МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В.Ломоносова
Факультет вычислительной математики и кибернетики

Отчет о выполненном задании «Ансамбли алгоритмов. Композиции алгоритмов для решения задачи регрессии»

По курсу «Практикум на ЭВМ»
Кафедры ММП ВМК МГУ

Студента 317 учебной группы
Феоктистова Дмитрия Дмитриевича

Москва, 17 декабря 2022

Содержание

Постановка задачи

В рамках данного задания предлагалось самостоятельно реализовать алгоритмы случайного леса и градиентного бустинга, а также провести ряд экспериментов, направленных на изучение зависимости скорости работы алгоритмов и качества предсказаний от структурных параметров моделей. Все эксперименты проводились на языке программирования Python на данных “House Sales in King County, USA”:

Предобработка данных

Для начала удалим признаки, которые не несут никакой полезной информации для нашей задачи – `id` и `zipcode`. Далее заметим, что в данных присутствуют координаты домов, в теории они несут полезную информацию, но ее надо предварительно извлечь. Для этого посчитаем геодезическое расстояние от домов до центра Сиэтла, как до ближайшего крупного города, в качестве центра возьмем долготу 47.6062 и широту -122.332, после вычисления расстояний удалим информацию о координатах. Затем разобьем информацию о дате продажи на три колонки: число, месяц и год; также посчитаем количество лет между продажей и последним ремонтом. На этом обработка данных окончена, теперь переведем выборку в `np.array` и разобьем ее на обучающую и валидационную выборки в соотношении $7 : 3$.

Случайный лес

Приступим к изучению случайного леса. Для начала посмотрим на влияние количества деревьев на `RMSE` и время обучения. Для этого выберем неограниченную максимальную глубину деревьев, размерность признакового пространства для каждого дерева равной $\frac{1}{3}$, а затем обучим случайный лес с количеством деревьев от 1 до 1000 с шагом один. Результаты эксперимента представлены на рис. 1 и рис. 2.

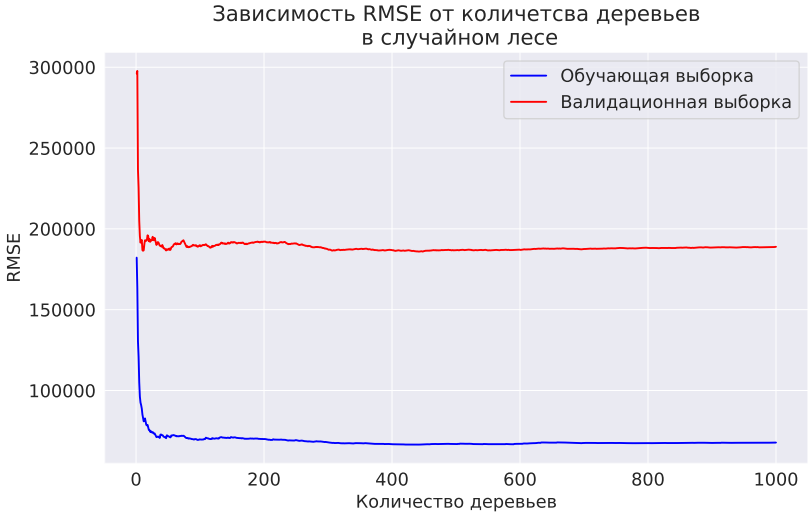

Рис. 1: Зависимость `RMSE` от количества деревьев в случайном лесе.

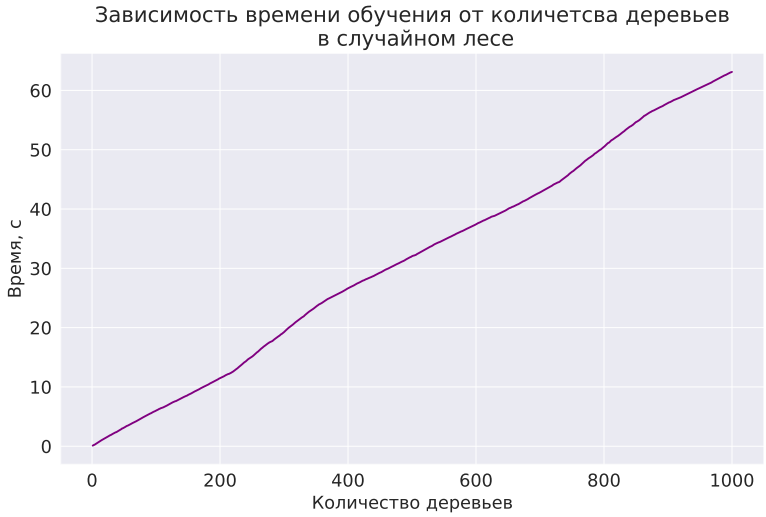

Рис. 2: Зависимость времени обучения от количества деревьев в случайном лесе.

Мы видим, что примерно при 300 деревьев графики обучения и валидации выходят на плато и дальше значения `RMSE` практически не меняются. Также отметим, что случайный лес не переобучается, так как не происходит увеличение `RMSE` на контрольной выборке при уменьшении потерь на обучающей. Все это позволяет в следующих экспериментах использовать леса с 1000 деревьев без опасения переобучения. Из рис. 2 видим, что время обучения случайного леса линейно зависит от количества используемых деревьев.

Теперь изучим влияние максимальной глубины деревьев. Для начала посмотрим на поведение при заданном ограничении. Для этого зафиксируем количество деревьев равное 1000, размерность признакового пространства для

каждого дерева равной $\frac{1}{3}$, а затем переберем ограничения на глубину от 1 до 29 с шагом 1. Результаты эксперимента представлены на рис. 3 и рис. 4.

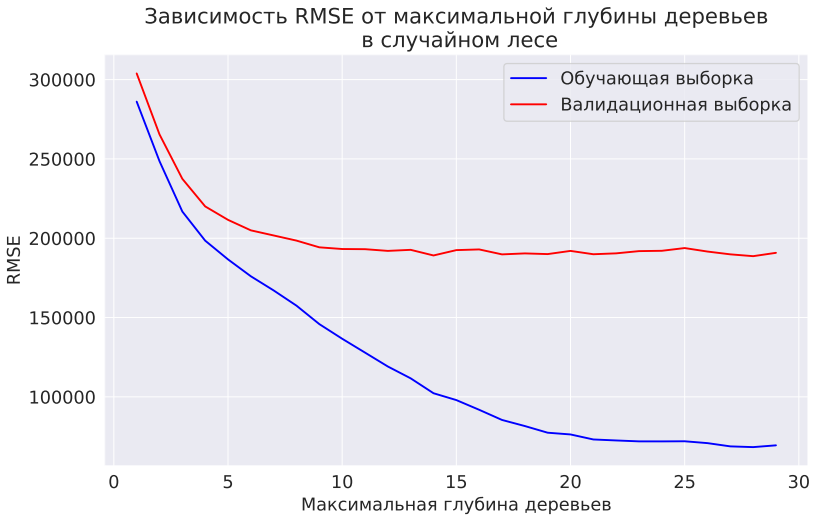

Рис. 3: Зависимость `RMSE` от максимальной глубины деревьев в случайном лесе.

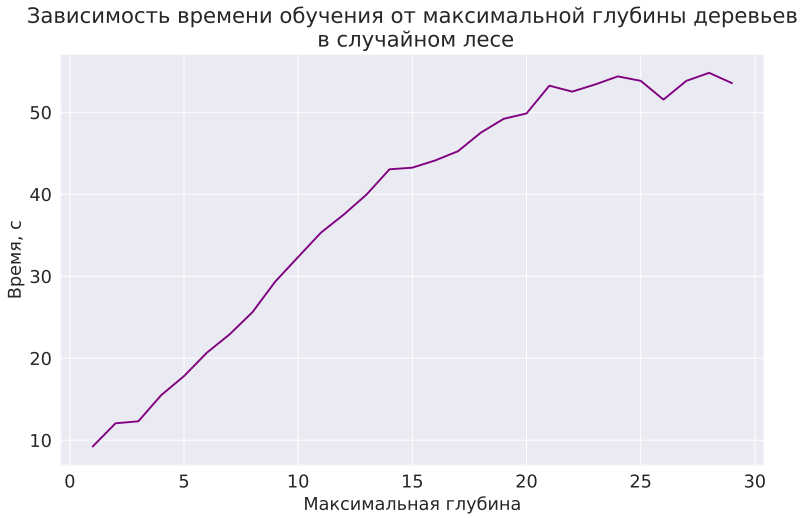

Рис. 4: Зависимость времени обучения от максимальной глубины деревьев в случайном лесе.

Мы видим следующее: при деревьях с глубиной до 5 `RMSE` падает с одинаковой скоростью как на валидационной выборке, так и на обучающей. После чего скорость убывания потерь на контрольной выборке замедляется, и уже на 11 деревьях `RMSE` выходит на плато, около которого и колеблется дальше. На обучающей же выборке такая картина наблюдается, начиная с 22 деревьев.

Отметим также, что несмотря на то, что с ростом максимальной глубины деревьев значение функции потерь на обучающей выборке падает, модель не теряет в обобщающей способности, что выражается в том, что RMSE на контрольной выборке не увеличивается, что еще раз подтверждает то, что случайный лес не переобучается. Если говорить о влиянии максимальной глубины деревьев на время обучения, то видим, что до ограничения в 20 деревьев зависимость похожа на линейную, после 20 уже тяжело назвать какой-то характер зависимости. Этот эффект можно обосновать тем, что появляется много деревьев, которые не используют доступную глубину полностью.

Теперь посмотрим на деревья неограниченной глубины. В этом случае RMSE на обучающей выборке составила 69084, на валидационной 190805, время обучения – 53 с. Эти результаты очень похожи на использование ограничения глубины равного 29, для которого соответствующие замеры равны 69438, 190795 и 53. Из чего можно сделать вывод, что при отсутствии ограничения на глубину по большей части не строятся деревья глубины, большей чем 29.

В качестве оптимального значения далее будем использовать максимальную глубину равную 11, так как на этом значении точность на валидационной выборке вышла на плато.

Осталось изучить последний параметр – размерность признакового пространства для каждого из деревьев. Для этого мы рассмотрим значения этого параметра от 0.1 до 1 с шагом 0.1. Результаты эксперимента приведены на рис. 5 и рис. 6.

Замечание. Количество признаков для каждого дерева вычисляется, как размерность признакового пространства, умноженная на количество признаков во всем датасете.

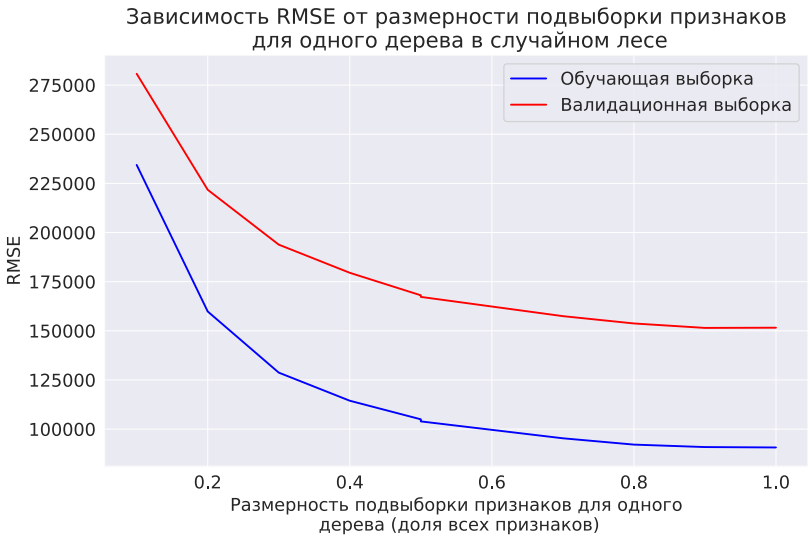

Рис. 5: Зависимость RMSE от размерности признакового пространства для одного дерева в случайном лесе.

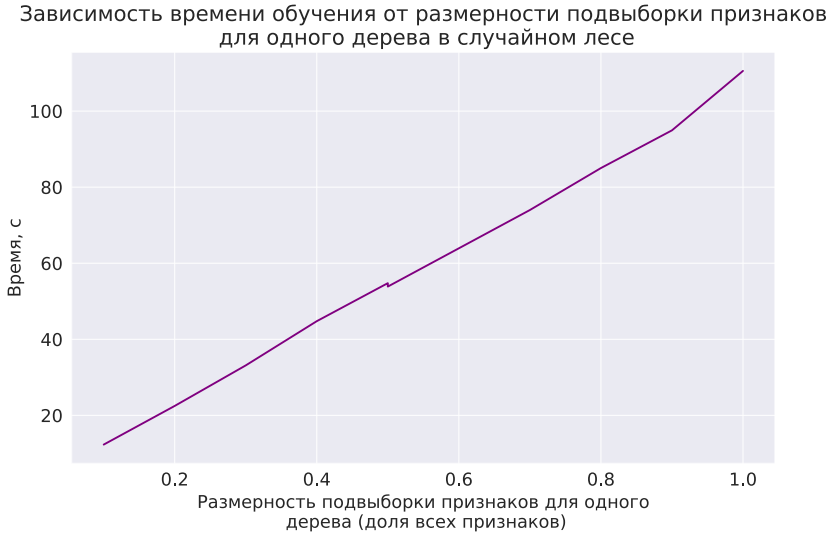

Рис. 6: Зависимость времени обучения от размерности признакового пространства для одного дерева в случайном лесе.

Мы видим, что увеличение данного параметра приводит к монотонному уменьшению потерь как на обучающей выборке, так и на контрольной. Это несколько странно, так как деревья начинают учиться на одних признаках, что приводит к большей корреляции базовых алгоритмов и, соответственно, увеличению составляющей разброса в ошибке ансамбля. Тем не менее, на практике результат оказался другой, что можно обосновать тем, что при увеличении размерности признакового пространства для одного дерева в этой задаче смещение убывает сильнее, чем возрастает разброс. Также мы видим, что зависимость времени обучения от этого параметра линейная.

Подведем итог. Лучшими параметрами для случайного леса оказались следующие – количество деревьев равное 1000, максимальная глубина равная 11 и размерность признакового пространства для одного дерева равная 1. При этих параметрах `RMSE` на обучающей выборке оказалась равной 90658, на контрольной – 151578, время обучения составило 110 секунд.

Градиентный бустинг

Теперь изучим градиентный бустинг. Для начала посмотрим на его поведение в зависимости от количества итераций (количества деревьев в ансамбле) и от темпа обучения. Для этого, как и в случае случайного леса, зафиксируем максимальную глубину деревьев неограниченной, а размерность признакового пространства для одного дерева равной $\frac{1}{3}$. Рассмотрим следующие значения темпа обучения: $10^{-5}, 10^{-4}, 10^{-3}, 0.01, 0.1$ и 1. Количество деревьев будем брать от 1 до 2000 с шагом 1. Изучим результаты экспериментов.

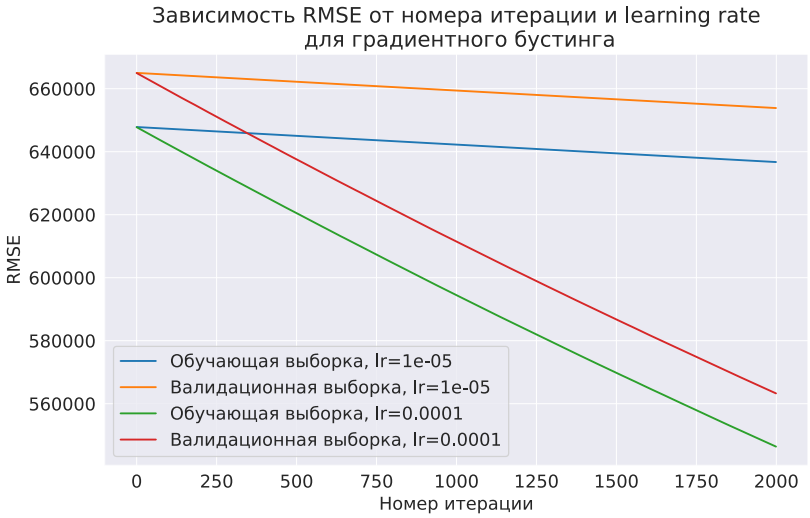

Рис. 7: Зависимость `RMSE` от `learning_rate` и количества деревьев в градиентном бустинге, часть 1.

На рис. 7 мы видим, что использование маленького темпа обучения приводит к тому, что алгоритм обучается сильно хуже, чем случайный лес, даже не смотря на большее количество базовых алгоритмов. Попробуем увеличить `learning_rate` (рис. 8).

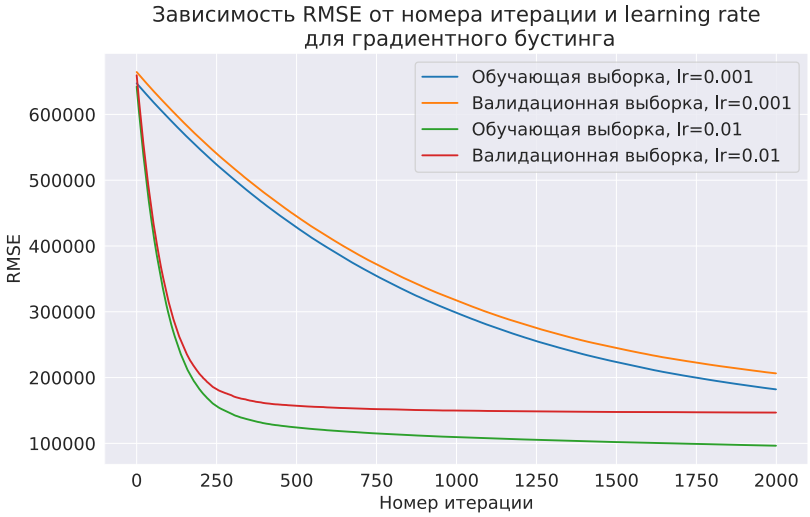

Рис. 8: Зависимость `RMSE` от `learning_rate` и количества деревьев в градиентном бустинге, часть 2.

При использовании темпа обучения, равного 0.001 мы видим, что `RMSE` на обучающей выборке падает с той же скоростью, что и на контрольной. При этом видно, что при заданных параметрах можно увеличить количество итераций.

При `learning_rate` равном 0.01 уже наступает легкое переобучение после 750 итераций. Давайте попробуем еще увеличить изучаемый параметр (рис. 9).

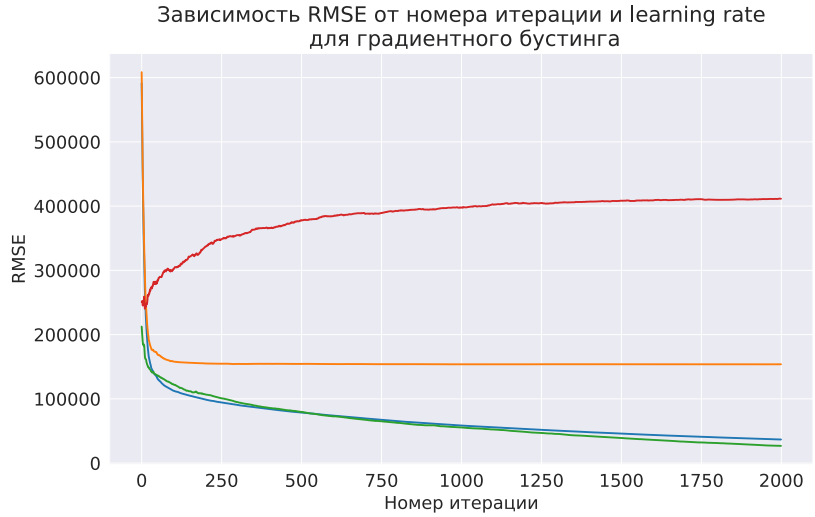

Рис. 9: Зависимость `RMSE` от `learning_rate` и количества деревьев в градиентном бустинге, часть 3.

Здесь мы видим, что при темпе обучения равном 0.1 ситуация похожа на то, что возникало при использовании данного параметра равного 0.01. А вот при `learning_rate` равном 1 уже наступает полноценное переобучение. Далее мы будем использовать модели с темпом обучения равным 0.001, так как его график выглядит наиболее безопасным с точки зрения переобучения для дальнейшей работы, количество деревьев возьмем равное 2000.

Также посмотрим на график времени работы в зависимости от исследуемых параметров.

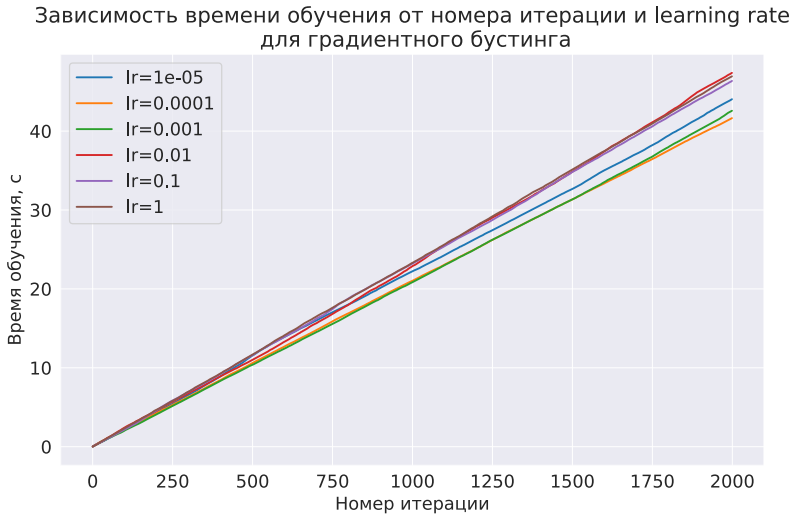

Рис. 10: Зависимость времени обучения от `learning_rate` и количества деревьев в градиентном бустинге.

На рис. 10 мы видим, что время обучения линейно зависит от количества деревьев. Темп обучения влияет на уровень наклона прямой, но явной зависимости времени обучения от этого параметра не наблюдается.

Перейдем к рассмотрению влияния максимальной глубины деревьев на `RMSE` и время обучения. Результаты эксперимента представлены на рис. 11 и рис. 12.

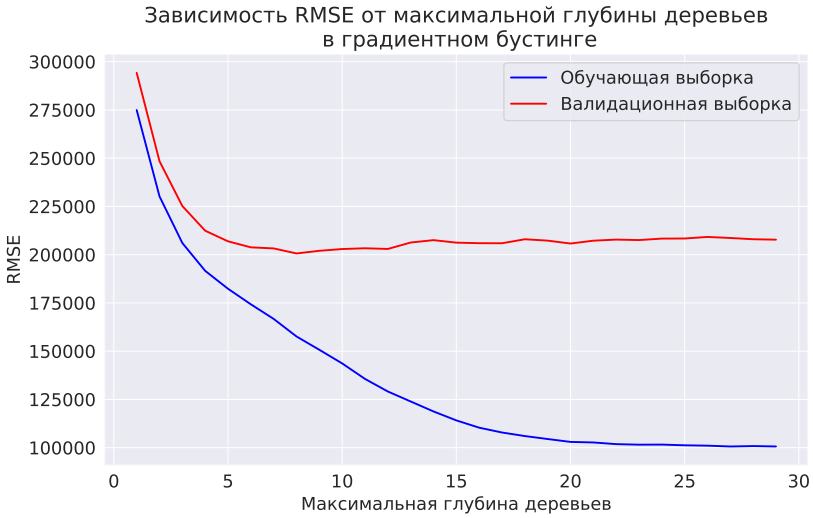

Рис. 11: Зависимость `RMSE` от максимальной глубины деревьев в градиентном бустинге.

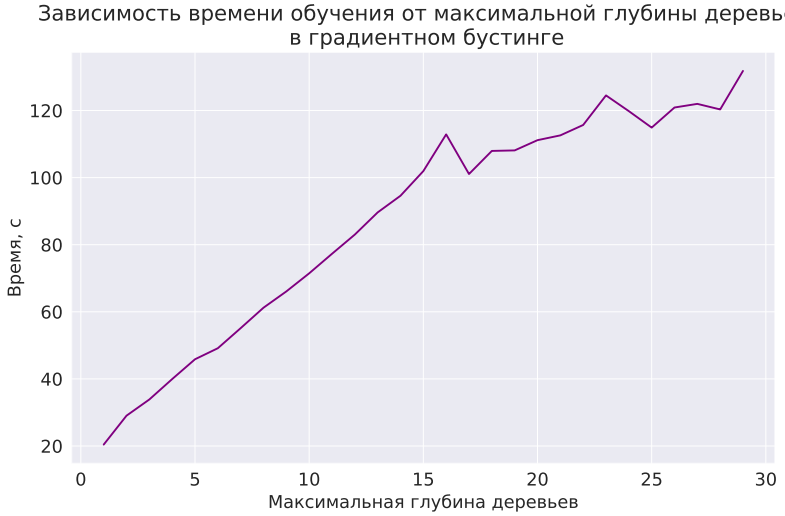

Рис. 12: Зависимость времени обучения от максимальной глубины деревьев в градиентном бустинге.

Мы видим картину аналогичную случайному лесу с той разницей, что вместо выхода на плато мы наблюдаем признаки переобучения, начиная с глубины равной 8. В качестве оптимального значения будем далее использовать максимальную глубину равную 7. Этот результат подтверждает то, что для градиентного бустинга следует использовать базовые алгоритмы с меньшей сложностью, чем для случайного леса. В зависимости времени обучения от данного параметра мы видим опять похожую картину: зависимость линейная, но после глубины 16 она начинает нарушаться. Обоснование этого явления аналогично тому, что было сказано для случайного леса.

Опять рассмотрим деревья с неограниченной максимальной глубиной. При этом значении параметра `RMSE` на обучающей выборке составила 100625, на контрольной 207515, а время обучения оказалось равным 156 секундам. Это поведение снова напоминает картину, наблюдаемую при максимальной глубине деревьев равной 29, при которой соответствующие замеры равны – 100612, 207772 и 131. Обоснование остается тем же.

Теперь перейдем к изучению влияния размерности признакового пространства одного дерева для градиентного бустинга. Рассмотрим те же значения, что и для случайного леса. Результаты представлены на рис. 13 и рис. 14.

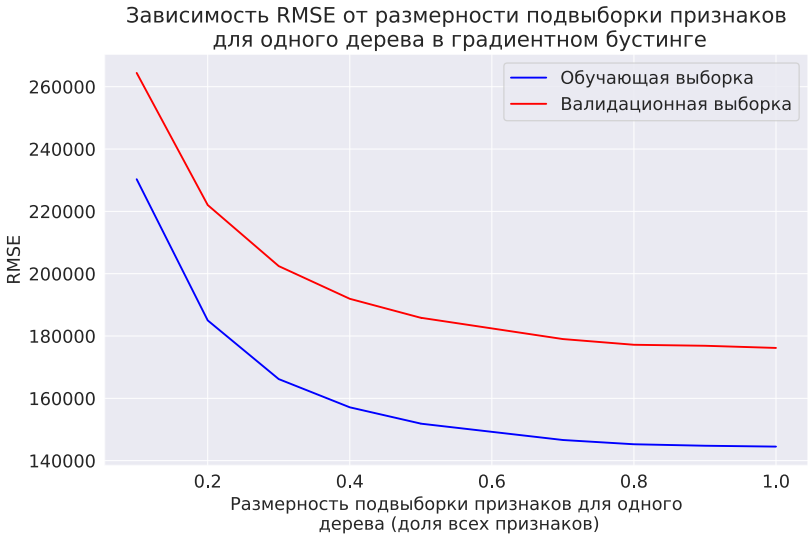

Рис. 13: Зависимость `RMSE` от размерности признакового пространства одного дерева в градиентном бустинге.

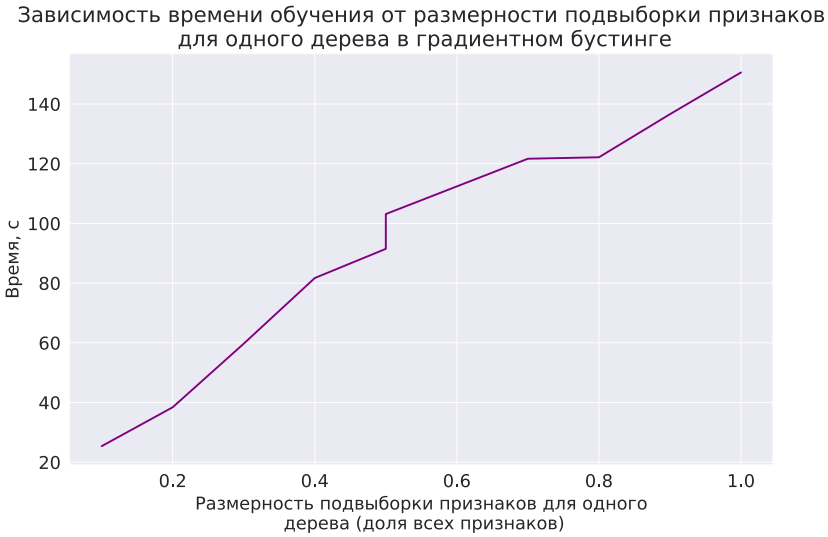

Рис. 14: Зависимость времени обучения от размерности признакового пространства одного дерева в градиентном бустинге.

Наблюдаемая картина аналогична тому, что мы видели в случайном лесе: при увеличении данного параметра `RMSE` на обоих выборках падает с примерно одинаковой скоростью. Время обучения также зависит линейно.

Подведем итог. Во-первых, обе модели ведут себя похожим образом при изменении структурных параметров, но градиентный бустинг имеет склонность к переобучению, поэтому для него надо аккуратнее подбирать параметры. Во-вторых, несмотря на то, что градиентный бустинг считается более сильной моделью, чем случайный лес, на рассматриваемых данных это казалось не так,

так как при оптимальных параметрах (количество деревьев – 2000, темп обучения – 0.001, максимальная глубина – 7, размерность признакового пространства для одного дерева – 1.0) `RMSE` для обучающей выборки составила 151872, для валидационной – 185854, время обучения оказалось равным 103. В-третьих, реализация случайного леса без параллельного обучения различных деревьев обучается дольше, чем градиентный бустинг, даже несмотря на то, что для него мы обучаем вдвое больше базовых моделей. Это явления частично обосновывается использованием более глубоких деревьев в случайном лесе.

Итог

В ходе выполнения задания были изучены и реализованы алгоритмы случайного леса и градиентного бустинга, а также было проведено их комплексное исследование в рамках работы над предсказанием стоимости домов.