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

Отчет о выполнении практической работы «Градиентные методы обучения линейных моделей. Применение линейных моделей для определения токсичности комментария»

Петрова Александра,
317 группа

Москва
2023

Оглавление

Введение 2
Предобработка данных 2
Случайный лес 2
- Размерность подвыборки признаков 2
- Максимальная глубина дерева 2
- Количество деревьев 2

Градиентный бустинг 3
- Скорость обучения 3
- Размерность подвыборки признаков 4
- Максимальная глубина дерева 4
- Количество деревьев 5

Заключение 5

Введение

Данное практическое задание посвящено исследованию ансамблей алгоритмов на примере использования случайного леса и градиентного бустинга для определения цены продажи дома.

Цель исследования - выявить зависимость ошибки и времени работы алгоритмов от гиперпараметров, выбрать лучший алгоритм для предсказания цены на жилье.

Предобработка данных

Данные были скачаны по ссылке. Датасет содержит цены продажи домов в округе Кинг, Вашингтон, в период с мая 2014 по май 2015 года, а также информацию о доме и времени покупки. Таблица состоит из 20 столбцов. Столбец price содержит целевую переменную - цену. Он был сохранен в отдельную переменную и удален из таблицы. Столбец date, в котором указана дата покупки дома, был преобразован в 5 столбцов: year, month, dayofyear, dayofmonth и dayofweek. Исходный столбец был удален. Данные были преобразованы в numpy.ndarray с вещественным типом данных. В итоге, матрица признаков содержала 21613 строк и 23 столбцов.

Выборка была разделена на обучающую и тестовую в соотношении 8:2. Обучающая выборка делилась на обучающую и валидационную (по ней подбирались лучшие параметры) подвыборки в соотношении 8:2.

Случайный лес

Исследовалась зависимость RMSE и времени работы алгоритма в зависимости от следующих параметров:

1. Размерности подвыборки признаков для одного деревьев: `feature_subsample_size`.
2. Максимальной глубины дерева: `max_depth`. Также разбирался случай, когда глубина дерева не ограничена.
3. Количества деревьев в ансамбле: `n_estimators`.

Результаты эксперимента приведены на Рис.1 - Рис.3.

Размерность подвыборки признаков

Параметр `feature_subsample_size` определяет максимальное количество признаков, которые могут быть учтены при выборе лучшего разделения в узле каждого дерева.

На Рис.1 показана зависимость RMSE и времени от размерности подвыборки признаков. График ошибки напоминает параболу, минимум достигается при `feature_subsample_size` = 15. Зависимость времени работы алгоритма от размерности подвыборки признаков линейная, что согласуется с теорией.

Максимальная глубина дерева

Из Рис.2 видно, что RMSE экспоненциально убывает при увеличении `max_depth`. При `max_depth` = 19 ошибка минимальна. Время возрастает линейно.

Количество деревьев

Рис.3 показывает, что RMSE убывает экспоненциально с ростом числа деревьев, время возрастает линейно, это связано с тем, что каждое дерево обучается независимо от других. Разница между ошибками после 300 дерева мала, поэтому это значение можно считать оптимальным.

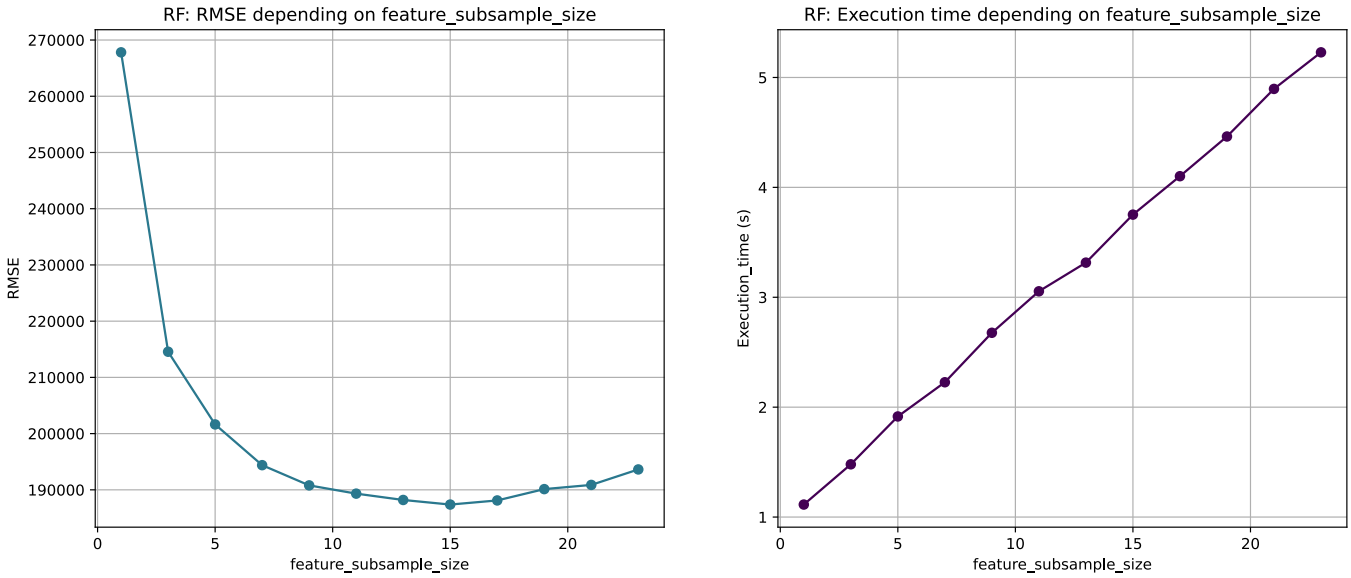

Рис. 1: Графики зависимости RMSE и времени от размерности подвыборки признаков для одного дерева для RF

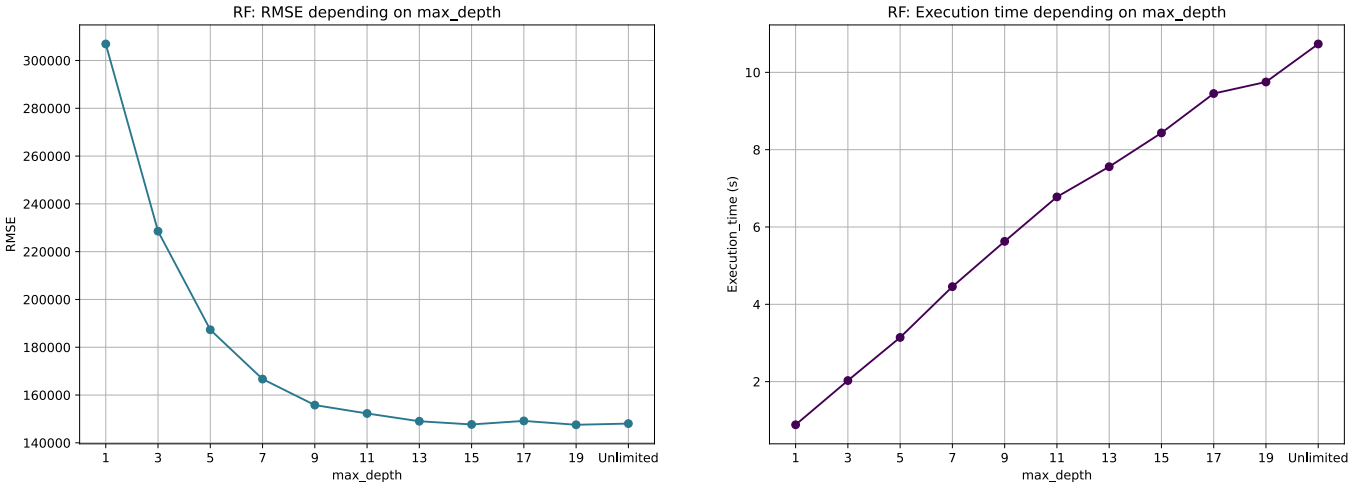

Рис. 2: Графики зависимости RMSE и времени от максимальной глубины дерева для RF

Градиентный бустинг

Исследовалась зависимость RMSE и времени работы алгоритма в зависимости от следующих параметров:

1. Скорости обучения: `learning_rate`.
2. Размерности подвыборки признаков для одного деревьев: `feature_subsample_size`.
3. Максимальной глубины дерева: `max_depth`.
4. Количества деревьев в ансамбле: `n_estimators`.

Результаты эксперимента приведены на Рис.4 - Рис.7.

Скорость обучения

Из графика на Рис.4 видно, что ошибка резко убывает в промежутке (0.001, 0.1). Минимум ошибки достигается при `learning_rate` $= 0.1$. Это может быть связано с тем, что при

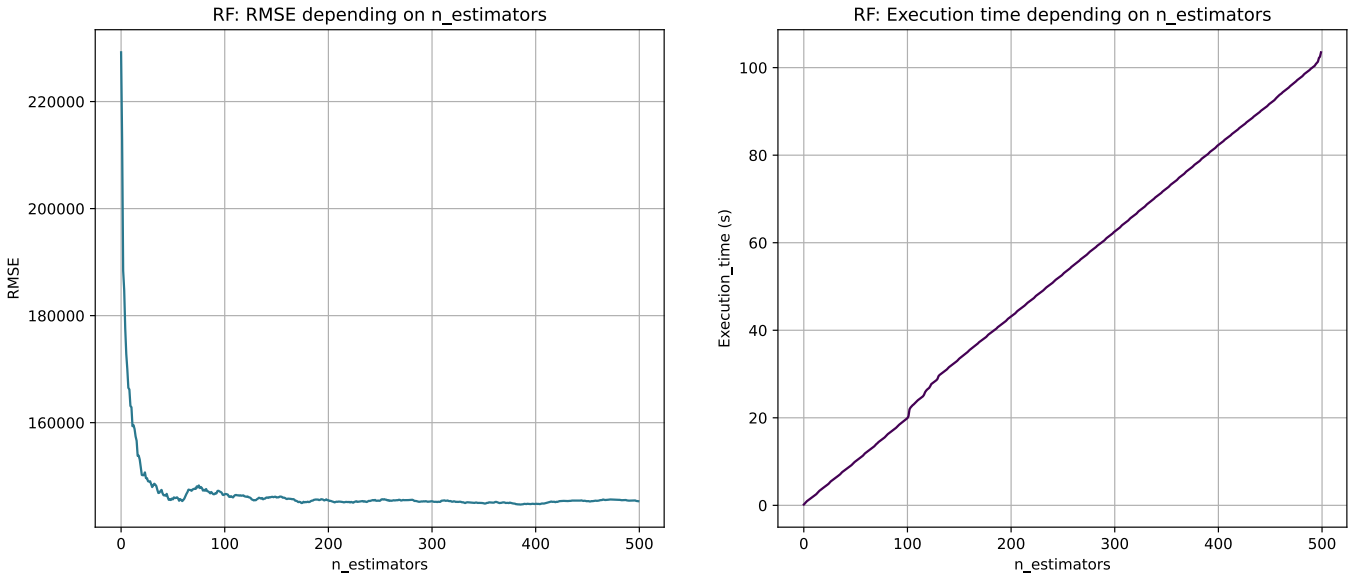

Рис. 3: Графики зависимости RMSE и времени от количества деревьев для RF

маленьких скоростях алгоритм не успевает сходиться, а при больших начинает переобучаться. Характерной зависимости времени от скорости обучения не наблюдается.

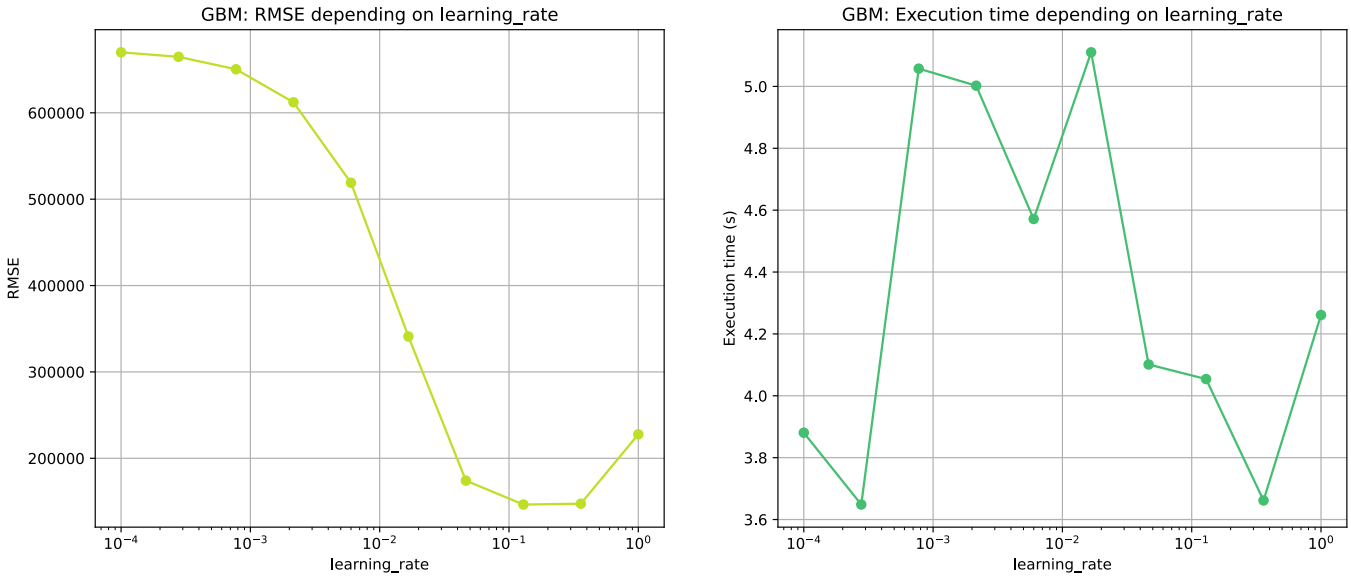

Рис. 4: Графики зависимости RMSE и времени от скорости обучения для GBM

Размерность подвыборки признаков

Из Рис.5 видно, что минимум ошибки достигается при `feature_subsample_size` = 13, при `feature_subsample_size` = 7 происходит скачок времени, далее оно растет линейно.

Максимальная глубина дерева

Из Рис.6 видно, что сначала ошибка резко уменьшается с ростом глубины. Минимум достигается при `max_depth` = 5. Далее ошибка начинает расти. Это связано с тем, что глубокие деревья переобучаются, когда исправляют ошибки друг друга. Поэтому, в градиентном бустинге используют неглубокие деревья. Время возрастает линейно.

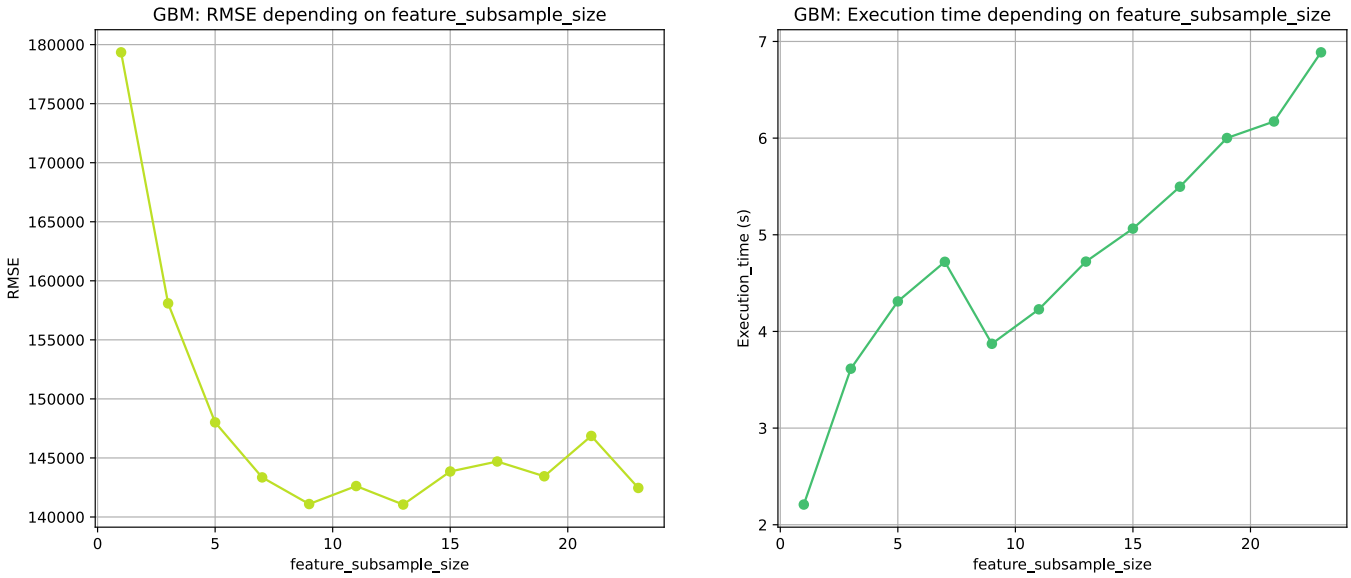

Рис. 5: Графики зависимости RMSE и времени от размерности подвыборки признаков для одного дерева для GBM

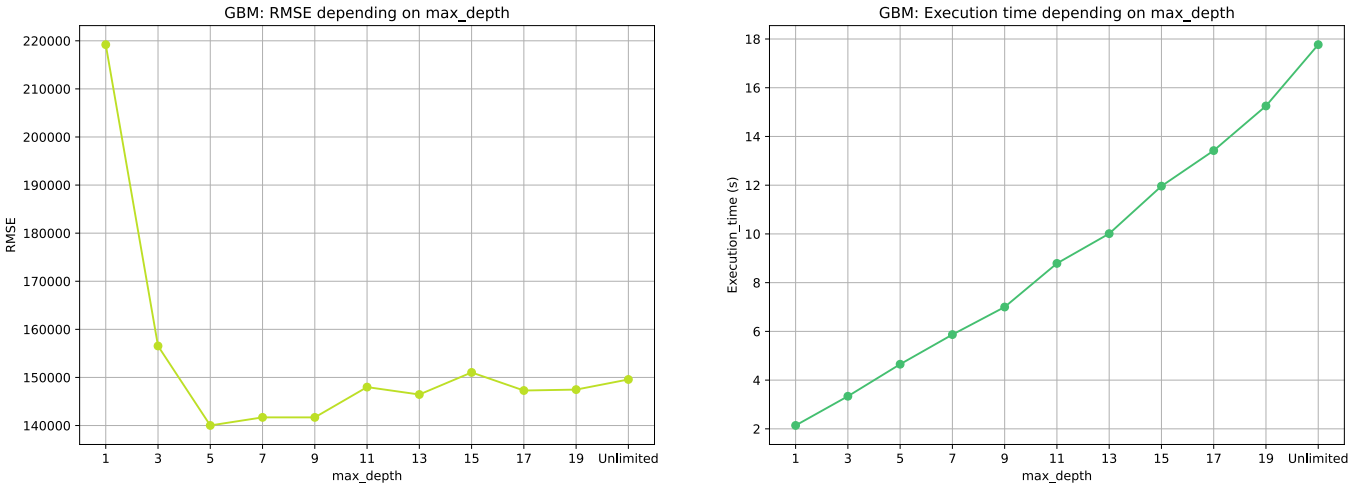

Рис. 6: Графики зависимости RMSE и времени от максимальной глубины дерева для GBM

Количество деревьев

Как видно по графику с Рис.7. RMSE убывает экспоненциально, а время возрастает по параболе, близкой к прямой. Это можно объяснить тем, что каждое новое дерево обучается на ошибках, оставшихся после предыдущих. После добавления дерева нужно пересчитать градиенты и обновить веса объектов. Этот процесс выполняется для каждого нового дерева, что может привести к кумулятивному увеличению времени обучения.

Оптимальное значение `n_estimators` = 300.

Заключение

В ходе работы были исследованы ансамбли алгоритмов на примере прогнозирования цен на жилье. Была выявлена зависимость RMSE и времени работы алгоритмов (случайного леса и градиентного бустинга) от таких гиперпараметров, как размерность пространства подвыборки для каждого дерева, максимальная глубина, количество деревьев, скорость обучения (для градиентного бустинга).

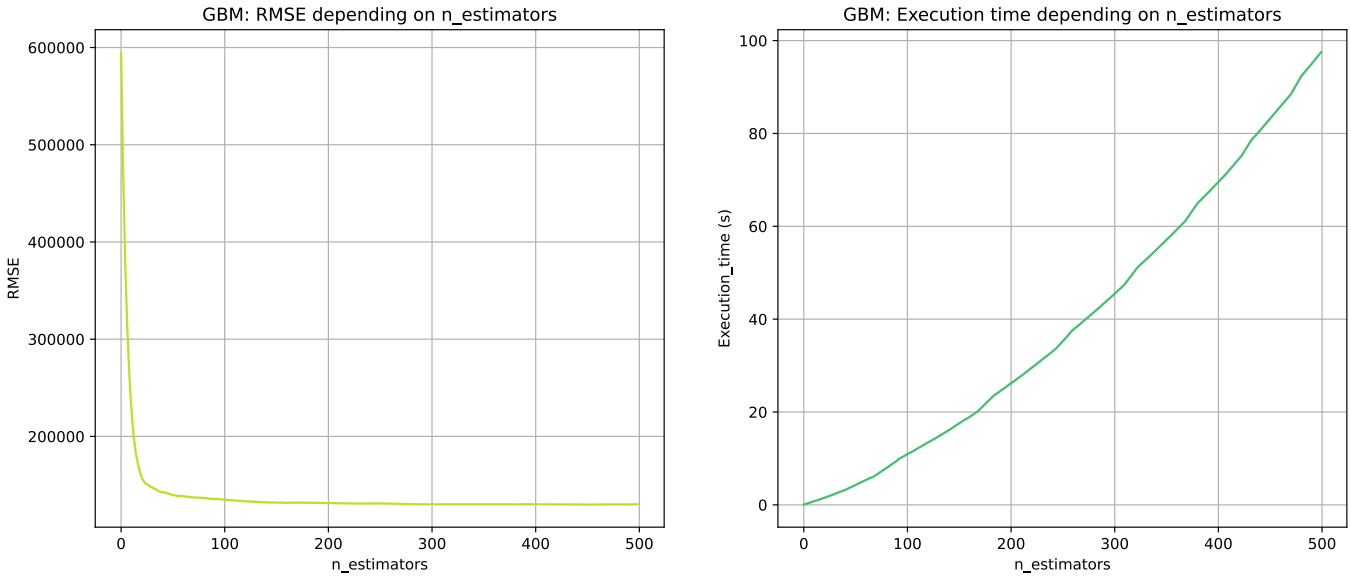

Рис. 7: Графики зависимости RMSE и времени от количества деревьев для GBM

Были выявлены следующие зависимости:

1. Для случайного леса время работы алгоритмов растет линейно с ростом размерности подвыборки. Для градиентного бустинга сначала происходит скачок, затем время растет линейно.

2. Оптимальная глубина деревьев в случайном лесе может быть не ограничена. В градиентном бустинге оптимальная глубина небольшая, иначе деревья переобучаются.

3. Для обоих алгоритмов выполнено: чем больше деревьев, тем меньше ошибка. В случае случайного леса, время обучения растет линейно с ростом числа деревьев в ансамбле. В градиентном бустинге время растет практически линейно.

4. Для градиентного бустинга оптимальная скорость обучения около 0.1.

Для задачи прогнозирования цены на дом лучше (с точки зрения RMSE) оказался градиентный бустинг (RMSE: градиентный бустинг - 133070, случайный лес -145244).