Московский государственный университет
имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математических методов прогнозирования

Отчёт по заданию 3.
Ансамбли алгоритмов. Веб-сервер. Композиции алгоритмов для решения задачи регрессии.

Выполнила:
Пронина Наталия
317 группа

Москва
2021

Содержание

Введение

В данном отчёте представлено описание экспериментов, которые были направлены на исследование двух моделей ансамблирования решающих деревьев: случайный лес (RF) и градиентный бустинг (GB). Эксперименты проводились на датасете с данными о недвижимости (House Sales in King County, USA).

Цель: предсказать цену недвижимости

Рассматривается задача регрессии с метрикой качества $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - a_i)^2}$, где N — размер выборки, y_i — истинное значение целевой переменной на i-м объекте, a_i — предсказанное.

Эксперименты

1) Предобработка текста

В датасете имеется только один признак типа object - это дата. Разобьём этот признак на три целочисленных - число, месяц и год, в итоге получается 21 признак. Удалим признак «id». Разделим выборку на обучающую и валидационную в отношении 7:3.

2) Случайный лес

Будем поочереди фиксировать лучшие найденные параметры и подбирать следующие. При исследовании зависимости от количества деревьев алгоритм обучался только один раз, так как в методе fit все исследуемые величины сохраняются в отдельные массивы.

2.1) Количество деревьев в ансамбле

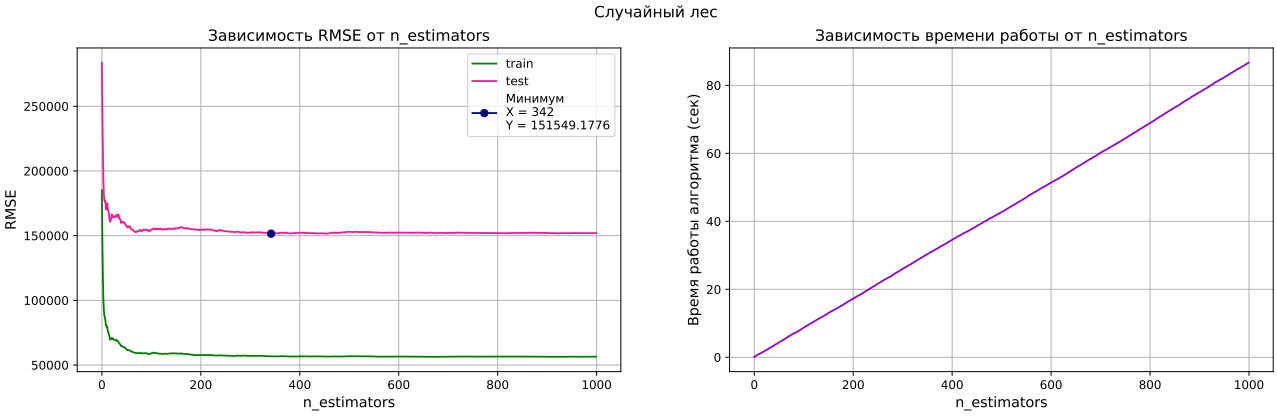

Рис. 1: Зависимость RMSE и времени работы от количества деревьев в ансамбле (RF)

Параметр $n\_estimators$ в RF регулирует количество деревьев, На графике 1 видно, что ошибка на обучающей выборке монотонно убывает. На валидации же немного возрастает при увеличении количества деревьев от $\sim$ 350, это свидетельствует о малом переобучении. Время обучения деревьев примерно одинаковое, поэтому было ожидаемо, что время обучения леса будет зависеть линейно от числа деревьев.

2.2) Размерность подвыборки признаков для одного дерева

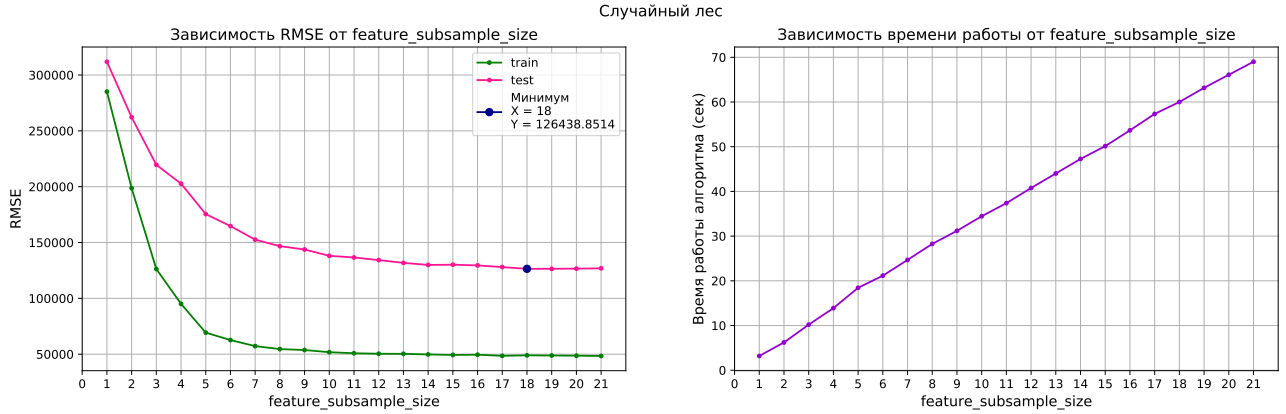

Рис. 2: Зависимость RMSE и времени работы от размерности подвыборки признаков для одного дерева (RF)

Параметр $feature\_subsample\_size$ определяет по скольким признакам будет обучаться каждое дерево. Из графика 2 делаем вывод, что выгоднее всего будет взять почти все признаки. В данном случаем так же наблюдается линейная зависимость времени от числа признаков.

2.3) Максимальная глубина дерева

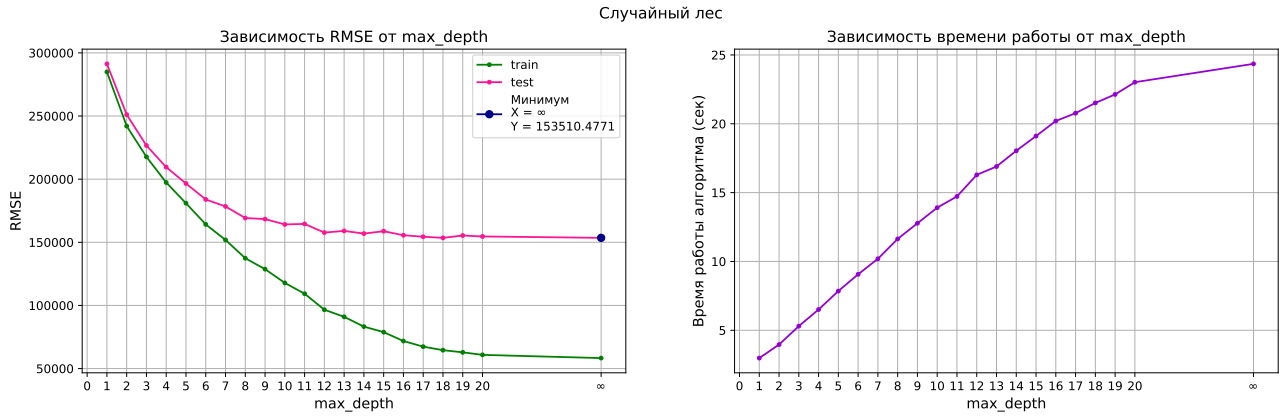

Рис. 3: Зависимость RMSE и времени работы от максимальной глубины каждого дерева (RF)

Параметр max_depth определяет глубину каждого дерева в ансамбле.

В данном эксперименте отдельно рассматривался случай неограниченной глубины (он условно обозначается «∞» на графике 3). Из теории мы знаем, что бэггинг лучше всего стоить на переобученных деревьях, что подтверждается нашим опытом.

Скорее всего, использование деревьев неограниченной глубины было бы невыгодно по времени для большего признакового пространства или для большей обучающей выборки, но в нашем случае время увеличивается всего на 4-8%.

Время возрастает чуть-чуть медленнее, чем линейно.

3) Градиентный бустинг

3.1) Количество деревьев в ансамбле

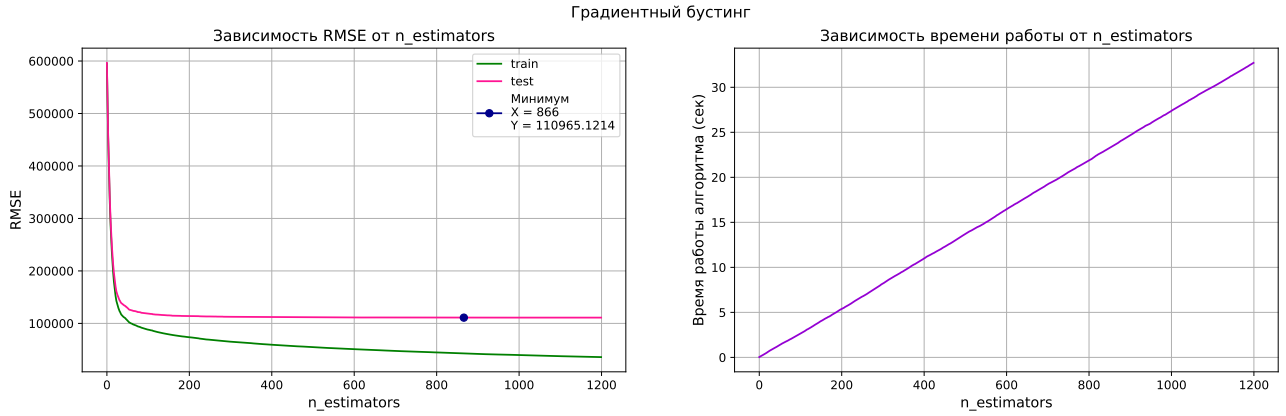

Рис. 4: Зависимость RMSE и времени работы от количества деревьев в ансамбле (GB)

Как и в модели случайного леса на графике 4 видно, что ошибка на обучающей выборке монотонно убывает. На валидации же практически не видно переобучения при увеличении количества деревьев от ~ 850. Из теории мы знаем, что бустинг переобучается только на громадном количестве деревьев, но такие эксперименты потребовали бы намного больше времени.

Время обучения деревьев примерно одинаковое, поэтому было ожидаемо, что время обучения леса будет зависеть линейно от числа деревьев.

3.2) Размерность подвыборки признаков для одного дерева

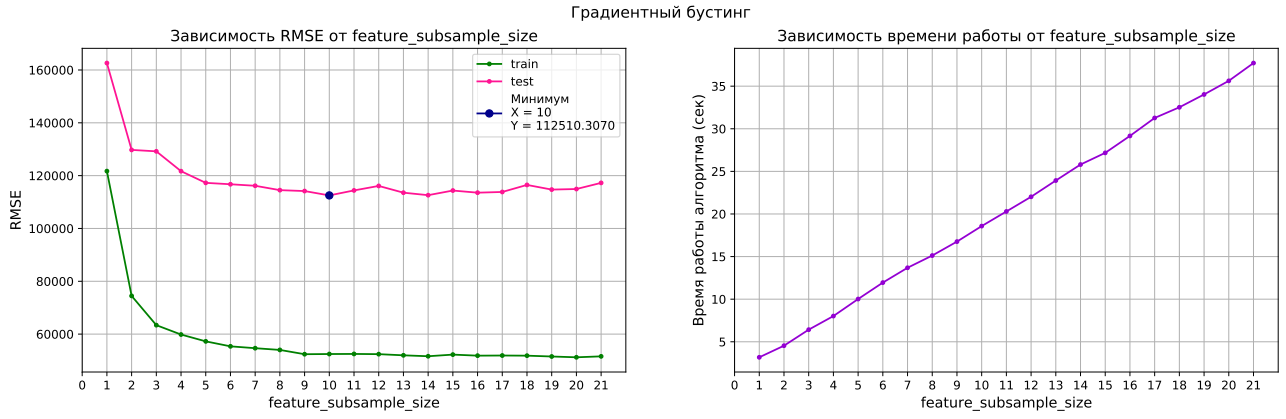

Рис. 5: Зависимость RMSE и времени работы от размерности подвыборки признаков для одного дерева (GB)

Видно, что графики 2 и 5 имеют разное поведение. В случае градиентного бустинга нет стремления к какой-либо асимптоте, экстремум легко определяется «на глаз». Зависимость времени так же линейная.

3.3) Максимальная глубина дерева

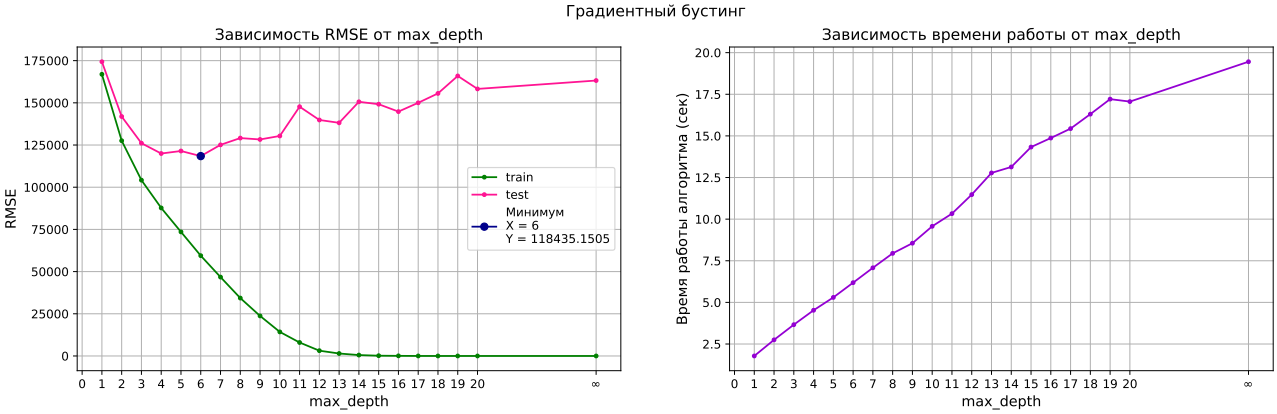

Рис. 6: Зависимость RMSE и времени работы от максимальной глубины каждого дерева (GB)

В градиентном бустинге обучение обычно проводится на большом количестве неглубоких деревьев, что сочетается с результатами 4 и 6 экспериментов. При увеличении глубины бустинг очень сильно переобучается, ошибка на обучающем выборке стремится к нулю.

Зависимость времени обучения тоже можно считать линейной, если не учитывать случай неограниченной глубины.

3.4) Выбор темпа обучения

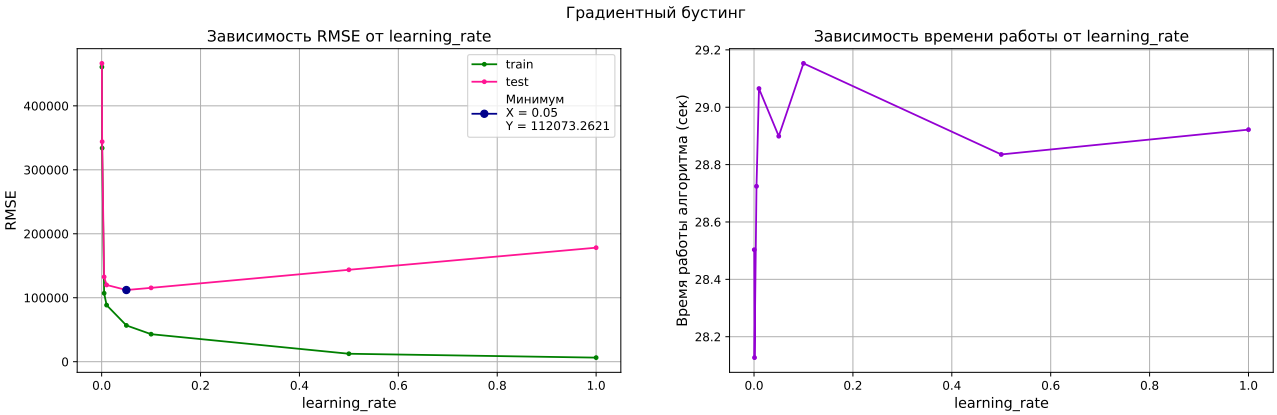

Рис. 7: Зависимость RMSE и времени работы от темпа обучения (GB)

В методе градиентного бустинга каждое следующее дерево зависит от предыдущих, пытается «исправить» их ошибки. Темп обучения (параметр $learning\_rate$) определяет, насколько быстро меняются деревья.

Из графика 7 видно, что от данного параметра больше всего зависит предсказание модели, на маленьких и больших значениях ошибка сильно возрастает.

Ранее по умолчанию использовалось значение $learning\_rate = 0.1$, что близко к оптимальному.

Сравнение моделей

Параметры	n_estimators	feature_subsample_size	max_depth	learning_rate
RF	350	8	∞	–
GB	850	10	6	0.05

Таблица 1: Сравнение оптимальных параметров

Алгоритмы	RMSE на валидации	Время (сек)
RF	153510.4771	25
GB	$112073,2621$	6.75

Таблица 2: Сравнение точности и времени

В заключение сравним ошибку и время работы на лучших параметрах. Из таблицы 2 видим, что градиентный бустинг имеет меньшую ошибку, а по скорости превзошёл RF в несколько раз.

Выводы

И в начале исследований было понятно, что бустинг будет давать лучшие предсказания, так как каждая следующая модель становится «лучше» всех предыдущих. Но из проведённых экспериментов видно, что бустинг даёт лучшее предсказание и за меньшее время, что показывает его очевидное превосходство над бэггингом.