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

Отчёт по заданию №3 практикума “Ансамбли алгоритмов. Веб-сервер. Композиции алгоритмов для решения задачи регрессии.”

Зыков Валерий
317 группа

Москва
2023

Содержание

Введение

В данном задании были рассмотрены такие алгоритмы композиции как “Случайный лес” и “Градиентный бустинг”.

Были проведены следующие этапы работы:

1. Реализация на языке Python алгоритмов “Случайный лес” и “Градиентный бустинг”.

2. Проведение экспериментов с датасетом, содержащем цены на продажу жилья в округе Кинг (США).

3. Реализация веб-сервера, который позволяет работать с реализованными моделями человеку, не знающему Python.

1 Эксперименты

1.1 Предобработка данных

Данные были загружены тут. Была удалена колонка “id”. Колонка “date” была преобразована к типу datetime, из нее были получены еще три колонки: “year”, “month” и “day”, сама же она была удалена. Итого получили 21613 объектов с 21 признаком. Данные были разбиты на обучающую и тестовую выборку в соотношении $7 : 3$.

1.2 Исследование поведения алгоритма “Случайный лес”

Изучим зависимость RMSE на отложенной (тестовой) выборке и время работы алгоритма “Случайный лес” в зависимости от следующих гиперпараметров:

- количество деревьев в ансамбле (параметр n_estimators);
- размерность подвыборки признаков для одного дерева (параметр feature_subsample_size);
- максимальная глубина дерева (может быть неограничена, параметр max_depth).

1.2.1 Количество деревьев в ансамбле

Фиксируем параметр feature_subsample_size равным трети признаков. Рассмотрим n_estimators $\in [1, 300]$ при неограниченной и ограниченной максимальной глубине max_depth и постоим график зависимости RMSE на тестовой выборке. График приведен на рис. 1.

Из него можно сделать следующие выводы:

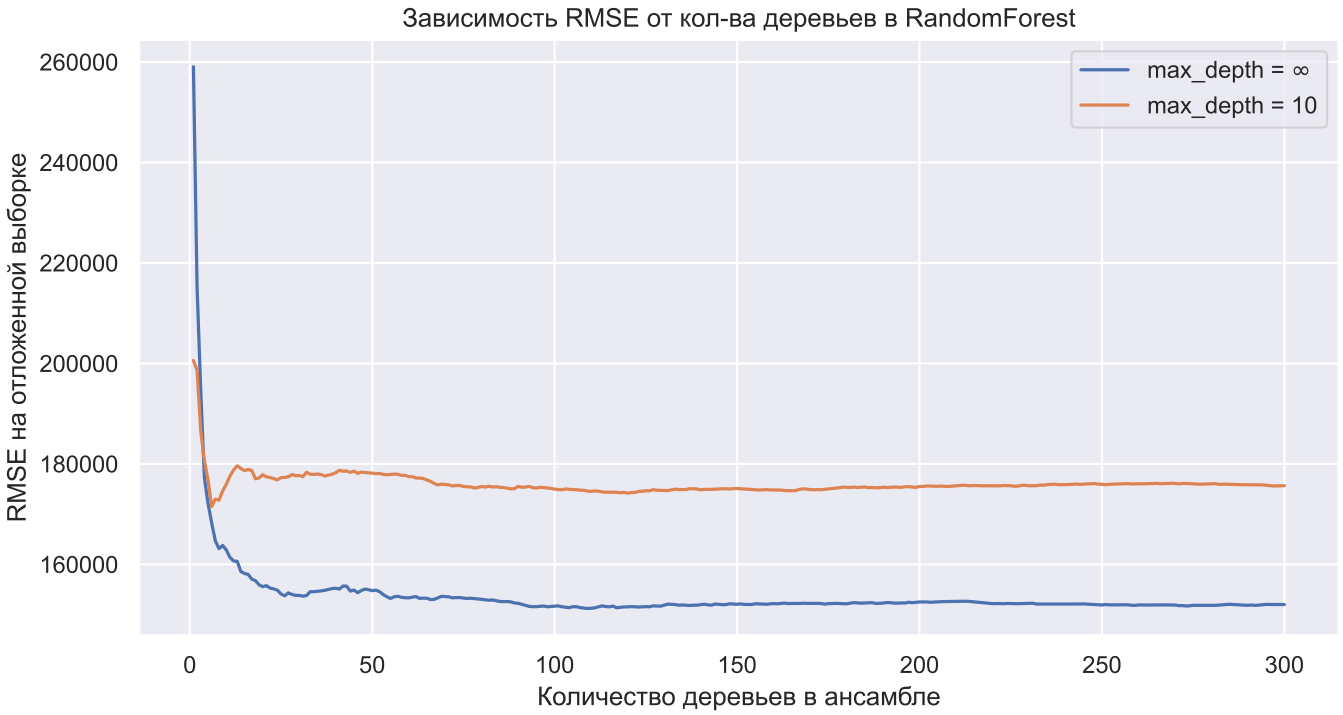

Рис. 1.

1. Случайный лес не переобучается с увеличением числа деревьев.

2. Ошибка выходит на асимптоту с ростом количества деревьев.

3. При n_estimators $\geqslant 100$ ошибка уже сильно не изменяется.

Фиксируем n_estimators $= 100$.

Время обучения и время предсказания, очевидно, линейно зависят от числа деревьев (следует из реализации алгоритма).

1.2.2 Размерность подвыборки признаков для одного дерева

В каждой вершине каждого дерева будем выбирать признак для разбиения из случайного подмножества признаков заданной мощности. Причем это подмножество будем делать новым для каждой вершины дерева. Это поможет повысить разнообразие базовых алгоритмов. В целом дерево будет обучено на всех признаках. Обучать базовый алгоритм на каком-то фиксированном для всех его вершин подмножестве признаков может оказаться плохой идеей, т.к. если у нас есть один или несколько признаков, значительно более хороших чем остальные, то эти признаки должны быть использованы при обучении каждого решающего дерева, иначе некоторые базовые алгоритмы могут оказаться слишком плохими по качеству.

Результаты перебора мощности случайных подмножеств отображены на рис. 2 и рис. 3.

Из этих графиков можно сделать следующие выводы:

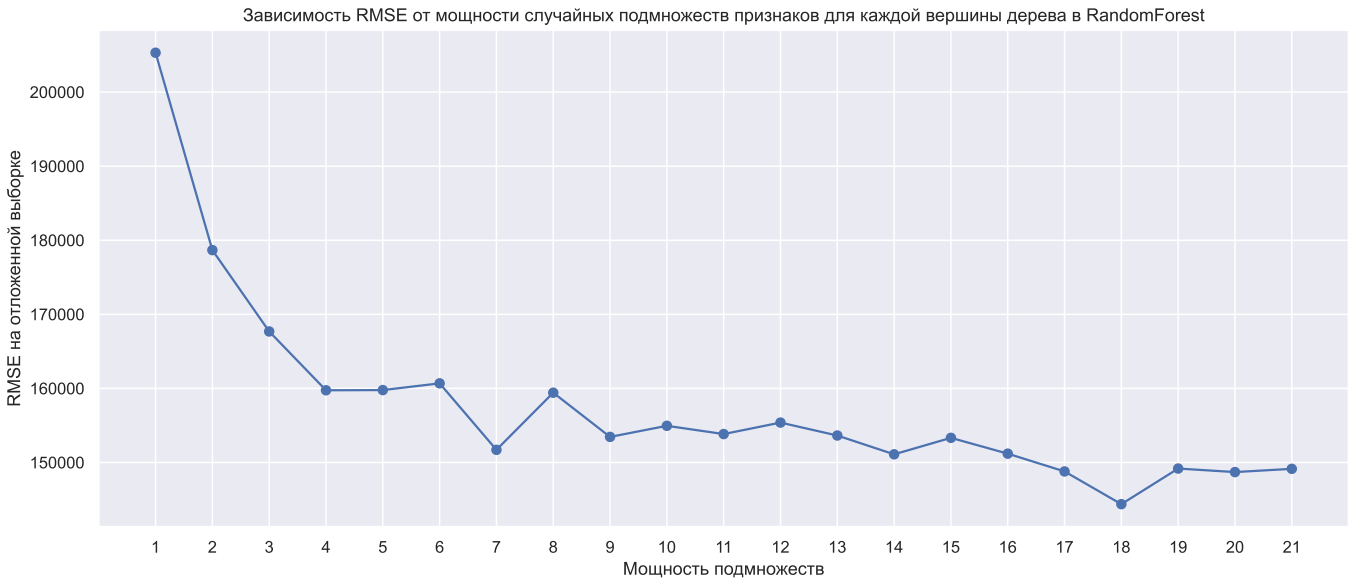

Рис. 2.

Рис. 3.

1. На графике RMSE виден тренд уменьшения ошибки с увеличением мощности (но минимум достигается в точке 18).
2. Чем больше мощность, тем больше время обучения (линейная зависимость).
3. Время предсказания не зависит от мощности.

Фиксируем feature_subsample_size = 18.

1.2.3 Максимальная глубина деревьев

Результаты перебора параметра max_depth отображены на рис. 4 и рис. 5. Из них следует, что:

1. С увеличением максимальной глубины деревьев RMSE уменьшается, время работы увеличивается.

2. При max_depth $\geqslant 50$ разница в RMSE и времени работы не заметна.

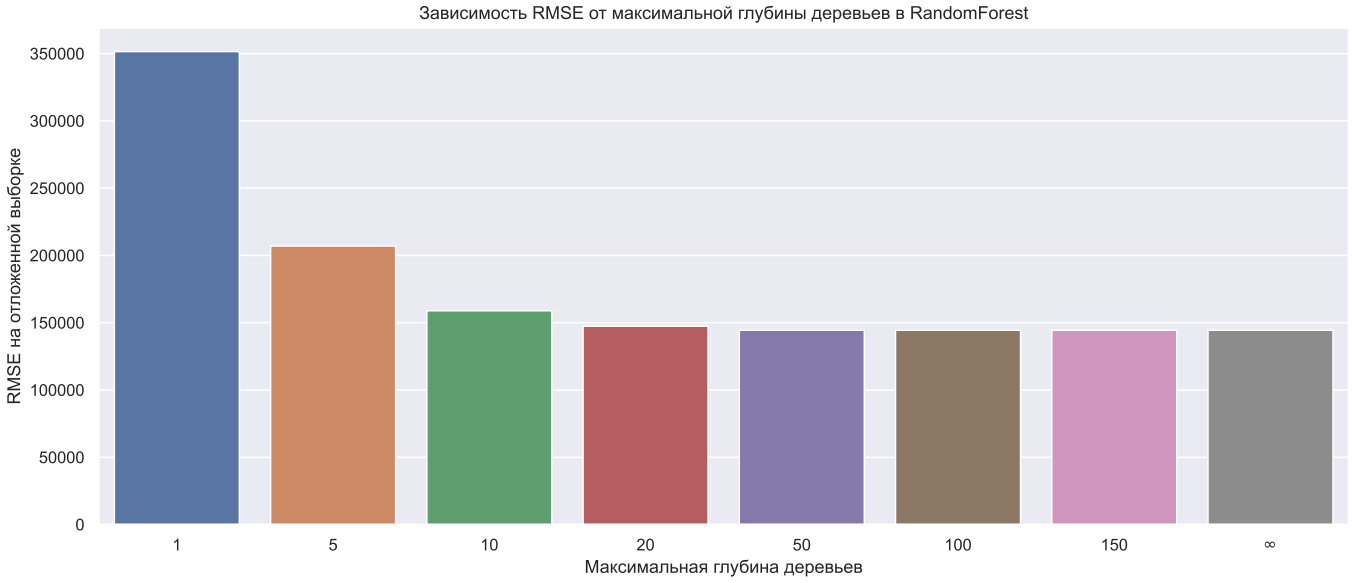

Рис. 4.

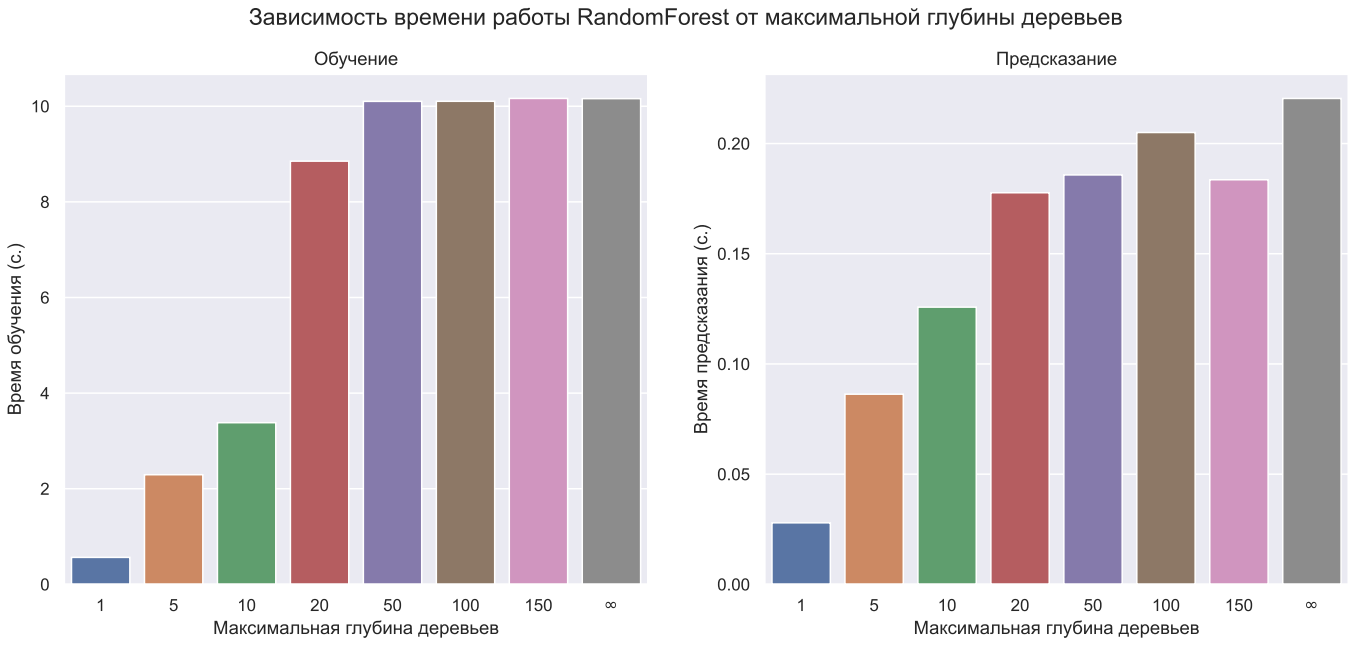

Рис. 5.

Фиксируем max_depth $= \infty$. При зафиксированных гипермараметрах ошибка RMSE у случайного леса на тестовой выборке оказалась равна **144373.02**.

1.3 Исследование поведения алгоритма “Градиентный бустинг”

Исследуем поведение градиентного бустинга при изменении параметров, рассмотренных для случайного леса, а также при изменении параметра learning_rate.

1.3.1 Количество деревьев в ансамбле

Результаты исследования зависимости ошибки RMSE градиентого бустинга на отложенной выборки от параметра n_estimators приведены на рис. 6.

Рис. 6.

Из этого графика можно сделать выводы, аналогичные случаю *случайного леса*. Время обучения и время предсказания также будут линейно зависеть от числа деревьев. Фиксируем n_estimators $= 100$.

1.3.2 Размерность подвыборки признаков для одного дерева

Результаты перебора мощности случайных подмножеств для градиентного бустинга отображены на рис. 7 и рис. 8.

Из этих графиков можно сделать следующие выводы:

Рис. 7.

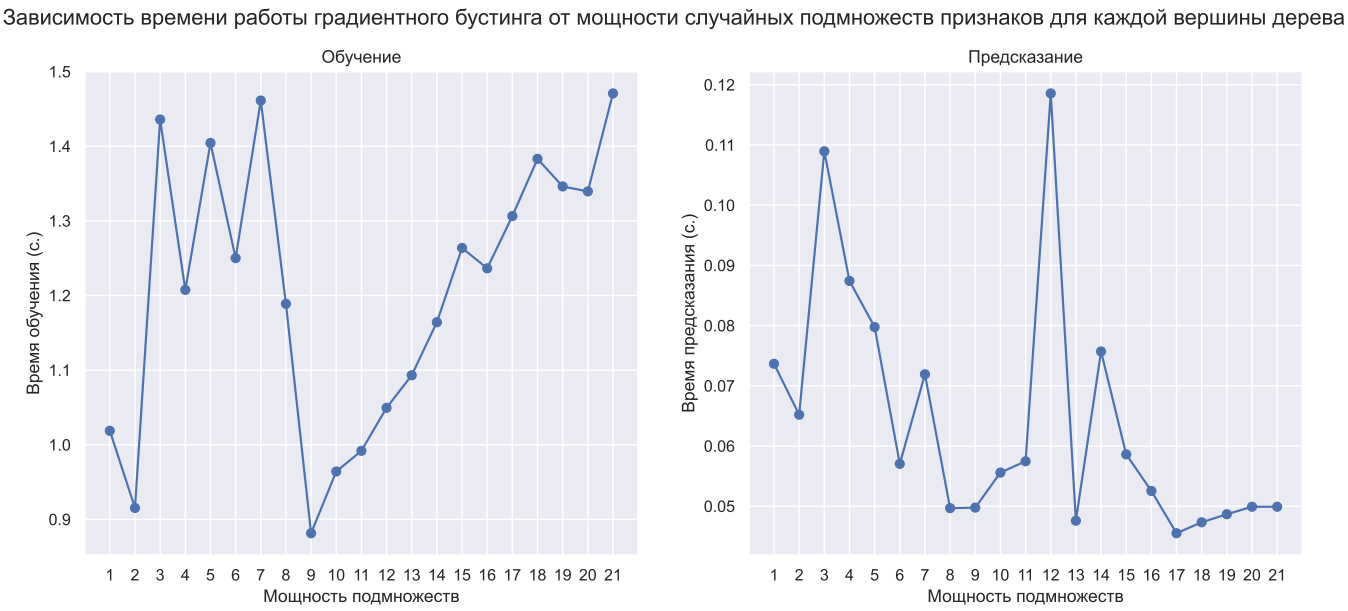

Рис. 8.

1. На графике RMSE виден тренд уменьшения ошибки с увеличением мощности (но минимум достигается в точке 20).

2. График времени обучения имеет “резкий провал” при мощности подмножеств = 9 (не известно почему), после чего наблюдается линейная зависимость времени работы от мощности подмножеств (больше мощность - больше время обучения).

3. Время предсказания не зависит от мощности.

Фиксируем feature_subsample_size = 20.

1.3.3 Максимальная глубина деревьев

Результаты перебора параметра max_depth для случая градиентного бустинга отображены на рис. 9 и рис. 10.

Из них следует, что:

1. RMSE как функция, зависящая от максимальной глубины, выпукла вниз и имеет минимум в точке 7.

2. С увеличением максимальной глубины деревьев время работы увеличивается.

Фиксируем max_depth = 7.

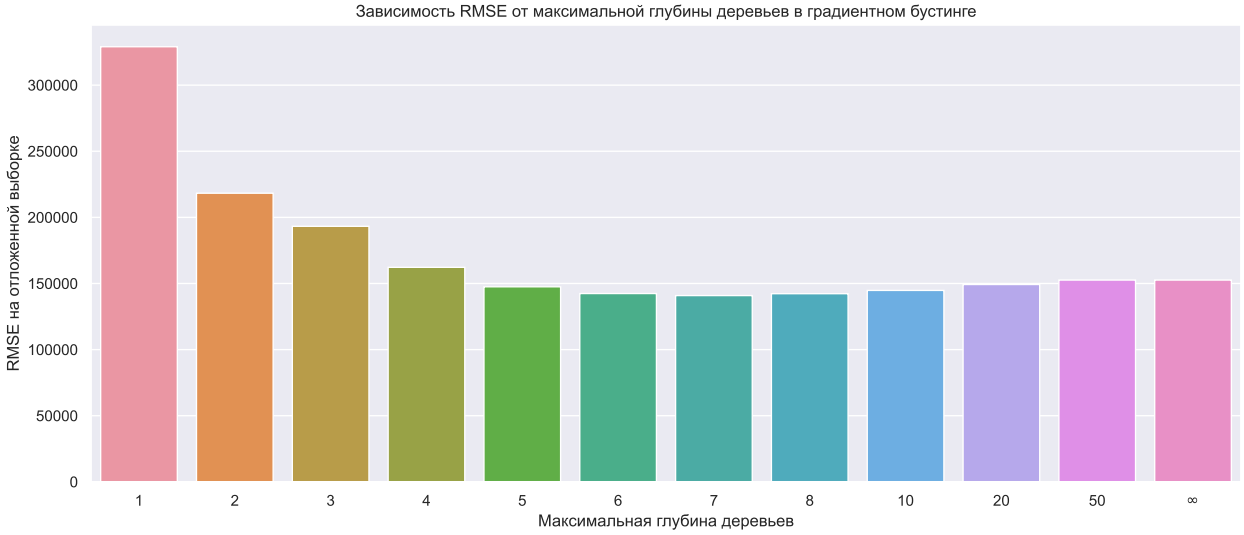

Рис. 9.

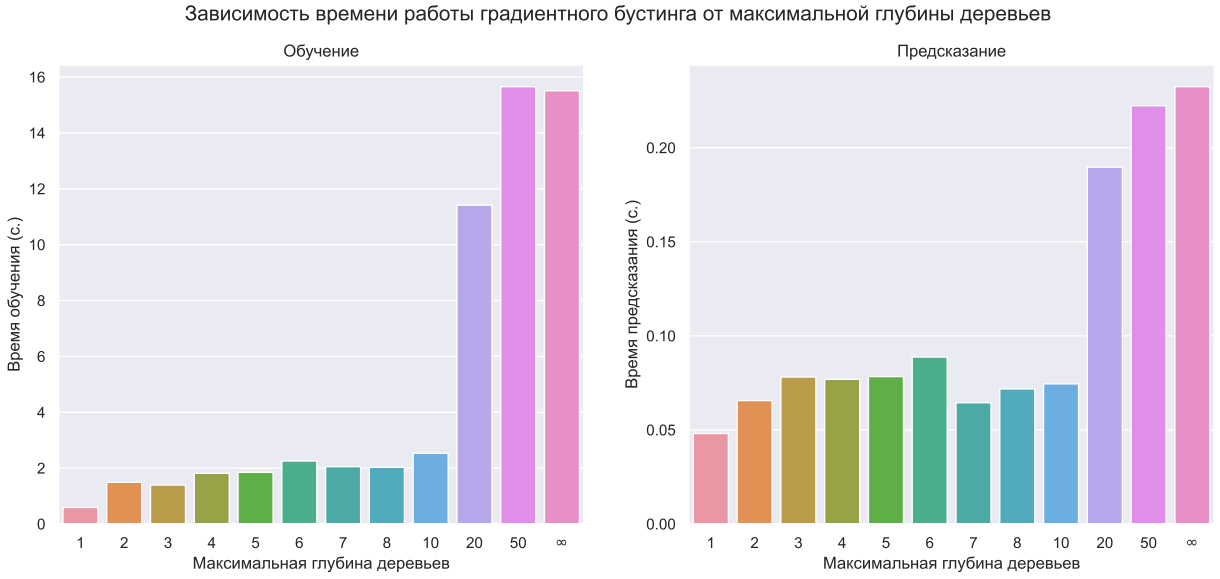

Рис. 10.

1.3.4 Параметр learning_rate

Результаты перебора параметра learning_rate отображены на рис. 11 и рис. 12.

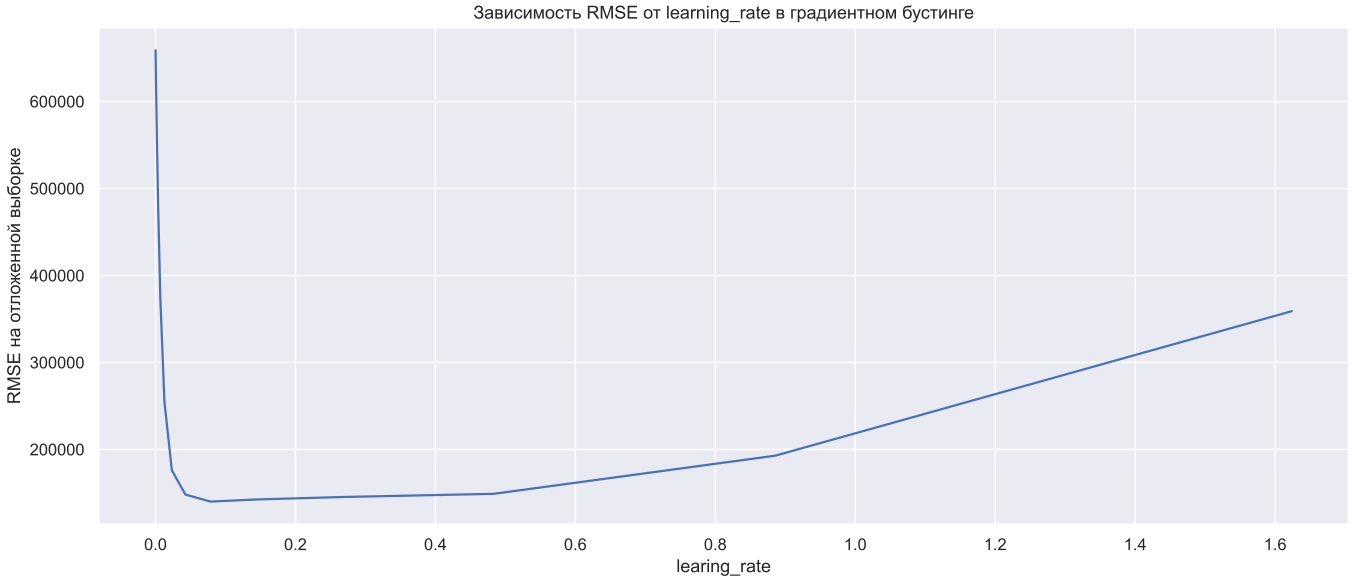

Рис. 11.

Рис. 12.

Из этих графиков следует, что:

1. RMSE как функция от learning_rate выпукла вниз и имеет минимум где-то в окрестности точки 0.07.

2. При слишком маленьком learning_rate ошибка большая.

3. При learning_rate > 1 RMSE резко возрастает.

4. Можно сказать, что время работы алгоритма не зависит от learning_rate.

Фиксируем learning_rate $= 0.1$. При обучении модели градиентного бустинга с зафиксированными параметрами получаем ошибку RMSE на отложенной выборке равную **140848.83**.

Заключение

Было рассмотрено поведение алгоритмов “Случайный лес” и “Градиентный бустинг” при разных параметрах.

При параметрах

$$n\_estimators = 100,\ feature\_subsample\_size = 18,\ max\_depth = \infty$$

RMSE для случайного леса на тестовой выборке = **144373.02**.

При параметрах

$$n\_estimators = 100,\ feature\_subsample\_size = 20,\ max\_depth = 7,$$

$$leaning\_rate = 0.01$$

RMSE для градиентого бустинга на тестовой выборке = **140848.83**, что меньше, чем у случайного леса на 3524.19 (на 2.5 %).

Итого, градиентный бустинг дал меньшую ошибку.

Список литературы

[1] Лекции Воронцова К. В.