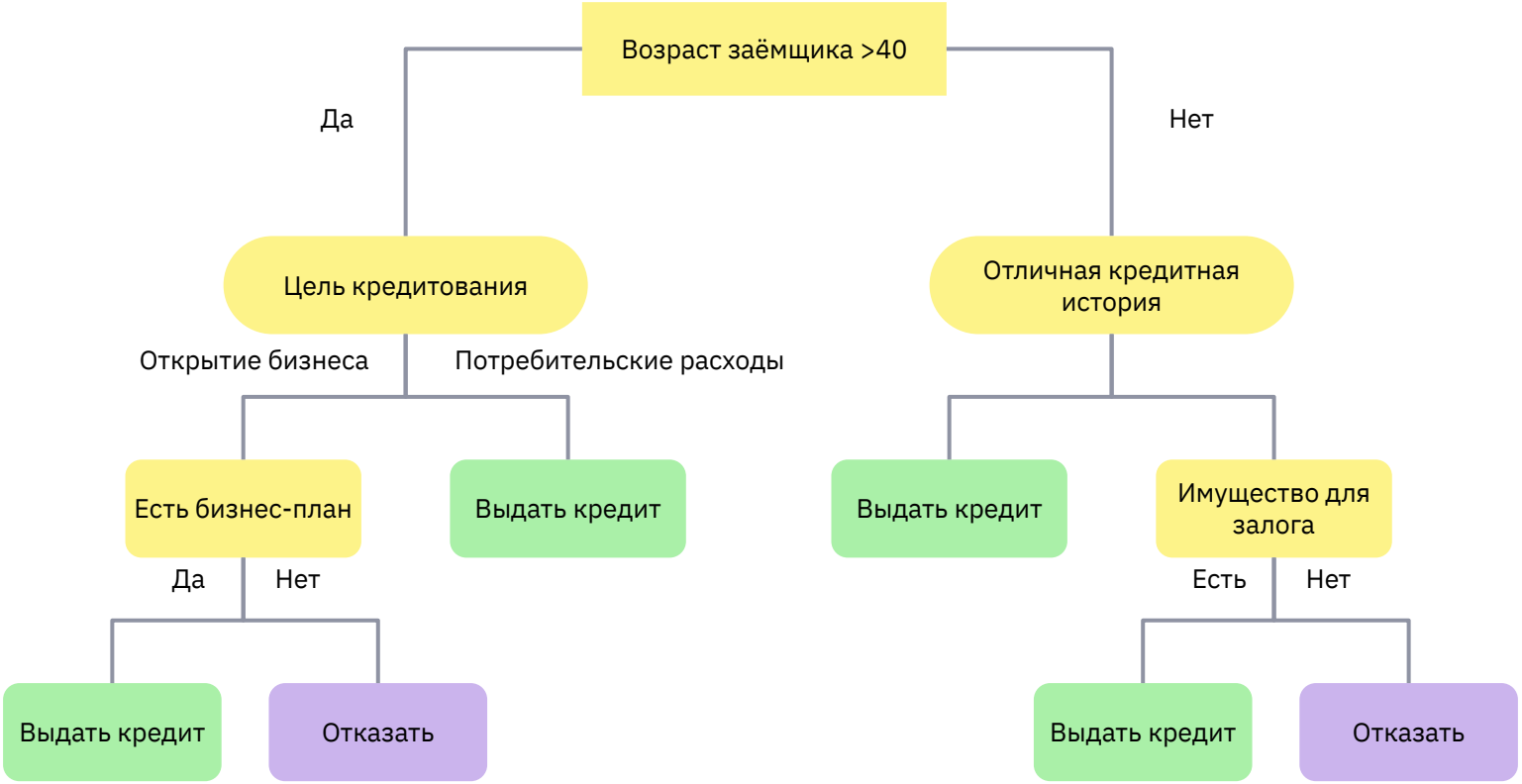
GeekBrains

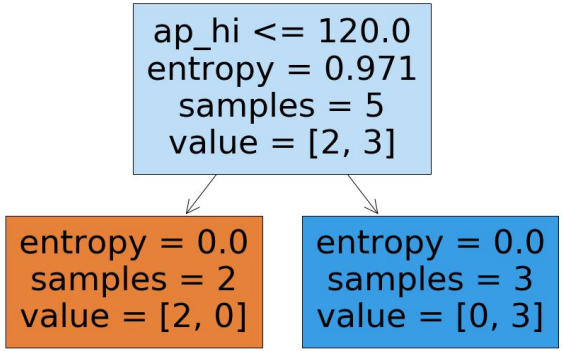
Деревья решений

Что будет на уроке сегодня

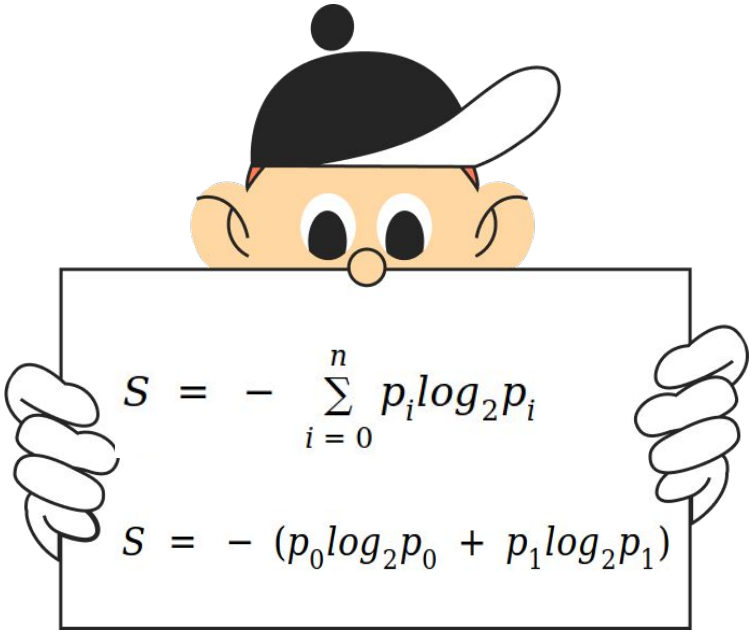
- Что такое дерево решений
- Как обучается модель «Дерево решений»


Возраст заёмщика >40
Да
Нет
Цель кредитования
Отличная кредитная история
Открытие бизнеса
Потребительские расходы
Есть бизнес-план
Выдать кредит
Выдать кредит
Имущество для залога
Да
Нет
Есть
Нет
Выдать кредит
Отказать
Выдать кредит
Отказать

id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	50	2	168	62.0	110	80	1	1	0	0	1	0
1	55	1	156	85.0	140	90	3	1	0	0	1	1
2	51	1	165	64.0	130	70	3	1	0	0	0	1
3	48	2	169	82.0	150	100	1	1	0	0	1	1
4	47	1	156	56.0	100	60	1	1	0	0	0	0



Энтропия Шеннона

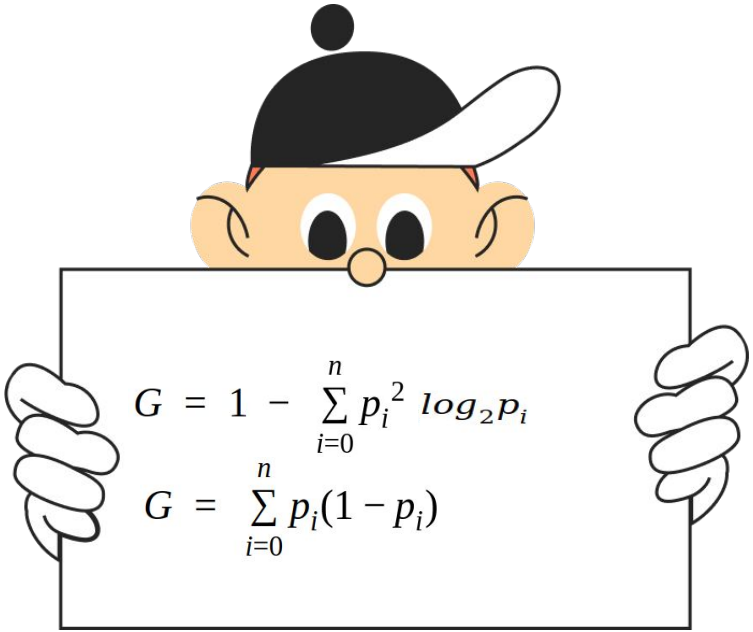


id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	50	2	168	62.0	110	80	1	1	0	0	1	0
1	55	1	156	85.0	140	90	3	1	0	0	1	1
2	51	1	165	64.0	130	70	3	1	0	0	0	1
3	48	2	169	82.0	150	100	1	1	0	0	1	1
4	47	1	156	56.0	100	60	1	1	0	0	0	0

$$S = -(p_0 log_2 p_0 + p_1 log_2 p_1)$$

$$S = -(\frac{2}{5} * log_2 \frac{2}{5} + \frac{3}{5} * log_2 \frac{3}{5}) = 0.97$$

Критерий Джинни



ap_hi <= 120.0
entropy = 0.971
samples = 5
value = [2, 3]

```
df[(df['ap_hi'] <= 120)]
```

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	active	cardio
id									
0	50	2	168	62.0	110	80	1	1	0
4	47	1	156	56.0	100	60	1	0	0

$$S = -\left(\frac{2}{2} * log_2 \frac{2}{2}\right) = 0$$

entropy = 0.0
samples = 2
value = [2, 0]

```
df[(df['ap_hi'] > 120)]
```

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	active	cardio
id									
1	55	1	156	85.0	140	90	3	1	1
2	51	1	165	64.0	130	70	3	0	1
3	48	2	169	82.0	150	100	1	1	1

$$S = -\left(\frac{3}{3} * log_2 \frac{3}{3}\right) = 0$$

entropy = 0.0
samples = 3
value = [0, 3]

age <= 50.0
entropy = 0.971
samples = 5
value = [2, 3]

```
df[df.age <= 50]
```

id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	active	cardio
0	50	2	168	62.0	110	80	1	1	0
3	48	2	169	82.0	150	100	1	1	1
4	47	1	156	56.0	100	60	1	0	0

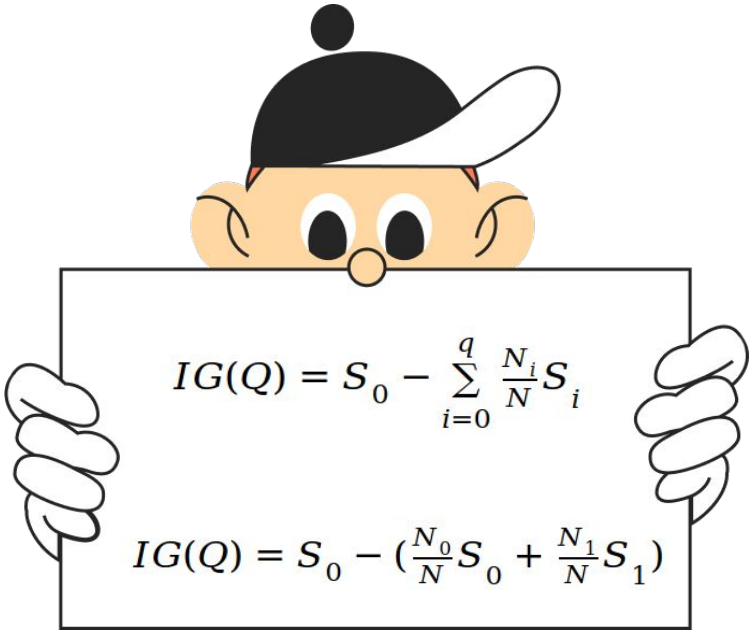
$$S = -\left(\frac{2}{3} * log_2 \frac{2}{3} + \frac{1}{3} * log_2 \frac{1}{3}\right) = 0.918$$

```
df[df.age > 50]
```

id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	active	cardio
1	55	1	156	85.0	140	90	3	1	1
2	51	1	165	64.0	130	70	3	0	1

$$S = -\left(\frac{2}{2} * log_2 \frac{2}{2}\right) = 0$$

Прирост информации (Information gain)



$$IG(Q) = S_0 - \sum_{i=0}^{q} \frac{N_i}{N} S_i$$

$$IG(Q) = S_0 - \left(\frac{N_0}{N} S_0 + \frac{N_1}{N} S_1\right)$$

age <= 50.0
entropy = 0.971
samples = 5
value = [2, 3]

```
df[df.age <= 50]
```

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	active	cardio
id									
0	50	2	168	62.0	110	80	1	1	0
3	48	2	169	82.0	150	100	1	1	1
4	47	1	156	56.0	100	60	1	0	0

$$S = -(\frac{2}{3} * log_2 \frac{2}{3} + \frac{1}{3} * log_2 \frac{1}{3}) = 0.918$$

```
df[df.age > 50]
```

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	active	cardio
id									
1	55	1	156	85.0	140	90	3	1	1
2	51	1	165	64.0	130	70	3	0	1

$$S = -(\frac{2}{2} * log_2 \frac{2}{2}) = 0$$

$$IG(Q) = 0.971 - (\frac{3}{5} * 0.918 + \frac{2}{5} * 0) = 0.42$$

age <= 48.0
entropy = 0.971
samples = 5
value = [2, 3]

```
df[df.age <= 48]
```

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	active	cardio
id									
3	48	2	169	82.0	150	100	1	1	1
4	47	1	156	56.0	100	60	1	0	0

```
df[df.age > 48]
```

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	active	cardio
id									
0	50	2	168	62.0	110	80	1	1	0
1	55	1	156	85.0	140	90	3	1	1
2	51	1	165	64.0	130	70	3	0	1

$$S = -(\frac{1}{2} * log_2 \frac{1}{2} + \frac{1}{2} * log_2 \frac{1}{2}) = 1$$

$$S = -(\frac{1}{3} * log_2 \frac{1}{3} + \frac{2}{3} * log_2 \frac{3}{3}) = 0.918$$

$$IG(Q) = 0.971 - (\frac{2}{5} * 1 + \frac{3}{5} * 0.918) = 0.02$$

weight <= 56.0
entropy = 0.971
samples = 5
value = [2, 3]

```
df[df.weight <= 56]
```

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	active	cardio
id									
4	47	1	156	56.0	100	60	1	0	0

```
df[df.weight > 56]
```

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	active	cardio
id									
0	50	2	168	62.0	110	80	1	1	0
1	55	1	156	85.0	140	90	3	1	1
2	51	1	165	64.0	130	70	3	0	1
3	48	2	169	82.0	150	100	1	1	1

$$S = -(\frac{1}{1} * log_2 \frac{1}{1}) = 0$$

$$S = -(\frac{1}{4} * log_2 \frac{1}{4} + \frac{3}{4} * log_2 \frac{3}{4}) = 0.811$$

$$IG(Q) = 0.971 - (\frac{1}{5} * 0 + \frac{4}{5} * 0.811) = 0.32$$

weight <= 63.0
entropy = 0.971
samples = 5
value = [2, 3]

```
df[df.weight <= 63]
```

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	active	cardio
id									
0	50	2	168	62.0	110	80	1	1	0
4	47	1	156	56.0	100	60	1	0	0

$$S = -\left(\frac{2}{2} * log_2 \frac{2}{2}\right) = 0$$

```
df[df.weight > 63]
```

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	active	cardio
id									
1	55	1	156	85.0	140	90	3	1	1
2	51	1	165	64.0	130	70	3	0	1
3	48	2	169	82.0	150	100	1	1	1

$$S = -\left(\frac{3}{3} * log_2 \frac{3}{3}\right) = 0$$

$$IG(Q) = 0.971 - \left(\frac{2}{5} * 0 + \frac{3}{5} * 0\right) = 0.971$$

age <= 50

$$IG(Q) = 0.971 - (\frac{3}{5} * 0.918 + \frac{2}{5} * 0) = 0.42$$

age <= 48

$$IG(Q) = 0.971 - (\frac{2}{5} * 1 + \frac{3}{5} * 0.918) = 0.02$$

weight <= 56

$$IG(Q) = 0.971 - (\frac{1}{5} * 0 + \frac{4}{5} * 0.811) = 0.32$$

weight <= 63

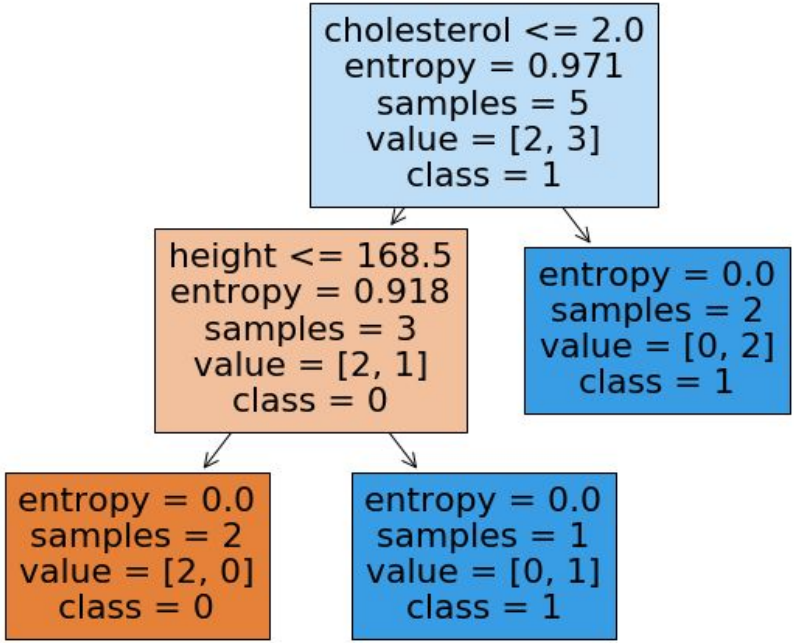
$$IG(Q) = 0.971 - (\frac{2}{5} * 0 + \frac{3}{5} * 0) = 0.971$$

Решающие деревья: критерии останова

- В вершине один объект
- В вершине объекты одного класса
- В вершину попало < n объектов
- Глубина превысила порог

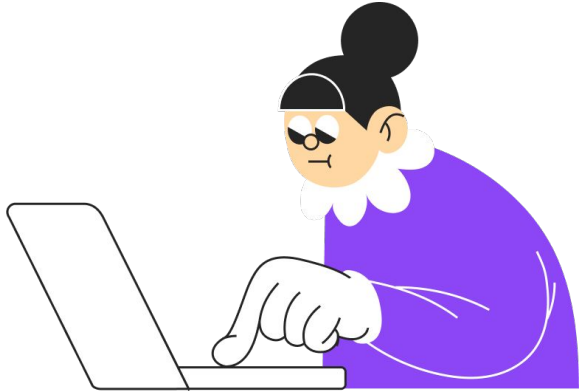
Решающие деревья: ответ

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	active	cardio
id									
0	50	2	168	62.0	110	80	1	1	0
4	47	1	156	56.0	100	60	1	0	0



Что мы сегодня узнали и чему научились

- Что такое дерево решений
- Как обучается модель «Дерево решений»

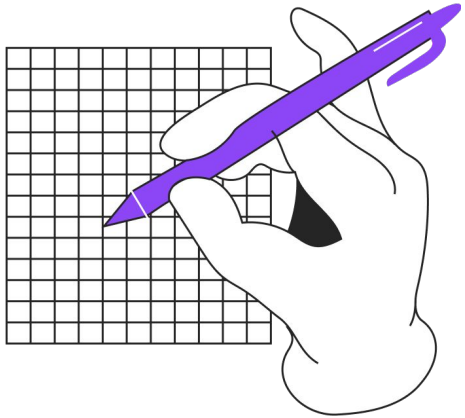


Дополнительные материалы

- Дерево решений для задачи регрессии
- Дерево решений для задачи классификации
- Критерии останова для дерева решений
- Метрика accuracy
- Переобучение/недообучение

Практическое задание

Практика



Вопросы?

Вопросы?

Вопросы?

Вопросы?