Отчет о выполненной работе "Композиции алгоритмов для решения задачи регрессии"

Махин Артем Александрович

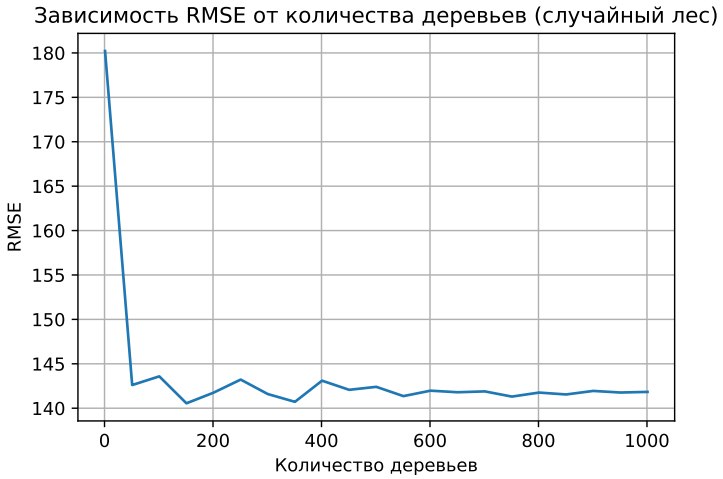
24 декабря 2020

1 Постановка задачи

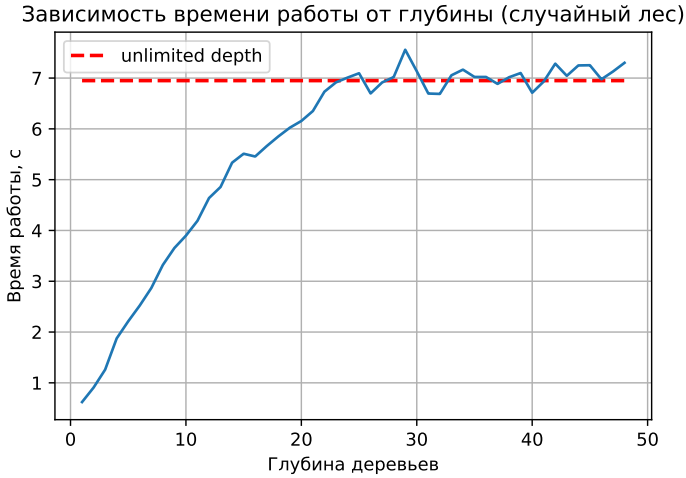
В качестве задачи предлагалось реализовать на языке Python два алгорима (случайный лес и градиентный бустинг) и применить их для прогнозирования цены дома. Так же найти зависимости между результатом работы алгоритмов от различных гиперпараметро, а так же скорость работы алгоритмов.

2 Случайный лес

Исследуем качество прогнозирования от количества деревьев (n_estimators). Далее во всех экспериментах будет использоваться функция потерь RMSE. Максимальную глубину возьмем 8. Количество деревьев будем перебирать от 1 до 1001 с шагом 50.

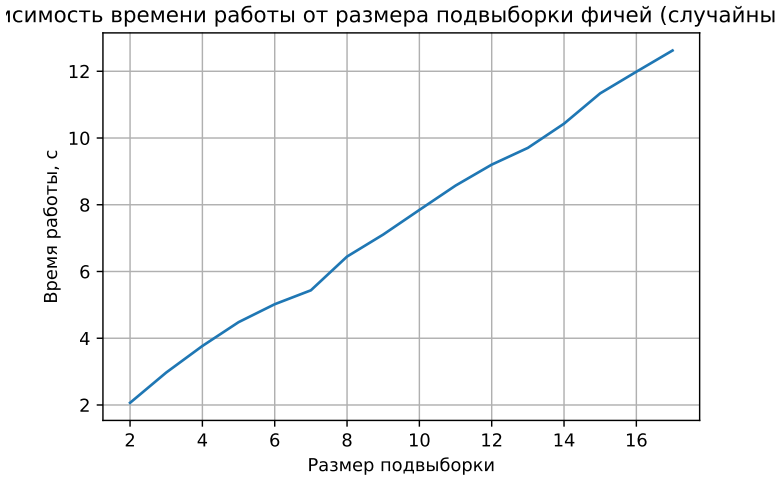


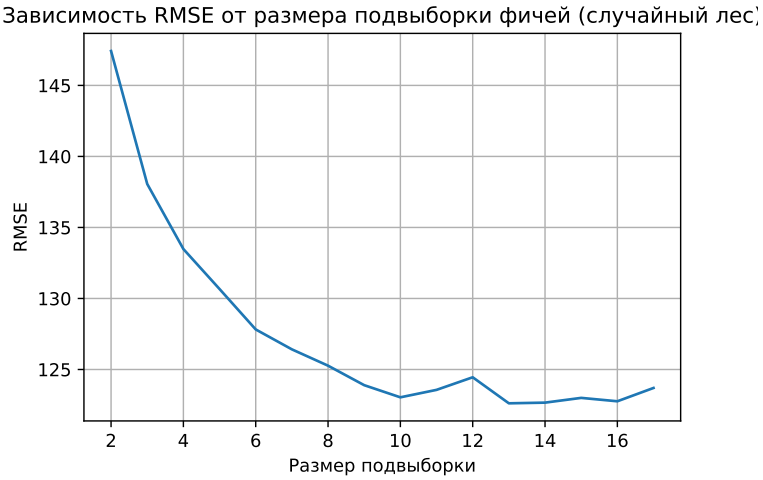
Как видно из графиков, ошибка перестает значительно уменьшаться после 150. Значит везде дальше будет использовать это значение как оптимальное. Время же работы увеличивается линейным образом.
Теперь исследуем зависимость от максимальной глубины. Рассмотрим варианты от 1 до 48 + неограниченный случай.



Как можно заметить, после значения 13 функция ошибки перестает значительно уменьшаться, значит в дальнейшем будем использовать это значение как оптимальное. До значения 23 время увеличивалось линейно, после же значение времени работы остается примерно постоянной. Связано это с тем, что модель не может построить еще более глубокие модели, поэтому все эти случаи примерно равны случаю неограниченной глубины (это можно видеть и по графику зависимости RMSE от глубины).

Исследуем теперь зависимости от размера набора признаков, по которому будет выбираться разбиения в вершинах деревьев. Перебираем значения от 2 до 17 (так как размерность признакового пространства - 17)



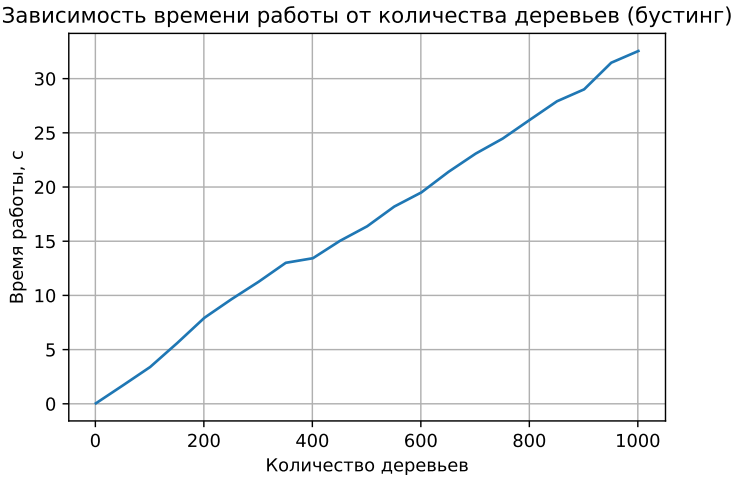


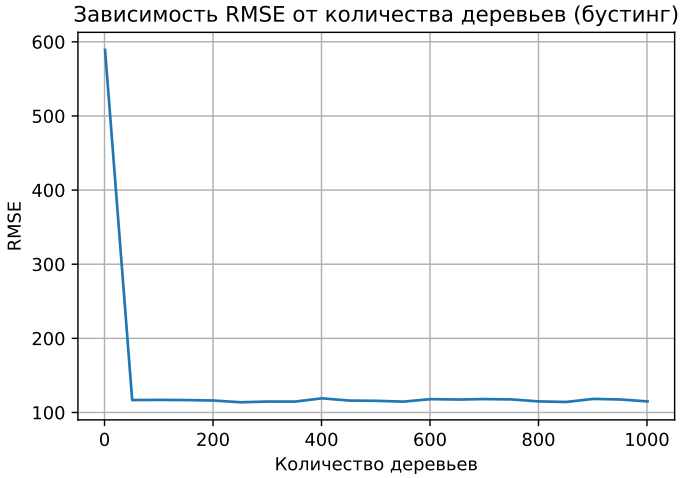
Видно, что оптимальное значение равно 13. Дальше функция ошибки перестает сильно уменьшаться. Так же отметим, что время работы алгоритма линейно зависит от размера подвыборки фичей, что вполне логично.

Отметим, что при параметрах, которые мы отыскали выше, смогли получить такую величину ошибки: 122.78

3 Градиентный бустинг

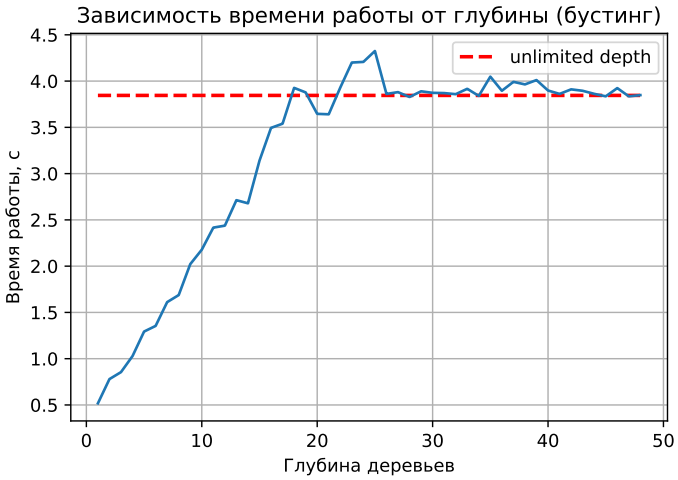
Исследуем качество прогнозирования от количества деревьев в градиентном бустинге. Выберем параметры max_depth равный 8, a learning_rate равный 0.1.

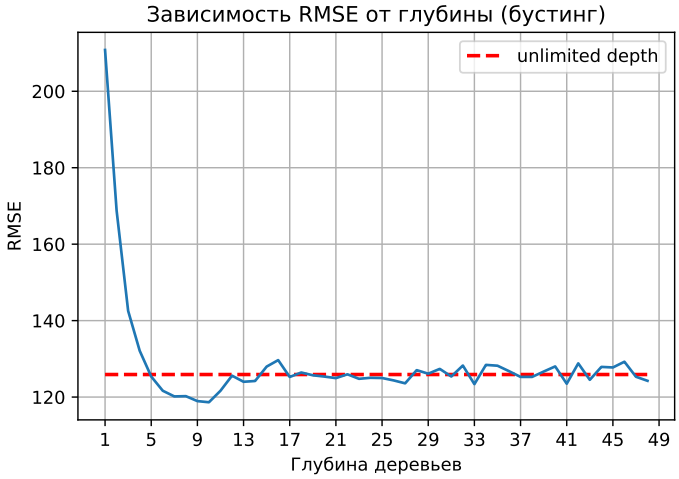




Видим, что время работы зависит линейно от количества деревьев. Функция ошибки после 50 перестает значительно уменьшаться, поэтому далее будем использовать именно это значение как оптиамльное.

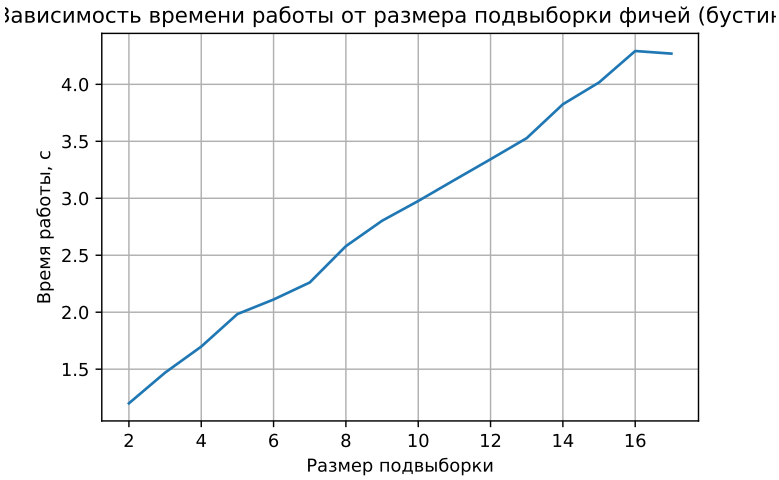
Рассмотрим теперь зависимости от максимальной глубины деревьев. Рассмотрим значения от 1 до 48 и неограниченный случай.

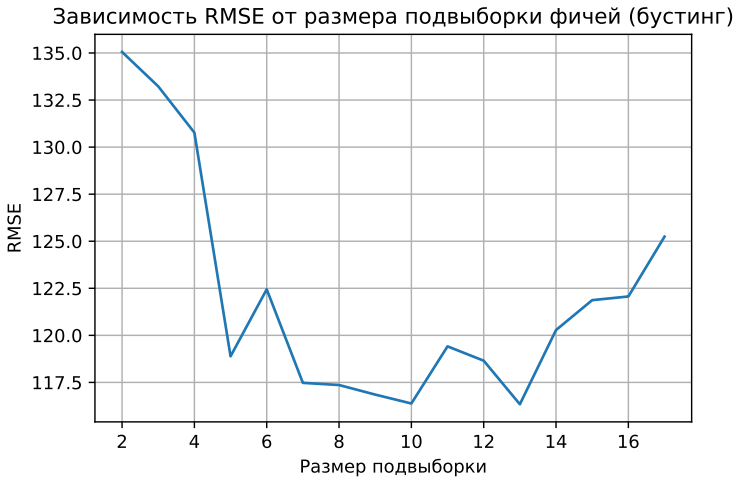




Заметим, что до значения 18 время растет линейно, дальше же выходит на константное значение. Это, так же как и в случае случайного леса, связано с тем, что при разбиении деревьев алгоритм не видит смысла делать более глубокие деревья, и хотя max_depth и большое, все равно глубина деревьев в среднем оказывается меньше этого значения. RMSE падает, но после значения максимальной глубины 9 немного увеличивается и прекращает сильно изменяться в дальнейшем. Поэтому выберем оптимальное значение равное 9 и будем использовать его дальше.

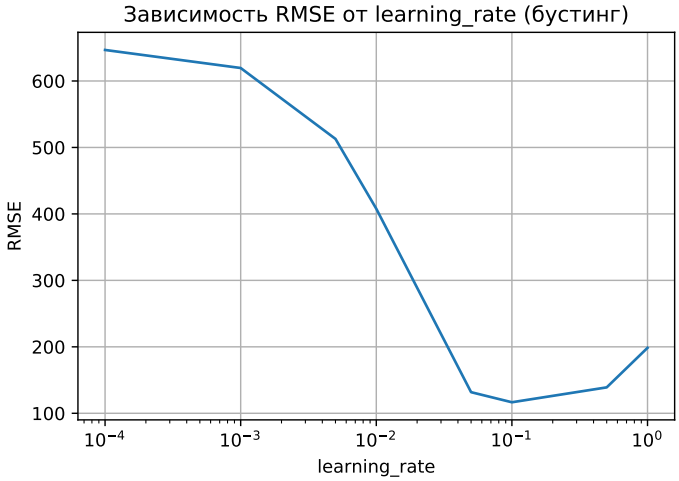
Исследуем теперь зависимости от размера набора признаков, по которому будет выбираться разбиения в верщинах деревьев.





Время работы алгоритма изменяется линейно. RMSE несколько уменьшается до значения 13, дальше же начинает увеличиваться. Вероятно, при большем значении происходит переобучение модели.

Посмотрим на зависимости от learning_rate.



Видим, что время сильно не менялось (2.7 - 3.0 секунд). Но минимум имеется в значении 0.05. Для того, чтобы можно было утверждать, что есть некая закономерность, необходимо проводить много экспериментов и усреднять значения, так как могут быть просто небольшие случайные изменения. RMSE же сначала убывает, достигает своего минимума в точке 0.1, а дальше начинает увеличиваться. Связано это с тем, что при маленьком значении алгоритм не успевает сходиться, а при большом возникает переобучение, которое так же неготивно сказывается на результат.

Отметим, что при параметрах, которые мы отыскали выше, смогли получить такую величину ошибки: 117.79 122.78

4 Выводы

В данной работе были показаны результаты экспериментов с алгоритмами случайного леса и градиентного бустинга, графики зависимости времени работы и функции ошибок RMSE в зависимости от параметров моделей, обученных на датасете с ценой жилья. Заметим, что при найденных нами параметрах в конкретной задаче градиентный бустинг показал результат несколько лучше (117.79 против 122.78).