Московский государственный университет имени М.В. Ломоносова

Факультет вычислительной математики и кибернетики
Кафедра Математических Методов Прогнозирования

Отчёт о выполнении задания №3: "Ансамбли алгоритмов. Композиции алгоритмов для решения задачи регрессии"

по курсу "Практикум на ЭВМ"

Прокудин Дмитрий
Сергеевич
317 группа

Москва
2022

Содержание

1 Введение

В данном задании осуществляется детальное ознакомление с алгоритмами композиций: случайным лесом и градиентным бустингом - проводится исследование зависимости точности моделей и времени обучения от их сложности на примере задачи регрессии с использованием датасета данных о продаже недвижимости. Для выполнения задания на языке Python пишутся собственные реализации случайного леса и градиентного бустинга на основе **DecisionTreeRegressor** из библиотеки scikit-learn.

2 Эксперименты

2.1 Предобработка данных

Эксперименты этого задания проводятся датасете данных о продаже недвижимости "House Sales in King County, USA". Пример датасета представлен в таблице 1. Всего в датасете 21 признак и 21613 записи.

Таблица 1: Пример первых 7 столбцов изначального датасета

	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot
0	7129300520	20141013T000000	221900.0	3	1.00	1180	5650
1	6414100192	20141209T000000	538000.0	3	2.25	2570	7242
2	5631500400	20150225T000000	180000.0	2	1.00	770	10000

Для дальнейшей работы с датасетом из него были извлечены три столбца: 'price' - как целевая переменная, 'id' и 'date' - как столбцы, имеющие не скалярный тип. Датасет из оставшихся 18 столбцов без дополнительных изменений был преобразован в формат numpy.ndarray, а затем разделён на обучающую и валидационную/отложенную выборки в отношении 7 : 3.

2.2 Исследование точности и времени работы алгоритмов

Для изучения зависимости точности алгоритмов, а также времени их работы от сложности модели проводится перебор следующих гиперпараметров моделей:

- Количество деревьев в ансамбле: $1 - 900$;
- Размерность подвыборки признаков для одного дерева: $[1, 6, 12, 18]$ - для случайного леса и $[6, 12, 18]$ - для градиентного бустинга;
- Максимальная глубина дерева: [1, 5, 10, 15, без_ограничения] - для случайного леса и [1, 5, 15, без_ограничения] - для градиентного бустинга;
- Learning_rate: $[0.1, 0.5, 1.0, 2]$ - только для градиентного бустинга.

Зависимость точности (на метрике RMSE) для случайного леса представлена на рис.1. Далее на графиках количество деревьев в ансамбле отмечено по оси x; в легенде 'fss' обозначает размер подвыборки признаков для каждого дерева, а 'm_d' - максимальную глубину каждого дерева (для градиентного бустинга параметр 'lr' отвечает за learning_rate).

Зависимость времени обучения для случайного леса представлена на рис.2. Замеры времени проводились при добавлении новых 100 деревьев в модель, то есть график построен по 10 точкам (аналогичным образом замеры проводились и для градиентного бустинга).

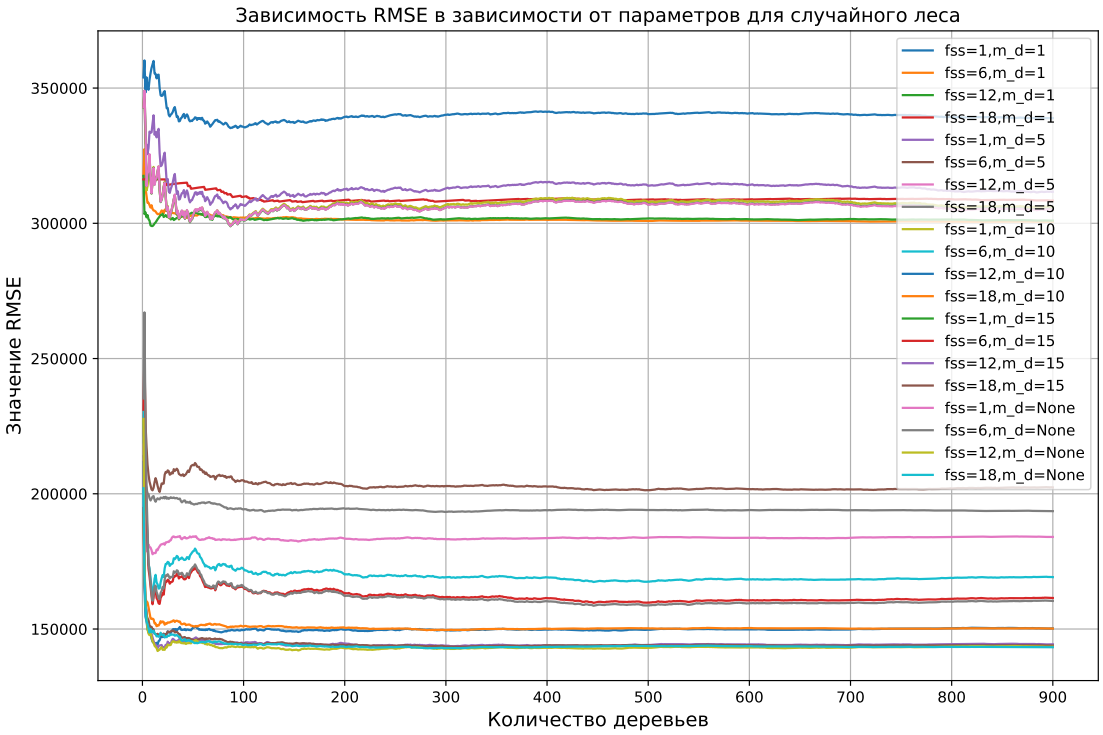

Рис. 1: Зависимость значений RMSE от параметров случайного леса.

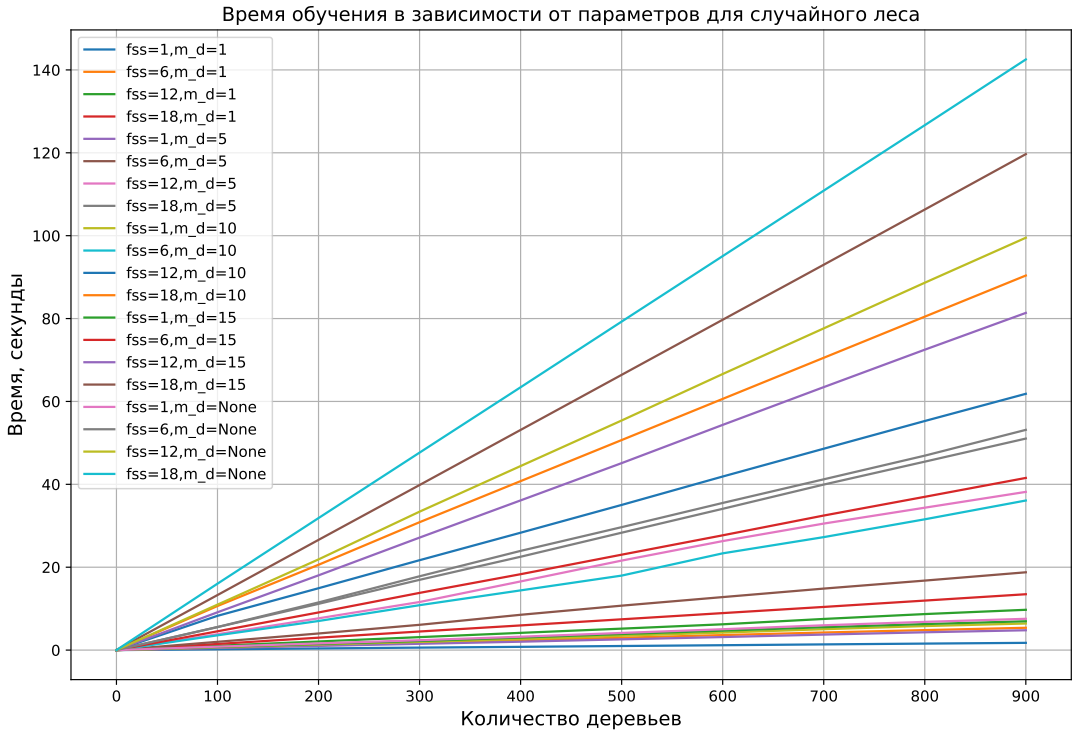

Рис. 2: Зависимость времени обучения от параметров случайного леса.

На основе результатов экспериментов для случайного леса можно сделать следующие выводы:

- Время обучения ансамбля линейно зависит от количества деревьев, при этом время обучения каждого дерева растёт при увеличении размера подвыборки признаков и глубины деревьев.

- Результаты ансамблей из деревьев малой глубины $(1, 5)$ существенно хуже результатов ансамблей с большей глубиной (значение RMSE больше в $1.5 - 2$ раза). Самые точные ансамбли построены с использованием деревьев без ограничения на глубину.

- Размер подвыборки признаков для каждого дерева оказывает значительно меньшее влияние на итоговую точность, чем глубина деревьев.

- При большом количестве деревьев в ансамбле добавление новых деревьев не оказывает существенный вклад на точность.

Зависимость точности (на метрике RMSE) для градиентного бустинга представлена на рис.3, 4. Большое значение learning_rate $= 2$ привело к тому, что алгоритмы не сошлись, а точность упала, поэтому на рис.5 представлены результаты измерений точности для learning_rate $= 0.1$ (результаты для остальных значений learning_rate имеют аналогичный вид и не представлены на графиках).

Зависимость времени обучения для градиентного бустинга представлена на рис.6 и рис.7.

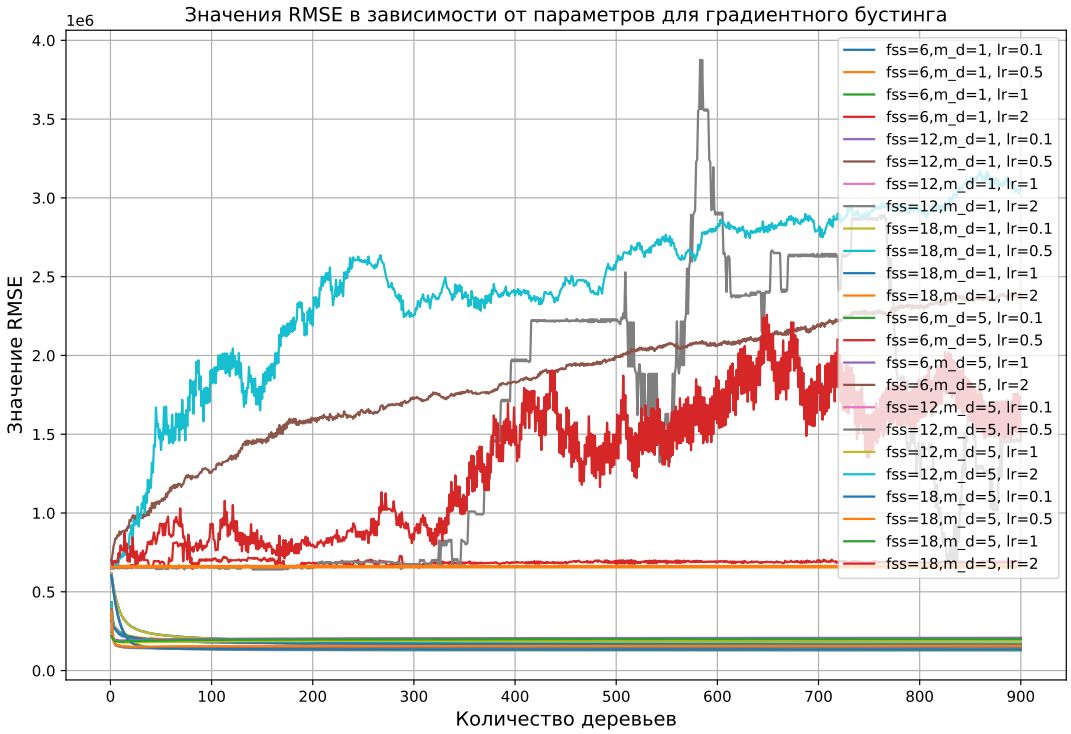

Рис. 3: Зависимость значений RMSE от параметров градиентного бустинга (первая половина параметров).

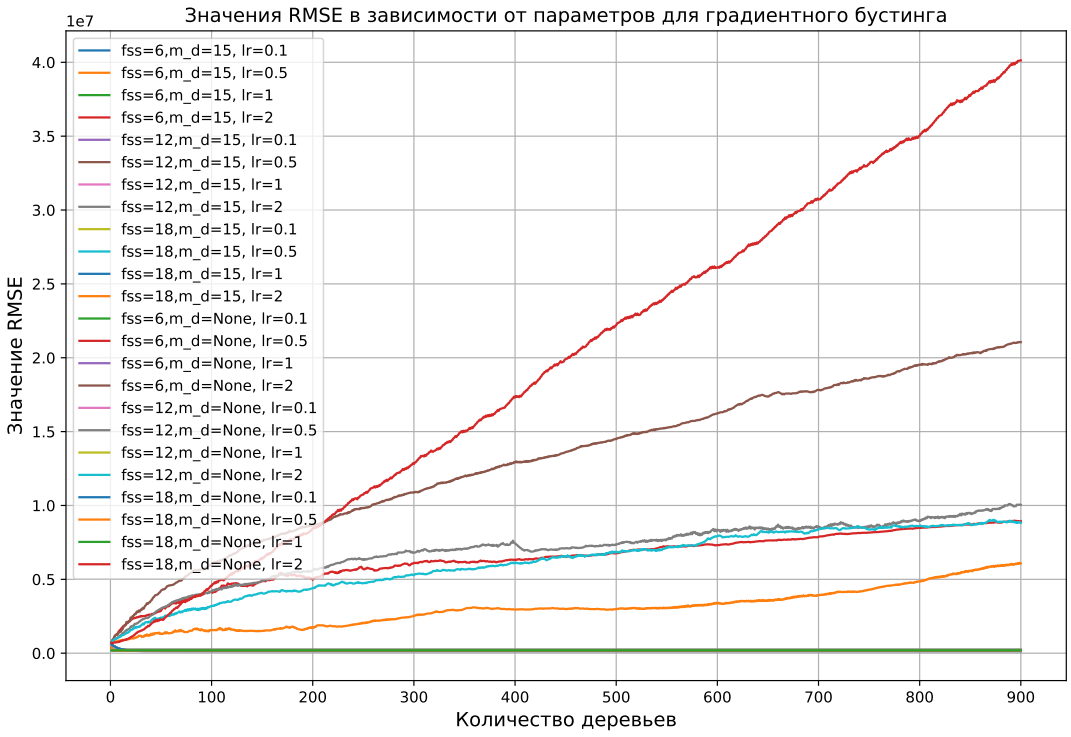

Рис. 4: Зависимость значений RMSE от параметров градиентного бустинга (вторая половина параметров).

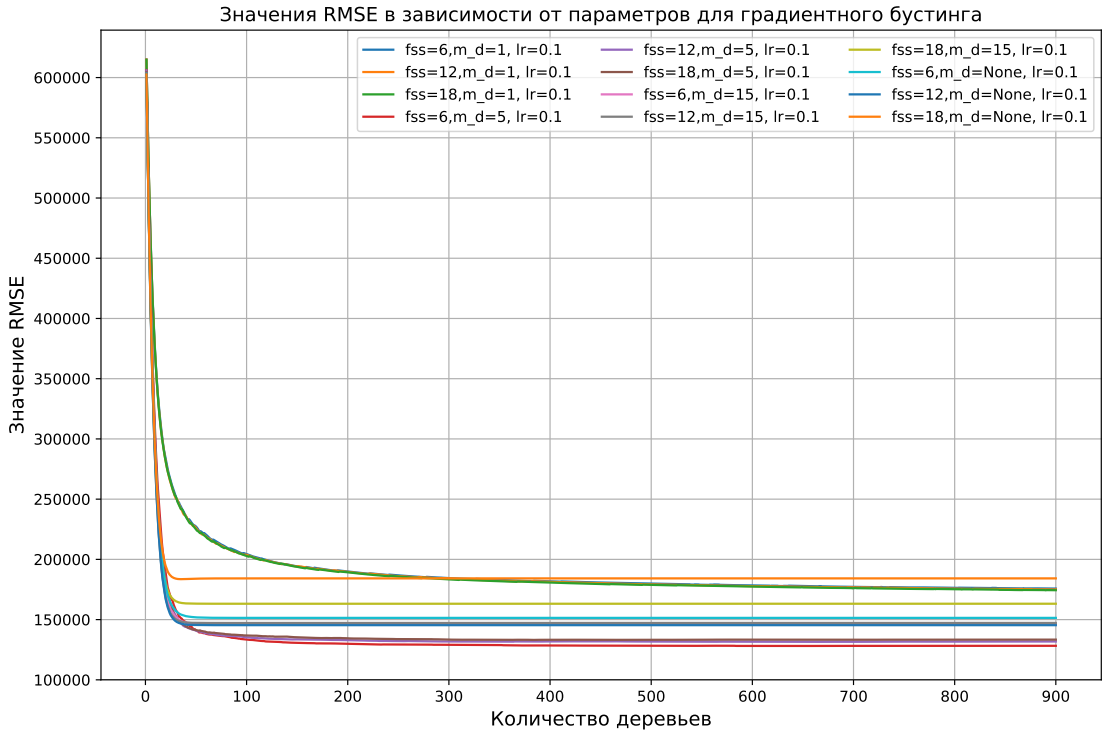

Рис. 5: Зависимость значений RMSE от параметров градиентного бустинга (для lr = 0.1).

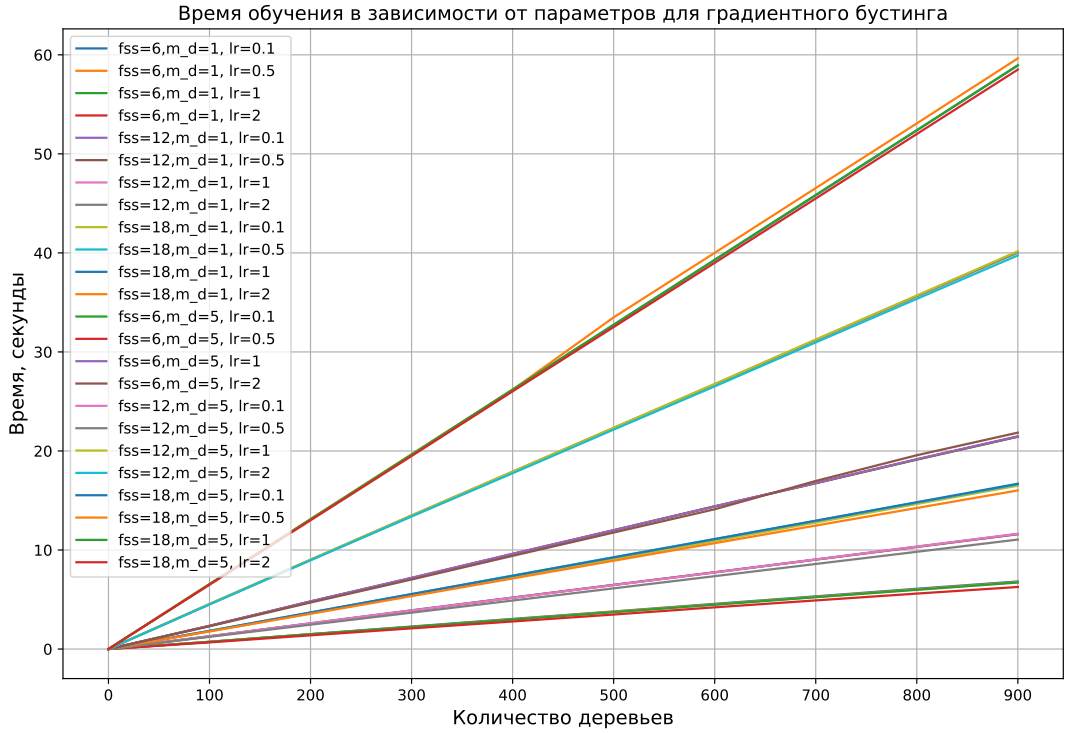

Рис. 6: Зависимость времени обучени от параметров градиентного бустинга (первая половина параметров).

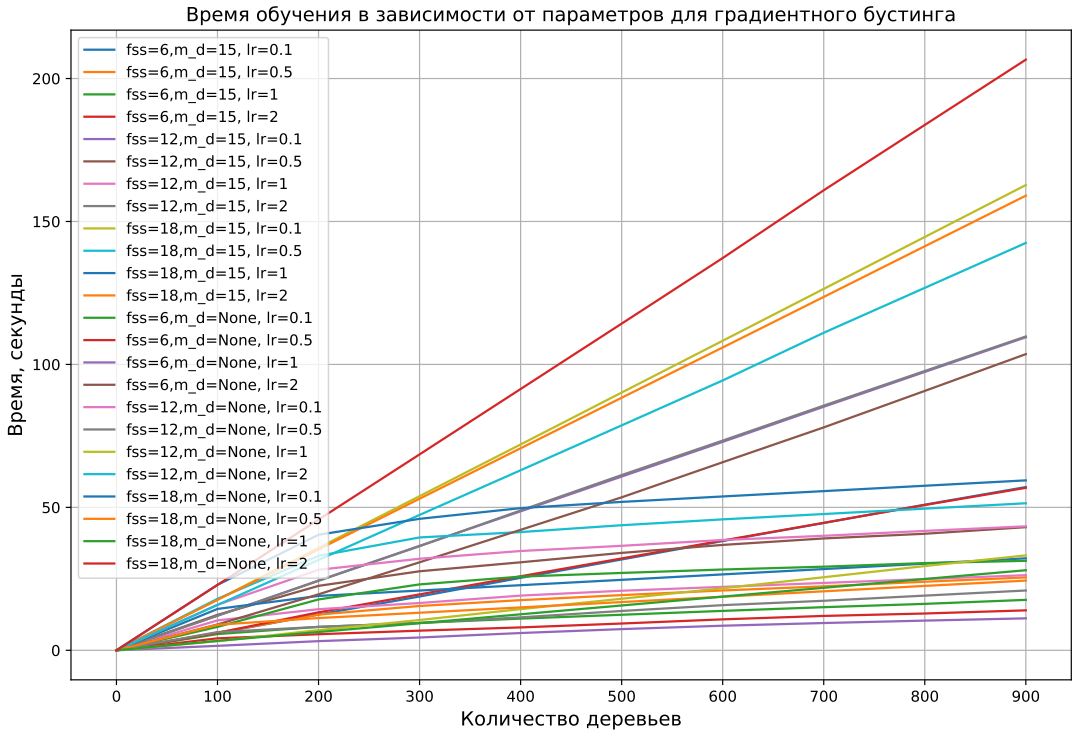

Рис. 7: Зависимость времени обучени от параметров градиентного бустинга (вторая половина параметров).

На основе результатов экспериментов для градиентного бустинга можно сделать следующие выводы:

- Для большинства комбинаций параметров время обучения зависело от количества деревьев линейно, однако есть примеры, где данная зависимость не наблюдается. При этом время обучения моделей градиентного бустинга оказалось больше времени обучения моделей случайного леса.
- Точность ансамблей зависит от глубины деревьев и размера подвыборки признаков в меньшей степени. Кривая точности имеет достаточно гладкий вид.
- Большое значение learning_rate (> 1) привело к 'расхождению' модели.
- Аналогично случайному лесу при большом количестве деревьев в ансамбле добавление новых деревьев не оказывает существенный вклад на точность.

3 Вывод

- В ходе выполнения данного задания было исследовано поведение точности и времени обучения ансамблей решающих деревьев в зависимости от сложности модели: время обучения и точность растут с увеличением глубины деревьев и размера подвыборки признаков.
- Градиентный бустинг показал результат лучше, чем случайный лес, но незначительно дольше обучался.
- Несмотря на небольшие различия в точности, обе модели хорошо подходят для решения данной задачи, время обучения отдельной модели невелико.
- Влияние гиперпараметров (количества деревьев, глубины и размера подвыборки признаков) на точность модели у случайного леса значительнее, чем у градиентного бустинга.