Практикум на ЭВМ.

Ансамбли алгоритмов. Веб-сервер.
Композиции алгоритмов для решения задач регрессии.

Ашинов Бислан Рамазанович
317 группа
Кафедра ММП ВМК МГУ

Москва 2021

1 Введение

В данном задании были реализованы ансамблевые алгоритмы регрессии над решающими деревьями, а именно случайный лес и градиентный бустинг. Необходимо было исследовать зависимость показателей работы алгоритмов в зависимости от различных параметров. Также надо было реализовать веб-сервер для взаимодействия с моделью.

2 Эксперименты

2.1 Предобработка

Данные взяты из соревнования. Колонка с датой сделки была разбита на две колонки - месяц и день, исходная колонка удалена. Также удалены колонки с id и ценой(это целевая переменная). Выборка разделена на обучающую и тестовую размерами 7:3 и преобразована в `numpy.ndarray`.

2.2 Случайный лес

В данном эксперименте исследуется поведение случайного леса в зависимости от различных параметров.
Обучим алгоритм с количеством деревьев 500 (будем смотреть на показатели в течении обучения, чтобы рассмотреть зависимость от количества деревьев). Максимальная глубина деревьев взята равной 5, а количество признаков, на которых обучается каждое дерево равно $\lfloor \frac{n}{3} \rfloor$(рекомендованные значения для регрессии [1]). Здесь и далее рассматриваются два показателя работы алгоритмов - ошибка RMSE на отложенной выборке и время обучения.

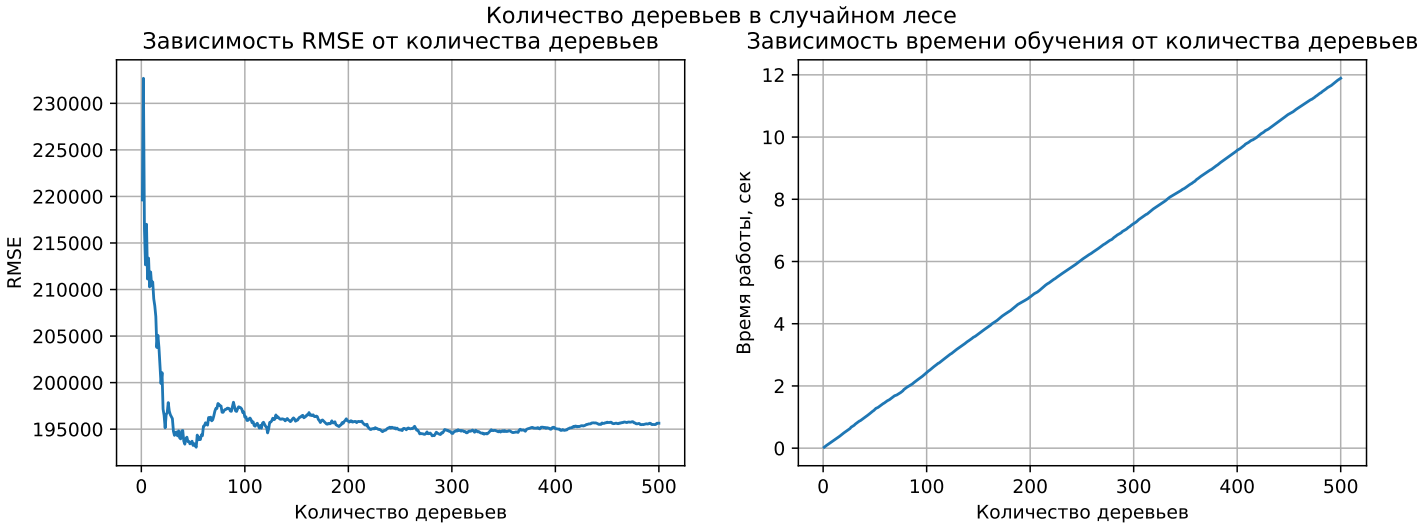

Рис. 1: Показатели обучения случайного леса в зависимости от количества деревьев. Были перебраны значения от 1 до 500. Максимальная глубина равна 5, размерность подвыборки признаков каждого дерева равна $\lfloor \frac{n}{3} \rfloor$

Время обучения ведет себя линейно. А RMSE при количестве деревьев больше 200 перестает сильно меняться(Рис.1). Это может быть связано с тем, что разброс перестает сильно меняться из-за большого количества деревьев, среди которых могут быть скоррелированные.

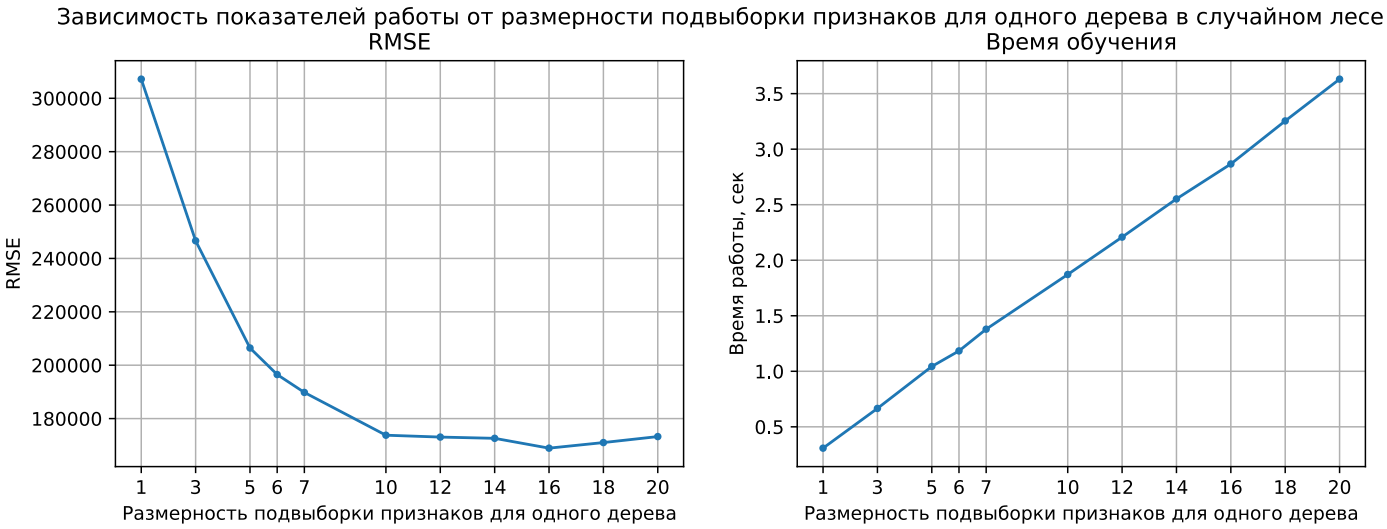

Рис. 2: Показатели обучения случайного леса в зависимости от количества признаков, на которых обучается каждое дерево. Перебранные значения подписаны на графике. Максимальная глубина равна 5, количество деревьев в ансамбле - 50.

На Рис.2 приведен график поведения алгоритма в зависимости от размерности подмножества признаков, на которых обучается каждое дерево. Время также растет линейно. RMSE сначала падает, а ближе к размерности общего признакового пространства начинает увеличиваться, так как деревья внутри леса становятся скоррелированными из-за обучения на одинаковом признаковом пространстве, из-за чего растет разброс.

Рис.3 показывает, что с увеличением глубины уменьшается ошибка алгоритма. Деревья внутри могут быть переобученными, но засчет ансамблирования уменьшается разброс конечного ответа. Если убрать ограничение на глубину дерева, время обучения резко возрастает, до этого рост имеет линейный характер.

Рис. 3: Показатели обучения случайного леса в зависимости от максимальной глубины деревьев. Перебранные значения подписаны на графике. Количество деревьев в ансамбле - 50, размерность подвыборки признаков каждого дерева равна $\lfloor \frac{n}{3} \rfloor$

2.3 Градиентный бустинг

В этом эксперименте исследуем поведение градиентного бустинга в зависимости от таких параметров, как количество деревьев в ансамбле, глубина деревьев, количество признаков для обучения каждого отдельного дерева, темп обучения. На Рис.4 показаны графики поведения

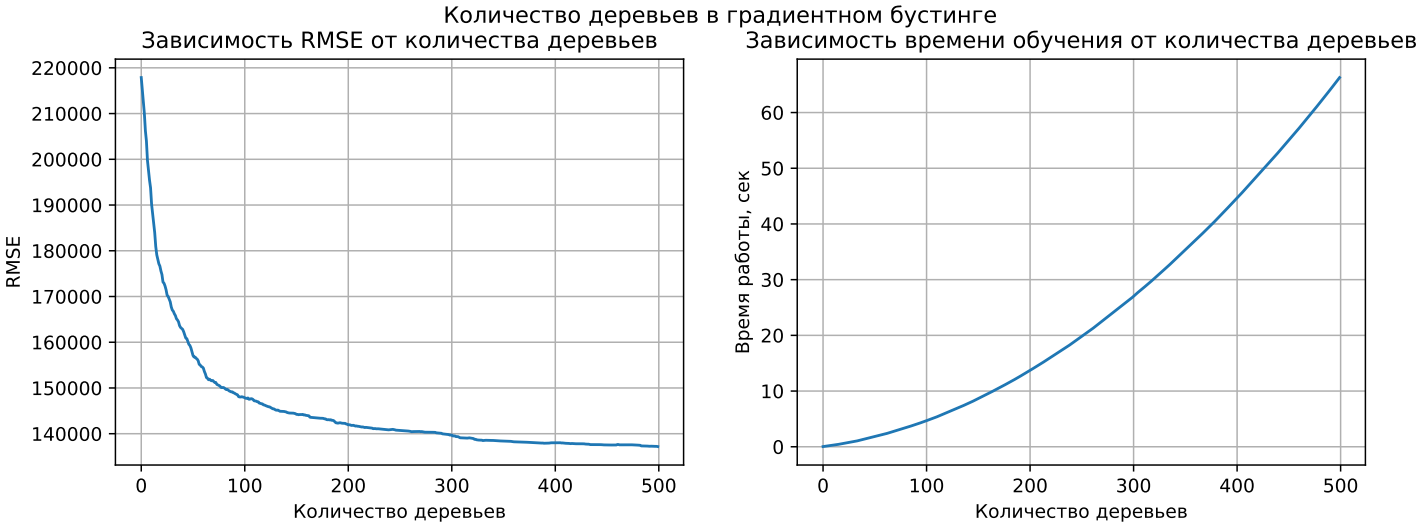

Рис. 4: Показатели обучения градиентного бустинга в зависимости от количества деревьев. Были перебраны значения от 1 до 500. Максимальная глубина равна 5, количество признаков $\lfloor \frac{n}{3} \rfloor$, темп обучения - 0.1.

градиентного бустинга при различном количестве деревьев. Глубина и темп обучения взяты по умолчанию - 5 и 0.1 соответственно. Как видно на графике, ошибка с увеличением деревьев уменьшается, переобучения не наблюдается. Однако если увеличить темп обучения(Рис.5), эффект переобучения появляется при большом количестве деревьев, ошибка растет. Рост времени здесь отличается от линейного: при небольшом количестве деревьев рост плавный, при большой скорость роста больше.

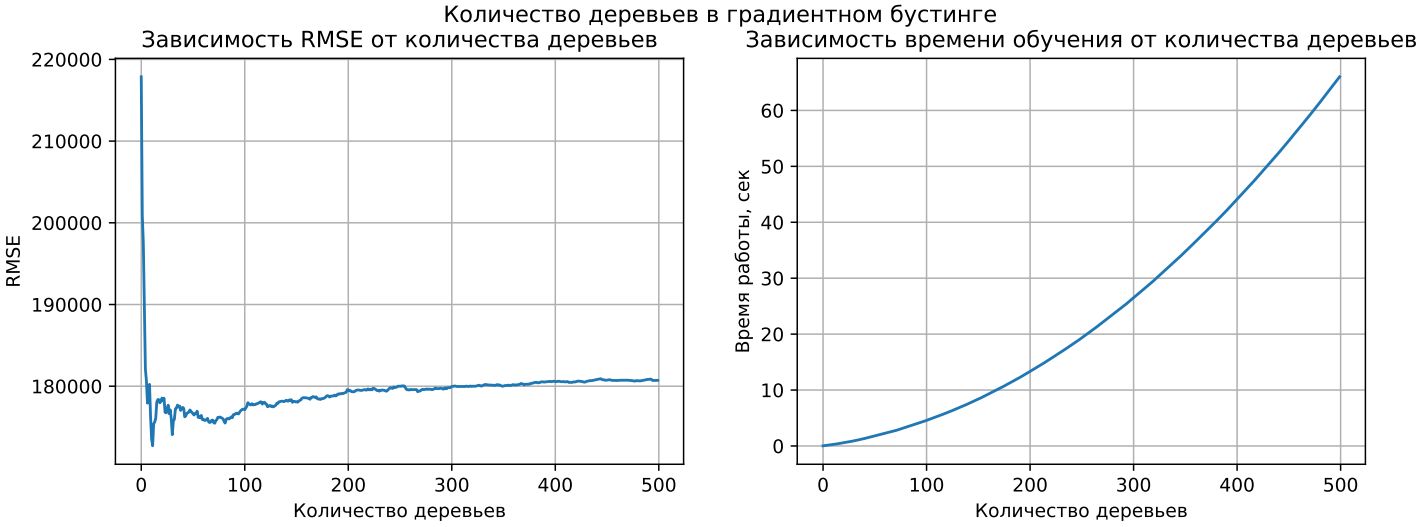

Рис. 5: Показатели обучения градиентного бустинга в зависимости от количества деревьев. Были перебраны значения от 1 до 500. Максимальная глубина равна 5, количество признаков $\lfloor \frac{n}{3} \rfloor$, темп обучения - 1.

На Рис.6 изображены графики поведения бустинга при разных размерностях подвыборки признаков. Зависимость не является какой-то явно выраженной, кажется что она случайная, но можно определенно сказать, что, как и в случайном лесе, при полном признаковом пространстве падает качество, что является следствием коррелированности отдельных деревьев. Рост времени имеет линейный характер.

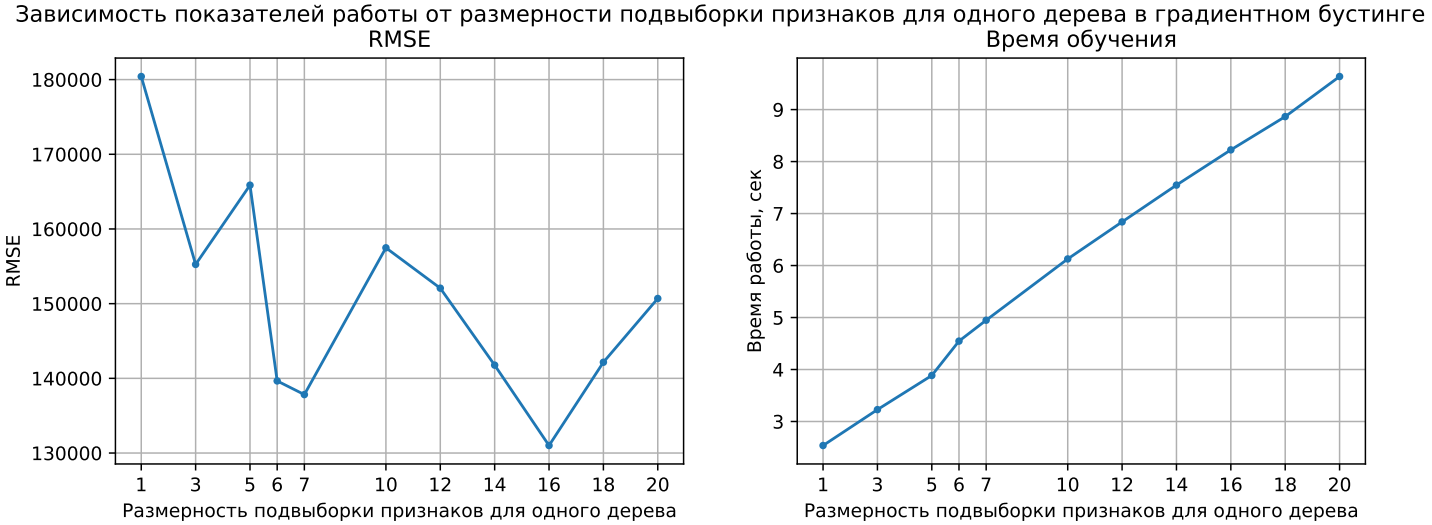

Рис. 6: Показатели обучения градиентного бустинга в зависимости от количества признаков, на которых обучается каждое дерево. Перебранные значения подписаны на графике. Количество деревьев в ансамбле - 100, максимальная глубина равна 5, темп обучения - 0.1.

Из Рис.7 видно, что время обучения при увеличении глубины ведет себя примерно как и в случае случайного леса: при снятии ограничения на глубину оно резко возрастает. А вот RMSE здесь ведет себя по-другому: при глубине, большей 5, показатель растет, что говорит о том, что модель больше подстраивается под обучающую выборку и переобучается.

Рис. 7: Показатели обучения градиентного бустинга в зависимости от максимальной глубины деревьев. Перебранные значения подписаны на графике. Количество деревьев в ансамбле - 100, количество признаков $\lfloor \frac{n}{3} \rfloor$, темп обучения - 0.1.

Последний исследуемый параметр - темп обучения. На Рис.8 видно, что оптимальное значения порядка 0.1. При значениях ближе к 1 модель работает хуже, но еще работает. При значении 2 она показывает очень большой RMSE, алгоритм не сошелся. При маленьких значениях алгоритм дает не такое хорошее качество, возможно требуется больше деревьев, чтобы алгоритм дообучился. Добавим деревьев в ансамбль - Рис.9. Здесь количество деревьев равно 500, видно, что ошибка на маленьких значениях темпа обучения стала меньше, оптимальным значением является $\alpha = 0.05$. Время обучения модели не имеет явно выраженный характер.

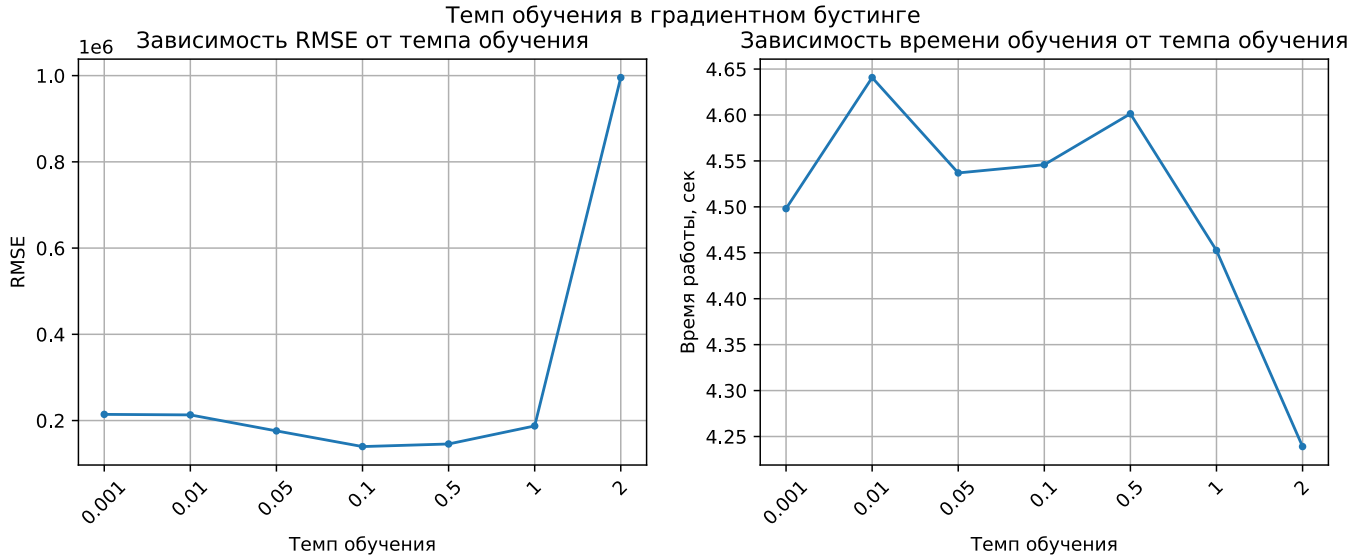

Рис. 8: Показатели обучения градиентного бустинга в зависимости от темпа обучения. Перебранные значения подписаны на графике. Количество деревьев в ансамбле - 100, количество признаков $\lfloor \frac{n}{3} \rfloor$, максимальная глубина - 5.

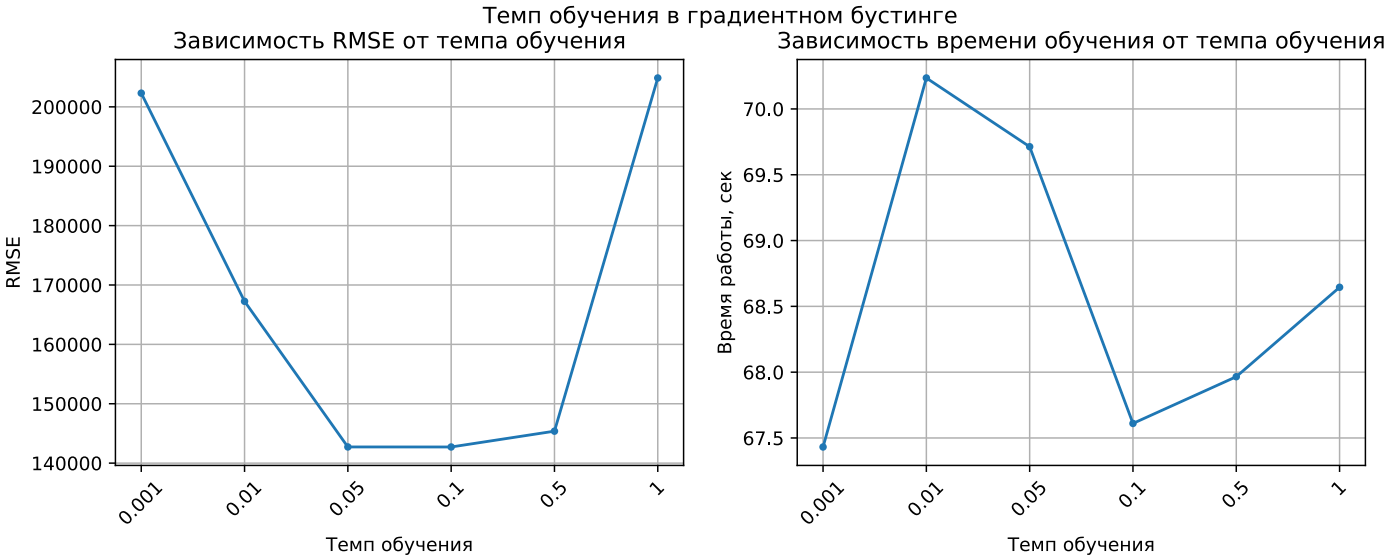

Рис. 9: Показатели обучения градиентного бустинга в зависимости от темпа обучения. Перебранные значения подписаны на графике. Количество деревьев в ансамбле - 500, количество признаков $\lfloor \frac{n}{3} \rfloor$, максимальная глубина - 5.

2.4 Вывод

Было рассмотрено два ансамблевых алгоритма над решающими деревьями - случайный лес и градиентный бустинг. Отличие эти алгоритмов в том, что случайный лес строит компоненты паралелльно, а бустинг - последовательно, исправляя ошибки уже построенной модели. Из-за этого наблюдается чувствительность бустинга к выбросам и склонность к переобучению. Случайный лес же усредняет ответы компонент, из-за чего разброс становится меньше. Также важным аспектом является размерность подвыборки признаков, для случайного леса важно, чтобы компоненты были некоррелированы, поэтому важно выбрать не очень большое значение этого параметра.

Литература

[1] Воронцов К.В. Линейные ансабли. -
http://www.machinelearning.ru/wiki/images/3/3a/Voron-ML-Compositions1-slides.pdf. - 2021.

[2] Воронцов К.В. Продвинутые методы ансамблирования -
http://www.machinelearning.ru/wiki/images/2/21/Voron-ML-Compositions-slides2.pdf. - 2021.