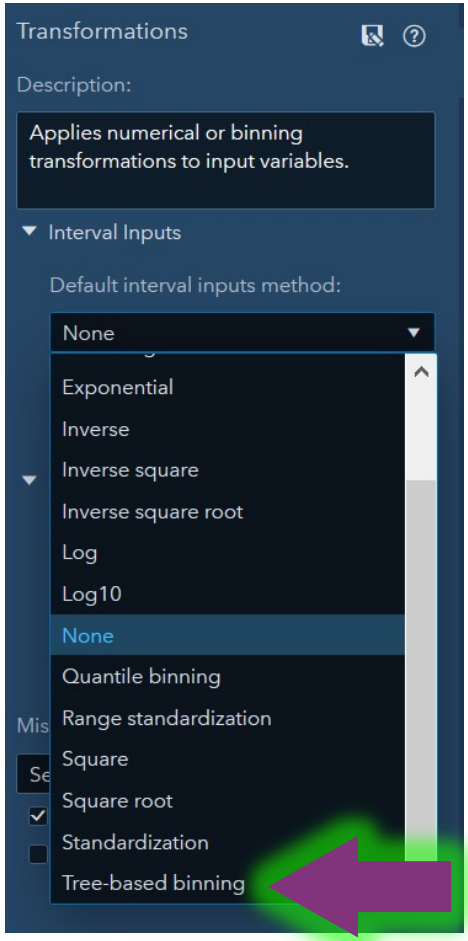


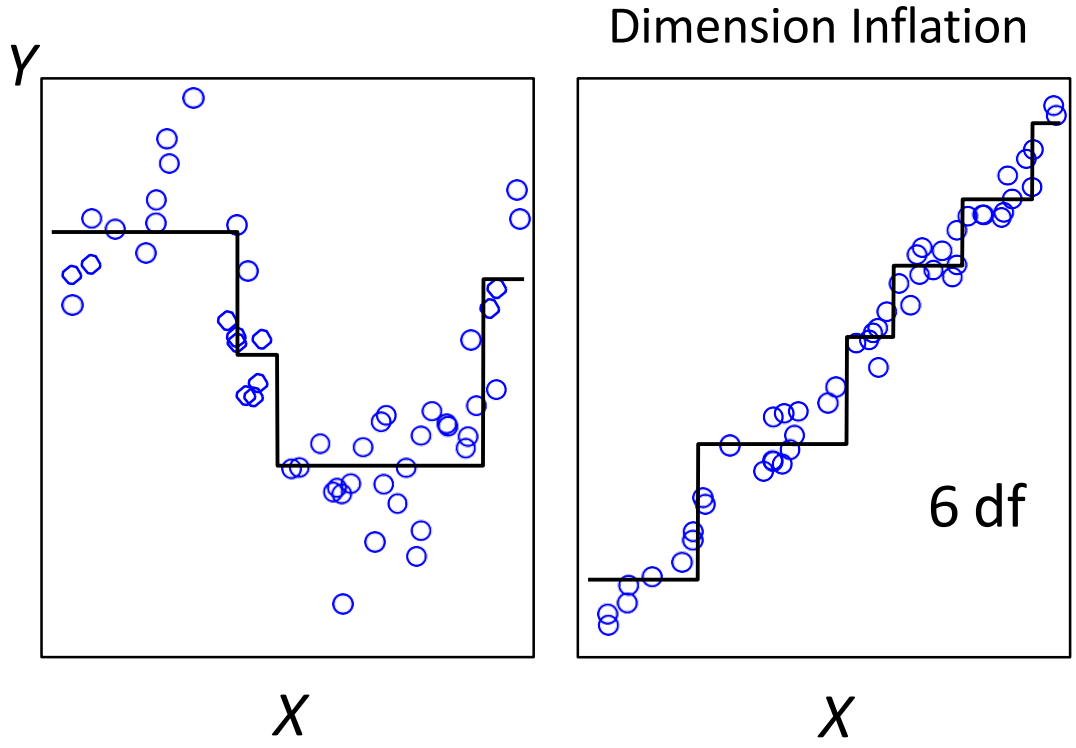
ВЫБОР ВАЖНЫХ / ЗНАЧИМЫХ ПЕРЕМЕННЫХ ДЛЯ ДРУГИХ МЕТОДОВ

Относительная важность



ДИСКРЕТИЗАЦИЯ НЕПРЕРЫВНЫХ ПЕРЕМЕННЫХ





Спасибо!

sas.com