Оценка взаимодействия признаков с помощью ансамблей консервативных деревьев

Громаков И.А.[1]

Ланге А.М.[2]

[1] МФТИ

[2] Сколковский институт науки и технологий (Сколтех)

Мотивация

- Поиск gene-gene взаимодействий (эпистазис)
- Поиск drug-drug взаимодействий (выявление синергических или антагонистических эффектов между различными препаратами)
- Поиск взаимодействия биомаркеров (комбинации клинических параметров, которые лучше предсказывают риск осложнений)

- X - пространство объектов
- Y - множество ответов
- $f_j = X \to D_j$, j = 1, 2, ..., p - признаки объектов
- $F = \{f_j \mid j \in \{1, 2, ..., p\}\}$ - множество признаков
- $\mathscr{F} \subseteq F$ - подмножество признаков, $|\mathscr{F}| \geq 2$
- $S(\mathscr{F})$ - мера взаимодействия $\mathscr{F}$

Требуется:

- Для всех выбранных подмножеств $\mathscr{F}$ определить, есть ли у них взаимодействие: $\forall k \leq k_0,\ k \in \{2, 3, ..., |F|\}$, где k_0 заданное ограничение, равное максимальной глубине деревьев
- Найти $T_k = \{\mathscr{F} \mid S(\mathscr{F}) \geq S_0^k,\ k = |\mathscr{F}|\}$
- Ранжировать степень взаимодействия у разлиных групп признаков $\mathscr{F} \in T_k$ в каждом найденном T_k

Пусть $a(\boldsymbol{x}) = a\left(x_1, x_2, \ldots, x_p\right)$ - функция, аппроксимирующая целевую функцию, а $(x_1, x_2, \ldots, x_p) - p$ признаков модели

Определение: вещественные признаки $\mathscr{F}$ будем называть *взаимодействующими*, если

$$E_x \left[\frac{\partial a(\boldsymbol{x})}{\prod_{x \in \mathscr{F}} \partial x} \right]^2 > 0$$

Определение: признаки $\mathscr{F}$ будем называть *не взаимодействующими*, если $a(\boldsymbol{x})$ можно представить в виде суммы $N \leq |\mathscr{F}|$ функций, каждая из которых не зависит от хотя бы одной переменной $x \in \mathscr{F}$:

$$a(x) = \sum_{k:\, x_k \in \mathscr{F}} f_{x_{-k}}(x_{-k}),$$

где x_{-k} означает все признаки кроме x_k

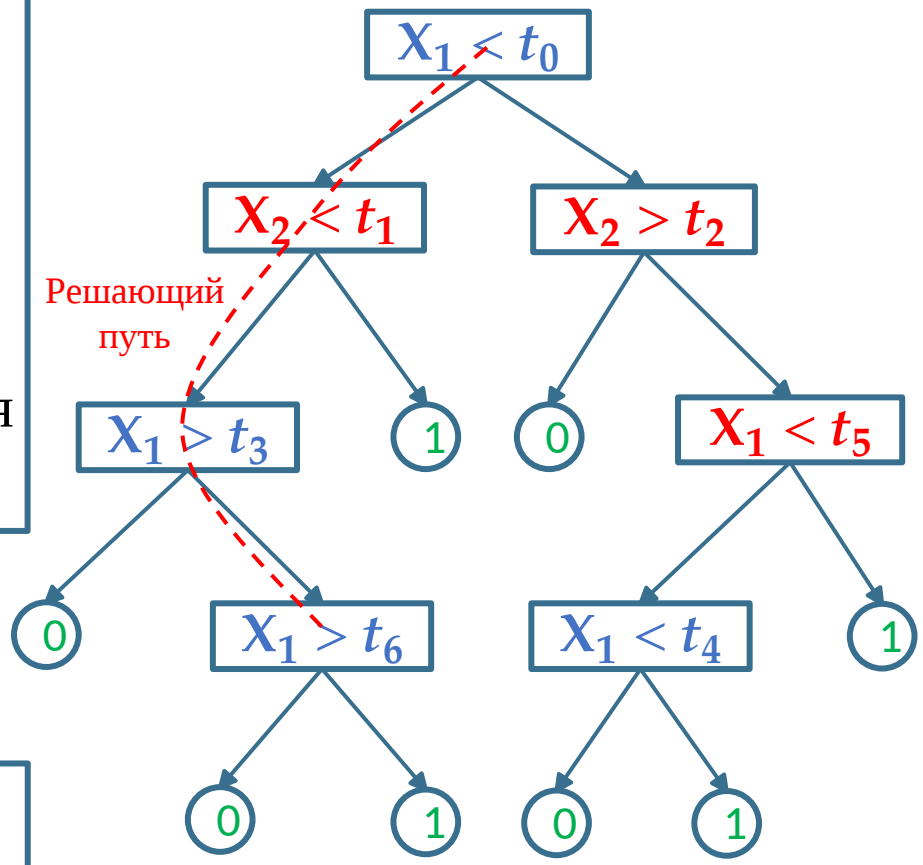
Решающее дерево:

Это алгоритм классификации $a(x)$, задающийся деревом с корнем $v_0 \in V$ и множеством внутренних вершин $V = V_{\text{внутр}} \cup V_{\text{лист}}$, в котором каждой внутренней вершине $v \in V_{\text{внутр}}$ сопоставлен дискретный признак (предикат) $\beta_v: X \to D_v$, правило перехода в дочерние вершины $S_v: D_v \to V$, и каждой листовой вершине $\forall v \in V_{\text{лист}}$ ставится в соответствие $y_v \in Y$

$$Gain(\beta, U) = \Phi(U) - \Phi(U \mid \beta),$$

где U – рассматриваемая выборка объектов

Даже если целевая функция аддитивна, то решающие пути все равно содержат разные признаки. Такие паттерны могут возникать как следствие случайных вариаций или рекурсивной природы алгоритмов построения дерева

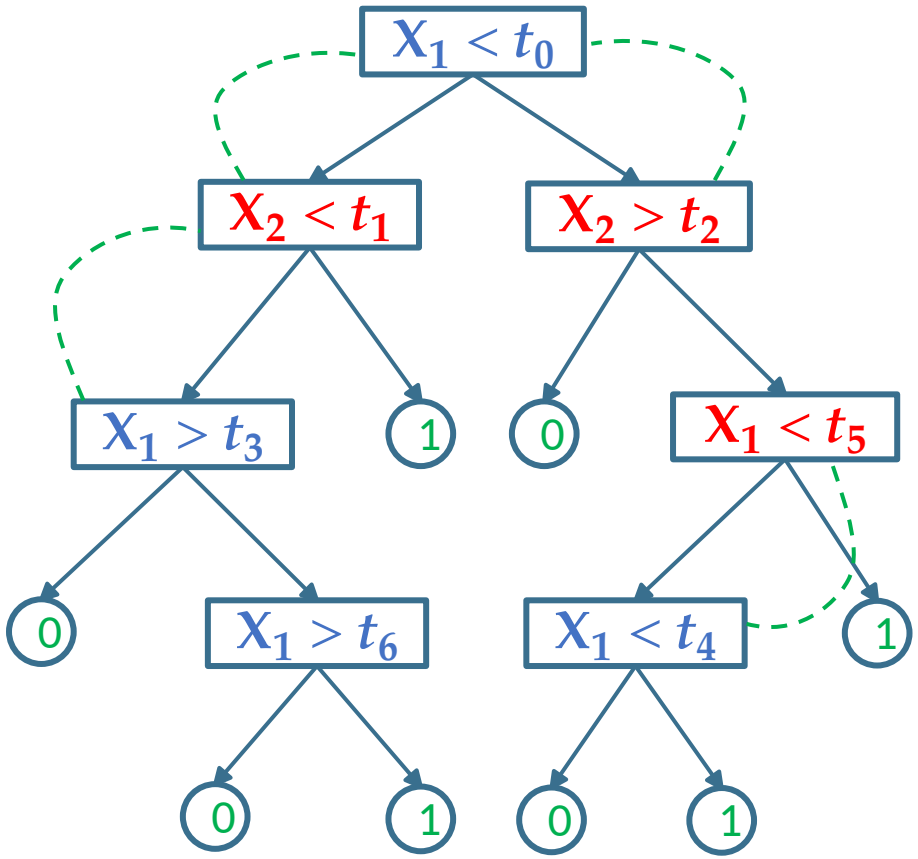


Консервативное решающее дерево:

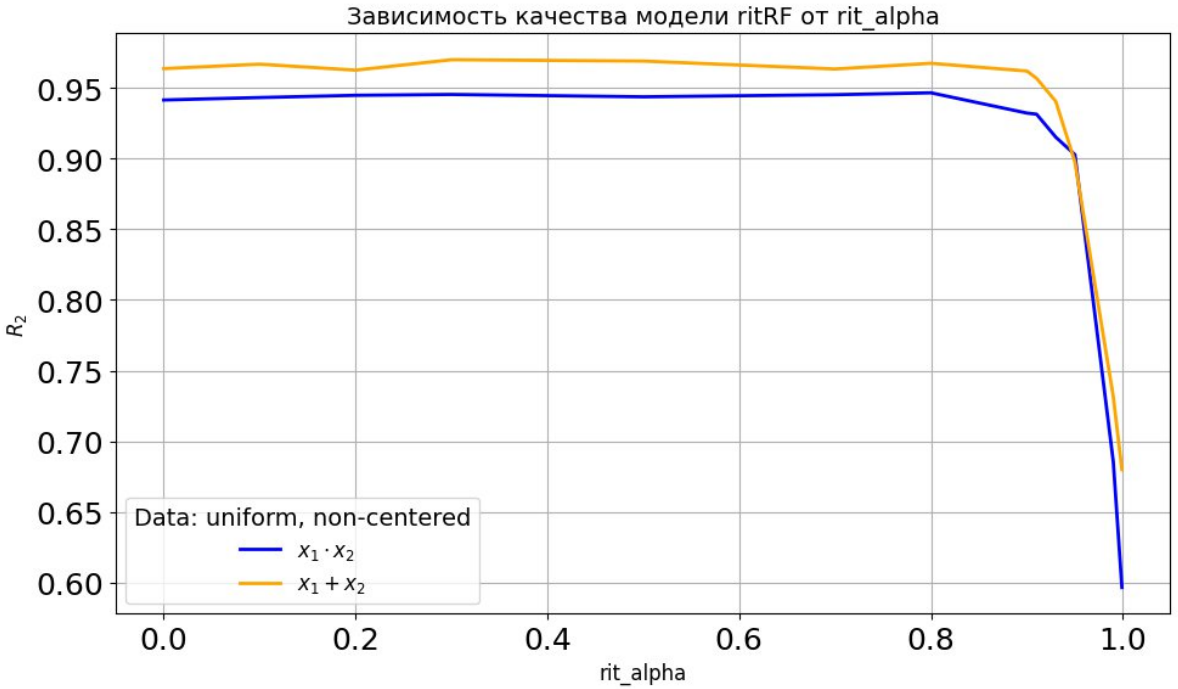
Это модификация CART, вводящая новый гиперпараметр *rit_alpha*[1] $\in(-\infty, 1)$. Использование *rit_alpha* дает преимущество при ветвлении тем признакам, которые уже были выбраны ранее в дереве

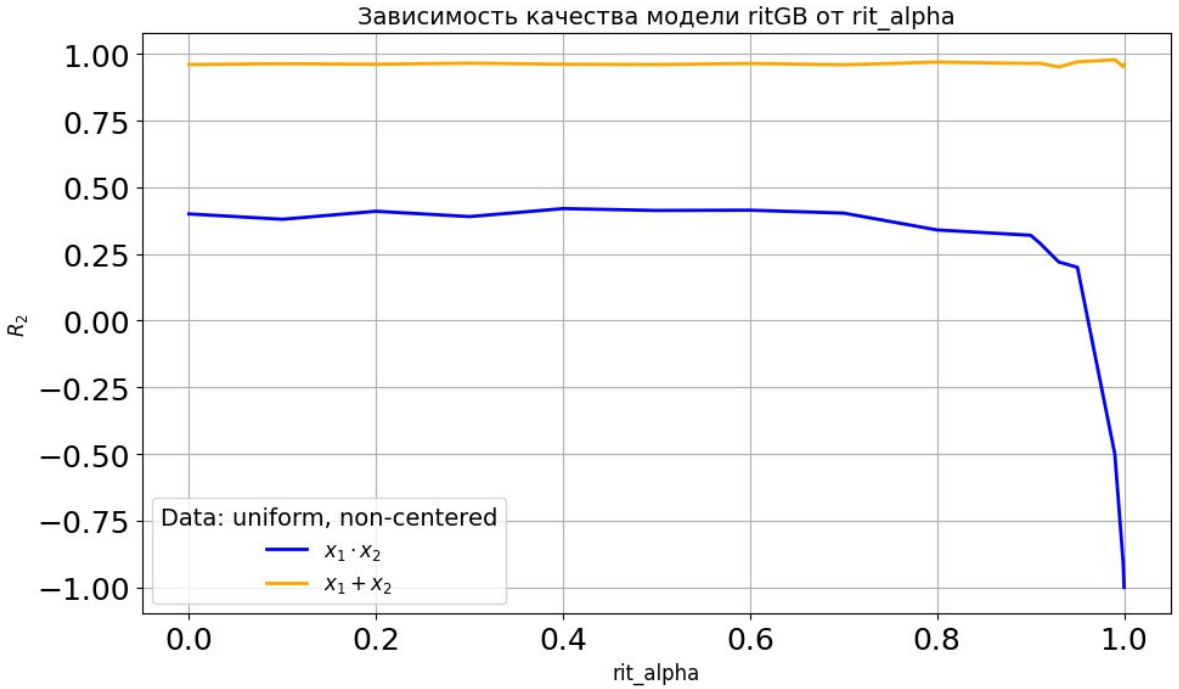
$$B = Gain(\beta_{f_{best}}, U)(1 - rit\_alpha)$$

Разные признаки в решающем пути – потенциальное взаимодействие



[1] RIT - Redundancy Insensitive Trees


Зависимость качества модели ritRF от rit_alpha
R_2
rit_alpha
Data: uniform, non-centered
$x_1 \cdot x_2$
$x_1 + x_2$


Зависимость качества модели ritGB от rit_alpha
R_2
rit_alpha
Data: uniform, non-centered
$x_1 \cdot x_2$
$x_1 + x_2$

Подбор гиперпараметра

Пусть

- $\mathscr{M}_\alpha = \{M_\alpha^{(1)}, \ldots, M_\alpha^{(N)}\}$ – множество N моделей градиентного бустинга, обученных с параметром $rit\_alpha = \alpha$ на случайных подвыборках обучающей выборки одинакового размера
- $score\,(M)$ – метрика качества модели M
- S_α - выборка с посчитанными метриками качества моделей

Формальная постановка задачи:

$$\begin{cases} H_0\text{: средние значения } score\left(M_{\alpha_0}^{(i)}\right) \text{ и } score\left(M_{-\infty}^{(j)}\right) \text{ совпадают,} \\ H_1\text{: качество при } \alpha_0 \text{ статистически значимо ниже, чем при } \alpha = -\infty \end{cases}$$

Для проверки используется односторонний критерий Манна-Уитни (U-test) на уровне значимости $\delta = 0.05$:

$$p = MannWhitneyU\left(S_{\alpha_0}, S_{-\infty}, alternative = 'less'\right)$$

Если $p > \delta$, то гипотеза H_0 **не отвергается**, и параметр α_0 может быть принят как допустимый, не ухудшающий качество модели

Предлагается подобрать гиперпараметры для N_{st} моделей с найденным *rit_alpha* и для каждой модели найти нормализованную статистику $S_n(M_\alpha,\mathscr{F})$, а затем усреднить результат:

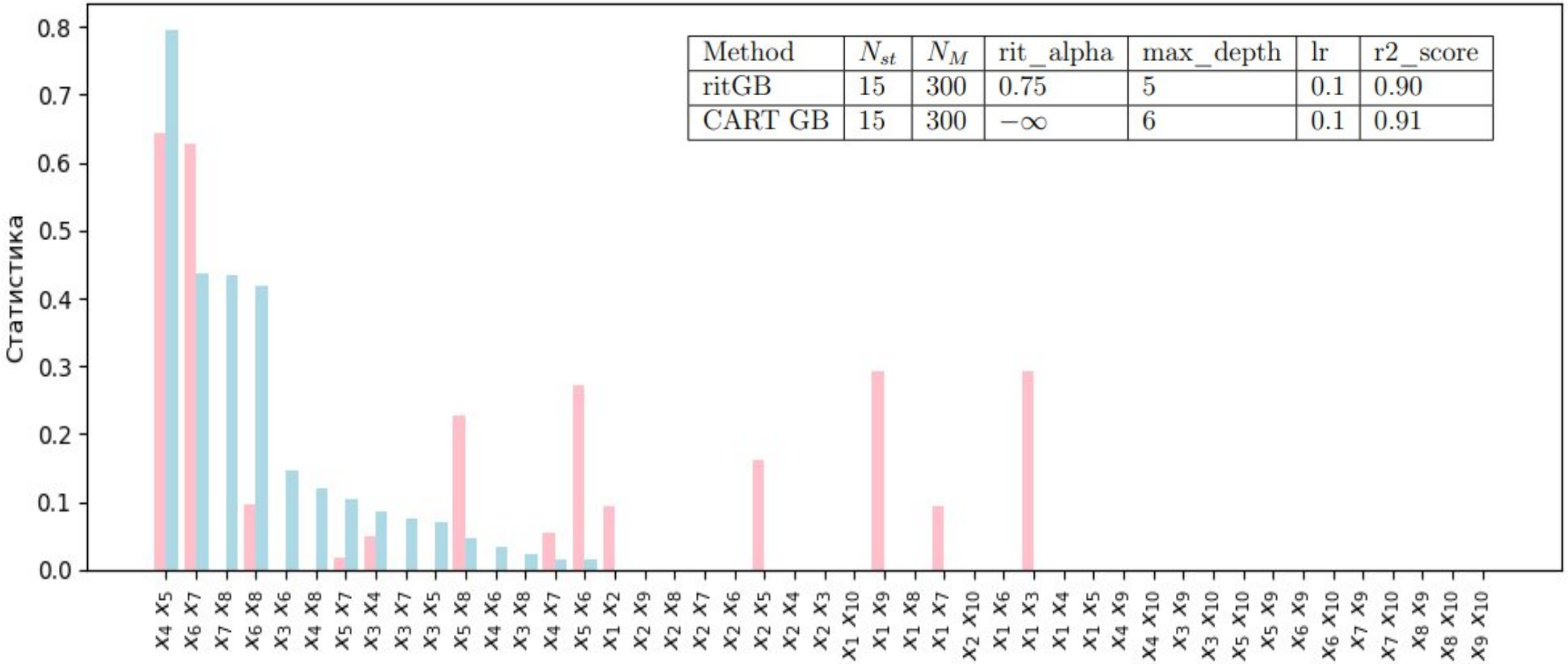
$$S(\mathcal{F}) = \frac{1}{N_{st}} \sum_{j=1}^{N_{st}} S_n(M_\alpha^j, \mathcal{F}) \qquad S_n(M_\alpha, \mathcal{F}) = \frac{S(M_\alpha, \mathcal{F})}{(\prod_{x \in \mathcal{F}} S^{|\mathcal{F}|}(x))^{1/|\mathcal{F}|}}$$

$$S^K(x_j) = \frac{1}{N_{M_\alpha}} \sum_{i \in M_\alpha} \frac{1}{|V^i_{\text{лист}}|} \sum_{v \in V^i_{\text{лист}}} \frac{N_v^i}{N^i} \frac{m_v^{i,K}(x_j)}{l_v^i}$$

$$S(M_\alpha, \mathcal{F}) = \frac{1}{N_{M_\alpha}} \sum_{i \in M_\alpha} \frac{1}{|V^i_{\text{лист}}|} \sum_{v \in V^i_{\text{лист}}} \frac{N_v^i}{N^i} \frac{n_v^i(\mathcal{F})}{l_v^i}$$

где l_v – длины путей от корня до листа v, $n_v(\mathscr{F})$ – индикатор множества уникальных признаков решающего пути с листовой вершиной v, $m_v^K(x)$ – индикатор наличия признака x в множестве уникальных признаков $\mathscr{F}$: $|\mathscr{F}| = K$ на решающем пути, N_v – количество объектов из обучающей выборки, которые попали в листовую вершину v, N – общее количество объектов в обучающей выборке, N_{M_α} – количество предикторов модели M_α

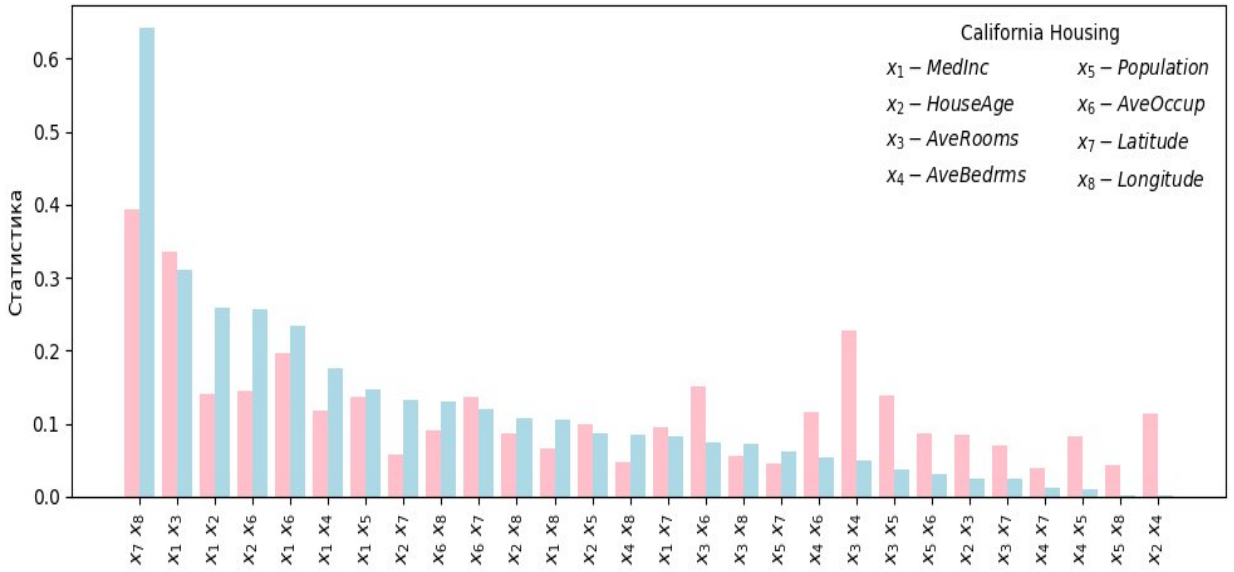
Синтетические данные



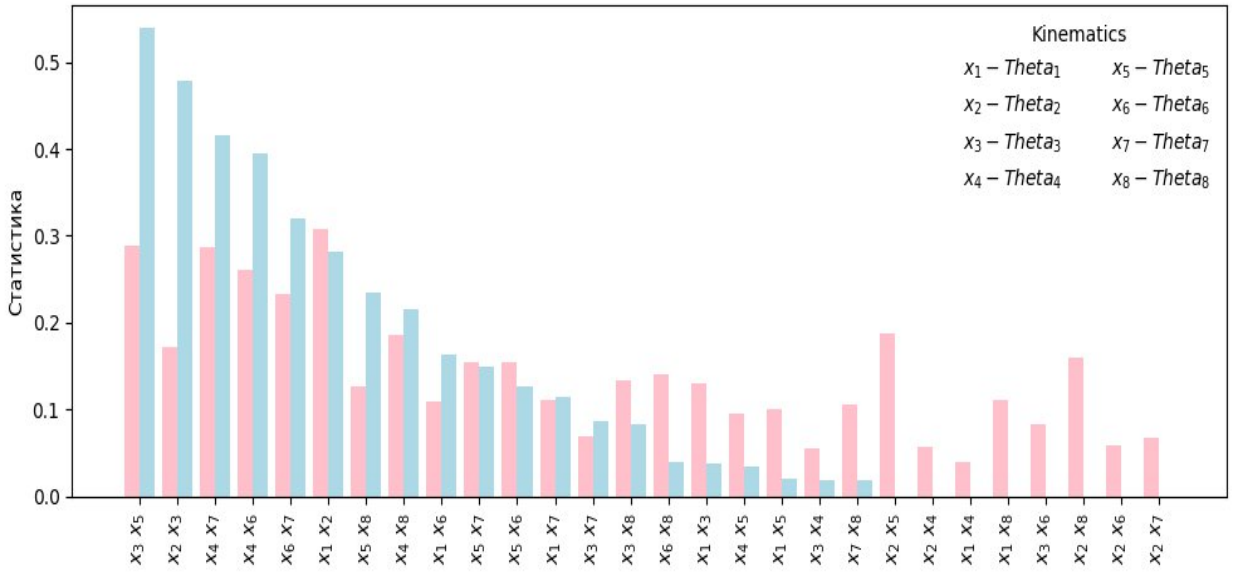
Method	N_{st}	N_M	rit_alpha	max_depth	lr	r2_score
ritGB	15	300	0.75	5	0.1	0.90
CART GB	15	300	$-\infty$	6	0.1	0.91

$$F(x) = 2x_1 + 3x_2^2 + sin(x_3) + 5x_4x_5 + 4x_6x_7x_8 + 0.5x_9 + \varepsilon$$

Реальные данные



Method	N_{st}	N_M	rit_alpha	max_depth	lr	r2_score
ritGB	20	500	0.95	5	0.1	0.83
CART GB	20	500	$-\infty$	5	0.1	0.84



Method	N_{st}	N_M	rit_alpha	max_depth	lr	r2_score
ritGB	10	100	0.95	6	0.1	0.76
CART GB	10	100	$-\infty$	7	0.1	0.78

Выводы

- Метод способен выявлять взаимодействия малых порядков (парные, тройные)
- Метод может быть использован как дополнение к комплексному анализу взаимодействий признаков