Практическое задание
Композиции алгоритмов для решения задачи регрессии

Кривонос Анна
317 группа

18 декабря 2022 г.

Оглавление

Глава 1

Введение

В данном отчете исследуются алгоритмы случайного леса и градиентного бустинга для задачи регрессии с функцией потерь RMSE. Для тестирования моделей используется дадасет данных о продажах недвижимости House Sales in King County, USA.

Глава 2

Эксперименты

Перед началом работы проведем предобработку данных. Удалим столбец 'id', так как он мало коррелирует с целевой переменнной. Также преобразуем столбец 'date' в Datetime и добавим на основе этого столбца три новых признака - день, месяц, год. После столбец 'date' удалим. Затем разделим данные на обучающую и валидационную выборки.

2.1 Исследование поведения алгоритма случайный лес.

Исследуем зависимость значения функции потерь на отложенной выборке и времени работы алгоритма в зависимости от следующих параметров:

- количество деревьев в ансамбле
- размерность подвыборки признаков для одного дерева
- максимальная глубина дерева

Рассмотрим следующие параметры :

размерность подвыборки признаков:	0.1	0.5	1	0.3
глубина дерева:	1	5	10	None

Количество деревьев рассмотри от 1 до 100

При увеличении числа деревьев значение RMSE начинает колебаться около одного числа(рис. 2.1). Наименьшее значение точности достигается, когда размерность подвыборки равна количеству признаков в датасете. Алгоритм на каждом шаге знает все признаки, поэтому лучше настраивается на данные. При глубине 1 и features_size=1 алгоритмы получаются похожими, поэтому точность почти не зависит от количества деревьев. Также из графика видно, что с ростом глубины значение RMSE растет, так как модель сильно подстраивается под данные и переобучается. При неограниченной глубине точность на подвыборке признаков в 10% оказывается лучше, чем на всей выборке признаков.

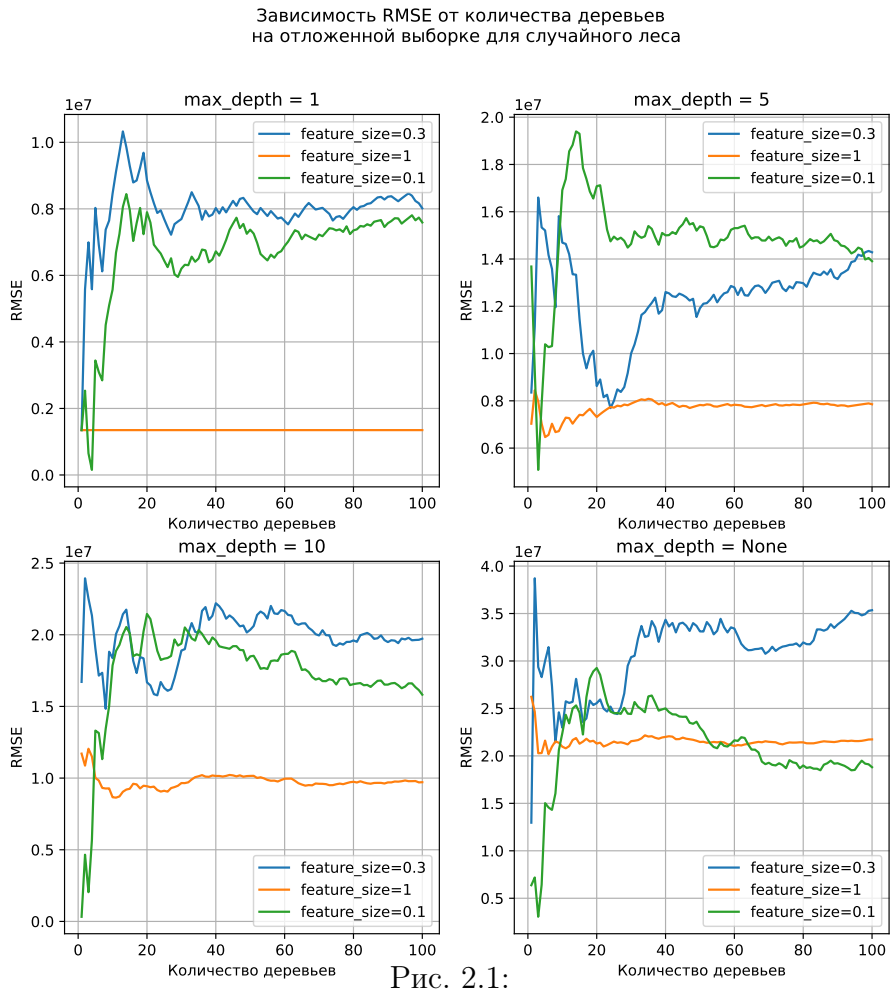
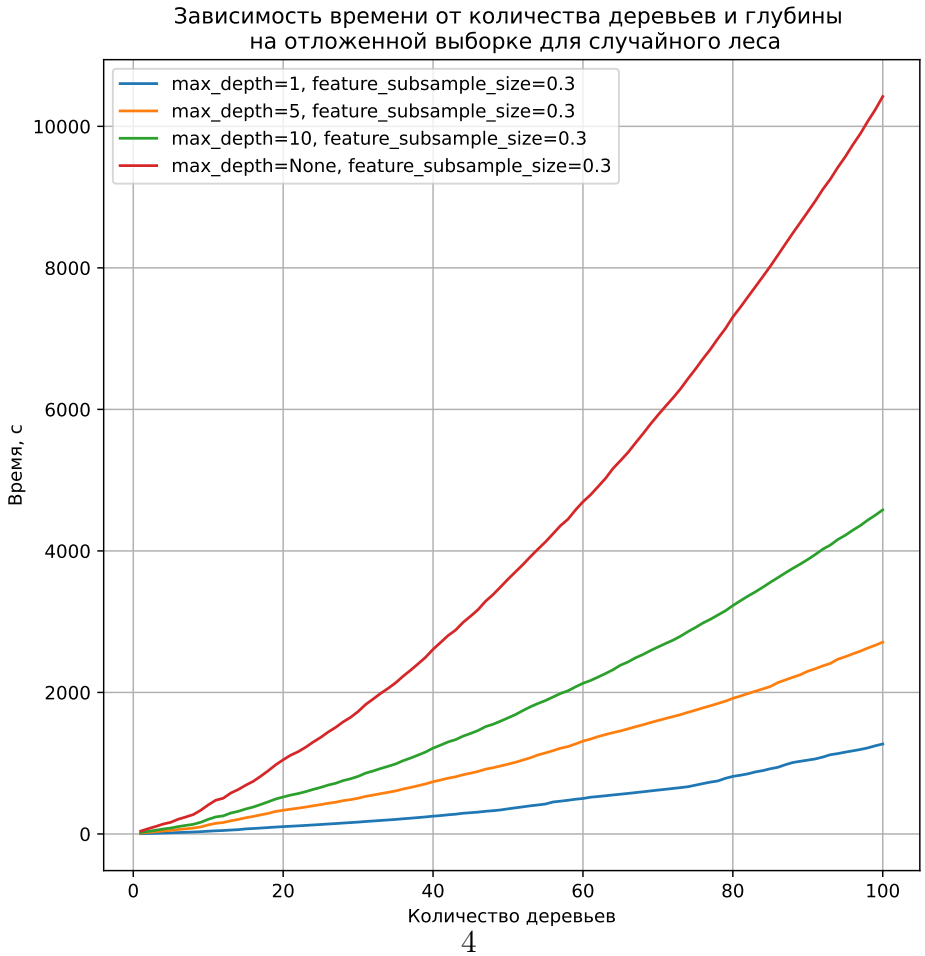
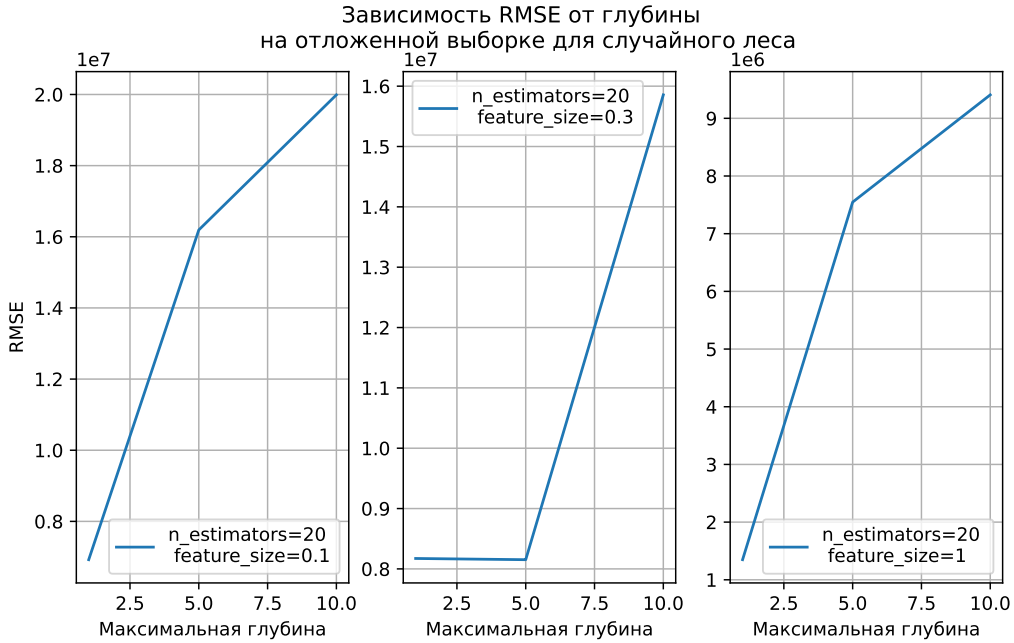


Рис. 2.1:

С ростом числа деревьев и максимальной глубины время работы алгоритма растет. Причем при неограниченной глубине время работы почти в два раза дольше.



Как видно из рисунка ниже, значение RMSE с ростом глубины растет, оптимальной является глубина меньше 5.



2.2 Исследование поведения алгоритма градиентный бустинг.

Исследуем зависимость значения функции потерь на отложенной выборке и времени работы алгоритма в зависимости от следующих параметров:

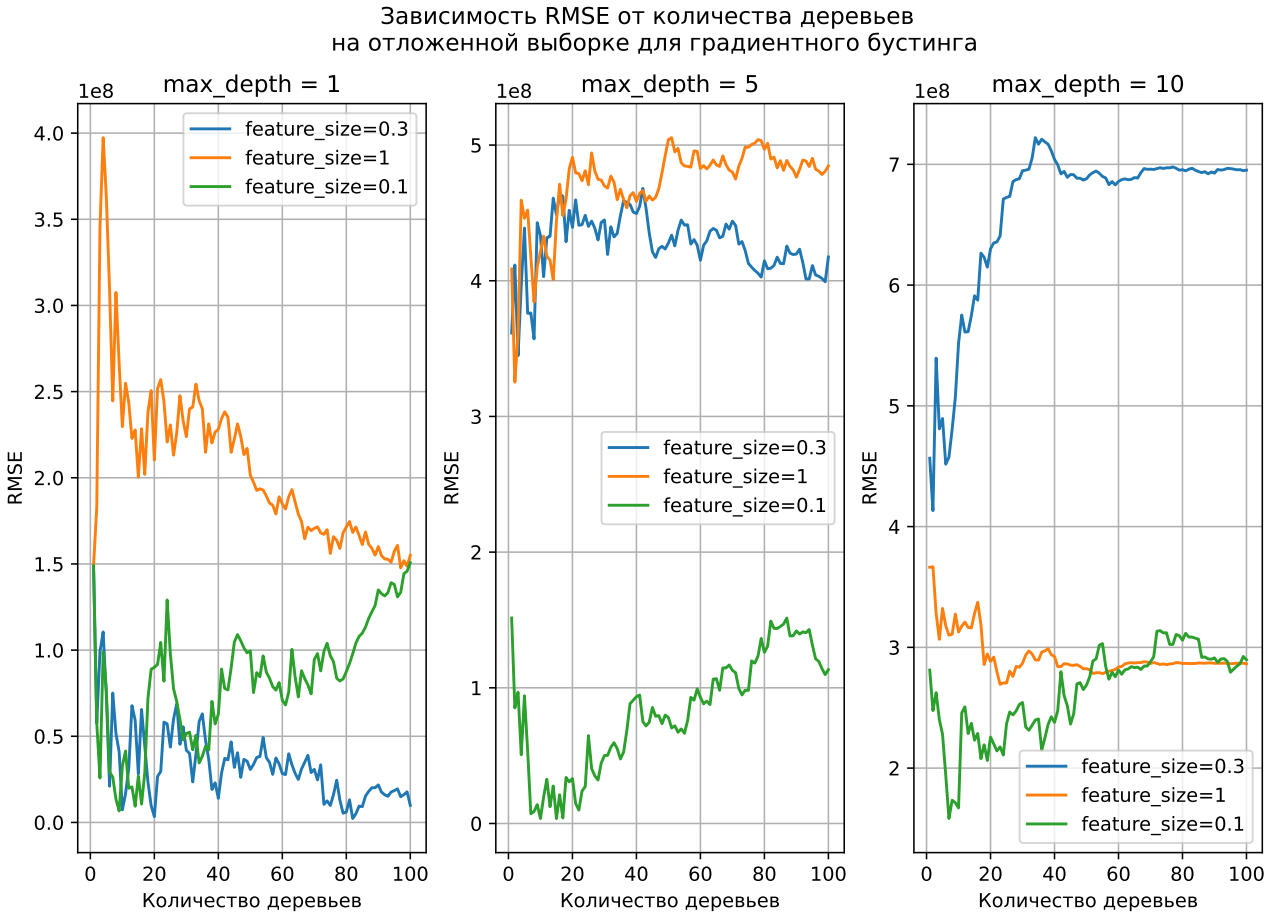
- количество деревьев в ансамбле
- размерность подвыборки признаков для одного дерева
- максимальная глубина дерева
- выбранный learning_rate

Рассмотрим следующие параметры :

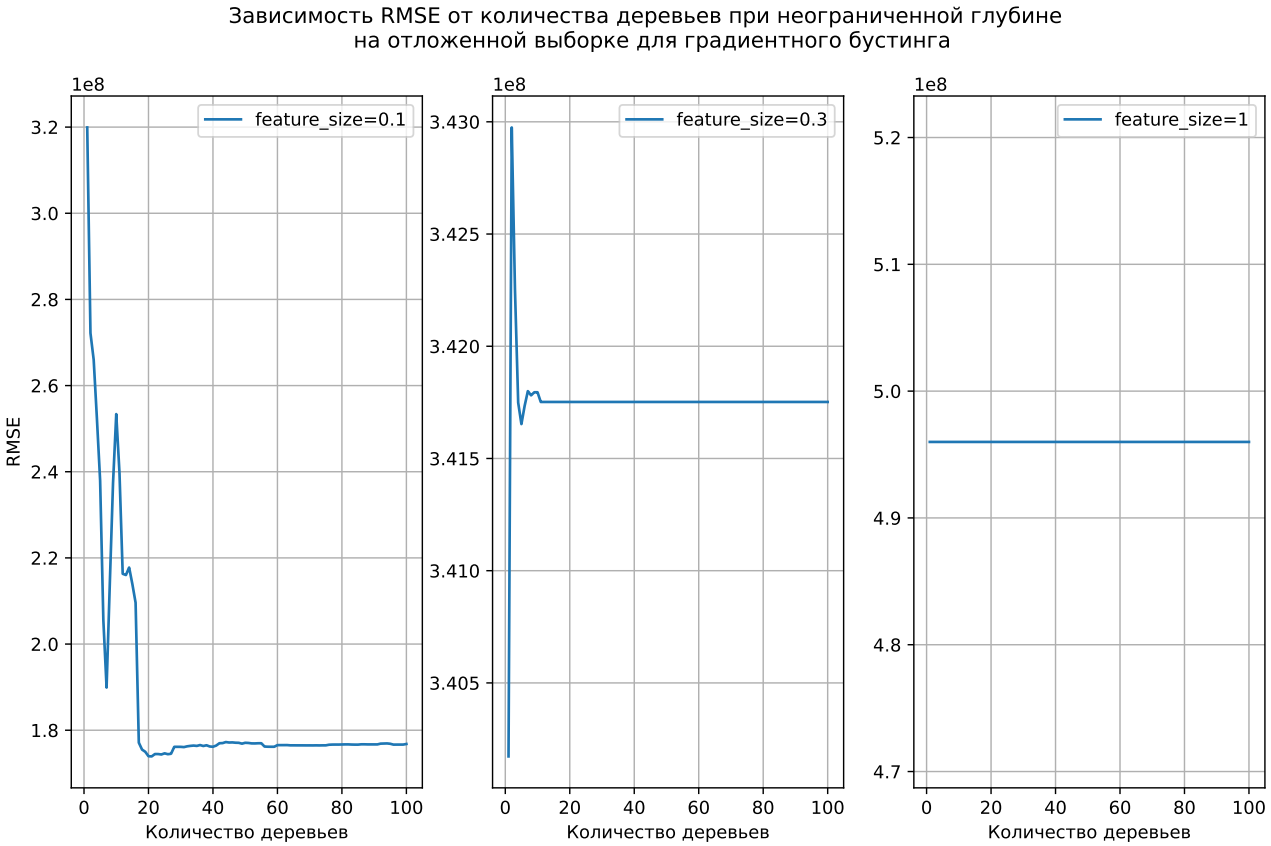
размерность подвыборки признаков:	0.1	0.5	1	0.3
глубина дерева:	1	5	10	None
learning_rate:	0.1	0.5	1	

Количество деревьев рассмотри от 1 до 100

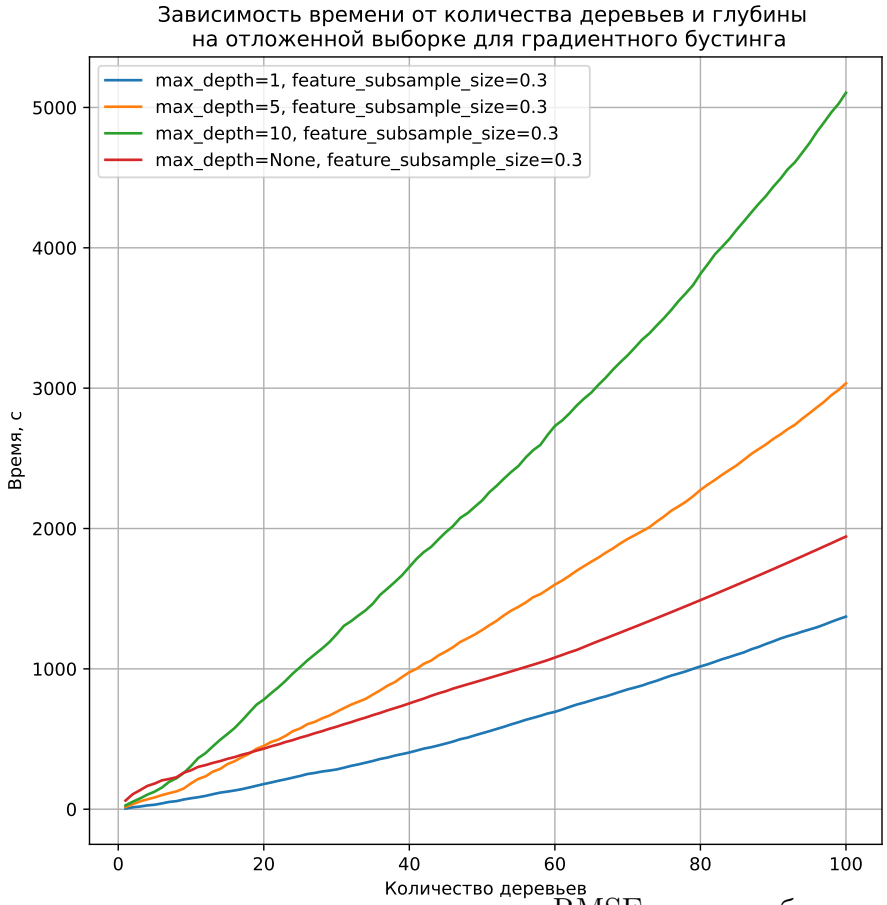
Как видно из рисунка значение RMSE сильно колеблется в зависимость от числа деревьев. С ростом глубины точность также становится хуже. При максимальной глубине 1 и размере подвыборки признаков 10-30% значение RMSE близки к нулю, но при большей глубине точность лучше при большей размерности подвыборки признаков. Это свзанно с тем, что с ростом глубины модель лучше настраивается, когда признаков больше.



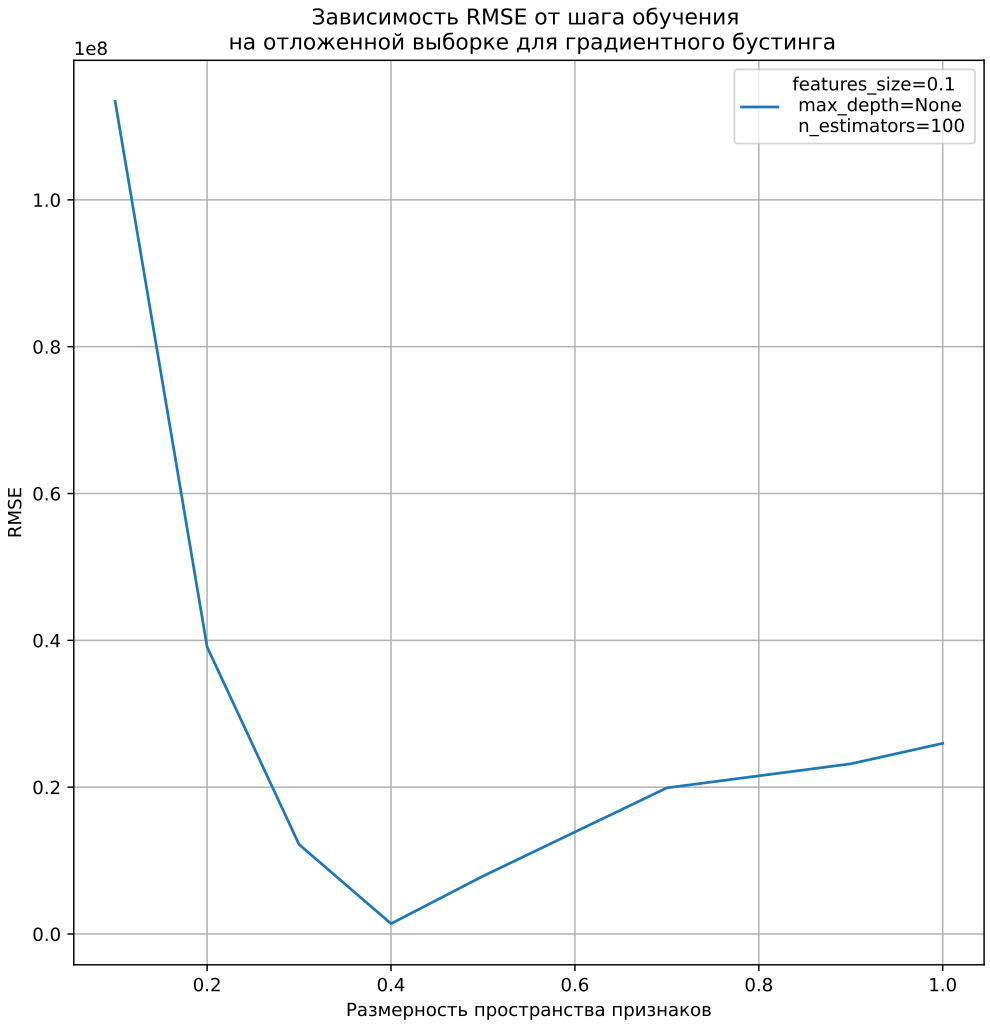
При неограниченной глубине RMSE больше и колеблется около одного значения. При этом на всем подмножестве признаков становится постоянным, так как алгоритмы получаются достаточно похожими.



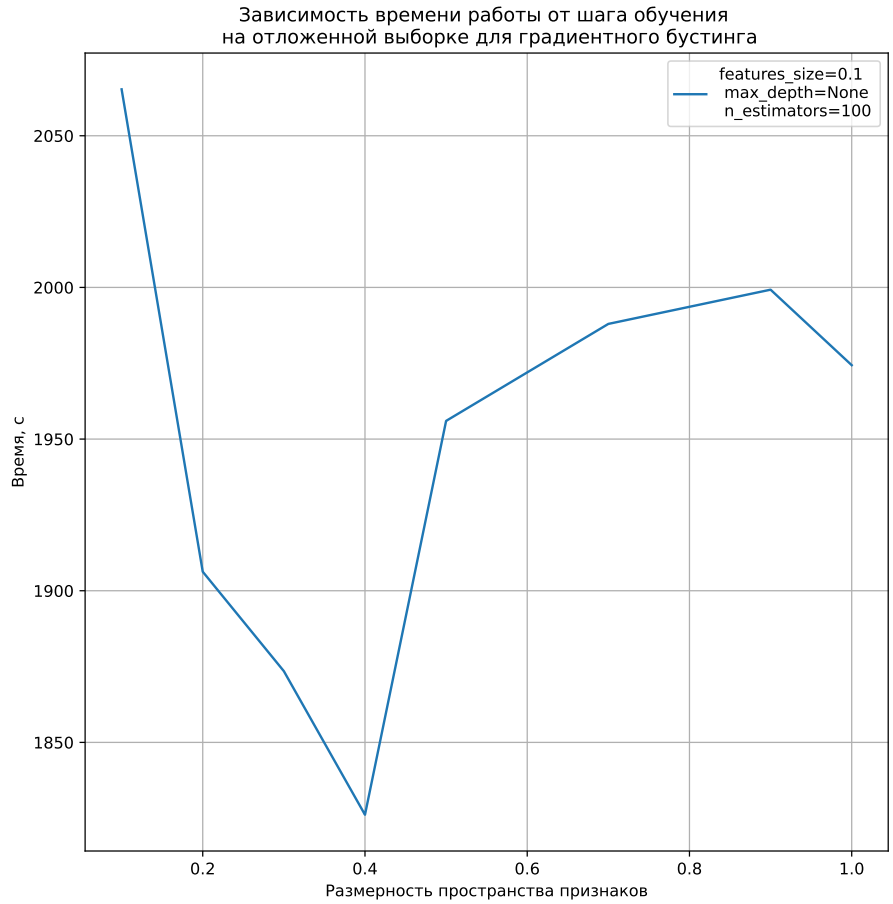
С ростом глубины и количества деревьев время работы также растет. При этом при неограниченной глубине алгорит работает быстрее, чем при глубине 5 или 10.



На рисунке ниже показана зависимость значения RMSE от шага обучения learning_rate. Как видно из графика наилучшая точность достигается при learning_rate $= 0.4$. При больших значенияпараметра точность становится хуже.



Наилучшее время работы также при значении learning_rate = 0.4.



Глава 3

Вывод

Алгоритмы случайного леса и градиентного бустинга существенно зависят от своих параметров. На наших данных наилучшее значение RMSE достигается при небольшой глубине деревьев. С ростом количества деревьев точность также как правило становится лучше, но при этом возрастает и время работы. Оптимальное значение параметра нужно подбирать в зависимости от ваших целей. Размерность пространства признаков также влияет на качество модели. В алгоритме случайный лес наилучшее точность достигается при большой размерности, в то время как в градиентном бустинге при малой. При неправильно подобранном параметре модель может переобучится.