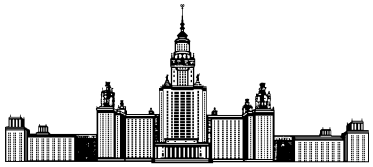
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

Отчет по заданию 3

по курсу «Практикум на ЭВМ 2021/2022»

Ансамбли алгоритмов. Веб-сервер.

Композиции алгоритмов для решения задачи регрессии.

Выполнил:

студент 3 курса 317 группы

Борисов Алексей Антонович

Москва, 2021

Содержание

1 Введение

Задание заключалось в ознакомлении с методами ансамблирования на примере случайного леса и градиентного бустинга. Применяя эти методы, необходимо было провести эксперименты на датасете данных о продажах недвижимости. Также важной частью задания было проектирование веб-интерфейса для взаимодействия с нашей моделью, представление решения в виде docker контейнера, а также ведение всего проекта в виде github репозитория.

2 Эксперименты

2.1 Предобработка данных

Прежде всего необходимо было провести предобработку данных. Был удален ненужный столбец id, представляющий собой просто некоторый идентификатор продажи. Из столбца date были извлечены следующие признаки: год, месяц, день недели, день в году, день в месяце. Далее были выделены категориальные признаки: zipcode, являющийся почтовым индексом, день недели, а также номер месяца. После этого данные были разбиты на обучение и контроль в отношении 7 к 3, а категориальные признаки были закодированы с помощью счетчиков со сглаживанием.

2.2 Случайный лес

Исследуем зависимость RMSE на отложенной выборке и время работы алгоритма случайного леса в зависимости от следующих факторов:

- Количество деревьев в ансамбле
- Максимальная глубина дерева
- Размерность подвыборки признаков для одного дерева

2.2.1 Зависимость от количества деревьев

Рассмотрим количество деревьев от 1 до 500. Для остальных параметров будем использовать значения по умолчанию: глубина деревьев неограничена, каждое дерево использует треть всех признаков. На рисунке 1 изображены графики, иллюстрирующие зависимость RMSE на отложенной выборке и время обучения в зависимости от количества деревьев в ансамбле. На левом графике рассмотрены значения количества деревьев от 3 до 500, так как при количестве деревьев 1 и 2 потери оказываются еще больше, и график становится менее презентативным. Видно, что с увеличением количества деревьев потери в среднем уменьшаются, и левый график выходит на асимптоту.

Нетрудно заметить, что наблюдается линейный рост времени обучения при увеличении числа деревьев в ансамбле.

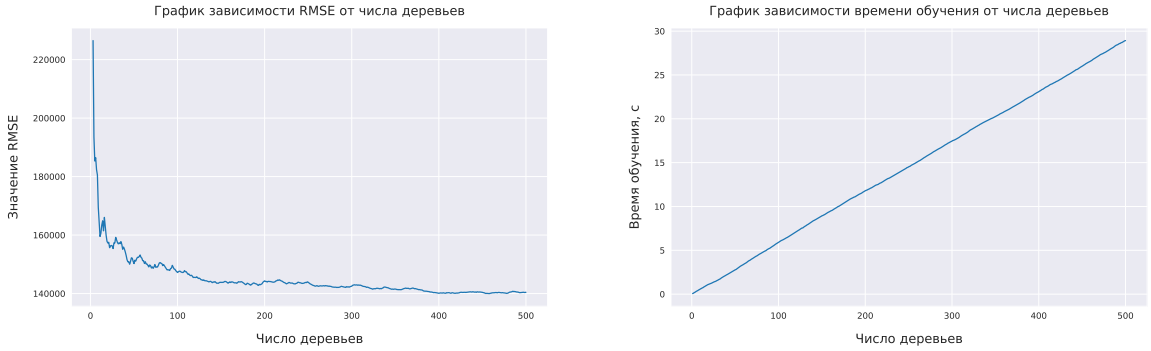


Рис. 1: Графики зависимости RMSE и времени обучения от числа деревьев для случайного леса.

2.2.2 Зависимость от максимальной глубины дерева

Зафиксируем число деревьев равное 400, размерность подвыборки признаков остается равной трети от всех признаков. Рассмотрим значения максимальной глубины дерева от 1 до 34, а также отдельно рассмотрим случай, когда глубина неограничена. На рисунке 2 изображены графики, иллюстрирующие зависимость RMSE на отложенной выборке и времени обучения от максимальной глубины дерева. Как мы видим, с увеличением глубины дерева потери на отложенной выборке уменьшаются и с некоторого момента начинают колебаться вокруг определенного значения около 140 000. Можно сделать вывод, что при использовании случайного леса на наших данных следует использовать достаточно глубокие деревья.

Анализируя зависимость времени обучения от максимальной глубины, заметим, что сперва наблюдается линейный рост времени обучения при увеличении максимальной глубины, но с некоторого момента время обучения также начинает колебаться около определенного значения, что может говорить о том, что на самом деле деревья получаются меньшей глубины, так как их рост заканчивается по другим критериям останова.

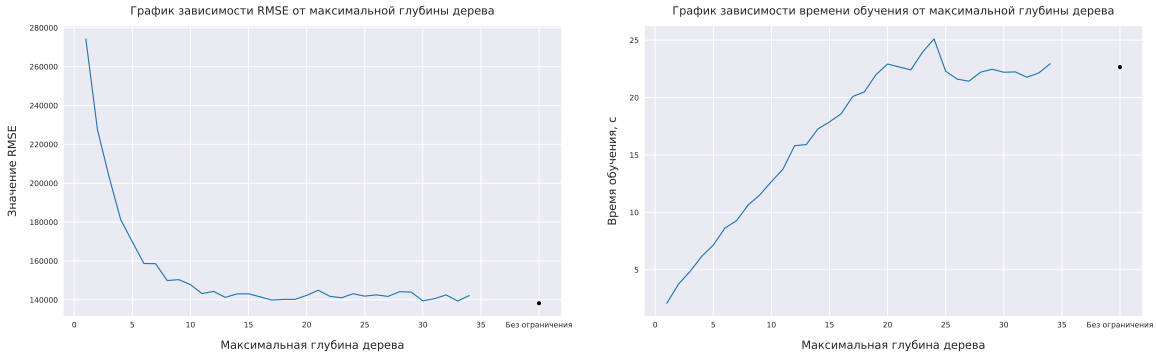


Рис. 2: Графики зависимости RMSE и времени обучения от максимальной глубины дерева для случайного леса.

2.2.3 Зависимость от размерности подвыборки признаков

Теперь исследуем зависимость RMSE и времени обучения от размерности подвыборки признаков в каждом дереве. Возьмем число деревьев равное 400, максимальную глубину деревьев ограничивать не будем. На рисунке 3 изображены графики, иллюстрирующие зависимость RMSE на отложенной выборке и времени обучения от размерности подвыборки признаков. Как мы видим, наименьшие потери достигаются при размерности подвыборки признаков равной 0.8 и 0.9. При использовании подвыборок размерности меньше 0.5 потери начинают резко увеличиваться. Это может свидетельствовать о том, что в датасете есть небольшое количество важных признаков, и модели, обучающиеся на подвыборках без этих важных признаков, обладают плохой предсказательной способностью.

Нетрудно заметить, что наблюдается линейный рост времени обучения при увеличении размерности подвыборки признаков. Это объясняется тем, что с ростом размерности подвыборки признаков линейно растет время перебора комбинаций признаков и порогов в каждой вершине дерева.

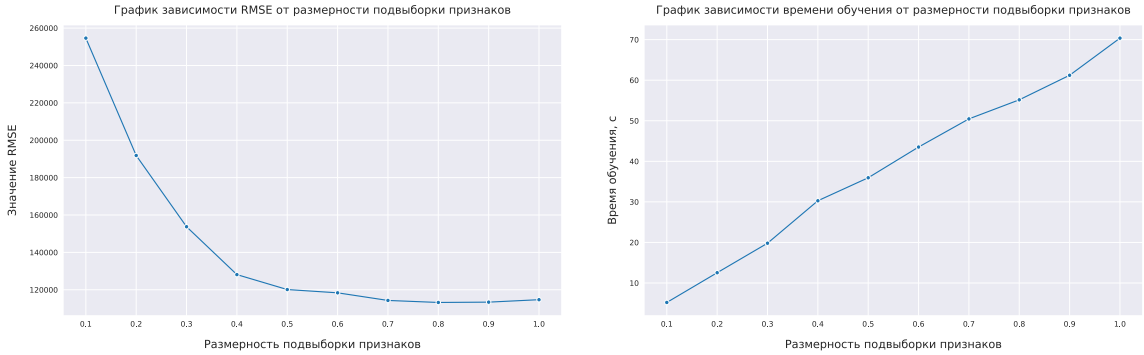


Рис. 3: Графики зависимости RMSE и времени обучения от размерности подвыборки признаков для случайного леса.

2.3 Градиентный бустинг

Исследуем зависимость RMSE на отложенной выборке и время работы алгоритма градиентного бустинга в зависимости от следующих факторов:

- Количество деревьев в ансамбле
- Максимальная глубина дерева
- Размерность подвыборки признаков для одного дерева
- Значение темпа обучения

2.3.1 Зависимость от количества деревьев

Будем использовать те же параметры, что и в случае случайного леса. Темп обучения равен 0.1. На рисунке 4 изображены графики, иллюстрирующие зависимость RMSE на отложенной выборке и время обучения в зависимости от количества деревьев в ансамбле. Как мы видим, с увеличением количества деревьев потери на отложенной выборке уменьшаются, причем здесь уменьшение очень плавное, тогда как для случайного леса график заметно дергался. Это можно объяснить тем, что в случае градиентного бустинга каждое новое дерево исправляет ошибки предыдущих. Также стоит отметить, что в случае градиентного бустинга график быстрее выходит на асимптоту, чем в случае случайного леса. Уже 100-200 деревьев дают значение функции потерь слабо отличающееся от бóльшего количества деревьев.

Как и для случайного леса наблюдается линейная зависимость времени работы от количества деревьев.

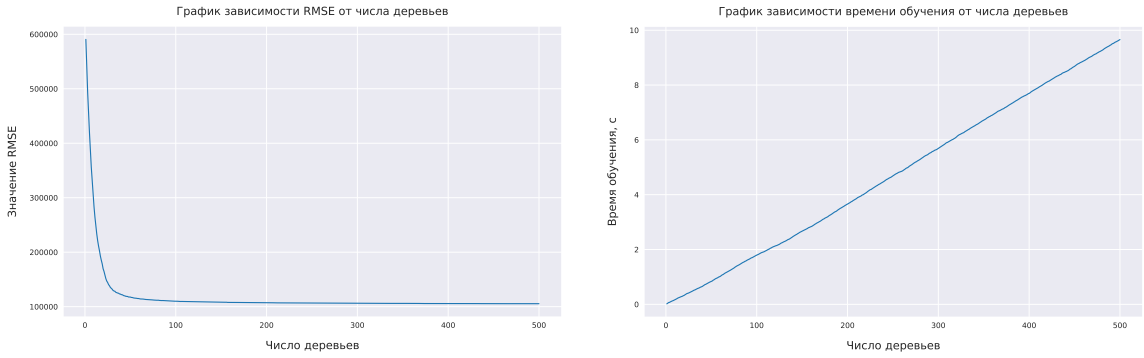


Рис. 4: Графики зависимости RMSE и времени обучения от количества деревьев для градиентного бустинга.

2.3.2 Зависимость от максимальной глубины дерева

Зафиксируем число деревьев равное 200, размерность подвыборки признаков остается равной трети от всех признаков. Рассмотрим значения максимальной глубины дерева от 1 до 34, а также отдельно рассмотрим случай, когда глубина неограничена. На рисунке 5 изображены графики, иллюстрирующие зависимость RMSE на отложенной выборке и времени обучения от максимальной глубины дерева. Как мы видим, в отличии от случайного леса для градиентного бустинга нет тенденции, что чем больше глубина тем лучше. Теперь лучшие результаты достигаются при максимальной глубине 4-7, а при 6 достигается наименьшее значение функции потерь. При увеличении максимальной глубины дерева значение потерь сперва возрастает, но с некоторого момента начинает колебаться вокруг определенного значения.

Анализируя зависимость времени обучения от максимальной глубины, заметим, что сперва наблюдается линейный рост времени обучения при увеличении максимальной глубины, но с некоторого момента рост прекращается, и график начинает колебаться около определенного значения, что как и для случайного леса может говорить о том, что на самом деле деревья получаются меньшей глубины, так как их рост заканчивается по другим критериям останова.

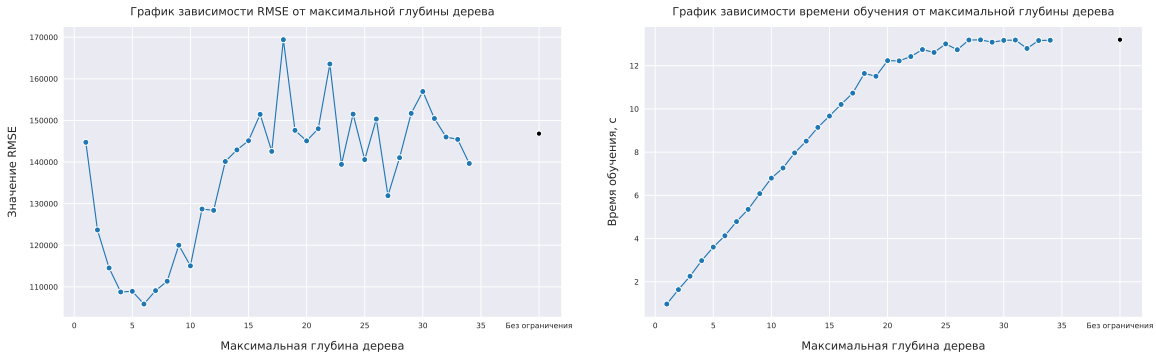


Рис. 5: Графики зависимости RMSE и времени обучения от максимальной глубины дерева для градиентного бустинга.

2.3.3 Зависимость от размерности подвыборки признаков

Теперь исследуем зависимость RMSE и времени обучения от размерности подвыборки признаков в каждом дереве. Возьмем число деревьев равное 200 и максимальную глубину дерева 6, так как эта глубина показала наилучшие результаты. На рисунке 6 изображены графики, иллюстрирующие зависимость RMSE на отложенной выборке и времени обучения от размерности подвыборки признаков. Как мы видим, наименьшие потери достигаются при размерности подвыборки признаков равной 0.5. Как и в случае случайного леса при достаточно малом размере подвыборки признаков значение функции потерь увеличивается, так как многие деревья обучаются на не самых полезных признаках.

Нетрудно заметить, что наблюдается линейный рост времени обучения при увеличении размерности подвыборки признаков. Как и для случайного леса это объясняется тем, что с ростом размерности подвыборки признаков линейно растет время перебора комбинаций признаков и порогов в каждой вершине дерева.

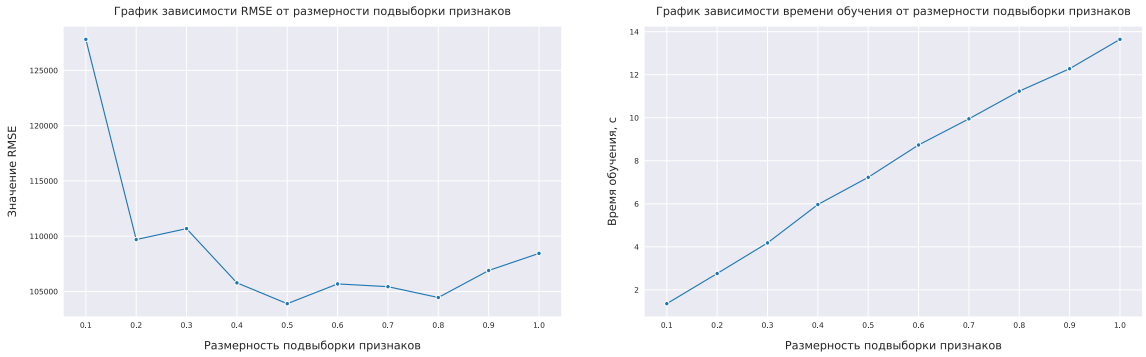


Рис. 6: Графики зависимости RMSE и времени обучения от размерности подвыборки признаков для градиентного бустинга.

2.3.4 Зависимость от значения темпа обучения

Теперь исследуем зависимость RMSE и времени обучения от темпа обучения. Возьмем число деревьев равное 200, максимальную глубину дерева 6, размерность подвыборки признаков 0.5. На рисунке 7 изображен график, иллюстрирующий зависимость RMSE на отложенной выборке от темпа обучения. Как мы видим, при маленьких значениях темпа обучения модель не успевает обучиться, чтобы хорошо предсказывать целевую переменную, а при больших значениях темпа обучения модели не удается попасть в достаточно малую окрестность оптимального решения. Наименьшие значения потерь достигаются при темпе обучения равном 0.0316 и 0.1.

После проведения нескольких запусков было выяснено, что время обучения, как и следовало ожидать, не зависит от темпа обучения.

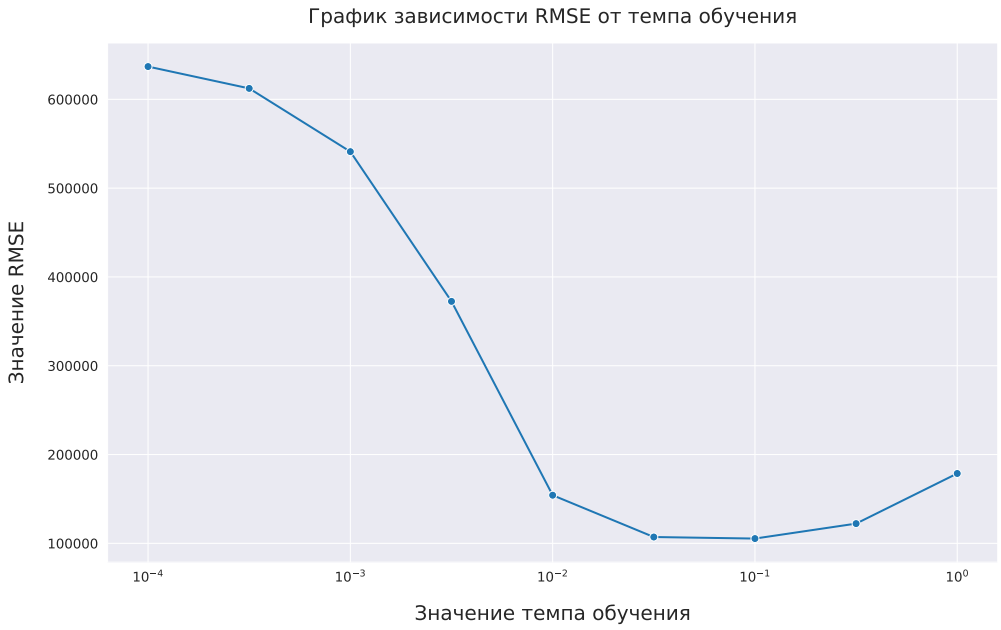


Рис. 7: График зависимости RMSE от темпа обучения для градиентного бустинга

2.4 Веб-сервер

Для взаимодействия с описанными выше моделями машинного обучения был написан небольшой веб-сервер. Более подробную информацию о нем можно найти в github репозитории.

2.5 Вывод

Были рассмотрены два различных подхода к созданию ансамблей: параллельный, на примере случайного леса, и последовательный, на примере градиентного бустинга. При проведении экспериментов уже были замечены некоторые отличия в этих подходах. Оба подхода обладают своими преимуществами и недостатками, например деревья для случайного леса можно учить одновременно, но их требуется больше чем в градиентном бустинге. В свою очередь градиентный бустинг зачастую дает лучшие результаты, но не предусматривает распараллеливание.