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

КОМПОЗИЦИИ АЛГОРИТМОВ ДЛЯ РЕШЕНИЯ ЗАДАЧИ РЕГРЕССИИ

ЗАДАНИЕ №3

Практикум 317 группы

ОТЧЕТ
о выполненном задании

студента 317 учебной группы факультета ВМК МГУ
Кузнецова Михаила Константиновича

Москва
2020

Введение

В данной работе рассматривается применение методов случайных лес и градиентный бустинг для задачи определения стоимости жилья. Исследуется влияние различных гиперпараметров на поведение моделей.

Эксперименты

Начальная обработка данных

Проведены следующие трансформации датасета:

1. удаление столбцов «index», «id», «date» из матрицы признаков
2. удаление столбца «index» из признаков целевого датасета

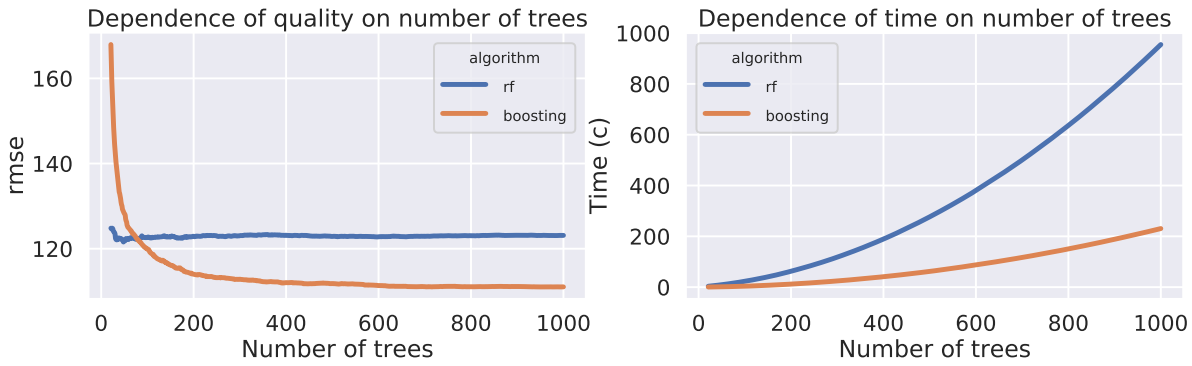
Влияние гиперпараметров

Исследуем поведение случайного леса и градиентный бустинга для задачи регрессии в зависимости от *n_estimators*, *feature_subsample_size*, *max_depth* и *learning_rate* для градиентного бустинга. Выбраны следующие дефолтные значения:

parameter	rf	boosting
n_estimators	75	75
feature_subsample_size	None	None
max_depth	None	5
learning_rate	–	0.1

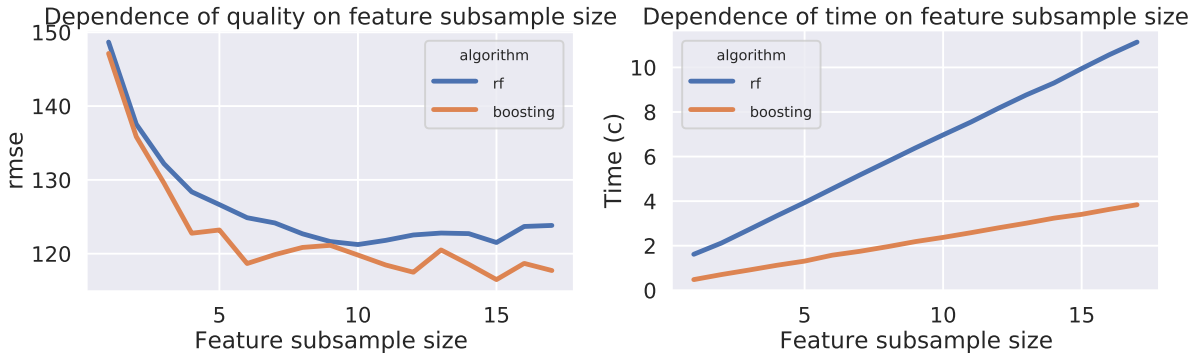
Таблица 1: Значения по умолчанию моделей

n_estimators



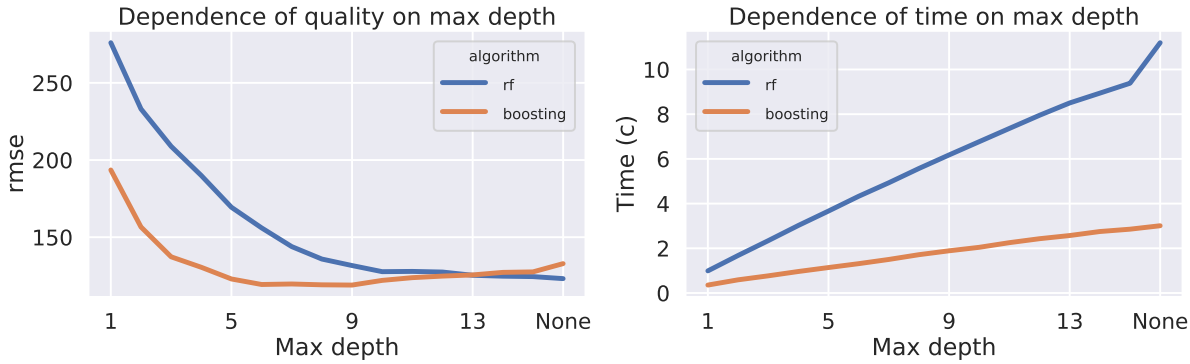
Качество постепенно растёт с увеличением количества деревьев в boosting методе, а в rf достигает максимума при 75 деревьев. Причиной этому может стать разница в значении *max_depth*. Заметим, что время обучения зависит приблизительно линейно от количества деревьев.

feature_subsample_size



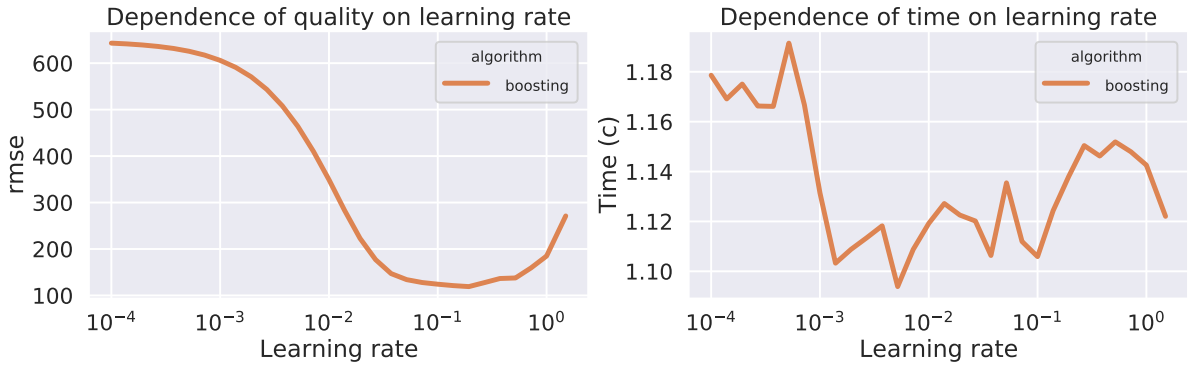
Наблюдается логичное улучшение качества при увеличении данного параметра. Линейная зависимость времени позволяет удобно подобрать параметр в определенных условиях задачи. Так как *max_depth* = None в rf, большой данный параметр не поможет построить различные деревья. В случае с boosting глубина не велика. Это даёт возможность постепенного увеличения качества.

max_depth



Отчётливо видно, что boosting над слишком простыми или сложными моделями даёт плохое качество. Есть линейная зависимость между временем и глубиной дерева.

learning_ rate



Маленький *learning_ rate* — не доходим до оптимума, большой — перескакиваем его. Аналогично с временем: маленький — слабо сдвигаем ошибки предыдущих ответов, следовательно, последующие модели будут сталкиваться с теми же проблемами, большой — приводит к исправлениям ответов с неправильным мультипликатором.

Выводы

В работе представлены экспериментальные результаты, касающиеся задачи регрессии для случайного леса и градиентного бустинга. Проведена сравнительная характеристика. В условиях данного эксперимента можно выделить следующие тезисы:

- при увеличении количества деревьев, с какого то момента rf не улучшает качество, а boosting наоборот
- $\frac{number\ of\ features}{2}$ является оптимальным *feature_ subsample_ size* для rf
- для обоих алгоритмов небольшая глубина является оптимальной
- *learning_ rate* имеет большое значение на сходимость метода