Урок 4. Алгоритм построения дерева решений.

2021

Мария Корлякова

Деревья решений

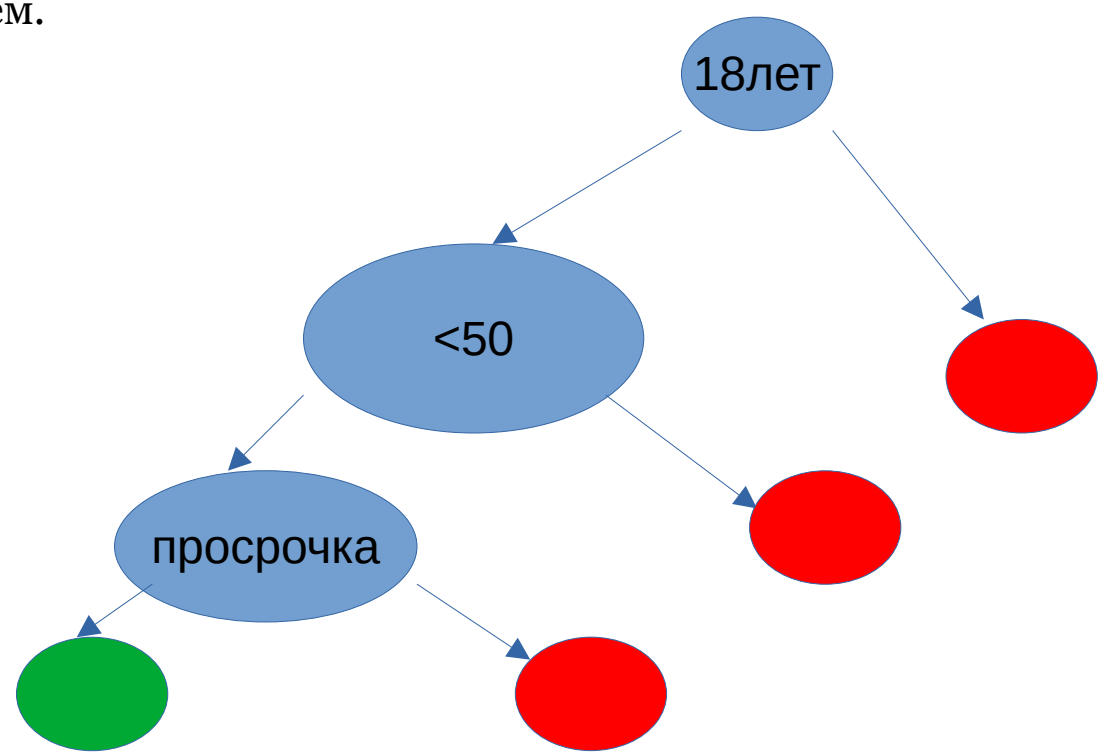
1. Старше ли клиент 18 лет? Если да, то продолжаем, иначе отказываем в кредите.

2. Превышает ли его заработок 50 тысяч рублей? Если да, то продолжаем, иначе отказываем в кредите.

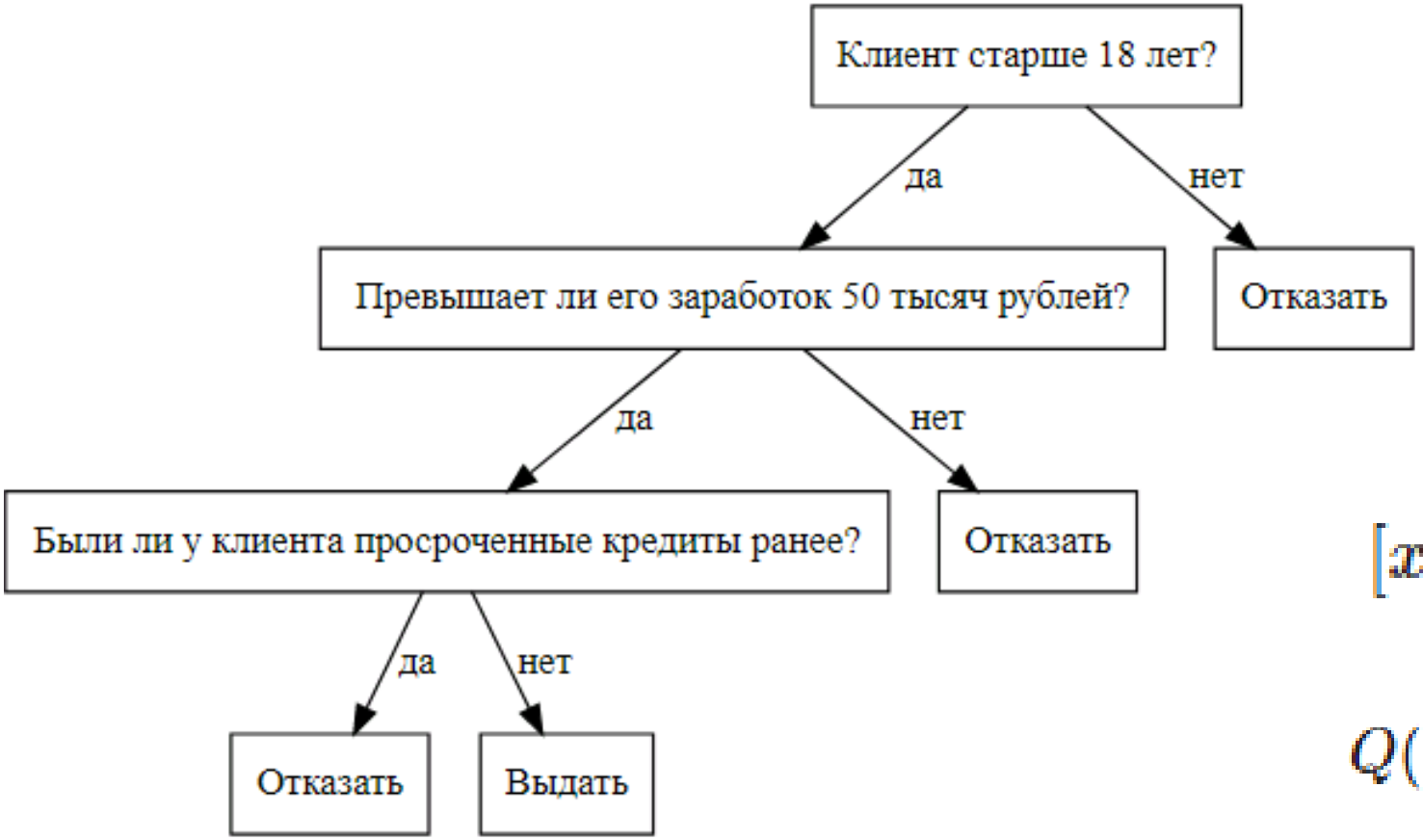
3. Были ли у клиента просроченные кредиты ранее? Если да, отказываем в кредите, иначе выдаем.

Деревья решений

1. Старше ли клиент 18 лет? Если да, то продолжаем, иначе отказываем в кредите.
2. Превышает ли его заработок 50 тысяч рублей? Если да, то продолжаем, иначе отказываем в кредите.
3. Были ли у клиента просроченные кредиты ранее? Если да, отказываем в кредите, иначе выдаем.



Классификация

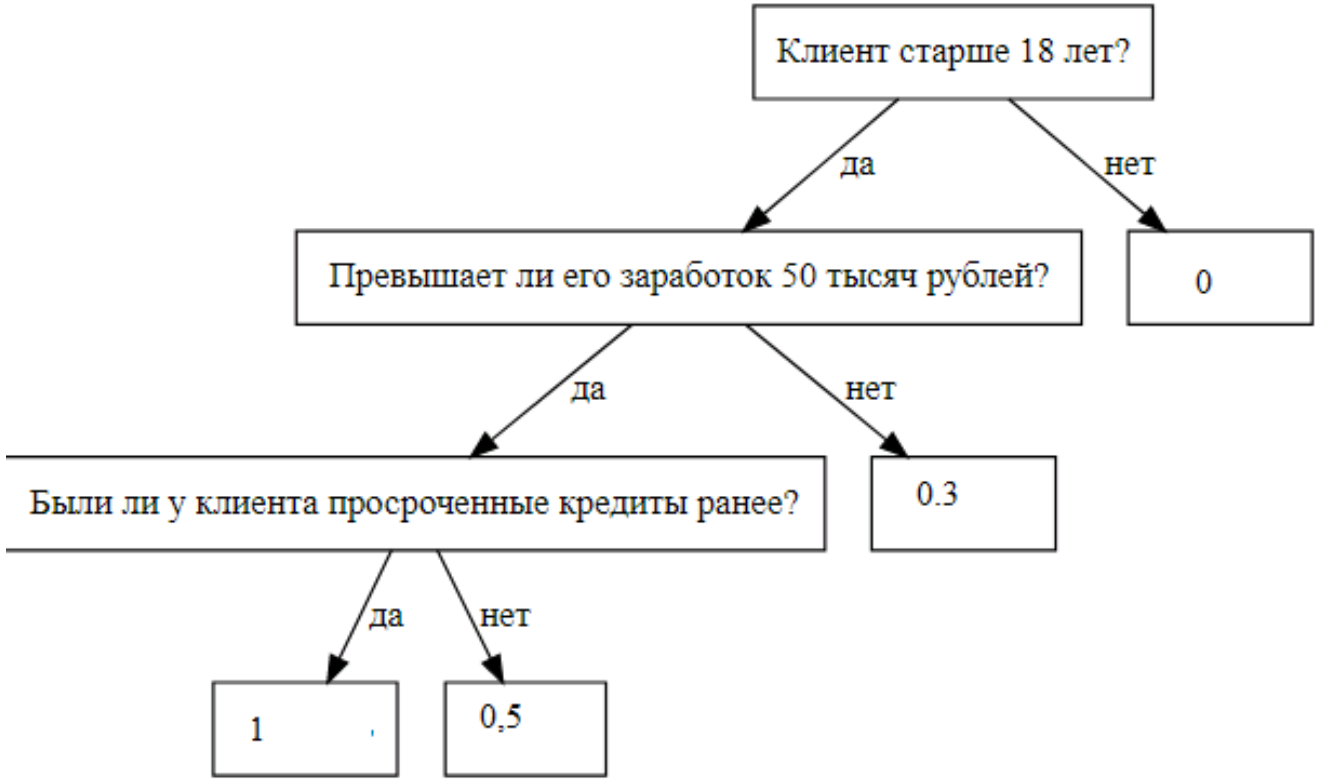


$$[x^j \leq t].$$

$$Q(X, j, t)$$

$$a_m = \operatorname{argmax}_{y \in Y} \sum_{i \in X_m} [y_i = y] \qquad a_{mk} = \frac{1}{|X_m|} \sum_{i \in X_m} [y_i = k].$$

Регрессия



$$[x^j \leq t].$$

$$Q(X, j, t)$$

$$a_m = \frac{1}{|X_m|} \sum_{i \in X_m} y_i.$$

Критерий информативности

Регрессия : $$H(X) = \frac{1}{X}\sum_{i \in X}(y_i - \bar{y}(X))^2,$$

Классификация:

- Вероятность верной классификации $$p_k = \frac{1}{|X|}\sum_{i \in X}[y_i = k].$$

- Джини $$H(X) = \sum_{k=1}^{K} p_k(1 - p_k);$$

- Энтропия Шеннона $$H(X) = -\sum_{k=1}^{K} p_k \log_2 p_k.$$

Критерии останова

- Ограничение максимальной глубины дерева.

- Ограничение максимального количества листьев.

- Ограничение минимального количества объектов в листе.

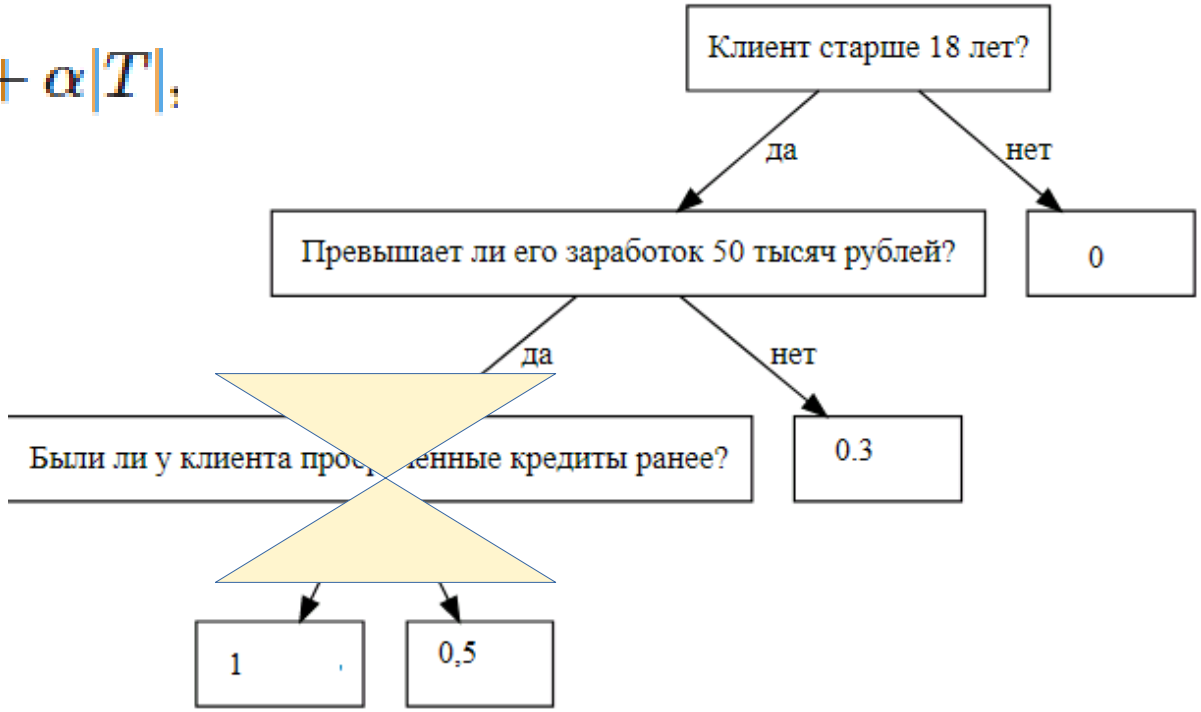
- Останов в случае, когда все объекты в листе относятся к одному классу.

- Требование улучшения функционала качества при разбиении на какую-то минимальную величину.

Обрезка

- Pruning

$$R_\alpha(T) = R(T) + \alpha|T|,$$

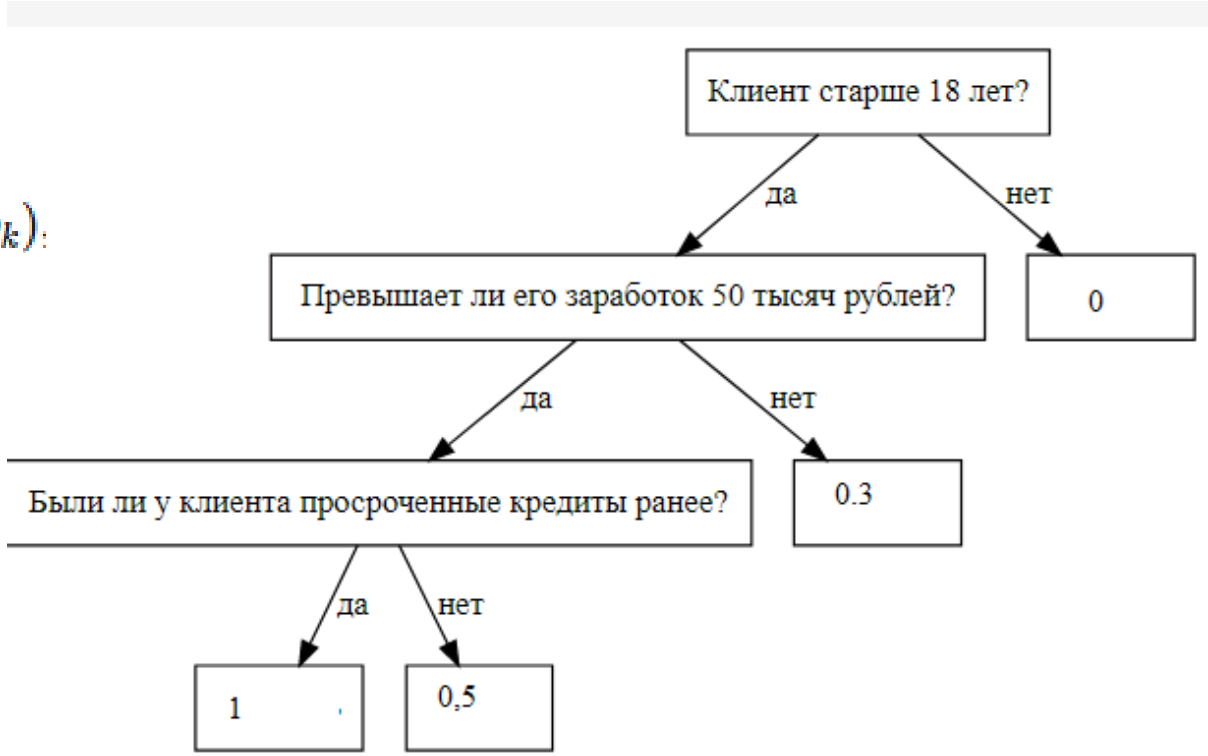


CART (Classification and regression trees)

- ID3
- C4.5

$$H(X) = \sum_{k=1}^{K} p_k(1 - p_k)$$

$$H(X) = 1 - \sum_{k=1}^{K} p_k^2.$$



Регрессия

- Overfiting
- Underfitin