Московский государственный университет имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики

Кафедра ММП, компьютерный практикум

Отчет о выполненном задании «Композиции алгоритмов»

Травникова Арина Сергеевна, 317 группа

Введение.

Задание заключается в реализации алгоритмов случайный лес и градиентный бустинг для решения задачи регрессии и исследовании их работы на датасете данных о продажах недвижимости. В процессе работы проведены различные эксперименты для сравнения эффективности и качества работы алгоритмов с разными параметрами и подбора оптимальных гиперпараметров. Также выполнено предсказание для тестовой выборки.

Постановка задачи композиции алгоритмов для решения задачи регрессии.

Пусть дана обучающая выборка с известными ответами $\{(x_i, y_i)\}_{i=1}^N$, где $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$, и выборка объектов $\{z_i\}_{i=1}^M$, $z_i \in \mathbb{R}^d$, для которых необходимо предсказать ответ.

Случайный лес.

Модель случайный лес заключается в объединении множества решающих деревьев $\{a_i(x)\}_{i=1}^t$ и предсказании ответа усреднением по всем прогнозам, полученным на отдельных деревьях множества:

$$a(x) = \frac{1}{t}\sum_{j=1}^{t} a_j(x).$$

Каждое дерево $a_j(x)$ алгоритма является глубоким и настраивается на бутстраповской псевдовыборке, которая генерируется случайным выбором объектов с повторениями из обучающей выборки. Подмножество признаков для построения каждого дерева также выбирается случайным образом.

Градиентный бустинг.

Главное отличие бустинга от случайного леса состоит в том, что базовые алгоритмы строятся не независимо, а каждый следующий настраивается жадным способом так, чтобы он исправлял ошибки предыдущих и повышал качество всего ансамбля.

Композиция для предсказания ответа определяется как сумма базовых алгоритмов: $a_t(x) = q\sum\limits_{j=1}^{t}\alpha_j b_j(x)$, где $b_j(x)$ - неглубокие деревья, α_j - некоторые параметры, $q \in (0,1)$ - скорость обучения.

Построение ансамбля методом градиентного бустинга представляет собой итеративный процесс, на первом шаге которого рассматривается константный алгоритм $a_0(x) = \frac{1}{N}\sum\limits_{j=1}^{N} y_j$. На шаге $t \geq 1$ решается оптимизационная задача для минимизации эмпирического риска:

$$\sum_{j=1}^{N} L(y_j, a_{t-1}(x_i) + \alpha b(x_i)) \to \min_{\alpha,b}, \quad L(y,z) = (y-z)^2.$$

На обучающей выборке она сводится к задаче оптимизации по вектору S:

$$F(s_1,\ldots,s_N) = \sum_{j=1}^{N} L(y_j, a_{t-1}(x_i) + s_i) \to \min_{S},$$

минимизирующий вектор $S = -\nabla F$. Алгоритм b_t и коэффициент α_t определяются соотношениями

$$b_t(x) = \operatorname*{argmin}_{b}\sum_{j=1}^{N}(b(x_i) - s_i)^2, \quad \alpha_t = \operatorname*{argmin}_{\alpha}\sum_{j=1}^{N} L(y_j, a_{t-1}(x_i) + \alpha b_t(x_i))$$

Итоговый алгоритм имеет вид: $a_t(x) = q\sum\limits_{j=1}^{t}\alpha_j b_j(x)$.

Число деревьев ансамбля и параметры каждого отдельного дерева: максимальная глубина, размерность подпространства признаков, коэффициент скорости обучения - являются гиперпараметрами модели и настраиваются для конкретной задачи в ходе работы.

Датасет для исследования состоит из данных о продажах недвижимости: 12096 объектов выделено для обучения и 5184 - для теста. Размерность пространства признаков - 19. Признаки включают в себя такие количественные характеристики недвижимости, как-то: количество комнат, номер этажа, географические координаты, а также id объекта и дату продажи. Для дальнейшей работы с данными дата переведена в секундное представление, а признак id отброшен, так как является уникальным для большинства объектов выборки. Задача регрессии заключается в предсказании цены продажи на некоторый объект недвижимости.

Эксперимент 1.

Эксперимент состоит в исследовании зависимости значения функции потерь RMSE и времени работы алгоритма случайного леса от различных настраеваемых гиперпараметров:

- $1 \leq n\_estimators \leq 1000$ - количество деревьев ансамбля
- $1 \leq feature\_subsample\_size \leq 18$ - размерность подвыборки признаков для одного дерева
- $1 \leq max\_depth \leq 15$- максимальная глубина дерева и случай, когда глубина неограничена

Ошибка алгоритма на тестовой выборке определяется формулой

$$RMSE = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(y_i - a(x_i))^2}.$$

Исследование зависимости времени работы и ошибки RMSE на тестовой выборке алгоритма случайного леса от числа деревьев в ансамбле проводится при заданных параметрах отдельных деревьев $max\_depth = 15$, $feature\_subsample\_size = 6$. Так как в алгоритме случайного леса велика доля рандомизации, для более корректного построения графиков проведено усреднение $RMSE$ по 3 независимым предсказаниям моделей с заданными параметрами. Результаты приведены на рис. 1.

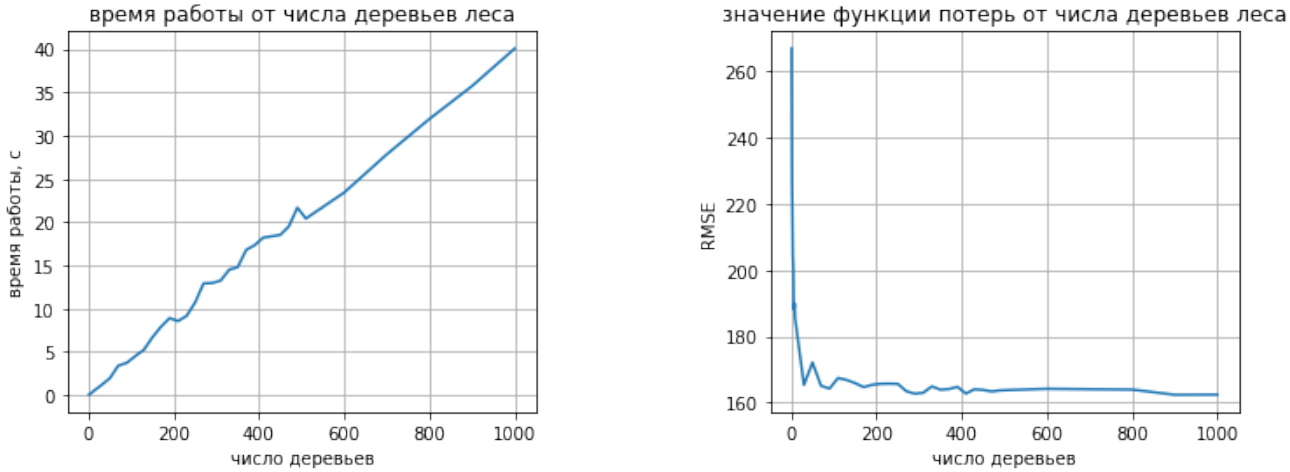

Рис. 1: Зависимость времени работы и ошибки RMSE на тестовой выборке алгоритма случайного леса от количества деревьев.

Графики показывают, что с увеличением числа деревьев в ансамбле значение функционала ошибок $RMSE$ уменьшается, а время работы алгоритма возрастает. Результат

подтверждается и теорией: так как все базовые деревья композиции имеют одинаковые гиперпараметры, обучаются независимо и последовательно, время работы линейно зависит от числа деревьев леса. При небольшом количестве деревьев $n\_estimators \leq 100$ $RMSE$ заметно уменьшается, однако к величине $n\_estimators = 500$ график выходит на асимптоту $RMSE = 163$, то есть можно говорить о сходимости.

Для проведения дальнейших экспериментов выбрано число деревьев 500: при таком количестве деревьев кривая $RMSE$ уже выходит на асимптоту, а время работы агоритма при этом значении функции ошибок минимально.

Исследование поведения работы случайного леса при различных значениях максимальной глубины каждого дерева ансамбля проводится при заданных $n\_estimators = 500$, $feature\_subsample\_size = 6$. Результаты приведены на рис. 2.

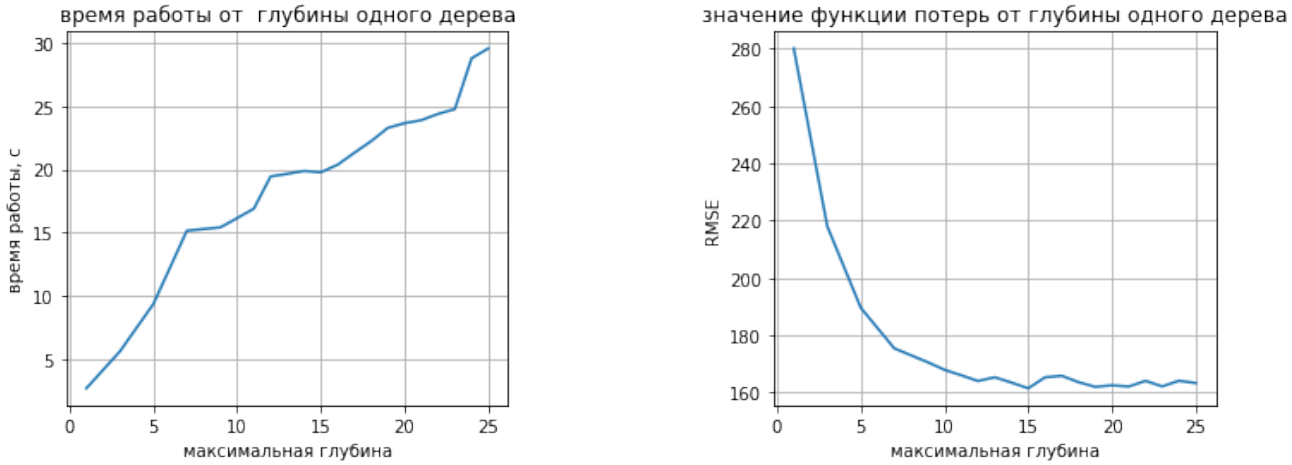

Рис. 2: Зависимость времени работы и ошибки RMSE на тестовой выборке алгоритма случайного леса от максимальной глубины деревьев.

При неограниченной глубине $RMSE = 162.8$. Результаты показывают, что с увеличением максимально допустимой глубины каждого дерева ансамбля значение функции потерь уменьшается, а время работы алгоритма увеличивается. Время работы становится больше из-за того, что для построения каждогого дерева модели требуется больше затрат, так как оно становится глубже.

Предсказание модели производится через усреднение ответов по всем деревьям случайного леса. Такой подход позволяет снизить дисперсию ошибки алгоритма, но не смещение, то есть для минимизации ошибки деревья ансамбля должны быть одновременно не похожими друг на друга и точными, соответственно глубокими.

Зависимость поведения метода случайного леса от размера подмножества признаков для каждого дерева исследуется при прочих фиксированных параметрах $max\_depth = 20$, $n\_estimators = 500$. Результаты работы показаны на рис. 3.

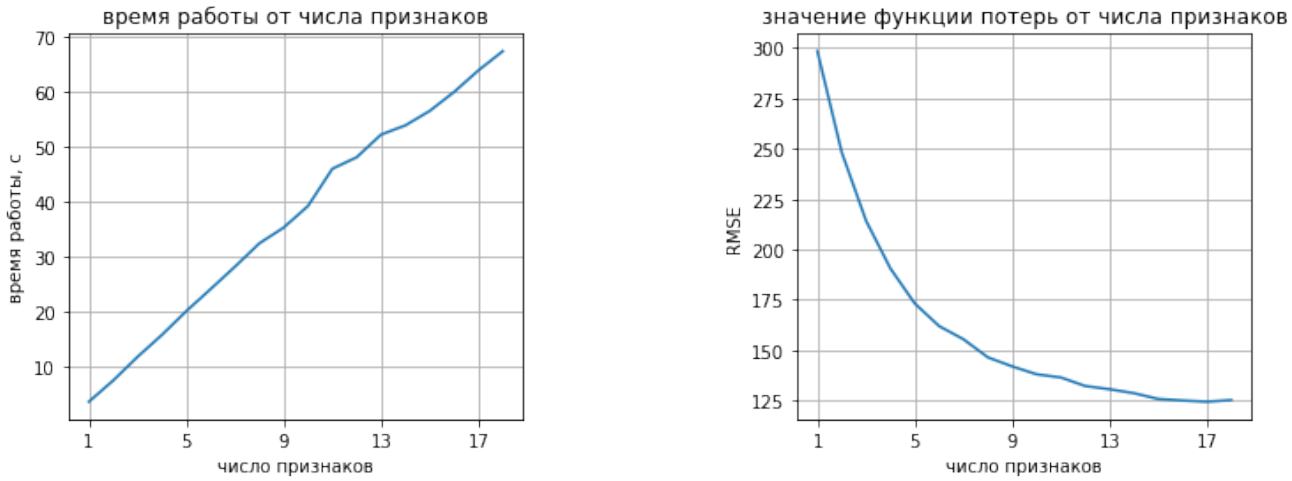

Рис. 3: Зависимость времени работы и ошибки RMSE на тестовой выборке алгоритма случайного леса от размера подмножества признаков для построения деревьев.

Эксперимент показывает: чем больше признаков выбирается для построения каждого отдельного дерева леса, тем меньше ошибка прогноза случайного леса и тем больше время его работы. Несмотря на то, что при увеличении размера подмножества признаков для обучения деревья становятся более однообразными, они все равно остаются непохожими из-за бутстраповских обучающих выборок. Можно предположить, что в таком случае дисперсия ошибки возрастает лишь незначительно, а смещение уменьшается, так как при большем наборе признаков сами деревья становятся точнее. Время работы увеличивается пропорционально количеству признаков по причине того, что для каждого дерева композиции увеличиваются размер обучающей и тестовой выборки, то есть требуются большие вычислительные затраты.

Применяя лучшие подобранные параметры: $n\_estimators = 1000$, $max\_depth = 20$, $feature\_subsample\_size = 17$ - для предсказания на тестовой выборке, получаем ошибку $RMSE = 124.95$.

Эксперимент 2.

Эксперимент состоит в исследовании зависимости значения функции потерь RMSE и времени работы алгоритма градиентного бустинга от различных настраеваемых гиперпараметров:

- $1 \leq n\_estimators \leq 1000$ - количество деревьев ансамбля
- $1 \leq feature\_subsample\_size \leq 18$ - размерность подвыборки признаков для одного дерева
- $1 \leq max\_depth < 10$ - максимальная глубина дерева и случай, когда глубина неограничена

- $learning\_rate \in (0, 1]$- скорость обучения

Исследование зависимости времени работы и ошибки RMSE на тестовой выборке алгоритма градиентного бустинга от числа деревьев в ансамбле проводится при заданных параметрах отдельных деревьев $max\_depth = 3$, $feature\_subsample\_size = 6, learning\_rate = 0.1$. Так как в модели велика доля случайности, для более корректного построения графиков проведено усреднение $RMSE$ по 3 независимым предсказаниям алгоритмов с заданными параметрами. Результаты приведены на рис. 4.

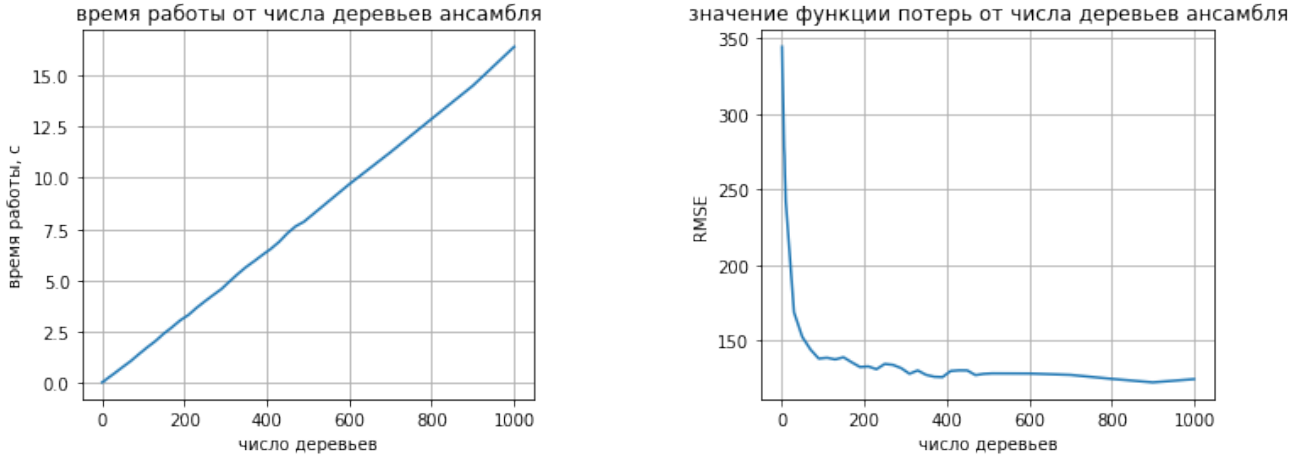

Рис. 4: Зависимость времени работы и ошибки RMSE на тестовой выборке алгоритма градиентного бустинга от количества деревьев в ансамбле.

Графики показывают, что с увеличением числа деревьев в ансамбле значение функционала ошибок $RMSE$ уменьшается, а время работы алгоритма линейно возрастает. При небольшом количестве деревьев $n\_estimators \leq 100$ ошибка быстро уменьшается, однако далее выходит на асимптоту и колеблется около некоторого значения $RMSE = 128$.

Для дальнейших экспериментов выбрано число деревьев 500, так как при таком количестве деревьев кривая функции ошибок приближается к асимптоте за минимальное временя работы.

Исследование поведения алгоритма градиентного бустинга при различных значениях максимальной глубины каждого дерева ансамбля проводится при заданных параметрах $n\_estimators = 500, feature\_subsample\_size = 6, learning\_rate = 0.1$. Результаты приведены на рис. 5.

С увеличением максимально допустимой глубины каждого дерева леса время работы алгоритма линейно возрастает. Функционал ошибки $RMSE$ достигает минимума в точке $max\_depth = 4$. С увеличиением $max\_depth > 4$ значение функции потерь возрастает, что объясняется эффектом переобучения, наступающим при усложнении модели - увеличении глубины деревьев. При неограниченной глубине $RMSE = 130.9$.

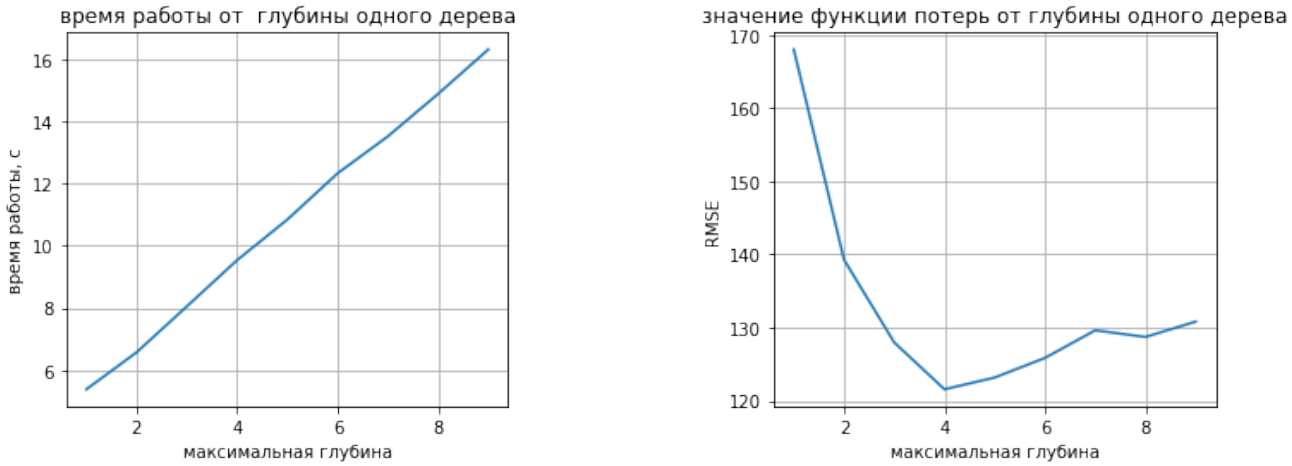

Рис. 5: Зависимость времени работы и ошибки RMSE на тестовой выборке алгоритма градиентного бустинга от максимальной глубины деревьев.

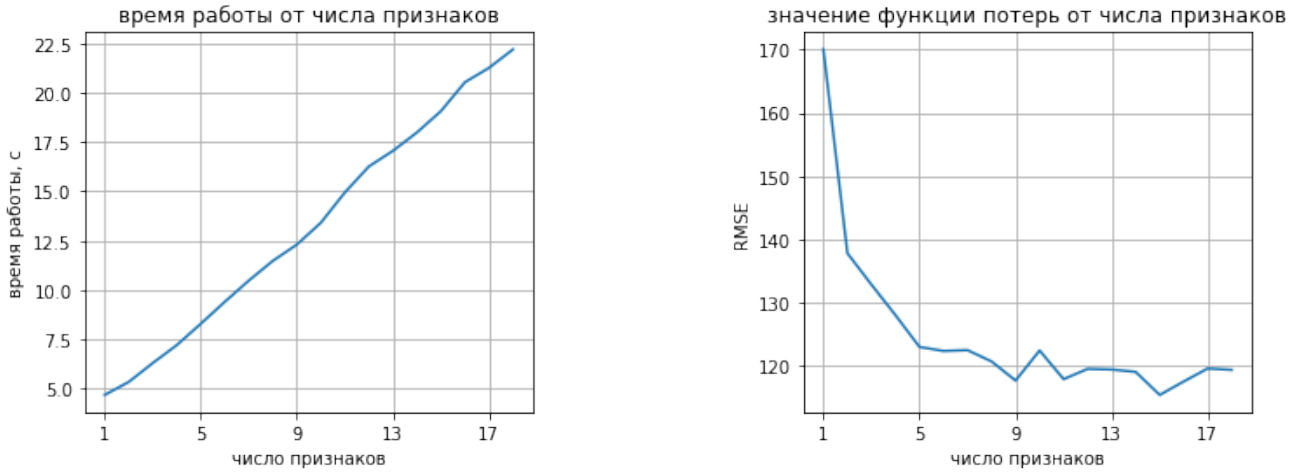

Рис. 6: Зависимость времени работы и ошибки RMSE на тестовой выборке алгоритма случайного леса от размера подмножества признаков для построения деревьев.

Зависимость поведения метода градиентного бустинга от выбора размера подмножества признаков для обучения кажлого дерева исследуется при прочих фиксированных параметрах $max\_depth = 4$, $n\_estimators = 500, learning\_rate = 0.01$. Результаты работы изображены на рис. 6.

Эксперимент показывает, что как и в случае случайного леса, наблюдается прямая зависимость времени работы алгоритма и обратная зависимость величины ошибки от количества признаков из-за усложнения модели и возможного переобучения.

Исследование зависимости ошибки RMSE и времение работы алгоритма градиентного

бустинга от значения параметра $learning\_rate$ проводится при остальных подобранных параметрах $n\_estimators = 500, max\_depth = 4, feature\_subsample\_size = 15$. Результаты работы изображены на рис. 7.

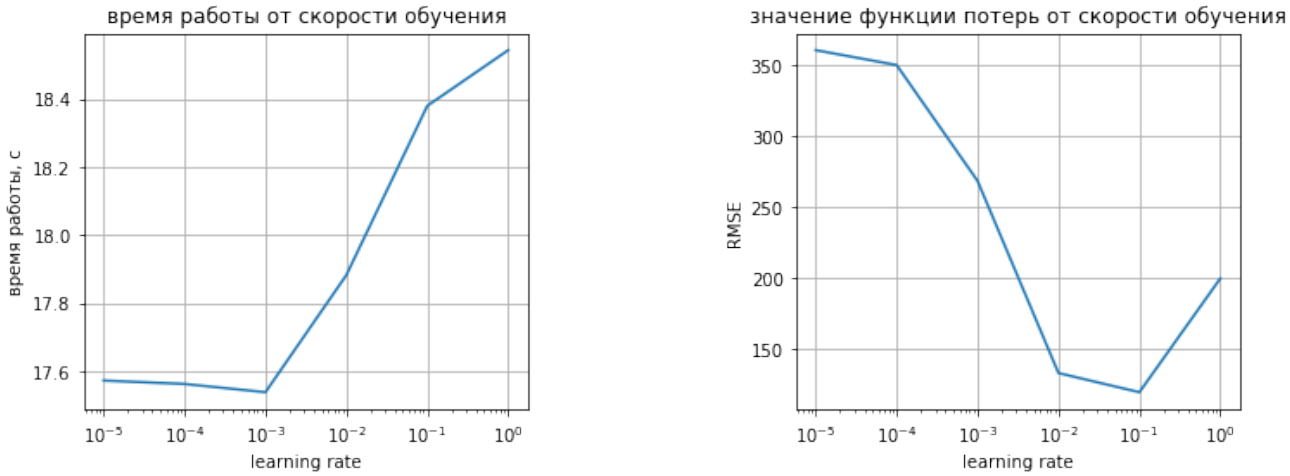

Рис. 7: Зависимость времени работы и ошибки RMSE на тестовой выборке алгоритма случайного леса от коэффициента $learning\_rate$.

По результатам эксперимента оптимальное значение $learning\_rate = 0.1$. С уменьшением параметра $learning\_rate < 0.1$ ошибка предсказания возрастает, так как на каждом шаге слишком слабо учитываются предыдущие алгоритмы, и их ошибка плохо исправляется. С увеличением $learning\_rate > 0.1$ ошибка также увеличивается: из-за сильной коррекции ошибочных предсказаний предыдущих алгоритмов может возникать переобучение.

Применяя лучшие подобранные параметры: $n\_estimators = 1000$, $max\_depth = 4$, $feature\_subsample\_size = 15, learning\_rate = 0.1$ - для предсказания на тестовой выборке, получаем ошибку $RMSE = 116.7$.

Вывод.

Проведенные эксперименты показали, что композиции алгоритмов с использованием некоторых эвристик дают хорошую точнось для задачи регрессии: случайный лес и градиентный бустинг помогают улучшить качество работы отдельно взятых деревьев, причем ошибка метода градиентного бустинга оказывается меньше.

Для оптимального решения важно правильно настроить гиперпараметры моделей, такие как: количество деревьев ансамбля, число признаков для построения каждого дерева, максимальная глубина - под конкретную задачу. Помимо повышения точности предсказания и уменьшения вемени работы, грамотно настроенные параметры также позволяют бороться с переобучением.