Отчет о провдении экспериментов, нацеленных на изучение поведения случайного леса и градиентного бустинга

Практикум на ЭВМ

Никифорова Мария Дмитриевна,
317 группа

Москва
Декабрь 2022

Содержание

1 Введение

Проводилось исследование поведения случайного леса и градиентного бустинга на датасете данных о продажах недвижимости House Sales in King County, USA. Цель алгоритомов - предсказание цен на недвижимость. Алгоритмы оценивались по времени работы и величине RMSE на тестовой выборке.

2 Предобработка данных

Исходная таблица была представлена в формате .csv. При парсинге данных была выполнена автоматическая конвертация колонки *date* в формат datetime, далее даты были разделены на 3 отдельных составляющих: число, месяц, год, каждая записана в отдельную колонку, исходная колонка *date* удалена. Был удален столбец *id*, отдельно выделена колонка с ответами *taget*. Итоговое количество признаков - 21. Затем данные были переведены в формат numpy.ndarray и разделены на обучающую и тестовую выборки в соотношении 1 к 3.

3 Эксперименты

Таргет лежит в диапазоне [75000, 7700000], его распределение выглядит следующим образом:

Метрика качества - RMSE.

3.1 Исследование поведения случайного леса

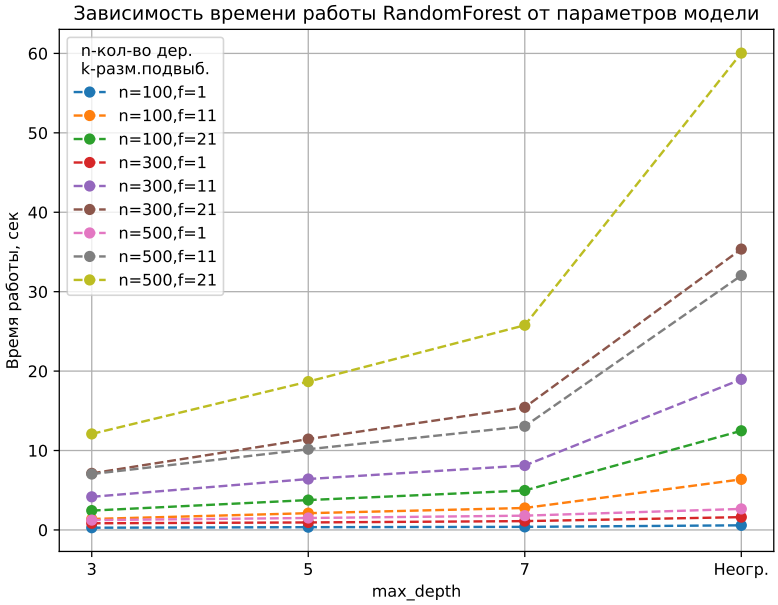
Для проведения экспериментов были выбраны следующие значения параметров:

- Количество деревьев в ансамбле: 100, 300, 500
- Размерность подвыборки признаков для одного дерева: 1, 10, 21

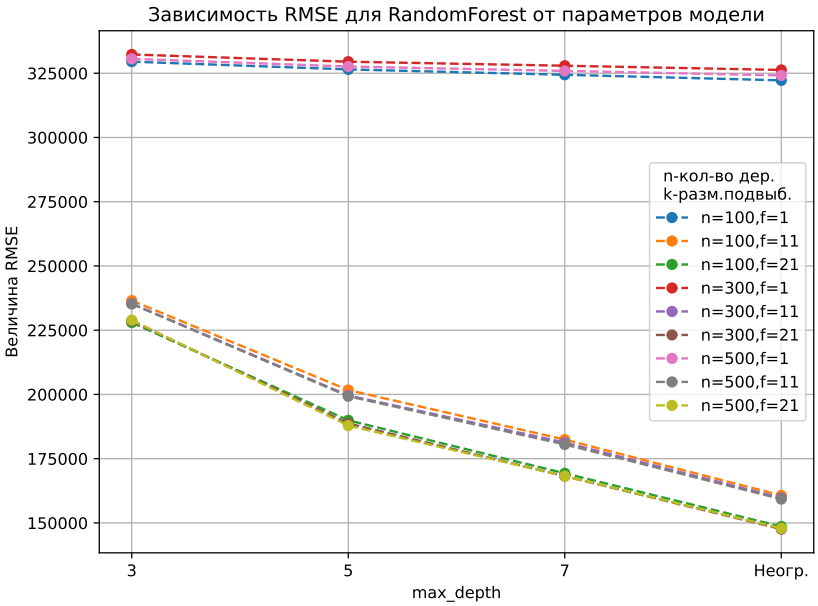
- Максимальная глубина дерева: 3, 5, 7, неограниченная

Был произведен полный перебор всех параметров. Наилучший результат показала модель при параметрах: количестве деревьев - 300, размерность подвыборки для одного дерева - 21, максимальная глубина дерева - неограниченный. RMSE при этом составила 147600, время работы - 35.4 сек.

Зависимость времени работы Random Forest от параметров модели выглядит следующим образом:



Время работы, ожидаемо, увеличивается вместе с увеличением сложности модели.
На графике зависимость RMSE от параметров модели можно увидеть группировку моделей при определенных параметрах:



Как видно, наиболее влиятельными параметрами являются размер подвыборки признаков и максимальная глубина дерева.

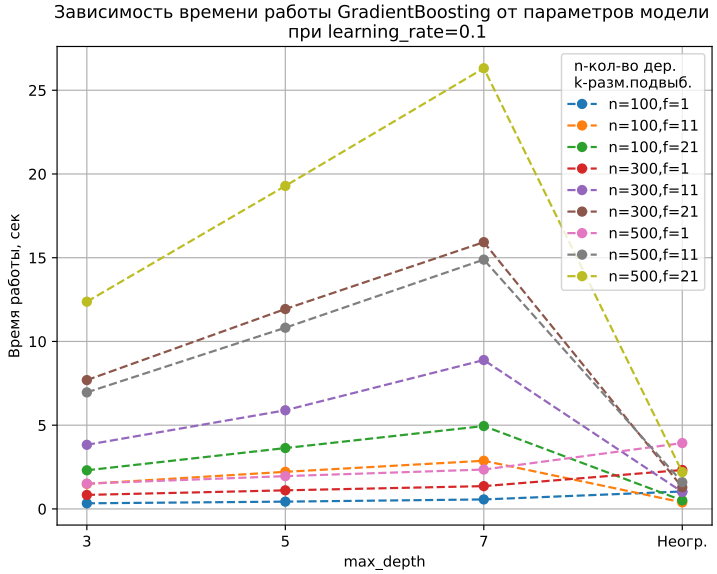
3.2 Исследование поведения градиентного бустинга

Для проведения экспериментов были выбраны следующие значения параметров:

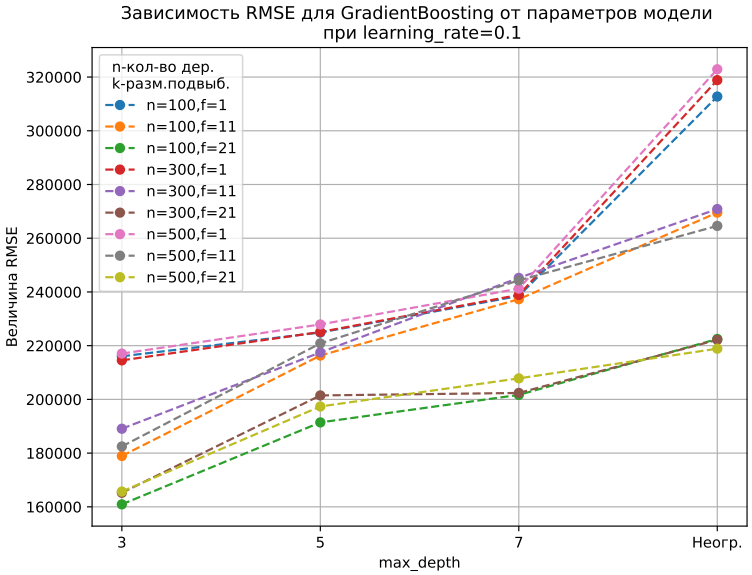
- Количество деревьев в ансамбле: 100, 300, 500
- Размерность подвыборки признаков для одного дерева: 1, 10, 21
- Максимальная глубина дерева: 3, 5, 7, неограниченная
- learning_rate: 0.05, 0.1, 0.2

Лучший результат показала модель при следующих параметрах: количество деревьев в ансамбле - 100, размерность подвыборки признаков для одного дерева - 21, максимальная глубина дерева - 3, learning_rate - 0.2. RMSE при таких параметрах составила 155558, время обучения - 2.3 сек.

Зависимость времени работы от параметров выглядит следующим образом:



При неограниченной глубине деревьев время работы уменьшается;



Таким образом, неглубокие деревья оказываются эффективнее.

4 Сравнительный анализ и выводы

Учитывая масштаб таргета, оба алгоритма показали свою применимость к задаче предсказания цен на недвижимость. Градиентный бустинг работает быстрее по сравнению со случайным лесом, однако его качество несколько хуже.

Тот факт, что обе модели лучше предсказывают данные в случае, когда используются все признаки, а не какое-либо их подмножество, скорее всего объясняется тем, что в нашей задаче признаков сравнительно мало(21).