Отчет о практическом задании «Ансамбли алгоритмов. Веб-сервер. Композиции алгоритмов для решения задачи регрессии».

Практикум 317 группы

Лапиков Владислав

Декабрь 2023

Содержание

1 Введение

Данное практическое задание посвящено исследованию ансамблей алгоритмов для решения задачи регрессии. В качестве моделей будет использоваться случайный лес и градиентный бустинг. Целью исследования является сравнение точности и времени работы модели в зависимости от параметров.

2 Пояснение к задаче

Дан датасет (House Sales in King County, USA), содержащий признаки различных домов «King Country» и цену на них. Необходимо реализовать алгоритм, позволяющий предсказывать цены на квартиры, опираясь на те же признаки. Необходимо проанализировать качество и время обучения в зависимости от различных параметров моделей.

3 Эксперименты

В данном разделе будут представлены результаты экспериментов, проведенных в рамках исследования.

3.1 Данные

Данные[1] представлены 21613 объектами с 21 признаком. Один из признаков, который находится в столбце с названием `price` является нашим таргетом.

3.2 Воспроизводимость результатов и вычислительные ресурсы

Для проведения экспериментов, мы будем использовать Python библиотеку `numpy` с зафиксированным `numpy.random.seed(42)`. Также для реализации ансамблевых алгоритмов использовался класс `DecisionTreeRegressor`.

Все вычисления производились на 2,6 GHz 6-ядерном процессоре Intel Core i7 и 32 ГБ 2667 MHz DDR4 оперативной памяти.

3.3 Предобработка данных

Из-за отсутствия информативности удалим из данных столбец `id`, а также для простоты нашей модели уберем столбцы `zipcode` (код зоны, где расположен дом), `lat` (широта), `long` (долгота). Так как мы не сможем адекватно оценивать эти параметры без дополнительной обработки.

Теперь разберемся с датами. В датасете присутствуют 3 колонки с датами - `date` (дата объявления), `yr_built` (год постройки дома), `yr_renovated` (год последнего ремонта дома, если ремонт не производился, то 0). Такие признаки будут не очень информативными для нашей модели, но мы можем преобразовать их в «возраст» дома и время до последнего ремонта (или, если он не производился, то возраст дома). Таким образом, запишем данные изменения в те же столбцы `yr_built` и `yr_renovated`, а столбец `date` удалим.

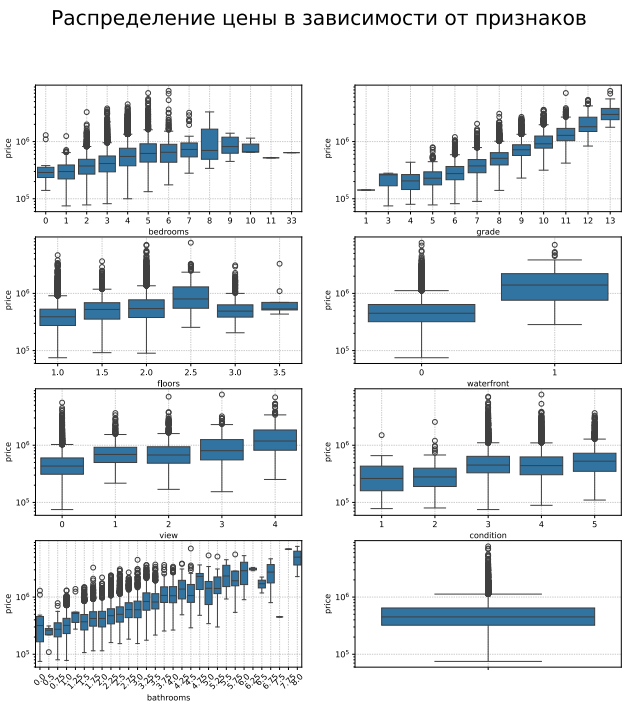

Рис. 1: Распределение цены в зависимости от значений некоторых столбцов

[1] https://www.dropbox.com/scl/fi/39eiebd61yc0cy5ckc7y5/kc_house_data.csv?rlkey=1419a4kcxp081ftjbkh7vtr51&dl=0

Теперь проведем анализ на выбросы. Как мы можем заметить на Рис. 1, у признака `bedrooms` присутствует предложение с 33 и 11 спальнями за очень даже приятные деньги. Скорее всего это были опечатки или ошибки системы. Две такие записи мы уберем из датасета.

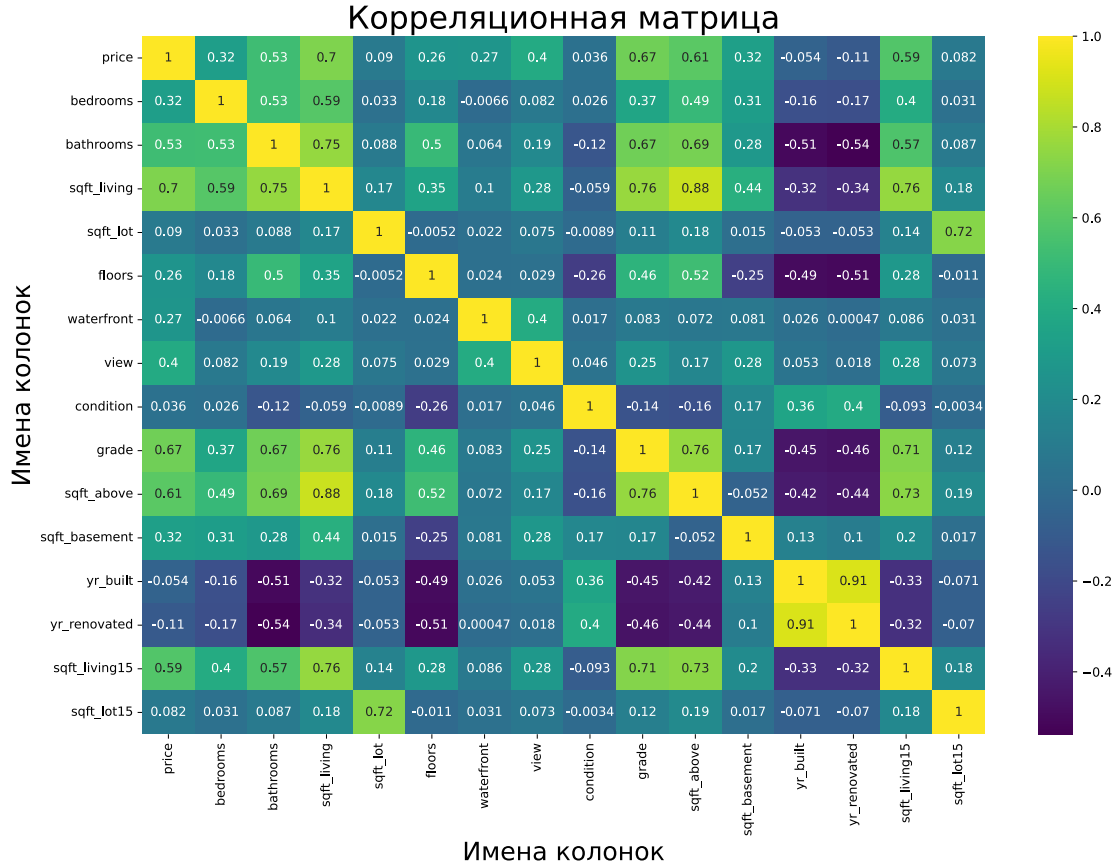

Рис. 2: Матрица корреляций для оставшихся данных

На Рис. 2 изображена корреляционная матрица для получившихся данных. Можем сразу заметить, что в данных присутствует много признаков, которые коррелируют между собой. Оставим из них только те, которые лучше коррелируют с нашим таргетом (колонка `price`). Данным действием, мы уберем колонки `yr_built`, `sqft_lot15`, `sqft_above`.

Оставшиеся данные мы разделим на тренировочную и тестовую выборки в соотношении 8/2.

Стоит упомянуть, почему при анализе данных на выбросы практически никакие данные не были удалены (за исключением двух объявлений). Это связано с тем, что на инференсе к нам будут приходить данные той же природы, что и имеется сейчас. А так как мы работаем сразу и с тестовой и с тренировочной выборками, то результат будет не объективен. Если же мы захотим провести полноценный анализ на выбросы, то лучше это делать после разбиения на тренировочную и тестовые выборки.

3.4 Анализ случайного леса

В данном разделе приведены выводы из исследования алгоритма случайный лес. А именно: будет показана зависимость метрики RMSE на отложенной выборке и время работы алгоритма

в зависимости от следующих факторов:

- количество деревьев в ансамбле
- размерность подвыборки признаков для одного дерева
- максимальная глубина дерева (а также случай, когда глубина неограничена)

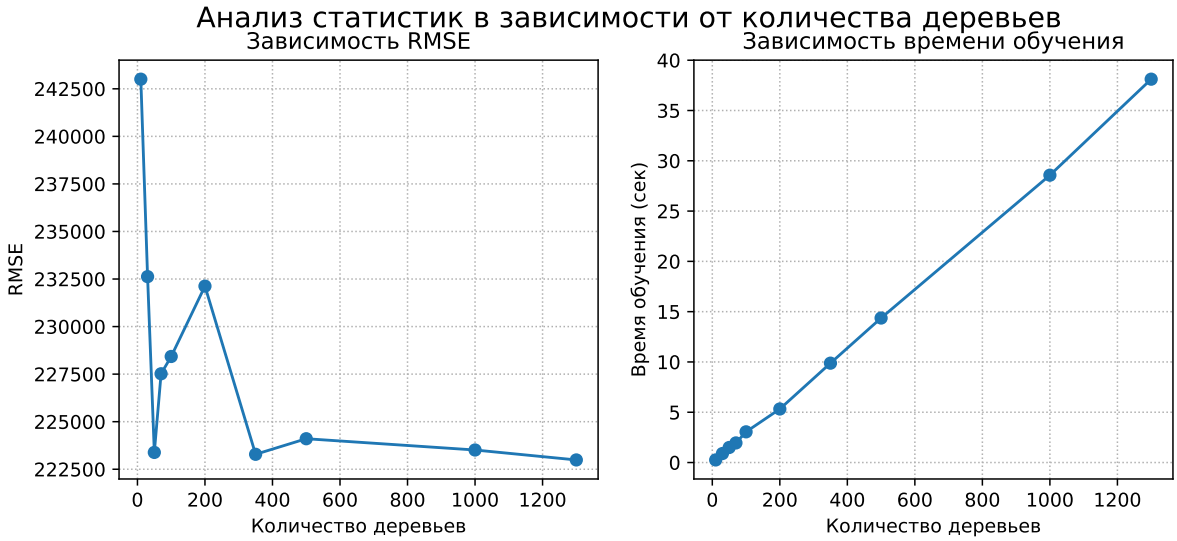

Рис. 3: Анализ зависимости статистик случайного леса от количества деревьев (максимальная глубина равна 20)

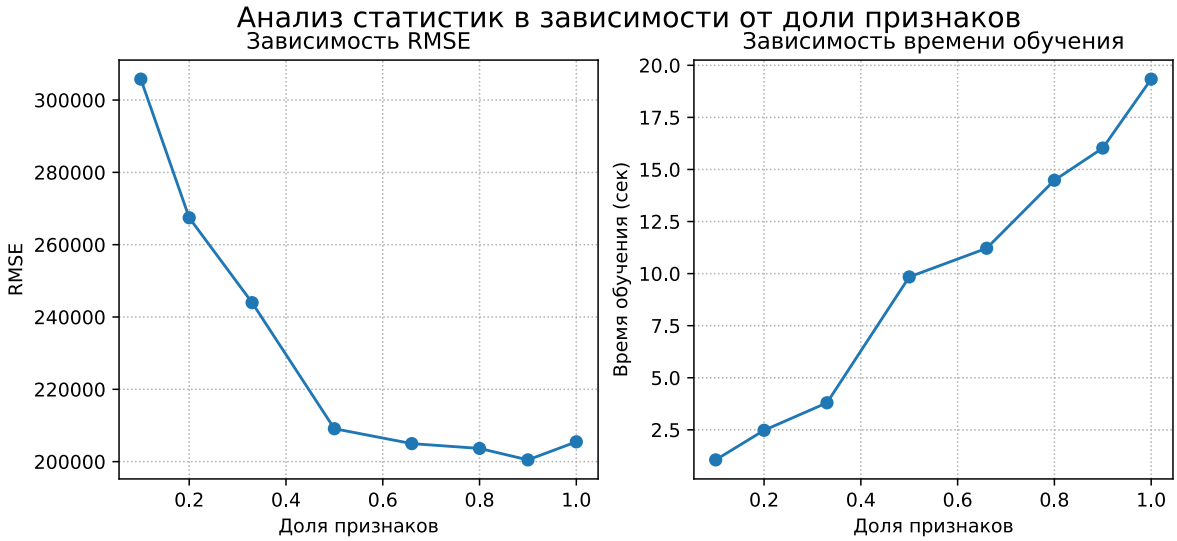

Рис. 4: Анализ зависимости статистик случайного леса от доли признаков для одного дерева (количество деревьев равно 200, а максимальная глубина равна 20)

На Рис. 3 изображена зависимость RMSE и времени обучения в зависимости от количества деревьев в ансамбле. Можем заметить интересное поведение: сначала лосс быстро падает, потом немного возрастает, а потом снова падает (ака. двойной спуск). При увеличении числа деревьев

лосс уходит в экстремум. Так как случайный лес - это алгоритм голосования, то модель неплохо обобщается на тестовые данные с увеличением числа деревьев. На графике времени обучения нет ничего необычного – время линейно зависит от количества деревьев.

На Рис. 4 изображена зависимость RMSE и времени обучения в зависимости от доли признаков для обучения относительно одного дерева ансамбля. Интересно, что при увеличении доли признаков качество только растет. Скорее всего в наших данных есть так называемые «золотые признаки», без которых или без совокопности которых алгоритм плохо обучается. На графике времени снова ничего особенного - время обучения растет, так как растет количество признаков.

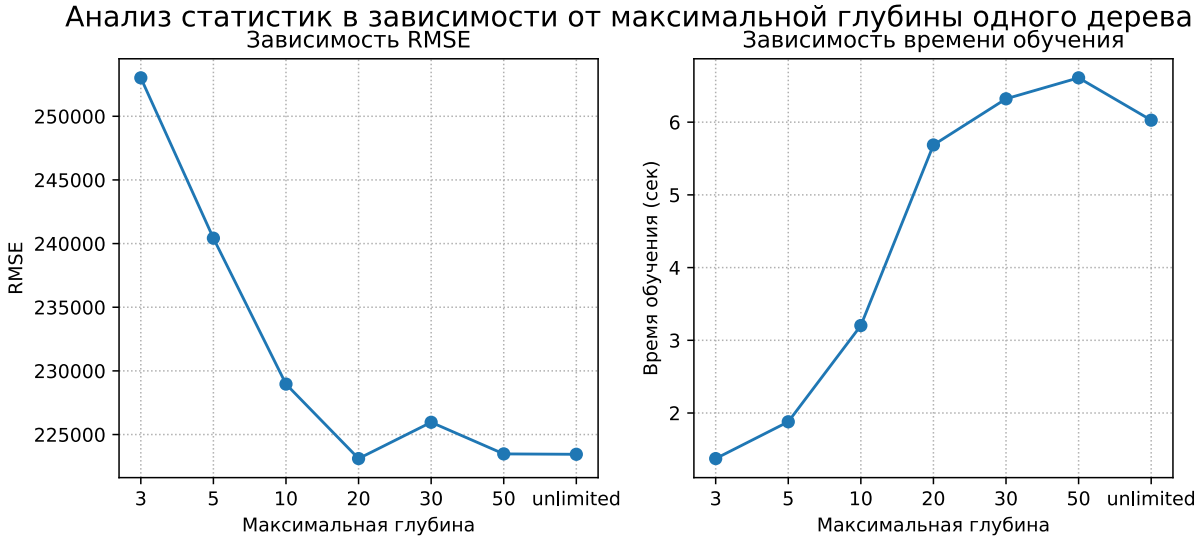

Рис. 5: Анализ зависимости статистик случайного леса от максимальной глубины для одного дерева (количество деревьев равно 200)

На Рис. 5 изображена зависимость RMSE и времени обучения в зависимости от максимальной глубины одного дерева ансамбля. Как это ни парадоксально, при увеличении глубины растет и качество. Каждое отдельное дерево будет иметь маленькое смещение и большую дисперсию, что в совокупности поможет сделать отличный алгоритм. На графике времени обучения снова ничего особенного.

3.5 Анализ градиентного бустинга

В данном разделе приведены выводы из исследования алгоритма градиентный бустинг. А именно: будет показана зависимость метрики RMSE на отложенной выборке и время работы алгоритма в зависимости от следующих факторов:

- количество деревьев в ансамбле
- размерность подвыборки признаков для одного дерева
- максимальная глубина дерева (а также случай, когда глубина неограничена)
- выбранный `learning_rate`

На Рис. 6 изображена зависимость RMSE и времени обучения в зависимости от количества деревьев в ансамбле. На графике можно увидеть переобучения - лосс на малых значениях

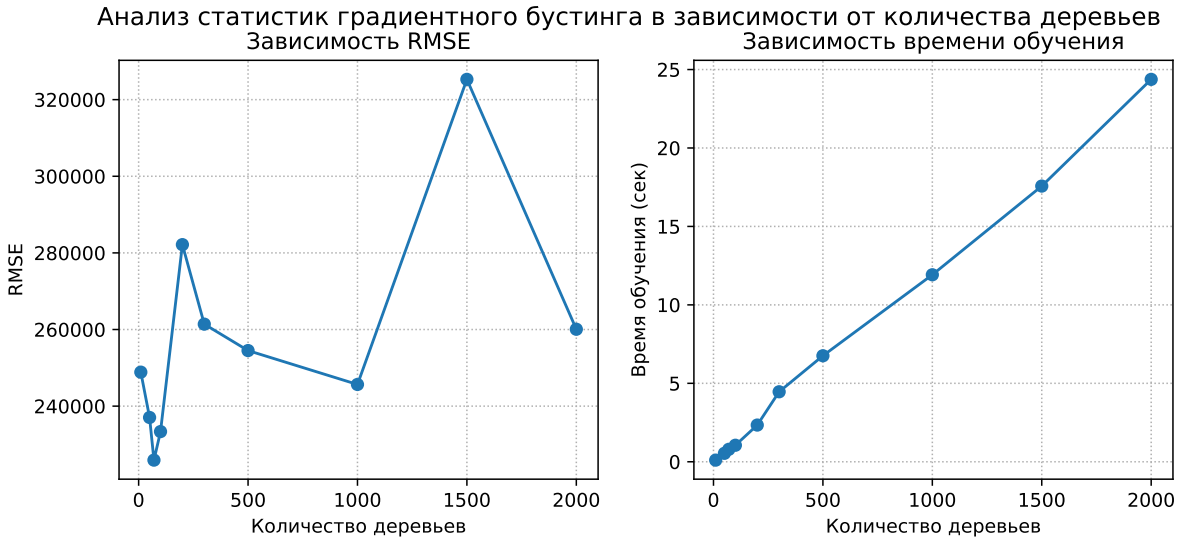

Рис. 6: Анализ зависимости статистик градиентного бустинга от количества деревьев (максимальная глубина равна 4)

достигает минимума, а потом возрастает. На больших значениях количества деревьев алгоритм слишком хорошо выучивает обучающую выборку и по сути подстраивается под нее. Однако, переобучение намного меньше заметно, чем в обычных линейных моделях. Также, как и в случайном лесе, на графике времени обучения нет ничего необычного – время линейно зависит от количества деревьев.

На Рис. 7 изображена зависимость RMSE и времени обучения в зависимости от доли признаков для обучения относительно одного дерева ансамбля. Судя по графику, качество градиентного бустинга не сильно зависит от количества признаков, но мы также, как и в случае случайного леса, можем заметить уменьшение ошибки при большом числе признаков. Объяснение этому явлению все те же «золотые признаки». На графике времени снова ничего особенного - время обучения растет, так как растет количество признаков.

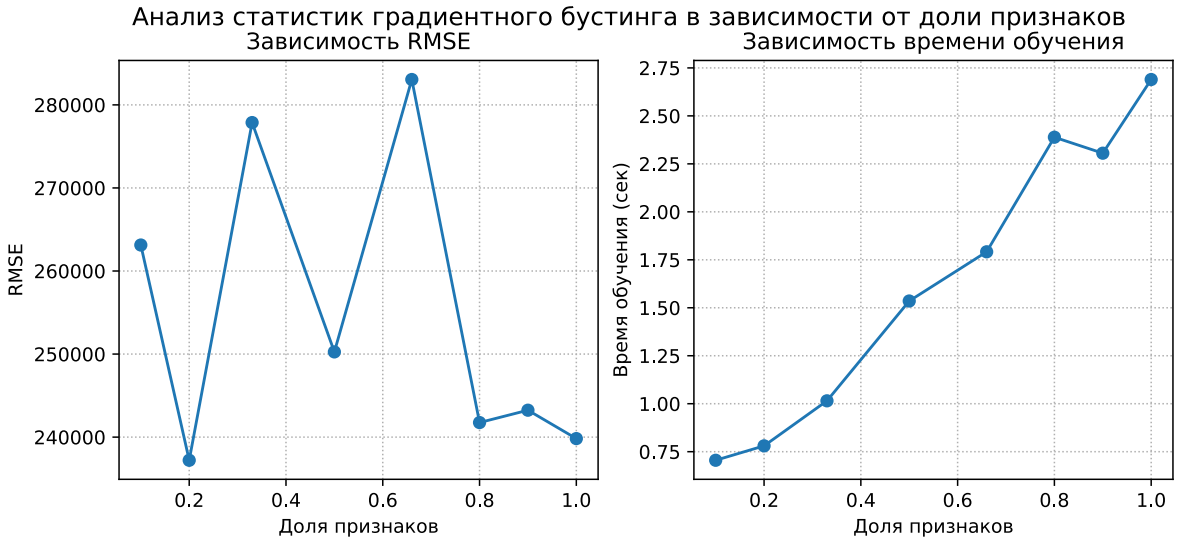

Рис. 7: Анализ зависимости статистик случайного леса от доли признаков для одного дерева (количество деревьев равно 100, а максимальная глубина равна 4)

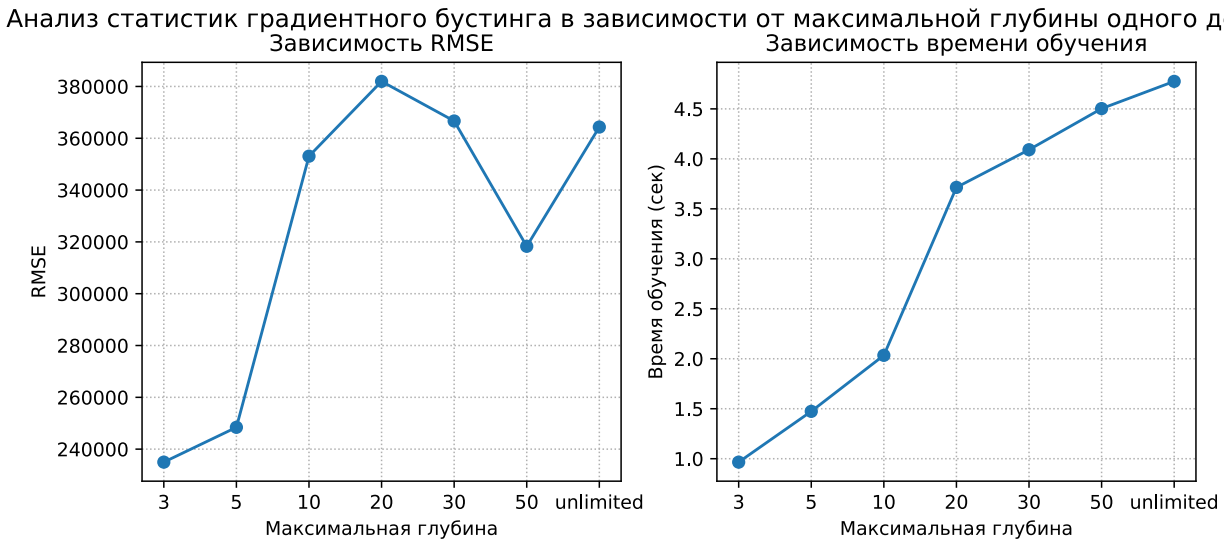

Рис. 8: Анализ зависимости статистик градиентного бустинга от максимальной глубины для одного дерева (количество деревьев равно 100)

На Рис. 8 изображена зависимость RMSE и времени обучения в зависимости от максимальной глубины одного дерева ансамбля. Здесь уже ситуация отличается от той, которая была в случайном лесе (Рис. 5). Бустинг на каждом шаге улучшает ответы предыдущего алгоритма, поэтому, если сразу брать сложные деревья, то алгоритм слишком сильно переобучится на тренировочные данные. На графике времени обучения снова ничего особенного.

На Рис. 9 изображена зависимость RMSE и времени обучения в зависимости от темпа обучения (`learning_rate`). Значения перебирались по логарифмической шкале от 10^{-5} до 1. Какой-то определенной закономерности обнаружено не было, но можно точно сказать, что темп обучения стоит подбирать для вашей конкретной задачи. Время обучения не сильно зависит от `learning_rate` .

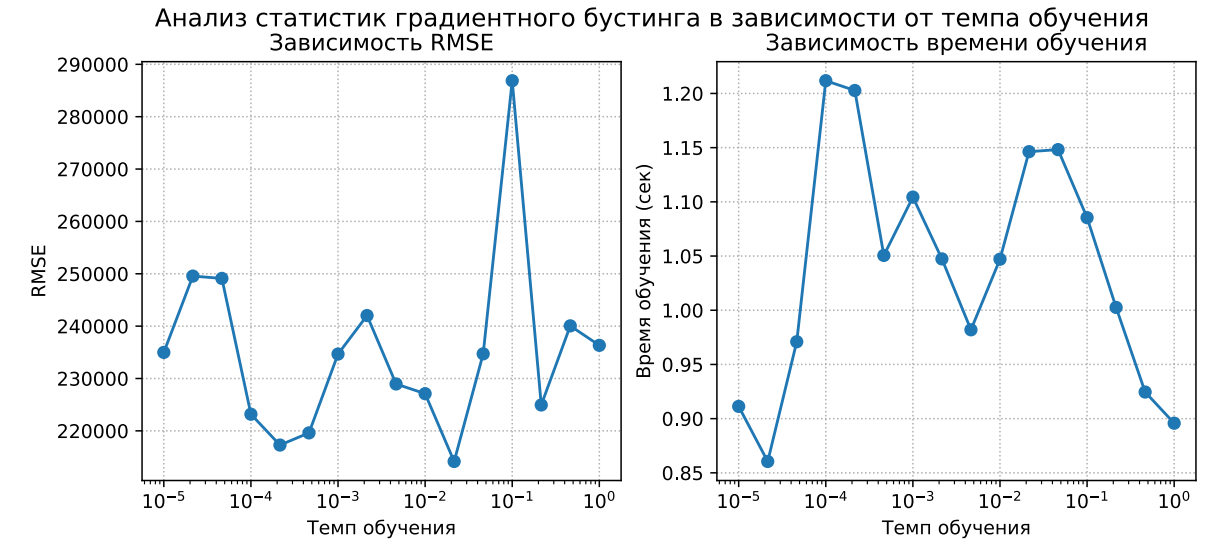

Рис. 9: Анализ зависимости статистик градиентного бустинга от темпа обучения (количество деревьев равно 100, максимальная глубина равна 3)

4 Выводы

В данной практической работе был проанализирован датасет «House Sales in King County, USA», реализованы и проанализированы ансамблевые алгоритмы на задачу регрессии - случайный лес и градиентный бустинг.