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

Практикум 317 группы, ММП ВМК МГУ.

Алексеев Илья Алексеевич.

Декабрь 2022.

Содержание

1 Введение.

2 Эксперименты.

2.1 Предобработка.

Из датасета убран признак `id`, принимающий значения типа `7129300520` или `2487200875`. Признак `date` формата `%Y%m%dT000000` заменён на три признака `year`, `month`, `day`. Затем признак `year` заменён на `year_2014` и `year_2015` с помощью OneHot кодирования. Признаки `month` и `day` преобразованы в `month_sin`, `month_cos`, `day_sin`, `day_cos` как циклические признаки. Весь датасет разделен на три выборки: обучающую, валидационную и тестовую (в соотношении 3:1:1).

2.2 Случайный лес.

Для десяти вещественных значений параметра `feature_subsample_size` в отрезке [0.001, 1] была обучена модель случайного леса со 100 деревьями с максимальной глубиной 40. Эксперимент был повторён три раза и результаты метрики RMSE были усреднены (рис. 1). Ошибка предсказания падает с ростом данного параметра вплоть до значения ≈ 0.667, затем начинает расти. Значит, малое число признаков не дает достаточно информации для точного предсказания, а слишком большое число делает деревья коррелированными, что повышает variance в формуле (????). Заметим, что разброс результатов трех экспериментов очень мал (error bar узкий, почти не заметен).

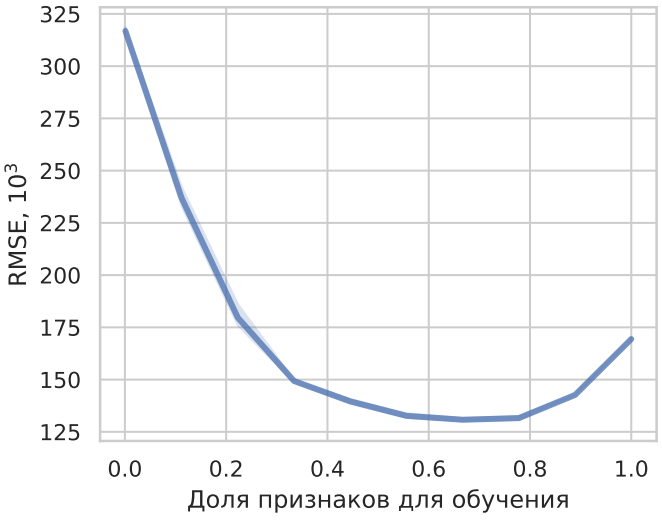
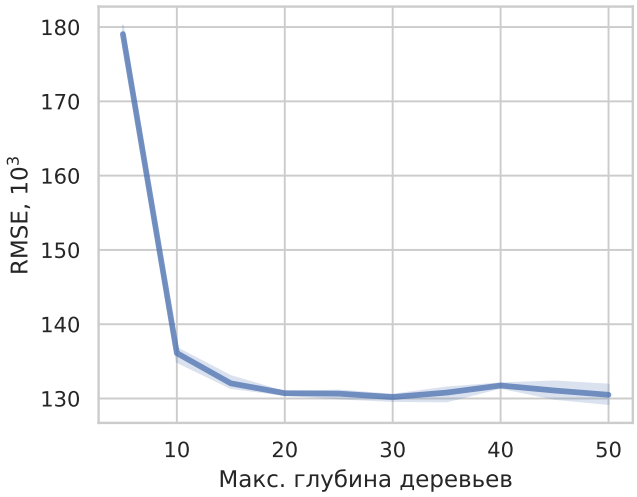


Рис. 1: Значения RMSE в зависимости от доли признаков, выбираемых для обучения деревьев.

Для значений параметра `max_depth` от 5 до 50 включительно с шагом 5 была обучена модель случайного леса со 100 деревьями и долей признаков, выбранной в прошлом эксперименте: 0.667. Эксперимент был повторён три раза и результаты метрики RMSE были усреднены (рис. 2a). Ошибка предсказания резко падает при изменении глубины с 5 до 10, затем продолжает медленно падать вплоть до глубины 30, затем выходит на плато. Причем на плато появляются различия (разброс на графике) для трех попыток эксперимента. Значит, при малой глубине деревьев модель слишком проста и дает слишком неточные предсказания, а при большой глубине возникает некоторая неустойчивость из-за чрезмерной сложности модели.

Для значений параметра `n_estimators` от 20 до 300 включительно с шагом 20 была обучена модель случайного леса с макс. глубиной деревьев и долей признаков, выбранными в предыдущих экспериментах: 30 и 0.667 соответственно. Эксперимент был повторён шесть раз и результаты метрики RMSE были усреднены (рис. 2b). На графике виден тренд на понижение ошибки RMSE с ростом числа деревьев.

(a) глубины деревьев



(b) числа деревьев

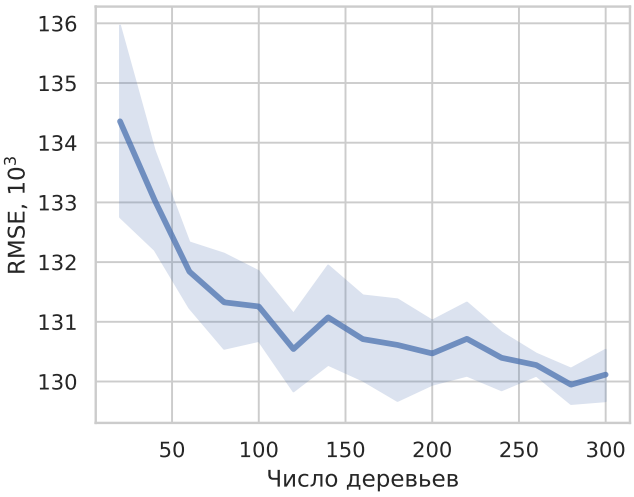


Рис. 2: Значение RMSE в зависимости от:

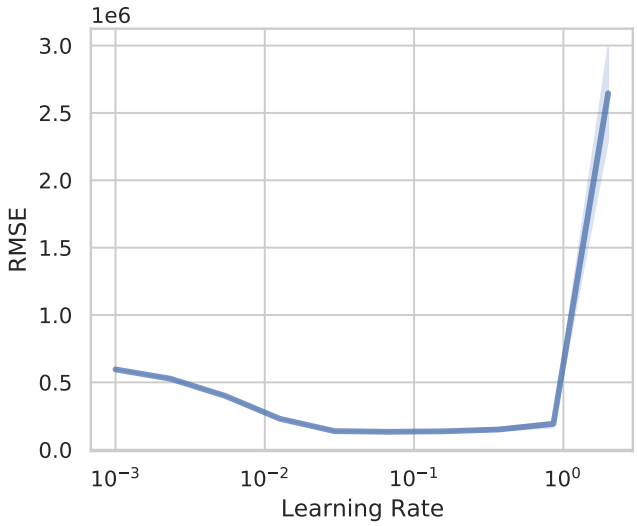
2.3 Градиентный бустинг.

Для десяти вещественных значений параметра `learning rate` от 0.001 до 2 включительно на логарифмической шкале была обучена модель градиентного бустинга со 100 деревьями с максимальной глубиной 40. Параметр `feature_subsample_size` фиксирован и равен 0.667. Эксперимент был повторён три раза и результаты метрики RMSE были усреднены (рис. 3a). При значении параметра больше 1 ошибка резко увеличивается, как и при размере шага сильно меньше 10^{-1}. Плато с адекватными значениями – это значения порядка 10^{-1}. Наименьшая ошибка достигнута при `learning_rate`, равном ≈ 0.158.

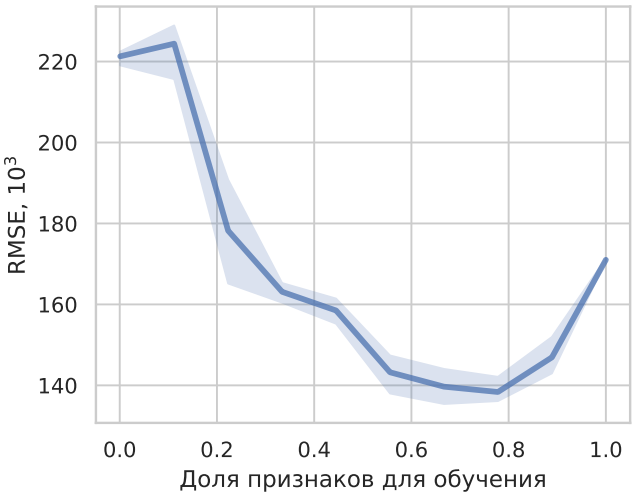
Для десяти вещественных значений параметра `feature_subsample_size` из отрезка $[0.001, 1]$ была обучена модель градиентного бустинга со 100 деревьями максимальной глубины 40. Параметр `feature_subsample_size` фиксирован и равен 0.667. Эксперимент был повторён три раза и результаты метрики RMSE были усреднены (рис. 3b). Ситуация аналогична случаю случайного леса, разве что разброс результатов разных попыток больше. Значение, при котором достигается минимум, равно 0.667.

Для всех значений параметра `max_depth` из отрезка $[1, 10]$ и для значений от 10 до 50 с шагом 10 была обучена модель градиентного бустинга со 100 деревьями и долей признаков, равной 0.667. Эксперимент был повторён три раза и результаты метрики RMSE были усреднены (рис. 3c). Лучшее качество достигается при глубинах, равных 5 и 8. Далее с увеличением глубины ошибка увеличивается. Значит, градиентный бустинг плохо работает в случае, когда базовые модели излишне усложнены.

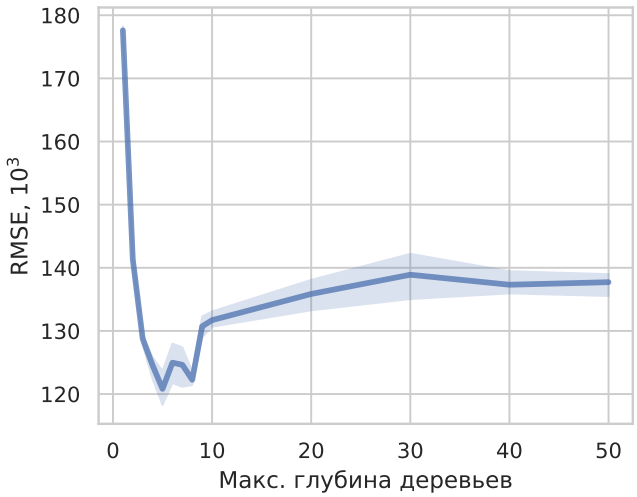
(a) шага обучения (`learning rate`)



(b) доли признаков, выбираемых для обучения деревьев



(c) глубины деревьев



(d) числа деревьев

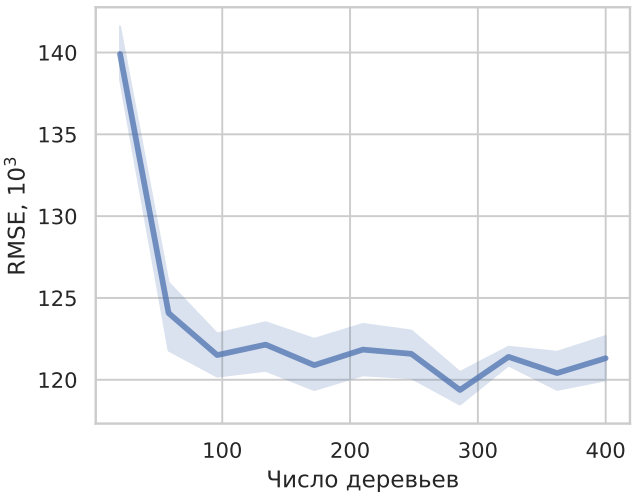


Рис. 3: Значение RMSE в зависимости от: