Композиции алгоритмов для решения задачи регрессии

Григорьев Илья, 317 группа

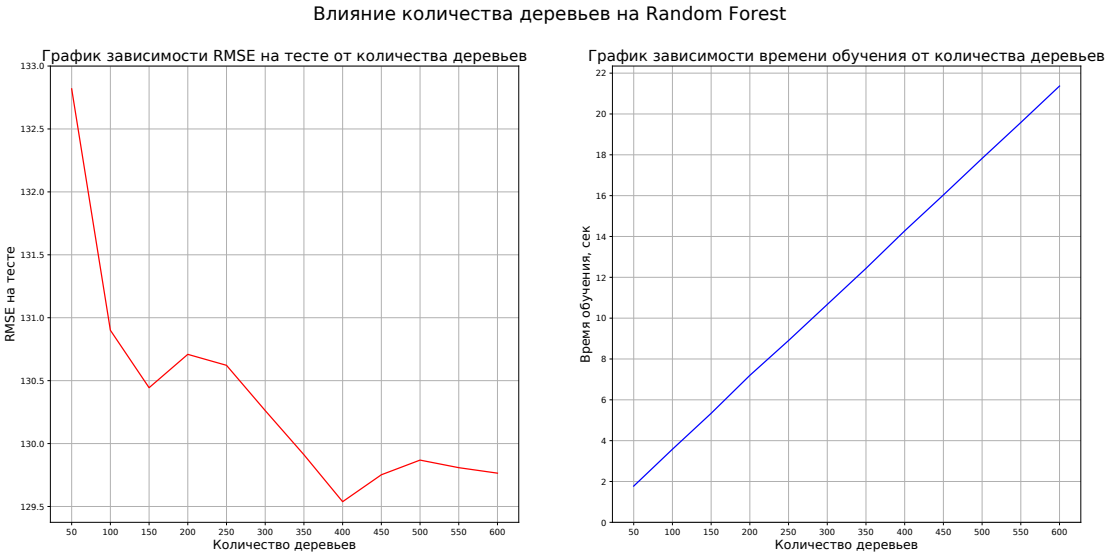
Декабрь, 2020

Введение

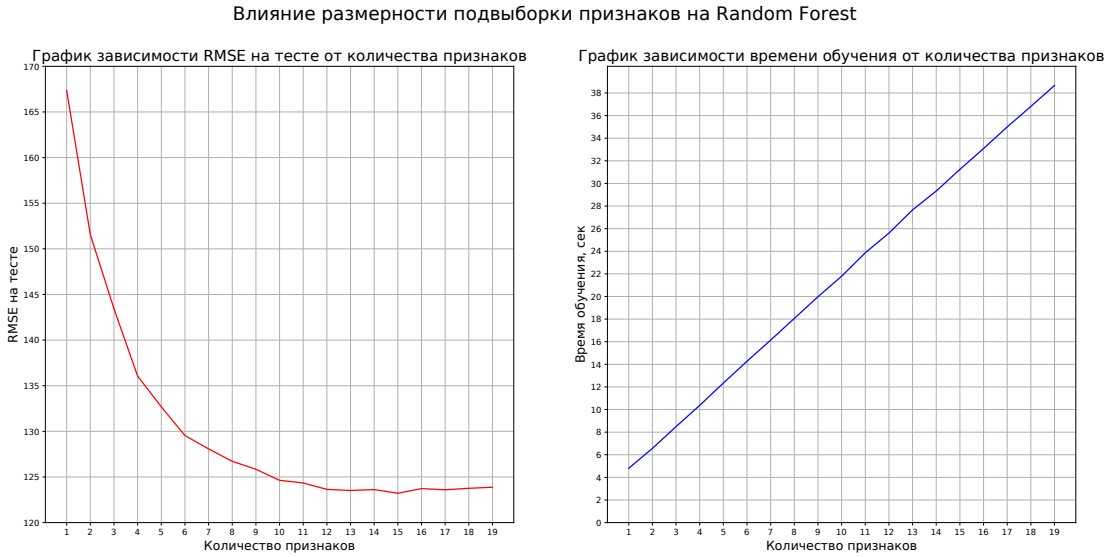
В данном практическом задании надо было реализовать алгоритмы Random Forest и градиентный бустинг для решения задачи регрессии. Далее были проведены эксперименты с этими ансамблевыми алгоритмами на датасете для предсказания цены на жилье.

Эксперимент 2

В этом эксперименте изучался алгоритм Random Forest и его поведение в зависимости от разных параметров. Выбирался какой-то параметр, и для различных значений этого параметра мы смотрели на качество предсказаний на отложенной выборке по метрике RMSE и на время обучения алгоритма.

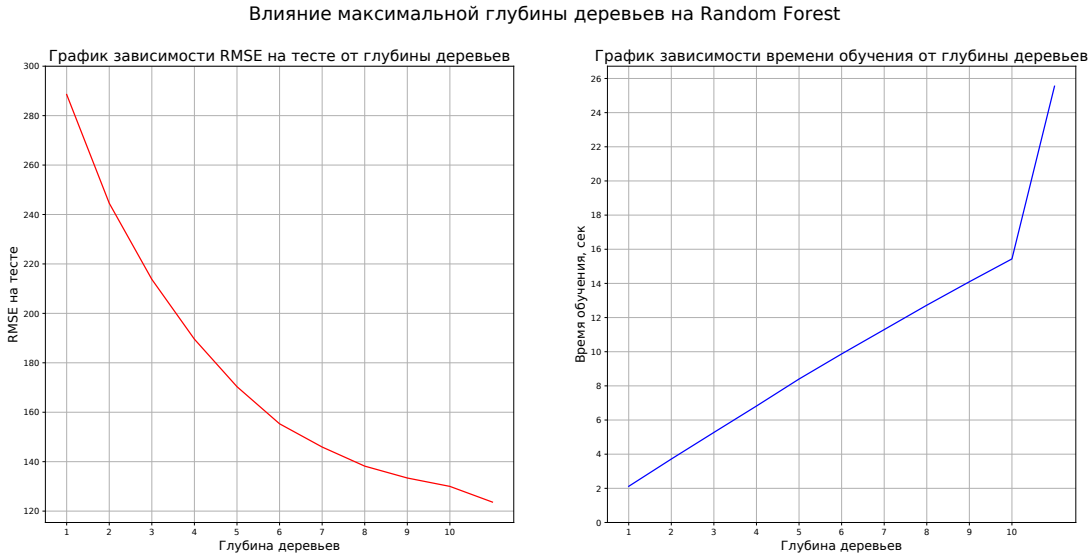


Случайный лес в отличие от бустинга не переобучается с ростом числа деревьев в ансамбле. В среднем RMSE на тесте падает при увеличении числа деревьев, но время обучения линейно растет.



Для регрессии принято брать размерность подвыборки признаков в 3 раза меньше, чем общее число признаков, то есть 6 или 7. По графику видно, что более оптимальными являются значения побольше для

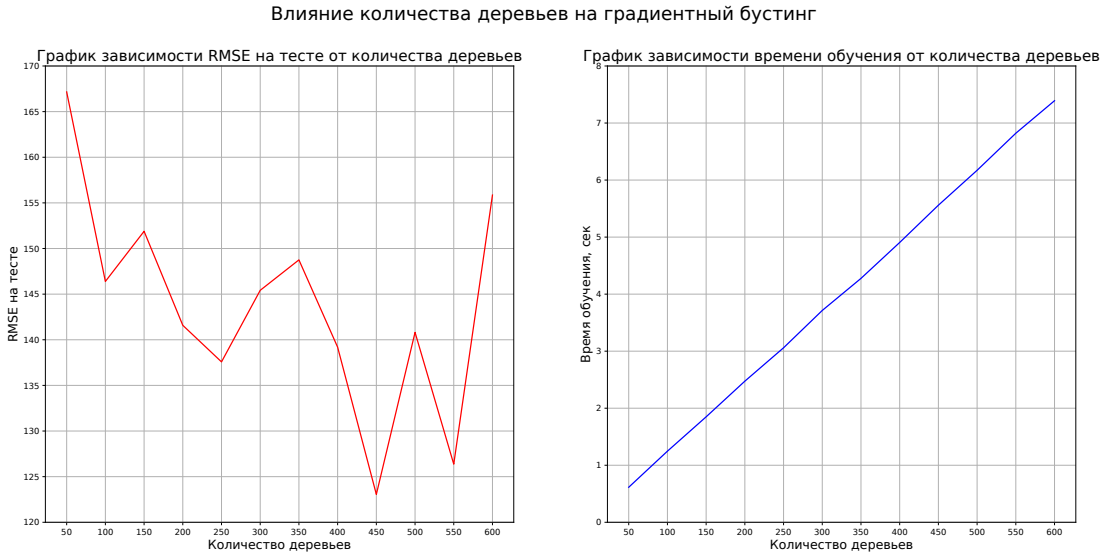
этого параметра, например 12. С увеличением этого параметра RMSE на тесте уменьшается до какого-то момента, а потом особо не изменяется. Время обучения линейно растет с ростом этого параметра.



Случайный лес – это бэггинг над решающими деревьями, а для бэггинга нужны мощные базовые модели. Поэтому с ростом максимальной глубины деревьев, RMSE на тесте уменьшается, а время обучения, очевидно, возрастает. Самый лучший вариант с точки зрения качества – когда глубина деревьев неограниченна.

Эксперимент 3

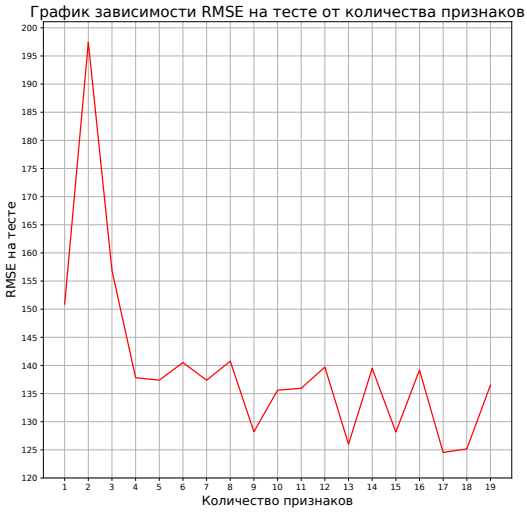
Теперь исследуем поведение градиентного бустинга в зависимости от различных параметров алгоритма.

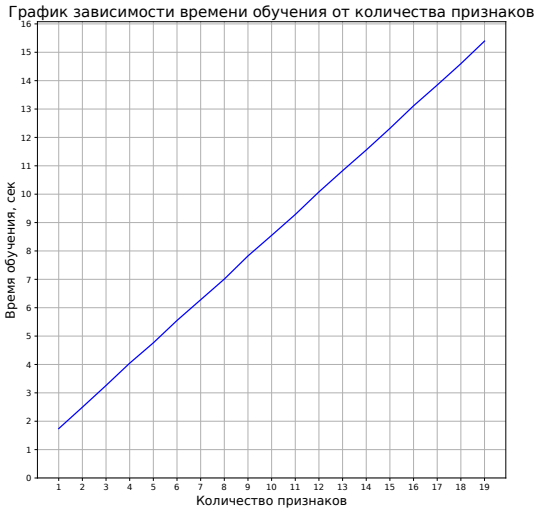


Градиентный бустинг может переобучаться при слишком большом количестве деревьев и неправильно подобранном learning_rate. Важно регулировать эти два параметра как одно целое. При увеличении количества деревьев, время обучения линейно растет.

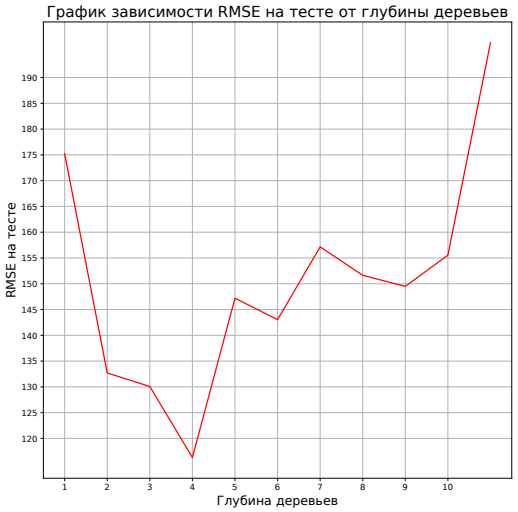
Оптимальные значения для размерности подвыборки признаков при построении разбиения в дереве лежат в диапазоне от 9 до 17 (всего признаков 19). Время обучения линейно растет с ростом этого параметра.

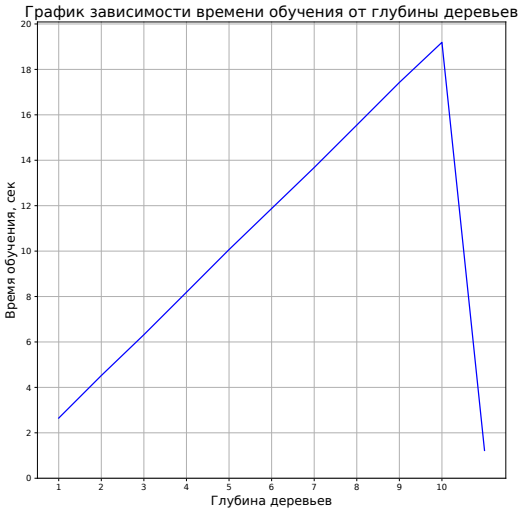
Влияние размерности подвыборки признаков на градиентный бустинг



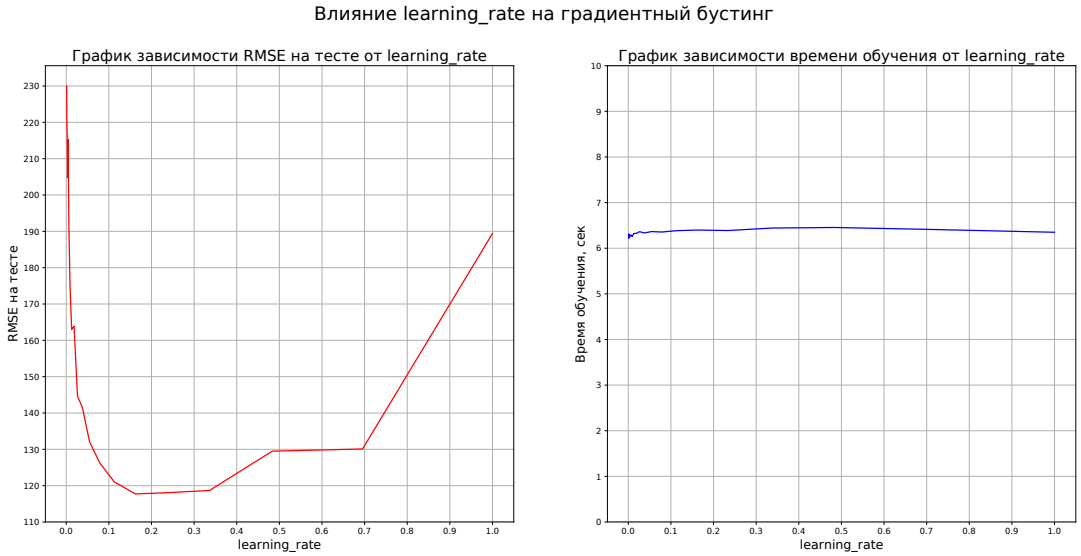


Влияние максимальной глубины деревьев на градиентный бустинг





Для градиентного бустинга не нужны мощные базовые модели, так как следующая модель исправляет ошибки всех предыдущих. Поэтому оптимальные значения для глубины деревьев лежат в диапазоне от 3 до 4. С ростом глубины, растет время обучения.



Параметр learning_rate перебирался по логарифмической шкале. Для n_estimators = 450 оптимальные значения learning_rate лежат в диапазоне от 0.15 до 0.25. При слишком маленьком значении этого параметра алгоритм недообучается, а при слишком большом – переобучается. На время обучения learning_rate не влияет.

Вывод

Ансамблевые алгоритмы случайный лес и градиентный бустинг отлично подходят для работы с разнородными данными. Бустинг обучается быстрее случайного леса и требует более простых базовых моделей. С точки зрения качества эти алгоритмы близки друг к другу, и нельзя сказать, что какой-то однозначно лучше. Надо выбирать алгоритм исходя из задачи.