Използване на силата на категорийните характеристики: Въведение в CatBoost

Изготвил: Симеон Христов

Общи понятия

Статия: [CatBoost: unbiased boosting with categorical features](#).

Според статията CatBoost е “a new gradient boosting toolkit”.

Името е акроним за Categorical Boosting и подчертава колко важна е обработката на категорийните характеристики за този алгоритъм.

Table 8: Comparison with baselines: logloss / zero-one loss, relative increase is presented in the brackets.

	CatBoost	LightGBM	XGBoost
Adult	**0.2695 / 0.1267**	0.2760 (+2.4%) / 0.1291 (+1.9%)	0.2754 (+2.2%) / 0.1280 (+1.0%)
Amazon	**0.1394 / 0.0442**	0.1636 (+17%) / 0.0533 (+21%)	0.1633 (+17%) / 0.0532 (+21%)
Click	**0.3917 / 0.1561**	0.3963 (+1.2%) / 0.1580 (+1.2%)	0.3962 (+1.2%) / 0.1581 (+1.2%)
Epsilon	**0.2647 / 0.1086**	0.2703 (+1.5%) / 0.114 (+4.1%)	0.2993 (+11%) / 0.1276 (+12%)
Appetency	**0.0715 / 0.01768**	0.0718 (+0.4%) / 0.01772 (+0.2%)	0.0718 (+0.4%) / 0.01780 (+0.7%)
Churn	**0.2319 / 0.0719**	0.2320 (+0.1%) / 0.0723 (+0.6%)	0.2331 (+0.5%) / 0.0730 (+1.6%)
Internet	**0.2089 / 0.0937**	0.2231 (+6.8%) / 0.1017 (+8.6%)	0.2253 (+7.9%) / 0.1012 (+8.0%)
Upselling	**0.1662 / 0.0490**	0.1668 (+0.3%) / 0.0491 (+0.1%)	0.1663 (+0.04%) / 0.0492 (+0.3%)
Kick	**0.2855 / 0.0949**	0.2957 (+3.5%) / 0.0991 (+4.4%)	0.2946 (+3.2%) / 0.0988 (+4.1%)

Теч на данни (Data Leakage)

Data Science No-No

Може да означава много ситуации.

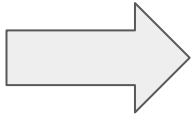
Ще го разгледаме в контекста на кодиране на променливи, които:

- се кодират спрямо стойността на целевата променлива
- и след това тези нови стойности се използват за нейното предсказване (предсказването на целевата променлива).

Трябва да избягваме теча на данни, т.к. води до пренагаждане (overfitting)!

One-Hot Encoding

Любим Цвят	Височина (м.)	Изплатил Кредит
Син	1.77	Да
Червен	1.32	Не
Зелен	1.81	Да
Син	1.56	Не
Зелен	1.64	Да
Зелен	1.61	Не
Син	1.73	Не



Любим Цвят Синьо	Любим Цвят Червен	Любим Цвят Зелен	Височина (м.)	Изплатил Кредит
1	0	0	1.77	1
0	1	0	1.32	0
0	0	1	1.81	1
1	0	0	1.56	0
0	0	1	1.64	1
0	0	1	1.61	0
1	0	0	1.73	0

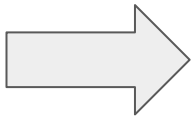
Използва се в XGBoost.

Проблеми:

- Голяма размерност при висока кардиналност
- Линейна зависимост: `Любим Цвят Синьо + Любим Цвят Червен + Любим Цвят Зелен = 1` (multicollinearity)

Label Encoding

Любим Цвят	Височина (м.)	Изплатил Кредит
Син	1.77	Да
Червен	1.32	Не
Зелен	1.81	Да
Син	1.56	Не
Зелен	1.64	Да
Зелен	1.61	Не
Син	1.73	Не



Любим Цвят	Височина (м.)	Изплатил Кредит
0	1.77	1
1	1.32	0
2	1.81	1
0	1.56	0
2	1.64	1
2	1.61	0
0	1.73	0

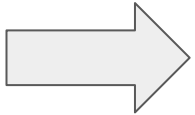
Решава проблемите на One-Hot Encoding.

Нови проблеми:

- Кой избира тези числа? Ние, т.е. кодирането е субективно и случайно.
- ”Фалшива” наредба:
  - Дърво на решенията с корен (`любим цвят < 0.5`) ще трябва да групира два класа.

Target Encoding

Любим Цвят	Височина (м.)	Изплатил Кредит
Син	1.77	1
Червен	1.32	0
Зелен	1.81	1
Син	1.56	0
Зелен	1.64	1
Зелен	1.61	0
Син	1.73	0



Любим Цвят	Височина (м.)	Изплатил Кредит
0.33	1.77	1
0	1.32	0
0.67	1.81	1
0.33	1.56	0
0.67	1.64	1
0.67	1.61	0
0.33	1.73	0

За да не избираме субективно числата, можем да използваме средната стойност на целевата променлива, за всяка от категориите (оттук идва и наименованието **Target Encoding**):

- средно за **Син** : 1 / 3 => 0.33
- средно за **Червен**: 0 / 3 => 0
- средно за **Зелен** : 2 / 3 => 0.67

Нов проблем: Уместността на числото е правопропорционална на броя наблюдения.

Weighted Target Encoding

$$\text{Weighted Mean} = \frac{n \times \text{Option Mean} + m \times \text{Overall Mean}}{n + m}$$

Любим Цвят	Височина (м.)	Изплатил Кредит
0.37	1.77	1
0.29	1.32	0
0.57	1.81	1
0.37	1.56	0
0.57	1.64	1
0.57	1.61	0
0.37	1.73	0

Син $\text{Weighted Mean} = \frac{3 \times 1/3 + 2 \times 3/7}{3 + 2} = 0.37$

Червен $\text{Weighted Mean} = \frac{1 \times 0/1 + 2 \times 3/7}{1 + 2} = 0.29$

Зелен $\text{Weighted Mean} = \frac{3 \times 2/3 + 2 \times 3/7}{3 + 2} = 0.57$

За да се справим с този проблем, използваме **претеглена средна стойност**.

`n` = брой наблюдения с тази стойност за категорията (напр. за **Червен**: `n = 1`).
`Option Mean` = средна стойност на целевата променлива за категорията.
`m` = хиперпараметър. Нека тук `m = 2`. `Option Mean` е ”по-важна” от `Overall Mean`, ако `n > 2`.
`Overall Mean` = средна стойност на целевата променлива.

Weighted Target Encoding

Така постигаме стойности, по-близки до средната (това зависи от m).

Имаме ли теч на данни?

Любим Цвят	Височина (м.)	Изплатил Кредит
0.37	1.77	1
0.29	1.32	0
0.57	1.81	1
0.37	1.56	0
0.57	1.64	1
0.57	1.61	0
0.37	1.73	0

Любим Цвят	Височина (м.)	Изплатил Кредит
0.33	1.77	1
0	1.32	0
0.67	1.81	1
0.33	1.56	0
0.67	1.64	1
0.67	1.61	0
0.33	1.73	0

K-Fold Target Encoding

`k = 2, m = 2`

$$\text{Weighted Mean} = \frac{n \times \text{Option Mean} + m \times \text{Overall Mean}}{n + m}$$

	Любим Цвят	Височина (м.)	Изплатил Кредит
А	0.22	1.77	1
	0.33	1.32	0
	0.42	1.81	1
	0.22	1.56	0
Б	0.42	1.64	1
	0.42	1.61	0
	0.5	1.73	0

Син (А)

$$\text{Weighted Mean} = \frac{1 \times 0/1 + 2 \times 1/3}{1 + 2} = 0.22$$

Син (Б)

$$\text{Weighted Mean} = \frac{2 \times 1/2 + 2 \times 2/4}{2 + 2} = 0.5$$

K-Fold Target Encoding

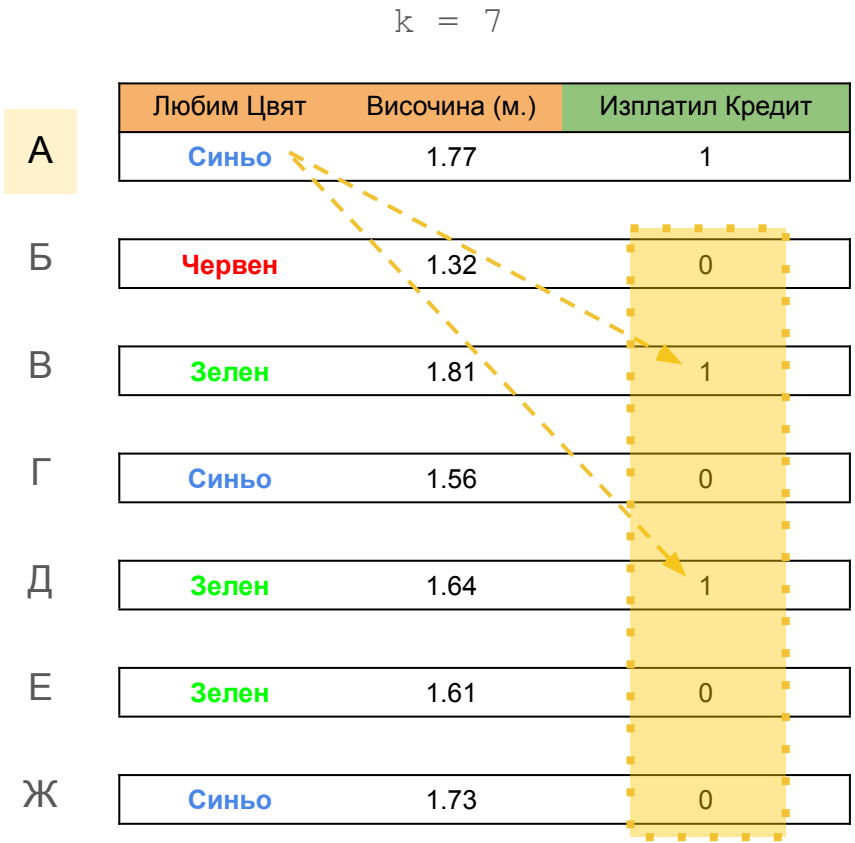
`k = 2, m = 2`

Любим Цвят	Височина (м.)	Изплатил Кредит
0.22	1.77	1
0.33	1.32	0
0.42	1.81	1
0.22	1.56	0
0.42	1.64	1
0.42	1.61	0
0.22	1.73	0

Резултати:

1. `Любим Цвят` вече е непрекъсната (числова) променлива.
2. **Намален** теч на данни, т.к. за кодирането на стойностите се използват данни от други наблюдения.

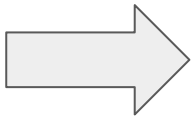
Leave-One-Out Target Encoding



Какво става, ако имаме само една стойност?

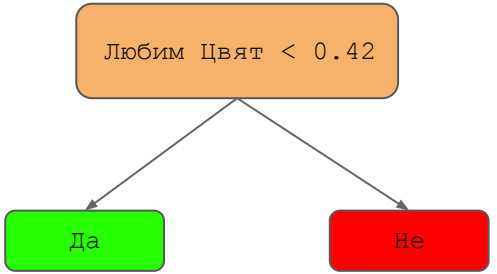
`k = 7, m = 2`

	Любим Цвят	Височина (м.)	Изплатил Кредит
А	Синьо	1.77	1
Б	Синьо	1.32	0
В	Синьо	1.81	1
Г	Синьо	1.56	0
Д	Синьо	1.64	1
Е	Синьо	1.61	0
Ж	Синьо	1.73	0



Любим Цвят	Височина (м.)	Изплатил Кредит
0.33	1.77	1
0.5	1.32	0
0.33	1.81	1
0.5	1.56	0
0.33	1.64	1
0.5	1.61	0
0.5	1.73	0

Теч на данни!
:(

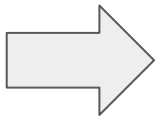


Ordered Target Encoding

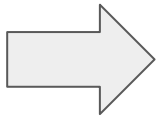
Нов метод, който е в основата на CatBoost.

CatBoost третира наблюденията като идващи *последователно*.

Любим Цвят	Височина (м.)	Изплатил Кредит
Син	1.77	Да
Червен	1.32	Не
Зелен	1.81	Да
Син	1.56	Не
Зелен	1.64	Да
Зелен	1.61	Не
Син	1.73	Не



Любим Цвят	Височина (м.)	Изплатил Кредит
Син	1.77	Да
Червен	1.32	Не
Зелен	1.81	Да
Син	1.56	Не
Зелен	1.64	Да
Зелен	1.61	Не
Син	1.73	Не



Любим Цвят	Височина (м.)	Изплатил Кредит
Син	1.77	Да
Червен	1.32	Не
Зелен	1.81	Да
Син	1.56	Не
Зелен	1.64	Да
Зелен	1.61	Не
Син	1.73	Не

. . .

Ordered Target Encoding

Любим Цвят	Височина (м.)	Изплатил Кредит
Син	1.77	1
Червен	1.32	0
Зелен	1.81	1
Син	1.56	0
Зелен	1.64	1
Зелен	1.61	0
Син	1.73	0

$$\text{CatBoost Encoding} = \frac{\text{OptionCount} + 0.05}{n + 1}$$

`n` = брой **получени** наблюдения с тази стойност за категорията.

`OptionCount` = брой **получени** наблюдения с тази стойност за категорията и 1 за целевата променлива.

Ordered Target Encoding

Любим Цвят	Височина (м.)	Изплатил Кредит
0.05	1.77	1
Червен	1.32	0
Зелен	1.81	1
Син	1.56	0
Зелен	1.64	1
Зелен	1.61	0
Син	1.73	0

$$\text{CatBoost Encoding} = \frac{\text{OptionCount} + 0.05}{n + 1}$$

$$\text{CatBoost Encoding} = \frac{0 + 0.05}{0 + 1} = 0.05$$

`n` = брой **предишни** наблюдения с тази стойност за категорията.

`OptionCount` = брой **предишни** наблюдения с тази стойност за категорията и 1 за целевата променлива.

Ordered Target Encoding

Любим Цвят	Височина (м.)	Изплатил Кредит
0.05	1.77	1
0.05	1.32	0
Зелен	1.81	1
Син	1.56	0
Зелен	1.64	1
Зелен	1.61	0
Син	1.73	0

$$\text{CatBoost Encoding} = \frac{\text{OptionCount} + 0.05}{n + 1}$$

$$\text{CatBoost Encoding} = \frac{0 + 0.05}{0 + 1} = 0.05$$

`n` = брой **предишни** наблюдения с тази стойност за категорията.

`OptionCount` = брой **предишни** наблюдения с тази стойност за категорията и 1 за целевата променлива.

Ordered Target Encoding

Любим Цвят	Височина (м.)	Изплатил Кредит
0.05	1.77	1
0.05	1.32	0
0.05	1.81	1
Син	1.56	0
Зелен	1.64	1
Зелен	1.61	0
Син	1.73	0

$$\text{CatBoost Encoding} = \frac{\text{OptionCount} + 0.05}{n + 1}$$

$$\text{CatBoost Encoding} = \frac{0 + 0.05}{0 + 1} = 0.05$$

`n` = брой **предишни** наблюдения с тази стойност за категорията.

`OptionCount` = брой **предишни** наблюдения с тази стойност за категорията и 1 за целевата променлива.

Ordered Target Encoding

Любим Цвят	Височина (м.)	Изплатил Кредит
0.05	1.77	1
0.05	1.32	0
0.05	1.81	1
0.525	1.56	0
Зелен	1.64	1
Зелен	1.61	0
Син	1.73	0

$$\text{CatBoost Encoding} = \frac{\text{OptionCount} + 0.05}{n + 1}$$

$$\text{CatBoost Encoding} = \frac{1 + 0.05}{1 + 1} = 0.525$$

`n` = брой **предишни** наблюдения с тази стойност за категорията.

`OptionCount` = брой **предишни** наблюдения с тази стойност за категорията и 1 за целевата променлива.

Ordered Target Encoding

Любим Цвят	Височина (м.)	Изплатил Кредит
0.05	1.77	1
0.05	1.32	0
0.05	1.81	1
0.525	1.56	0
0.525	1.64	1
Зелен	1.61	0
Син	1.73	0

$$\text{CatBoost Encoding} = \frac{\text{OptionCount} + 0.05}{n + 1}$$

$$\text{CatBoost Encoding} = \frac{1 + 0.05}{1 + 1} = 0.525$$

`n` = брой **предишни** наблюдения с тази стойност за категорията.

`OptionCount` = брой **предишни** наблюдения с тази стойност за категорията и 1 за целевата променлива.

Ordered Target Encoding

Любим Цвят	Височина (м.)	Изплатил Кредит
0.05	1.77	1
0.05	1.32	0
0.05	1.81	1
0.525	1.56	0
0.525	1.64	1
0.683	1.64	0
Син	1.73	0

$$\text{CatBoost Encoding} = \frac{\text{OptionCount} + 0.05}{n + 1}$$

$$\text{CatBoost Encoding} = \frac{2 + 0.05}{2 + 1} = 0.683$$

`n` = брой **предишни** наблюдения с тази стойност за категорията.

`OptionCount` = брой **предишни** наблюдения с тази стойност за категорията и 1 за целевата променлива.

Ordered Target Encoding

Любим Цвят	Височина (м.)	Изплатил Кредит
0.05	1.77	1
0.05	1.32	0
0.05	1.81	1
0.525	1.56	0
0.525	1.64	1
0.683	1.64	0
0.35	1.73	0

$$\text{CatBoost Encoding} = \frac{\text{OptionCount} + 0.05}{n + 1}$$

$$\text{CatBoost Encoding} = \frac{1 + 0.05}{2 + 1} = 0.35$$

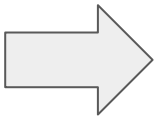
`n` = брой **предишни** наблюдения с тази стойност за категорията.

`OptionCount` = брой **предишни** наблюдения с тази стойност за категорията и 1 за целевата променлива.

Catboost

При всяко създаване на дърво CatBoost
първо разбърква наблюденията.

Любим Цвят	Височина (м.)
Син	1.56
Червен	1.32
Зелен	1.81
Син	1.77
Зелен	1.64



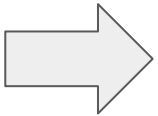
Любим Цвят	Височина (м.)
Зелен	1.81
Син	1.56
Син	1.77
Червен	1.32
Зелен	1.64

Catboost

След това прилага Ordered Target Encoding върху всяка характеристика с поне 3 различни стойности.

- Категории с по-малко от 3 различни стойности се кодират с 0 и 1.
- Ако предсказваме непрекъсната променлива, то тя се дискретизира.

Любим Цвят	Височина (м.)	Интервал
Зелен	1.81	1
Син	1.56	0
Син	1.77	1
Червен	1.32	0
Зелен	1.64	1



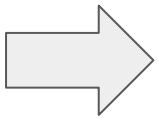
Любим Цвят	Височина (м.)	Интервал
0.05	1.81	1
0.05	1.56	0
0.025	1.77	1
0.05	1.32	0
0.525	1.64	1

Catboost

Колоната **Интервал** е необходима само за кодирането, т.ч. ще я скрием.
Добавят се две колони:

- **Предсказание**: Резултат от спускане по дървото. По подразбиране - `0`.
- **Грешка** (residual): `Истинска стойност - Предсказана стойност`

Любим Цвят	Височина (м.)	Интервал
0.05	1.81	1
0.05	1.56	0
0.025	1.77	1
0.05	1.32	0
0.525	1.64	1



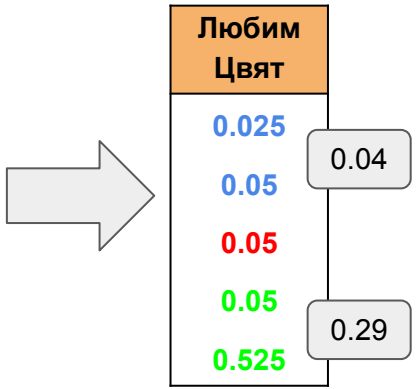
Любим Цвят	Височина (м.)	Предсказание	Грешка
0.05	1.81	0	1.81
0.05	1.56	0	1.56
0.025	1.77	0	1.77
0.05	1.32	0	1.32
0.525	1.64	0	1.64

Catboost

Строим пън на база `Любим Цвят`.

За улеснение тук, ще създадем само пън и като крайно дърво.

Любим Цвят	Височина (м.)	Предсказание	Грешка
0.05	1.81	0	1.81
0.05	1.56	0	1.56
0.025	1.77	0	1.77
0.05	1.32	0	1.32
0.525	1.64	0	1.64

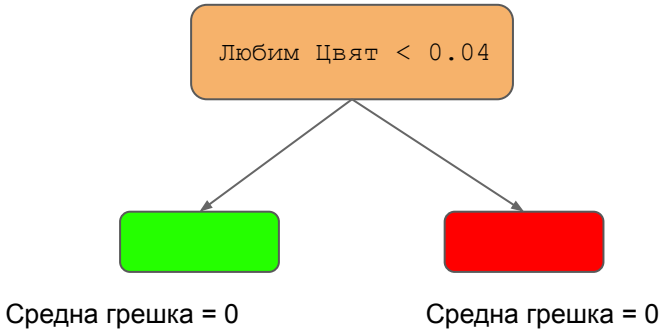


Catboost

Добавяме още една колона **Средна грешка**, която съхранява средната грешка на листо при добавяне на даден пример в него.

Средната грешка на листо е сумата на грешките в това листо, разделена на броя им.

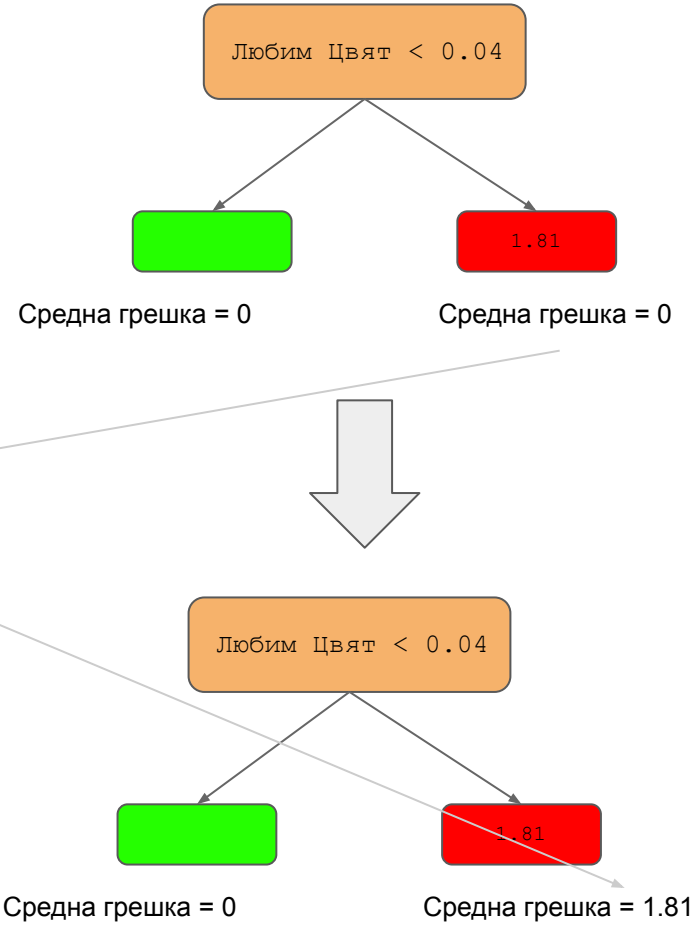
Любим Цвят	Височина (м.)	Предсказание	Грешка	Средна грешка
0.05	1.81	0	1.81	0
0.05	1.56	0	1.56	0
0.025	1.77	0	1.77	0
0.05	1.32	0	1.32	0
0.525	1.64	0	1.64	0



Catboost

Новата средна грешка за листото се пресмята след като старата стойност се запише в таблицата.

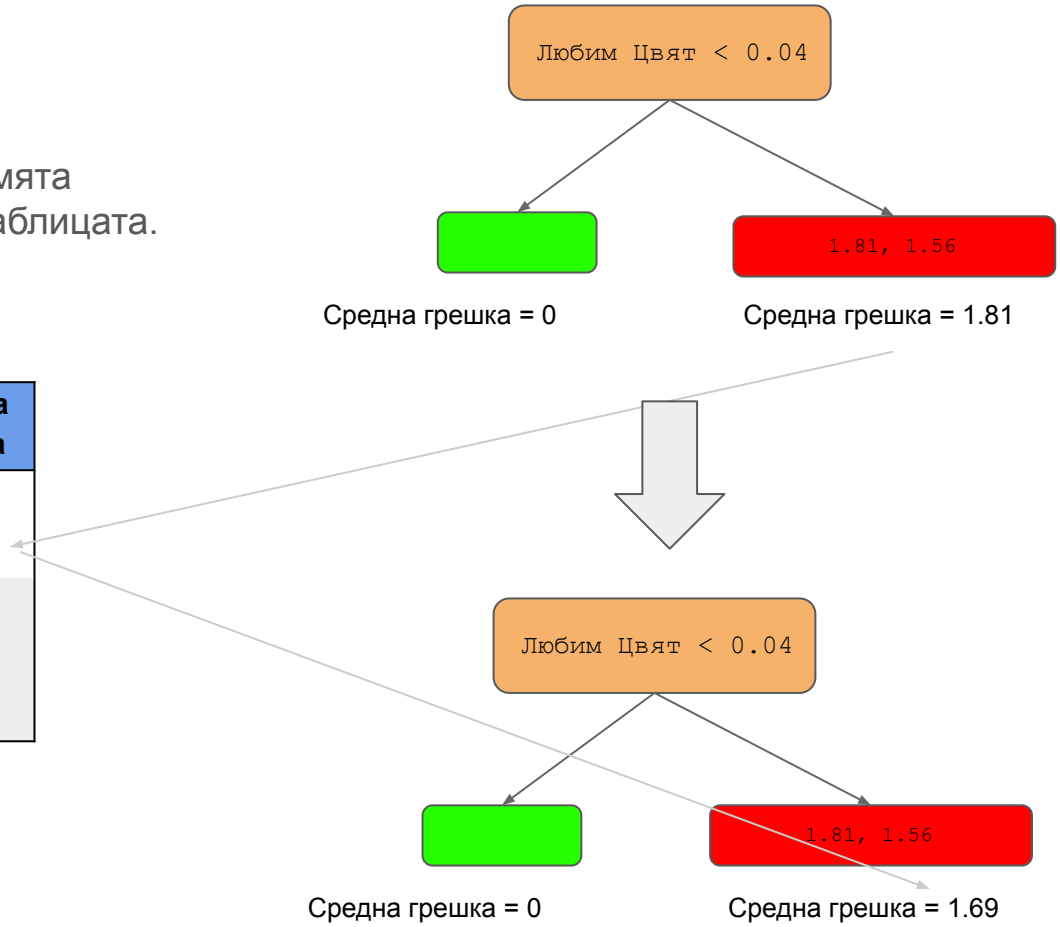
Любим Цвят	Височина (м.)	Предсказание	Грешка	Средна грешка
0.05	1.81	0	1.81	0
0.05	1.56	0	1.56	0
0.025	1.77	0	1.77	0
0.05	1.32	0	1.32	0
0.525	1.64	0	1.64	0



Catboost

Новата средна грешка за листото се пресмята след като старата стойност се запише в таблицата.

Любим Цвят	Височина (м.)	Предсказание	Грешка	Средна грешка
0.05	1.81	0	1.81	0
0.05	1.56	0	1.56	1.81
0.025	1.77	0	1.77	0
0.05	1.32	0	1.32	0
0.525	1.64	0	1.64	0



Catboost

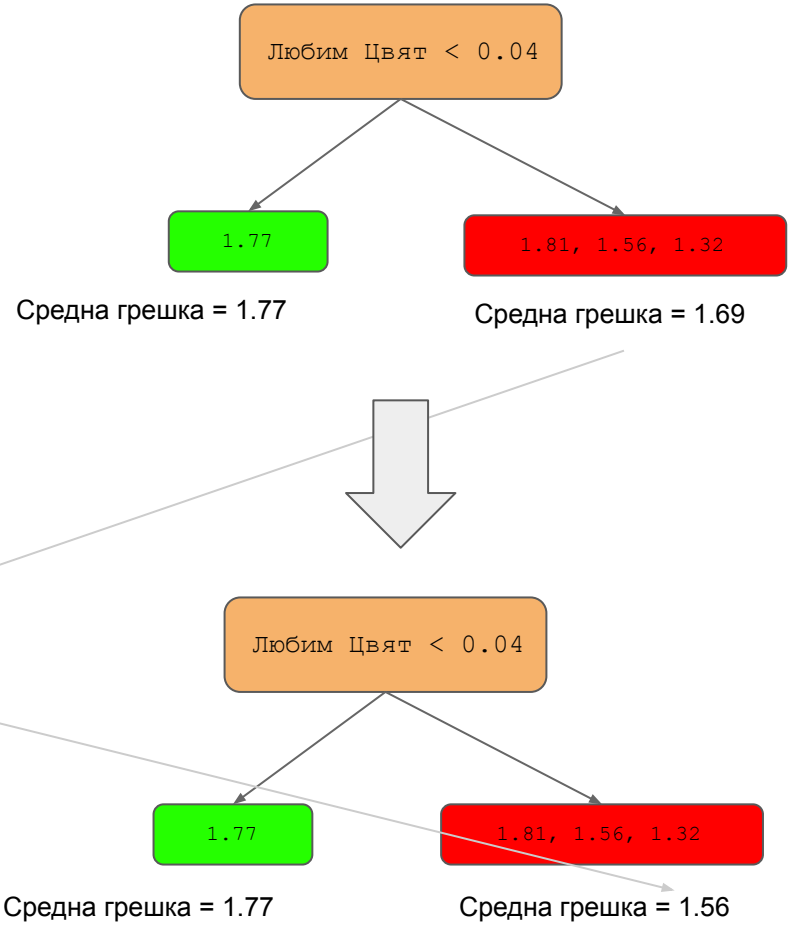
Новата средна грешка за листото се пресмята след като старата стойност се запише в таблицата.

Любим Цвят	Височина (м.)	Предсказание	Грешка	Средна грешка
0.05	1.81	0	1.81	0
0.05	1.56	0	1.56	1.81
0.025	1.77	0	1.77	0
0.05	1.32	0	1.32	0
0.525	1.64	0	1.64	0

Catboost

Новата средна грешка за листото се пресмята след като старата стойност се запише в таблицата.

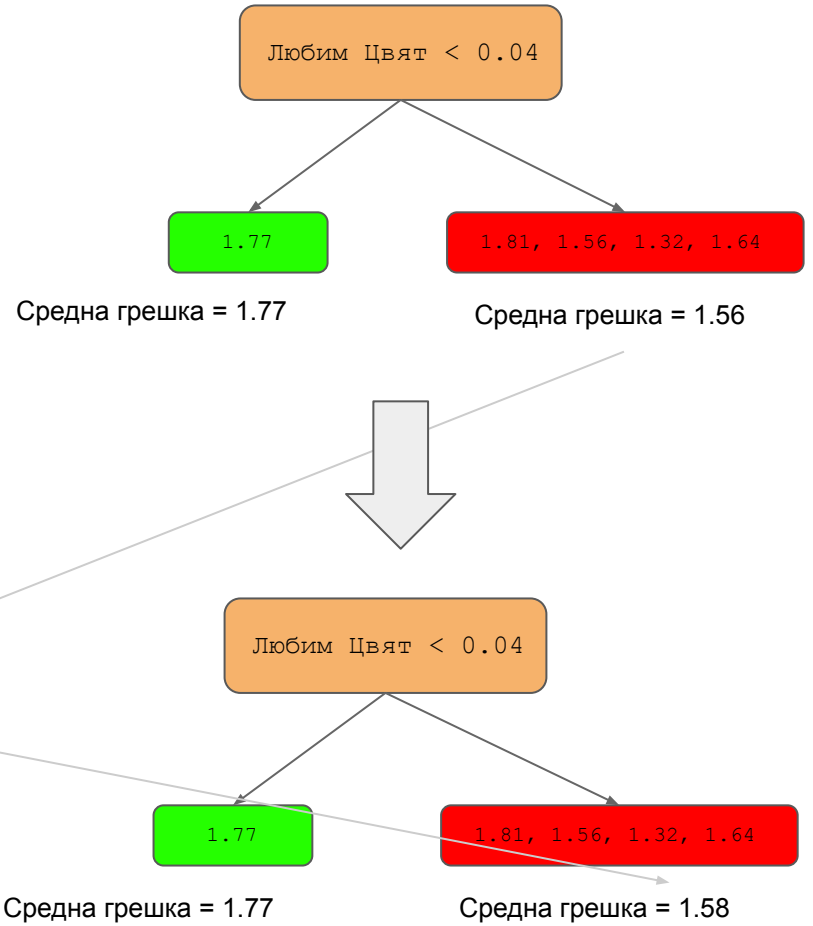
Любим Цвят	Височина (м.)	Предсказание	Грешка	Средна грешка
0.05	1.81	0	1.81	0
0.05	1.56	0	1.56	1.81
0.025	1.77	0	1.77	0
0.05	1.32	0	1.32	1.69
0.525	1.64	0	1.64	0



Catboost

Новата средна грешка за листото се пресмята след като старата стойност се запише в таблицата.

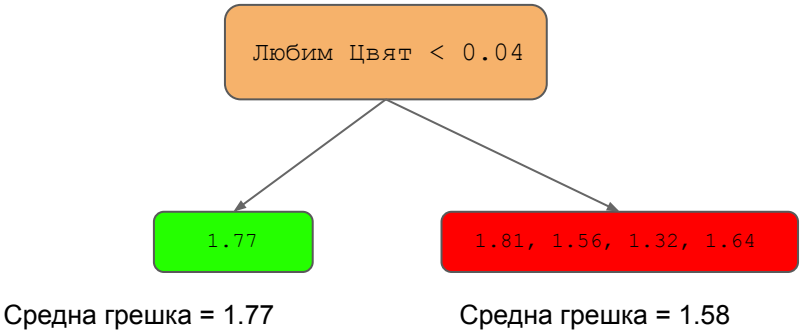
Любим Цвят	Височина (м.)	Предсказание	Грешка	Средна грешка
0.05	1.81	0	1.81	0
0.05	1.56	0	1.56	1.81
0.025	1.77	0	1.77	0
0.05	1.32	0	1.32	1.69
0.525	1.64	0	1.64	1.56



Catboost

Използва се косинусова близост на двете грешки за изчисляване доколко това разделяне е оптимално.

Косинусова Близост $= \dfrac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$

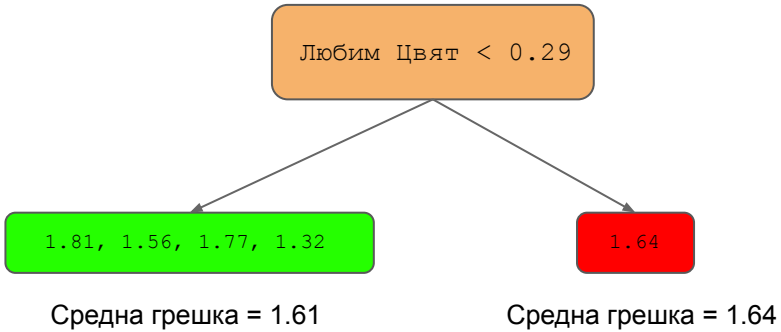


			A	B
Любим Цвят	Височина (м.)	Предсказание	Грешка	Средна грешка
0.05	1.81	0	1.81	0
0.05	1.56	0	1.56	1.81
0.025	1.77	0	1.77	0
0.05	1.32	0	1.32	1.69
0.525	1.64	0	1.64	1.56

Косинусова Близост за 0.04 = **0.71**

Catboost

По същият начин пресмянаме косинусовата близост за всички разбивки.



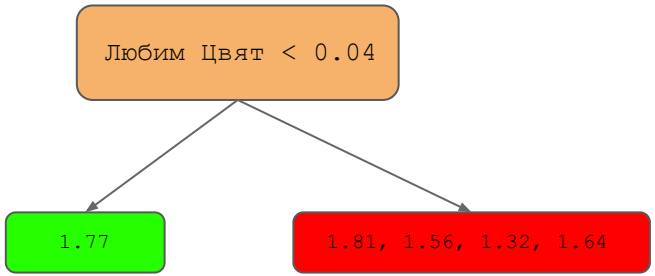
			A	B
Любим Цвят	Височина (м.)	Предсказание	Грешка	Средна грешка
0.05	1.81	0	1.81	0
0.05	1.56	0	1.56	1.81
0.025	1.77	0	1.77	1.69
0.05	1.32	0	1.32	1.71
0.525	1.64	0	1.64	0

Косинусова Близост за 0.29 = **0.79**

Catboost

Избираме разбивката, която има най-висока косинусова близост.
При повече характеристики трябва да се сравняват и разбивките, породени от тях.

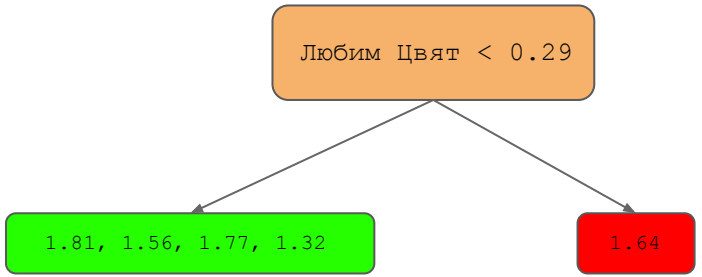
Косинусова Близост за 0.04 = **0.71**



Средна грешка = 1.77

Средна грешка = 1.58

Косинусова Близост за 0.29 = **0.79**



Средна грешка = 1.61

Средна грешка = 1.64

Как отново избягваме теч на данни

При изграждането на дърво отново третираме данните, все едно ги получаваме *последователно*. Това означава, че за всяко наблюдение стойността в **Грешка** не се използва при изчисляване на **Средна грешка**. Следователно **Грешка** за текущо наблюдение не се използва и при изчисляването на новата стойност за **Предсказание** за текущо наблюдение.

Така се избягва теч на данни от текущото наблюдение.

Любим Цвят	Височина (м.)	Предсказание	Грешка	Средна грешка
0.05	1.81	0	1.81	0
0.05	1.56	0	1.56	1.81
0.025	1.77	0	1.77	0
0.05	1.32	0	1.32	1.69
0.525	1.64	0	1.64	1.56

Catboost

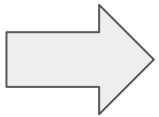
Сега ще построим второ дърво (в този случай, пън).
Преди това ще обновим предказанията и ще занулим грешката.

Предсказание = Предсказание + (Скорост на обучение * Средна грешка)

Нека Скорост на обучение = 0.1

Резултатите не са добри, но
са по-добри от началните.

Любим Цвят	Височина (м.)	Предсказание	Грешка	Средна грешка
0.05	1.81	0	0	0
0.05	1.56	0	0	1.81
0.025	1.77	0	0	0
0.05	1.32	0	0	1.69
0.525	1.64	0	0	1.56



Любим Цвят	Височина (м.)	Предсказание	Грешка	Средна грешка
0.05	1.81	0	0	
0.05	1.56	0.18	0	
0.025	1.77	0.17	0	
0.05	1.32	0.17	0	
0.525	1.64	0	0	

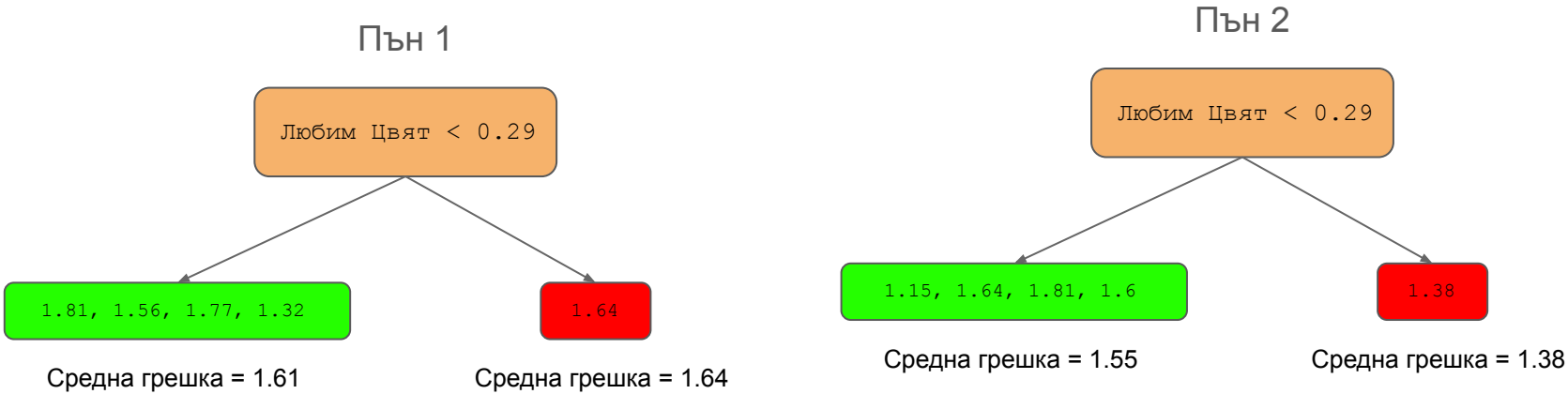
Catboost

Пресмятаме новата грешка, връщаме стойностите на категориите и процесът се повтаря отново.

Любим Цвят	Височина (м.)	Предсказ ание	Грешка	Средна грешка
Зелен	1.81	0	1.81	
Син	1.56	0.18	1.38	
Син	1.77	0.17	1.6	
Червен	1.32	0.17	1.15	
Зелен	1.64	0	1.64	

Как Catboost предсказва нови данни?

Нека имаме две крайни дървета (пънове).

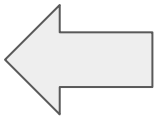


Любим Цвят	Височина (м.)
Син	???

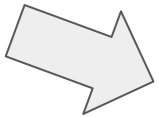
Как Catboost предсказва нови данни?

Използваме данните, които имаме, за да закодираме категорията.

Любим Цвят	Височина (м.)	Интервал
0.05	1.81	1
0.05	1.56	0
0.025	1.77	1
0.05	1.32	0
0.525	1.64	1



Любим Цвят	Височина (м.)
Син	???

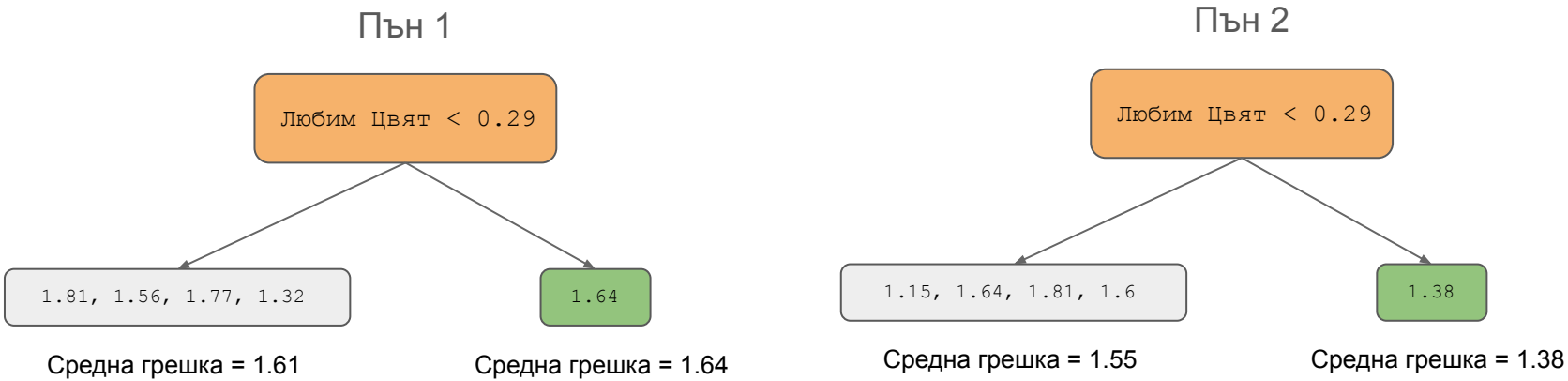


Любим Цвят	Височина (м.)
0.35	???

Как Catboost предсказва нови данни?

Спускаме се по дърветата и използваме формулата

`Предсказание = Скорост на обучение * (сума от средните грешки) .`



Любим Цвят	Височина (м.)
0.35	0.3

`0.1 * (1.64 + 1.38) = 0.3`