Отчет по заданию №3

"Композиции алгоритмов для решения задач регрессии".

Градиентный бустинг и случайный лес

Косарев Евгений,
студент 3 курса ВМК МГУ
кафедра ММП,
2020г. Декабрь

0. Введение

В ходе выполения практического задания было реализовано и протестировано 2 алгоритма: градиентного бустинга и случайного леса, а именно GradientBoostingMSE и RandomForestMSE. Все эксперименты проводились на датасете данных о продаже недвижимости. Данные были разделены на обучающую (80%) и тестовую (20%) выборки. Цель задания - пронализировать зависимость значений метрики RMSE и времени выполеннния от определнных гиперпараметров. Далее представлены экспериментальные данные по каждому из алгоритмов.

1. Введение в эксперимент

Для корректного проведения эксперимента из датасета были удалены следующие столбцы:

- index - порядковый номер записи в датасете
- id - уникальный номер записи, не содержит информации о цене недвижимости
- date - дата совершения обращения по имущественному вопросу. Было сделано преположение, что сезонность и год продажи не влияют на таргет

Над каждым алгоритмом была проведена серия экспериментов. Если не оговорено иное, то стандартными параметрами для **GradientBoostingMSE** являются:

- max_depth = None – максимальная глубина решающих деревьев, если None, то она неограничена.
- n_estimators = 20 – число решающих деревьев в ансамбле
- learning_rate = 0.1 – шаг обучения
- feature_subsample_size = None – число признаков, что будут использованы в построении каждого дереве ансамбля, если None, то используются все признаки.

Аналогично, для **RandomForestMSE** стандартный набор параметров:

- max_depth = None – максимальная глубина решающих деревьев, если None, то она неограничена.
- n_estimators = 20 – число решающих деревьев в ансамбле
- feature_subsample_size = None – число признаков, что будут использованы в построении каждого дереве ансамбля, если None, то используются все признаки.

2.Исследование алгоритма RandomForestMSE

В каждом из экспериментов будет исследована зависимость времени и метрики RMSE от перебираемых параметров. Здесь ансамбль - набор независмых алгоритмов, предсказание - усреднение предсказаний каждого алгоритма.

Зависимость от n_estimators

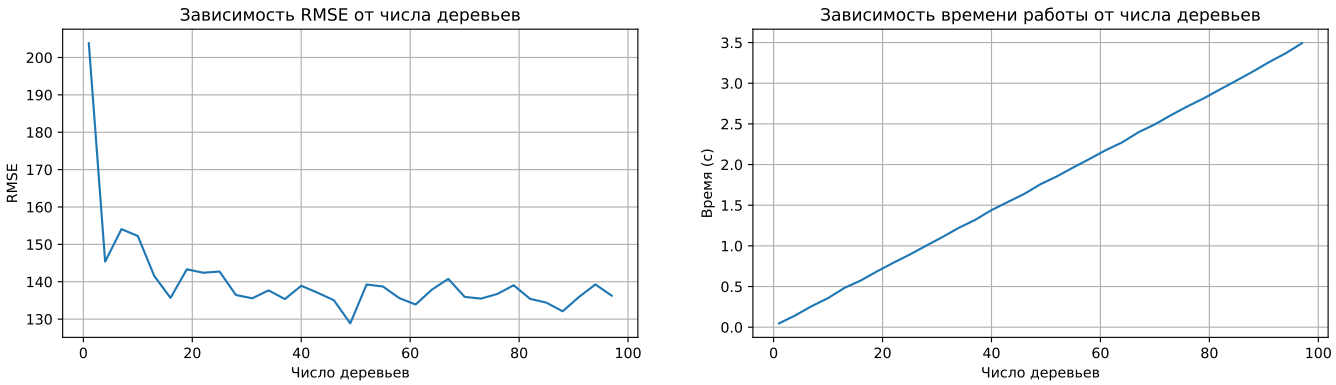

Рис. 1:
Графики зависимости времени работы и RMSE в зависимости от числа деревьев (параметра n_estimators)

Видно, что при параметре увеличении числа деревьев ошибка уменьшается, однако при значениях n_estimators ≥ 20 график функции потерь колеблется вокруг константного значения. При этом, время работы алгоритма линейно возрастает в зависимости от числа деревьев.

Зависимость от feature_subsample_size

Тестирование проводилось на датасете с 16 признаками.

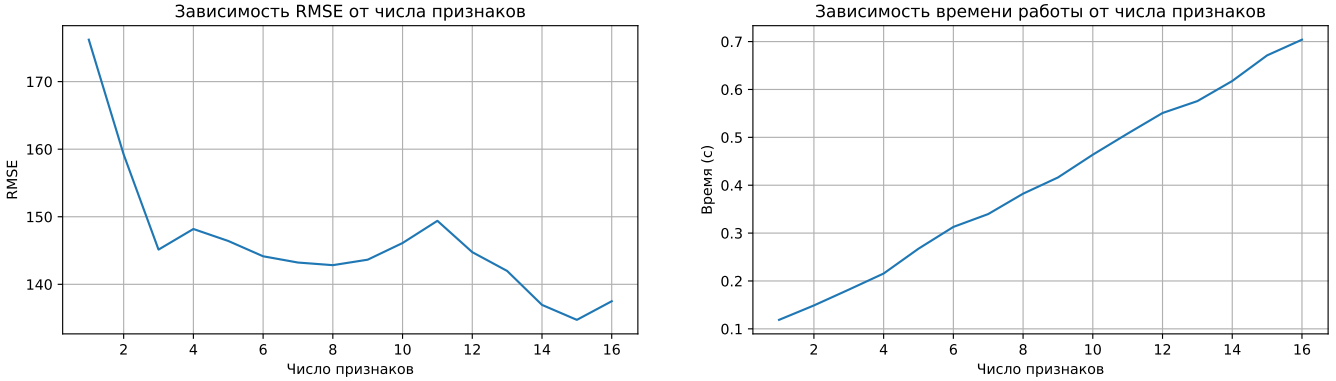

Рис. 2:
Графики зависимости времени работы и RMSE в зависимости от числа признаков (параметра feature_subsample_size)

Затраченное время линейно возрастает в зависимости от числа признаков. Ошибка же тем меньше, чем больше признаков использует модель. Однако случается, что при добавлении новых признаков качество незначительно падает. Это может быть связано с тем, что в модели

появляются менее информативные признаки и усложняют предсказания, но вцелом работает предположение, что чем больше признаков, тем лучше качество.

Зависимость от max_depth

Рассматривается ограничение на глубину дерева, точка "без ограничений" выделена отдельным маркером.

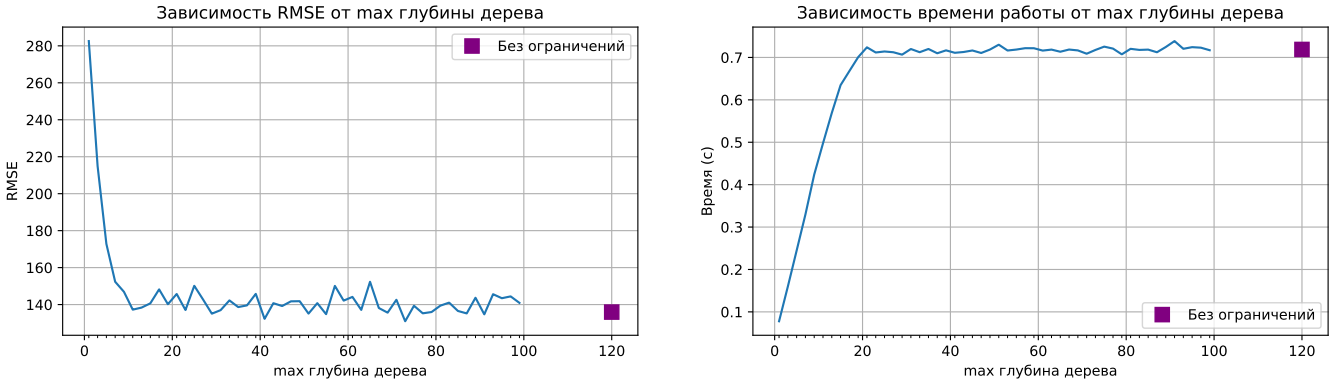

Рис. 3:
Графики зависимости времени работы и RMSE в зависимости от глубины дерева (параметра max_depth)

При небольшой глубине дерева качество работы ансамбля низкое, оно повышается до глубины дерева равной 15, а дальше RMSE колеблется вокруг константы. Время работы же растет с увеличением глубины дерева и выходит на константу при глубине 20. Это связано с тем, что алгоритм не строит более глубокие деревья на выбранном датасете.

Вывод

На выбранных данных подтверждены теоретические выводы о работе RandomForestMSE. Ошибка уменьшается с ростом глубины, и числа деревьев, но существует порог, когда она начинает колебаться вокруг некоторого значения. Увеличение числа признаков уменьшает ошибку, но это верно лишь в случае, когда все они - информативные. Добавление шумовых признаков может ухудшить результат предсказания. Время работы увеличивается при росте каждого из параметров, кроме ограничения на глубину деревьевв силу невозможности строить неограниченные деревья.

3. Исследование алгоритма GradientBoostingMSE

Отличие от предыдущего алгоритма в том, что при построении ансамбля на каждой новой итерации алгоритм учитывает ошибки предсказания на предыдущей. Предсказание строится в зависимости от коэффициентов вхождения каждого алгоритма в итоговый ансамбль.

Зависимость от n_estimators

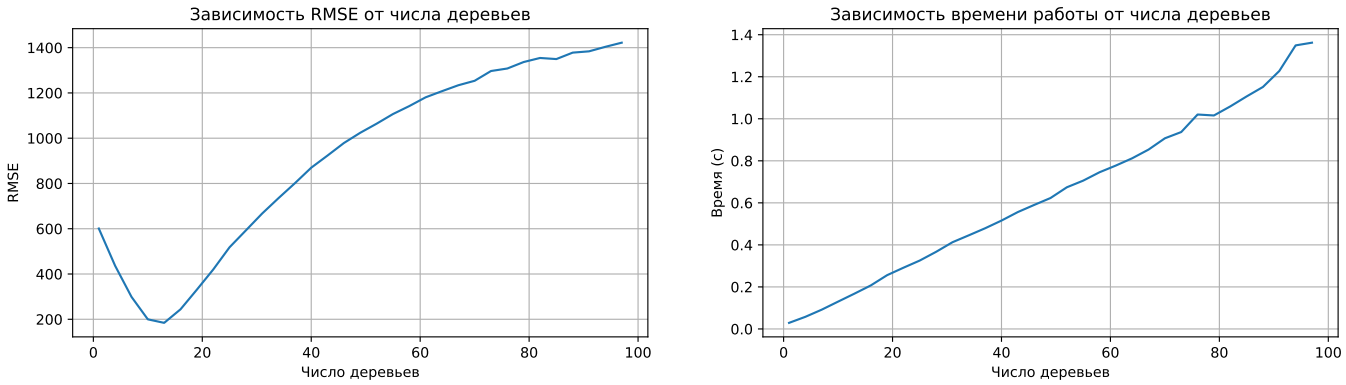

Рис. 4:
Графики зависимости времени работы и RMSE в зависимости от числа деревьев (параметра n_estimators)

График демонсстрирует, что время растет линейно, однако ошибка имеет более интересный характер зависимости. Оптимамальное число деревьев равно 17, дальше ошибка растет. Это свидетельствует о переобучении ансамбля.

Зависимость от feature_subsample_size

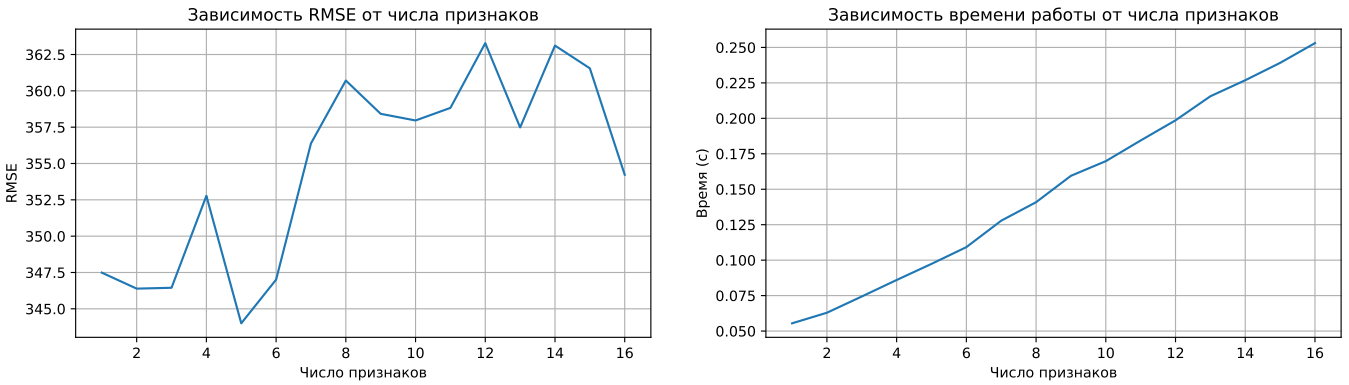

Рис. 5:
Графики зависимости времени работы и RMSE в зависимости от числа признаков (параметра feature_subsample_size)

Лучшее число признаков оказалось равно 5. Это может быть связано с тем, что существует зависимость между итерациями построения ансамбля и разной информативностью признаков. Время растет пропорционально числу деревьев.

Зависимость от max_depth

Аналогично, рассматривается ограничение на глубину дерева, точка "без ограничений" выделена отдельным маркером.

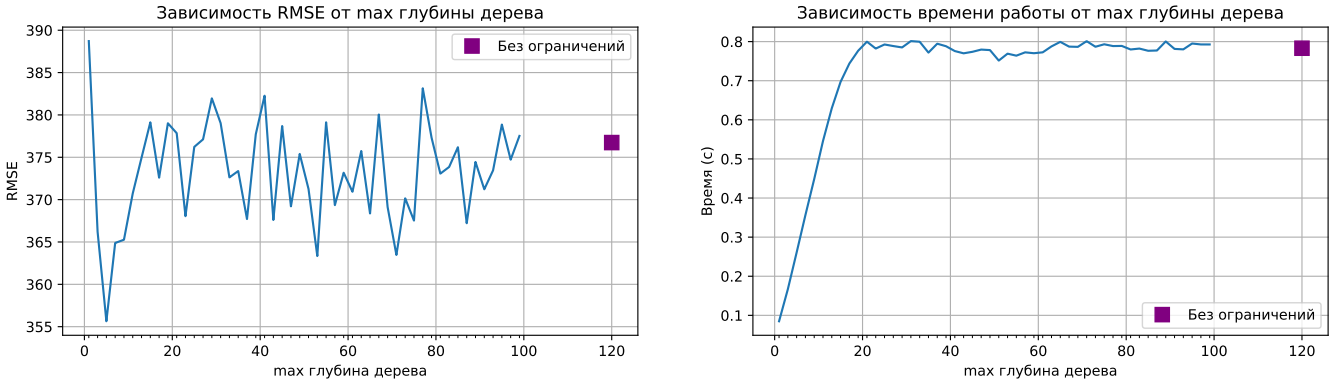

Рис. 6:
Графики зависимости времени работы и RMSE в зависимости от глубины (параметра max_depth)

Лушчая глубина дерева равна 5, далее качество повышается и колеблется вокруг константы. Время работы так же растет до определенного момета, поведение аналогично с экспериментом у алгоритма RandomForestMSE.

Зависимость от learning_rate

Графики представлены ниже. На них видно, что оптимальное значение параметра равно 0.1. Если оно меньше, то ансамбль недообучается, если выше, то переобучается. Время работы колеблется примерно у одного значения, но лучшим можно считать его в случае значения leraning_rate, равного 0.1.

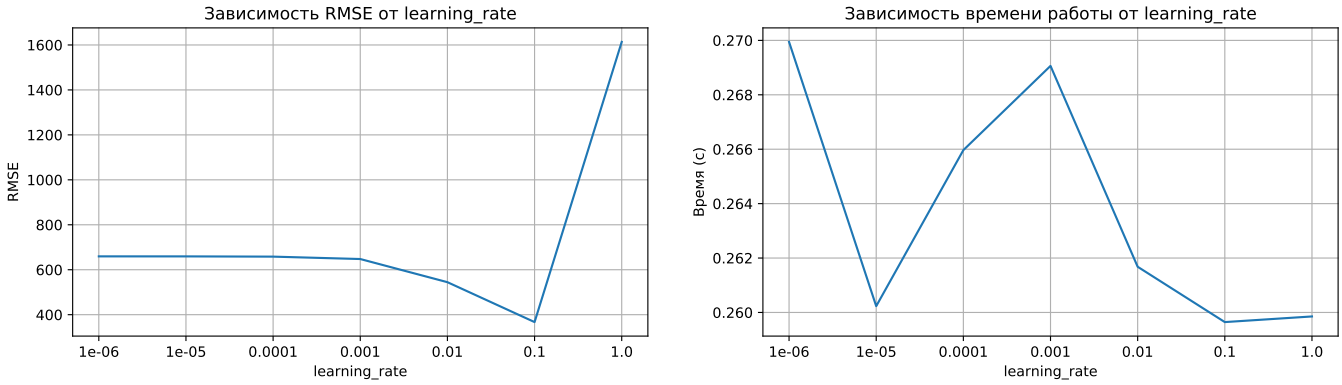

Рис. 7:
Графики зависимости времени работы и RMSE в зависимости от параметра learning_rate

Вывод

Анализ графиков позволяет сделать вывод, что поведение многих зависимостей градиентного бустинга схоже со случайным лесом. Однако, это не всегда так. Увеличение числа деревьев может привести к переобучению, как и увеличение параметра learning_rate, увеличение числа признаков не всегда дает улучшение в качестве предсказания. Можно считать, что значение параметра learning_rate не влияет на скорость работы алгоритма.

Ссылки на dockerhub и github

Вся реализация сервера расположена по ссылке: GitHub.
Образ докера расположен по ссылке: DockerHub.