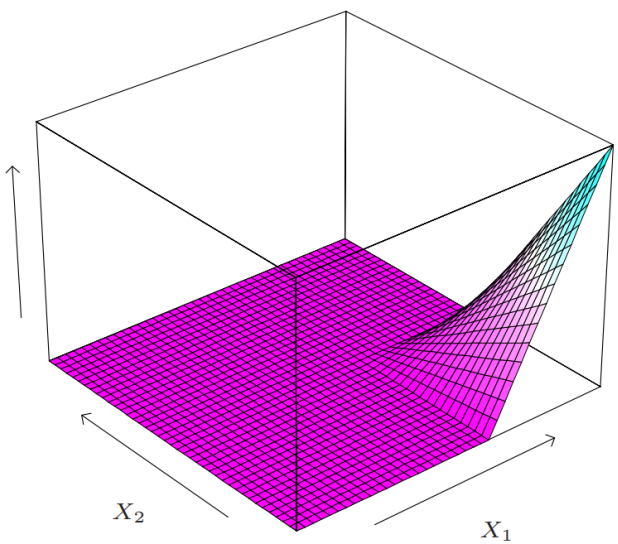
Multivariate Adaptive Regression Splines

$D = \{x_i, y_i\}_{i=1}^N$ – набор данных. $f(x^1, x^2, ..., x^p)$ аппроксимируем $g(x^1, x^2, ..., x^p)$. Качество аппроксимации оценивается с помощью RMSE:

$$RMSE(D) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(g(x_k) - y_k)^2}$$

$$g(x) = \sum_{m=0}^{M} a_m \cdot B_m(x), \text{ где } B_m(x) = \prod_{k=1}^{K_m} b_{k,m}.$$

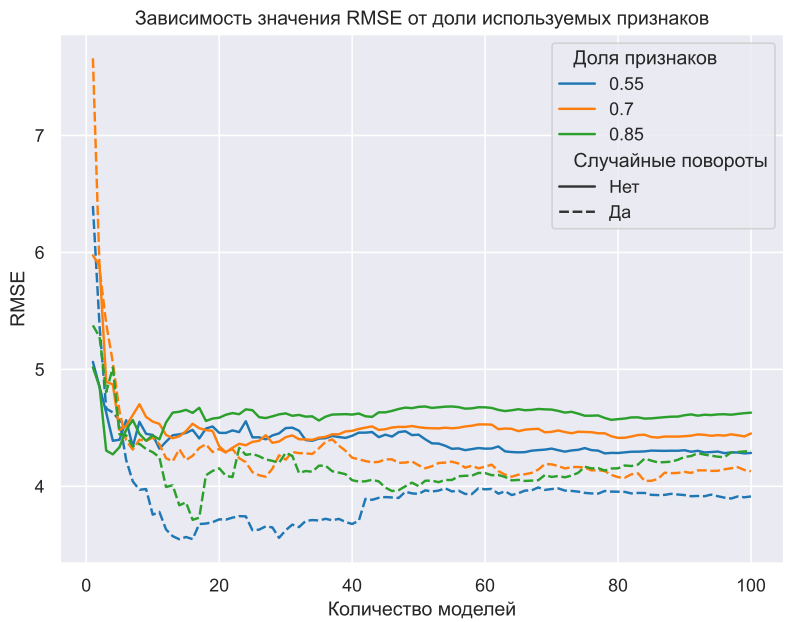
$b_{k,m}$ равен либо $max\{+(x-t),\ 0\}$, либо $max\{-(x-t), 0\}$.

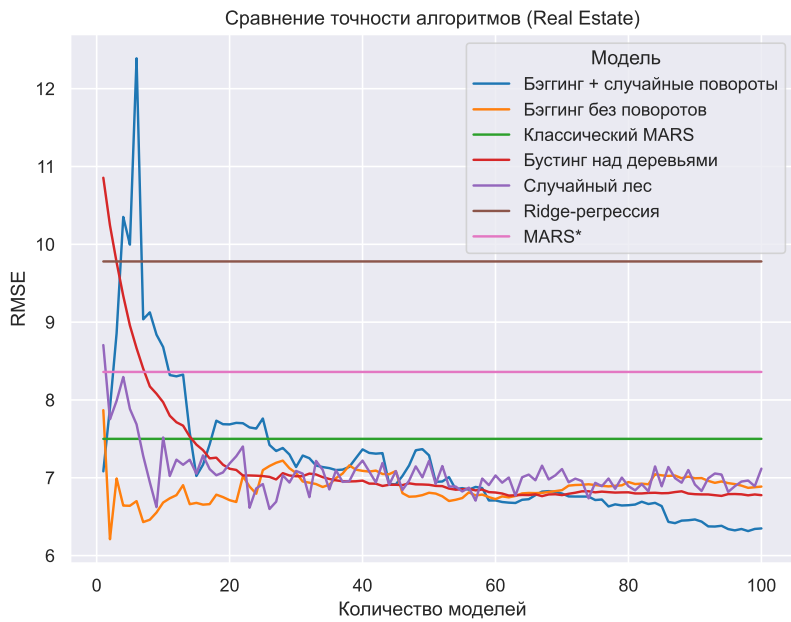


Идея: Создать ансамбль $M(x) = \frac{1}{T}\sum_t^T a_t(x)$, где у каждой модели вход умножен на случайную ортогональную матрицу. $\overline{x} = Qx$, где Q сгенерирована из распределения Хаара.

Результаты применения

На первом графике показана зависимость значения RMSE от доли признаков, использующихся в бэггинге. На втором графике изображено значение RMSE на валидационной выборке у различных алгоритмов.





Таким образом, использование случайных поворотов позволяет добиться лучшего качества.

Выводы

В таблице приведены результаты работы алгоритмов на тестовой выборке.

Модель\Набор данных	BH	CH	RE	DD
Классический MARS	3.09	0.66	5.27	53.5
MARS*	3.32	0.67	6.08	57.2
Бэггинг без случайных поворотов	**2.43**	0.63	4.92	53.1
Бэггинг + случайные повороты	2.47	0.62	**4.48**	**52.2**
Ridge-регрессия	5.41	0.75	6.81	58.3
Случайный лес	2.77	0.63	5.06	54.1
Бустинг над деревьями	2.51	**0.61**	4.85	54.3