Отчет о выполненной работе по теме "Ансамбли алгоритмов. Веб-сервер. Композиции алгоритмов для решения задачи регрессии."

Бессонов Аркадий

Факультет вычислительной математики и кибернетики

Московского государственного университета им. М.В. Ломоносова

Кафедра математических методов прогнозирования, группа 317.

beccohov.a@yandex.ru

12 декабря 2021 г.

Аннотация

В данной работе проводится исследование параметров алгоритмов случайного леса и градиентного бустинга над деревьями. Оба алгоритма предварительно были реализованы с использованием библиотек numpy, scipy, sklearn (для использования деревьев). Проводится анализ влияния каждого параметра на точность регрессии, а также время обучения модели. Все эксперименты сопровождаются графиками, улучшающими восприятие. Для удобства также фиксируются наиболее оптимальные параметры при исследовании влияния других. Все эксперименты проведены на данных о продажах недвижимости «House Sales in King County, USA». В отчете представлена экспериментальная часть работы, техническая часть реализации модели имеется в репозитории.

1 Введение

Алгоритм случайного леса является одним из немногих универсальных алгоритмов. Универсальность заключается, во-первых, в том, что он хорош во многих задачах. Во-вторых, в том, что есть случайные леса для решения задач классификации, регрессии, кластеризации, поиска аномалий, селекции признаков и многого другого. Случайный лес - это множество решающих деревьев. В задаче регрессии их ответы усредняются, в задаче классификации принимается решение голосованием по большинству. При этом каждое дерево в ансамбле леса должно быть хорошим и как можно меньше зависеть от остальных деревьев для снижения разброса. Для бустинга ситуация обратная, каждое дерево в ансамбле должно быть достаточно простым, чтобы не было переобучения базового алгоритма.

2 Предобработка данных

В предложенных для рассмотрения данных содержится много категориальных признаков, а также лишних признаков. Из набора данных были исключены признаки «date» и «id», поскольку признак «id» нерепрезентативен, а «date» содержит слишком мало дат (данные всего за два года). В качестве категориальных признаков рассматривались те признаки, которые нельзя упорядочить по значениям. Например, «zipcode», «view», «condition», «grade». Эти признаки характеризуют общую оценку недвижимости и её местоположение («zipcode»). Заметим, что в данной работе не было проведено генерации признаков, поскольку целью является демонстрация влияния параметров на поведение алгоритма. Поэтому, например, географические координаты не были преобразованы в расстояния. Для кодирования категориальных признаков использовался target-encoding со сглаживанием (по тренировочной выборке), поскольку при его использовании размерность признакового пространства остается той же, а результаты не хуже, чем при использовании One-Hot кодирования.

3 Результаты экспериментов

3.1 Алгоритм «Случайный лес»

Исследование зависимости от количества деревьев в ансамбле

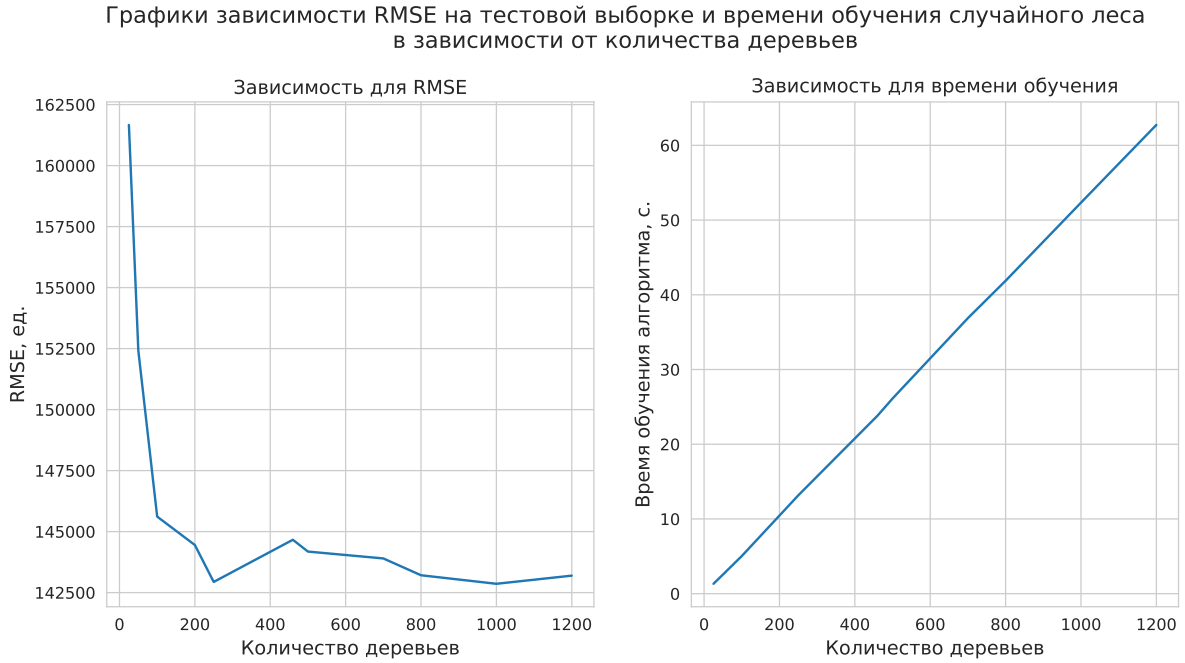

Рис. 1: Зависимость RMSE и времени работы алгоритма случайного леса от количества деревьев

В данном эксперименте фиксировался размер подвыборки (около 60%), а также максимальная глубина дерева (равная 8). Выше (рис. 1) показан график полученных результатов. Видно, что в промежутке до 250 деревьев наблюдается значительный прирост в точности регрессии, далее наблюдается выход на плато. При этом время

обучения линейно растет по количеству деревьев, поэтому оптимальным в данном случае будет 250 деревьев.

Исследование зависимости от размерности пространства признаков

При исследовании влияния размерности подвыборки исследуются различные значения параметра feature_subsample_size при фиксированных остальных значениях параметров (250 деревьев и максимальная глубина 8). Полученные результаты отражены на графике ниже (рис. 2). Заметим, что в данной реализации применяется

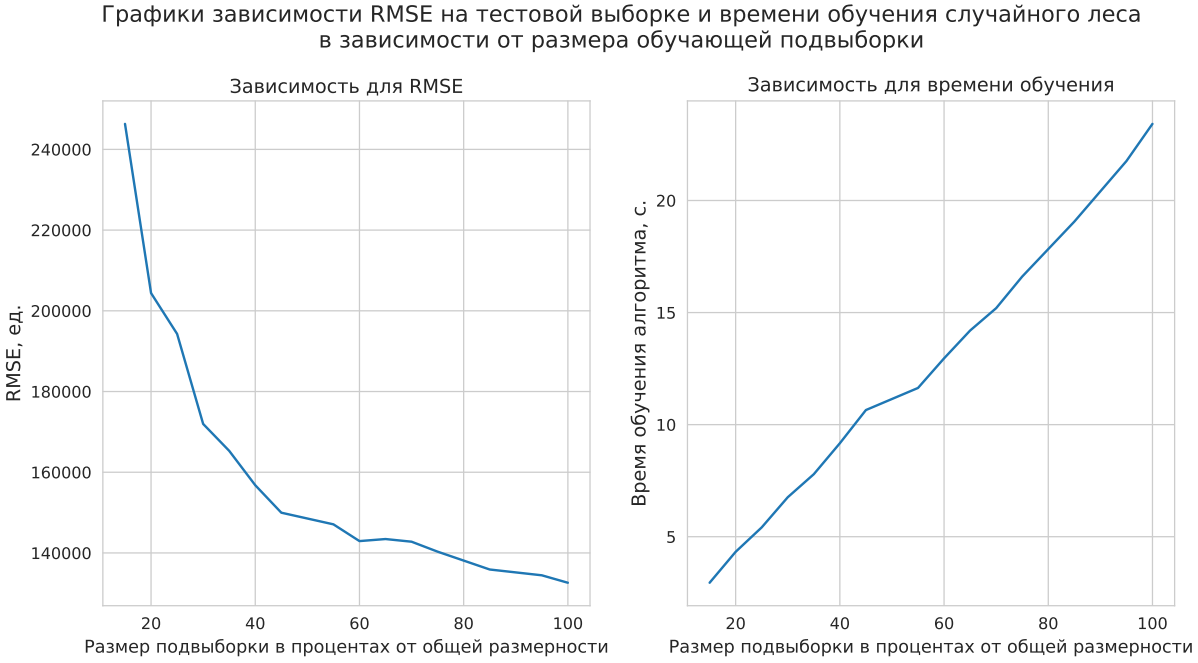

Рис. 2: Зависимость RMSE и времени работы алгоритма случайного леса от размерности подвыборки признаков

способ выбора с возвращением, как это обычно принято. Видно, что, как и ожидалось, при увеличении размерности признакового пространства время обучения увеличивается линейно, а точность регрессии возрастает (каждый алгоритм в композиции становится более качественным при обучении на большей размерности). Для удобства размер подвыборки приведен в процентах от исходной выборки (равной 18), чтобы была видна аналогия с значением параметра feature_subsample_size. Из графиков видно, что оптимальное значение этого параметра порядка 0.6, поскольку время обучения мало, а дальнейшее уменьшение ошибки не такое большое.

Исследование зависимости от максимальной глубины дерева

В данном эксперименте фиксировано количество деревьев (250) и размерность подпространства признаков (0.6). Исследуется влияние ограничения максимальной глубины дерева на ошибку регрессии и время обучения алгоритма случайного леса. График полученных результатов приведен ниже (рис. 3).

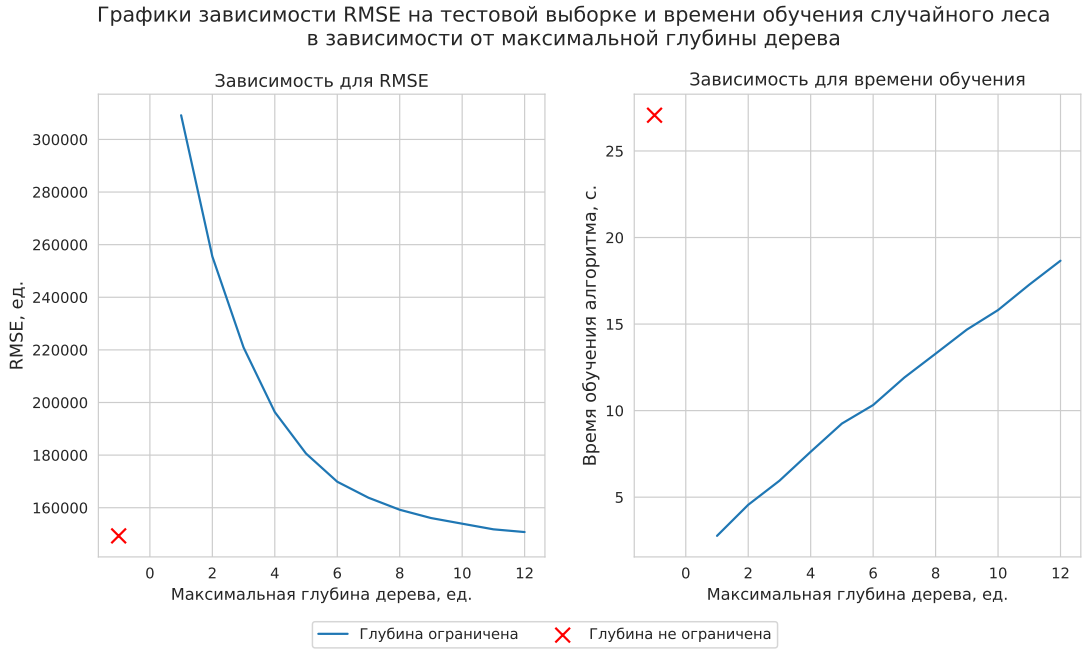

Рис. 3: Зависимость RMSE и времени работы алгоритма случайного леса от максимальной глубины деревьев

Видно, что с увеличением глубины дерева качество на тестовой выборке растет. Это происходит потому, что настройка базового алгоритма становится точнее. Отметим, что случай неограниченной глубины отмечен красным крестиком на графике, по горизонтальной оси это соответствует глубине, равной -1. Видно также, что при обучении неограниченного дерева достигается наименьшая ошибка, но время обучения достаточно велико.

3.2 Алгоритм градиентного бустинга

Исследование зависимости от количества деревьев в ансамбле

Теперь проведем аналогичные эксперименты для алгоритма градиентного бустинга. Как и в предыдущем разделе, при исследовании параметра будем использовать фиксированные значения остальных параметров. Полученные результаты отражены

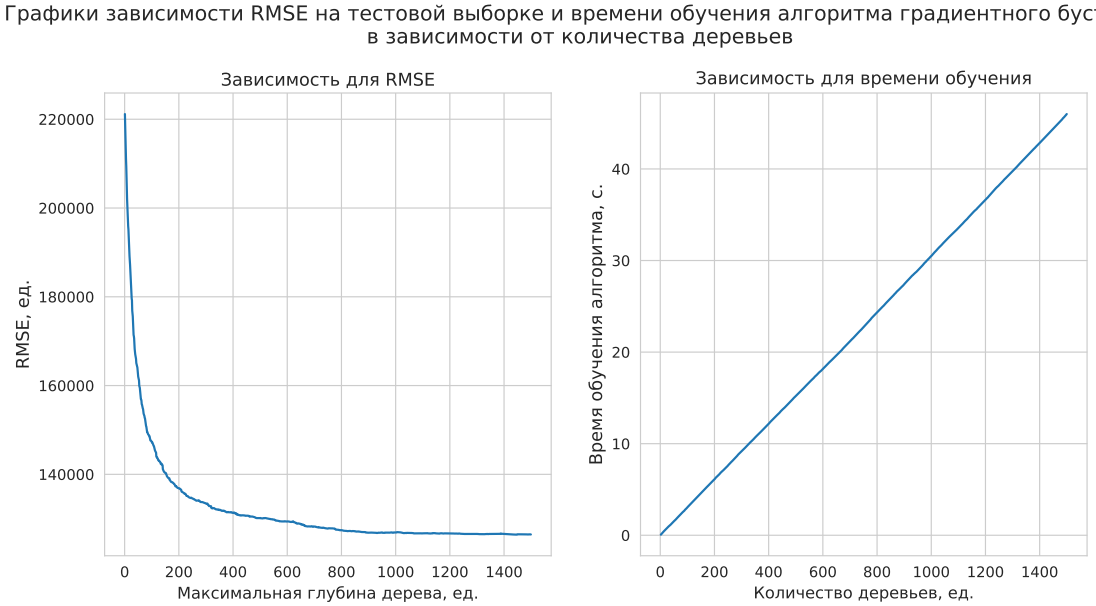

Рис. 4: Зависимость RMSE и времени работы алгоритма бустинга от количества деревьев

на графике выше (рис. 4). Заметим, что для получения графиков переиспользовались результаты, полученные на предыдущем шаге бустинга. Видно, что начиная примерно с 800 деревьев уменьшения ошибки регрессии не происходит, т.е алгоритм выходит на плато. При этом время обучения монотонно растет линейно. При проведении эксперимента использовалась наибольшая глубина каждого дерева, равная трем, а также размер подвыборки признаков 0.6.

Исследование зависимости от размерности пространства признаков

При проведении этого эксперимента аналогично использовалось 800 деревьев глубины не более 3. Результаты эксперимента приведены на графике (рис. 5). Заметим, что аналогично соответствующему эксперименту для случайного леса, наблюдается уменьшение ошибки при возрастании размерности признакового пространства, а также увеличение времени обучения.

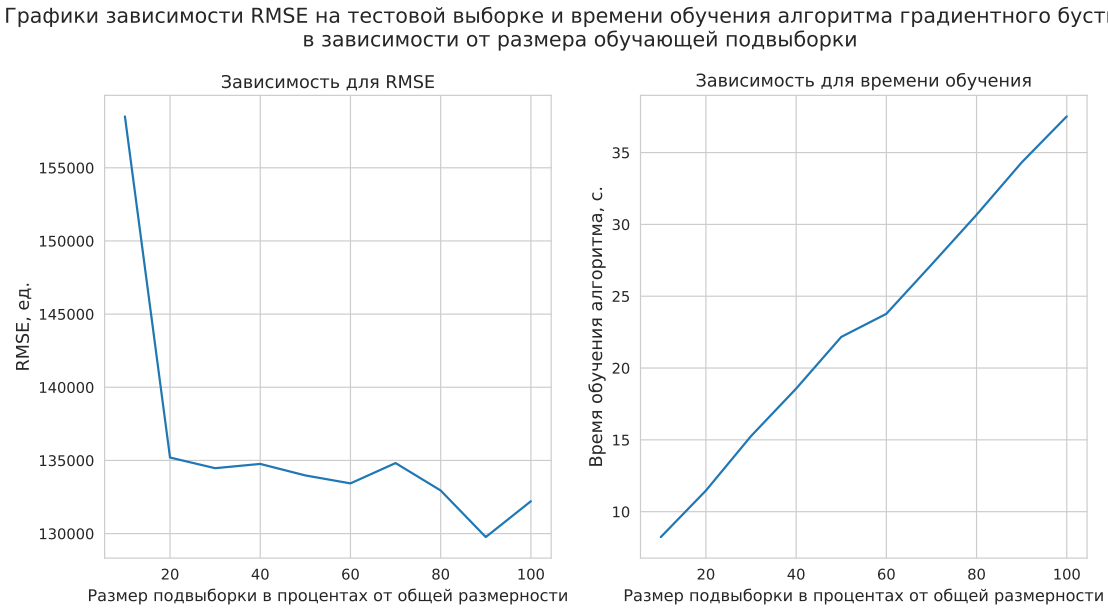

Рис. 5: Зависимость RMSE и времени работы алгоритма бустинга размерности подвыборки признаков (в % от размерности исходной выборки)

Исследование зависимости от максимальной глубины дерева

В данном эксперименте используется 800 деревьев, а также размер подвыборки признаков 0.6. При этом отдельно исследован случай неограниченности высоты дерева (отмечен красным крестиком на графике рис. 6). Из графика видно, что наилучшие результаты достигаются на неглубоких деревьях - глубины не больше 3-4. Это соответствует теоретическим ожиданиям и практическим рекомендациям.

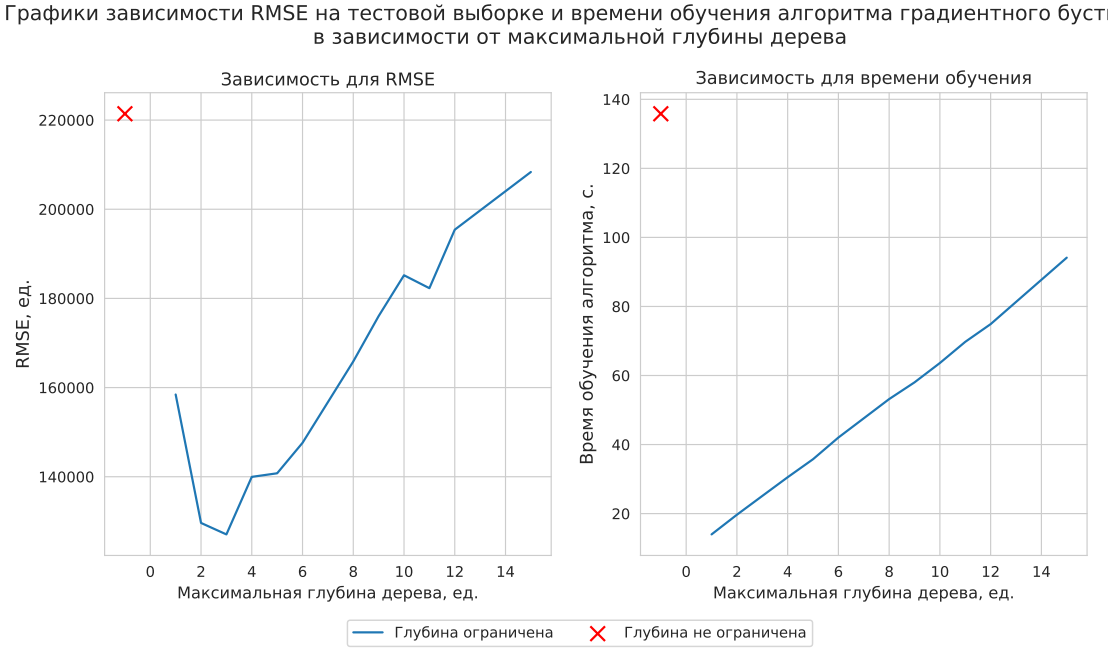

Рис. 6: Зависимость RMSE и времени работы алгоритма бустинга размерности подвыборки признаков (в % от размерности исходной выборки)

Исследование зависимости от темпа обучения

В градиентном бустинге каждая добавляемая модель имитирует антиградиент, внося такой вклад в ансамбль, чтобы суммарная ошибка уменьшилась. Поэтому используется темп обучения для регулирования величины шага в направлении антиградиента. Исследуем влияние темпа обучения на работу алгоритма. Полученные результаты приведены на графике ниже (рис. 7). Видно, что алгоритм постепенно выходит на плато и при дальнейшем увеличении темпа обучения заметной разницы в точности регрессии нет – для слишком больших темпов обучения вновь добавляемые алгоритмы только исправляют ошибки от предыдущего добавленного алгоритма из-за большого темпа обучения.

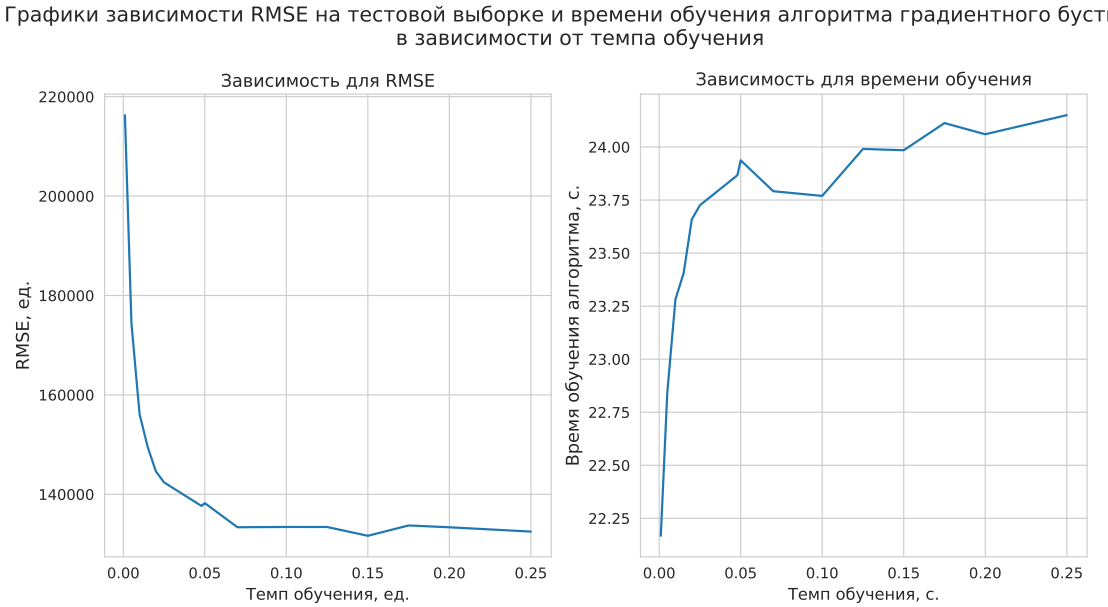

Рис. 7: Зависимость RMSE и времени работы алгоритма бустинга размерности подвыборки признаков (в % от размерности исходной выборки)

4 Вывод

В данной работе было проанализировано два алгоритма, основанных на ансамблировании - случайный лес и градиентный бустинг. Оба алгоритма довольно универсальны и частно используются на практике в различных областях машинного обучения. В отличие от случайных деревьев, в бустинге увеличение числа деревьев не всегда приводит к уменьшению ошибки - как правило, зависимость унимодальная (см. рис. 6). В целом, на всех графиках отражено влияние параметров алгоритмов на работу алгоритмов. Конечно, данная работа является обзорной, но в то же время, довольно наглядной. Заметим, что в веб-приложении реализованы интерактивные графики с помощью plotly и дополнительно сделан автоматическое оповещение в github issue с помощью github actions (см. Actions в репозитории).

5 Список использованных ресурсов

1. Профессиональный информационно-аналитический ресурс, посвященный машинному обучению - http://www.machinelearning.ru.
2. А.Г. Дьяконов «Введение в анализ данных и машинное обучение».
3. Блог А.Г. Дьяконова «анализ малых данных» dyakonov.org.

Содержание