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

Практикум 317 группы, ММП ВМК МГУ.

Аристеев Олег Алексеевич.

Декабрь 2024.

Содержание

1 Вступление

Ансамбли алгоритмов широко применяются в задачах машинного обучения благодаря своей способности повышать точность прогнозов и устойчивость к переобучению. В данном отчете представлены результаты практического задания, посвященного использованию ансамблевых методов для решения задачи регрессии — прогнозирования цен на недвижимость в King County, USA.

Целью работы также было сравнение двух популярных ансамблевых методов: случайного леса и градиентного бустинга. Были выполнены следующие шаги:

- Исходный датасет был подвергнут предварительной обработке, включая преобразование категориальных признаков, создание новых признаков на основе географических данных и анализ корреляции признаков с целевой переменной.
- Были реализованы и настроены модели случайного леса и градиентного бустинга, с целью определения оптимальных гиперпараметров для каждого метода.
- Результаты работы моделей были сравнены по метрике RMSE (Root Mean Squared Error) на тестовой выборке, что позволило выявить наиболее эффективный метод для данной задачи.

2 Экспериментальная часть

В экспериментальной части исследования алгоритмы случайного леса и градиентного бустинга были реализованы на основе классического алгоритма дерева решений из библиотеки `sklearn.tree.DecisionTreeReg` Реализации классов GradientBoostingMSE и RandomForestMSE были разработаны с нуля, что позволило глубже понять принципы работы ансамблевых методов и их внутреннюю механику.

Исходный код реализаций доступен на GitHub, где представлены комментарии к каждому этапу реализации.

2.1 Предобработка данных

Исходный датасет содержит данные о продажах недвижимости в King County, USA. Всего объектов в выборке – 21 613. Целевой переменной является цена недвижимости. Каждый объект описывается 20-ю признаками (не считая таргет), среди которых: дата продажи, количество ванных комнат и спален, внутренняя площадь квартиры, площадь земельного участка и другие. Данные были разделены на обучающую и тестовую в отношении 8:2.

В признаковом описании объекта есть долгота и широта. Было принято решение выбрать условный «центр» города, где находится самая дорогая недвижимость, и для каждого объекта рассчитать расстояние до этого центра – `distance_to_center`.

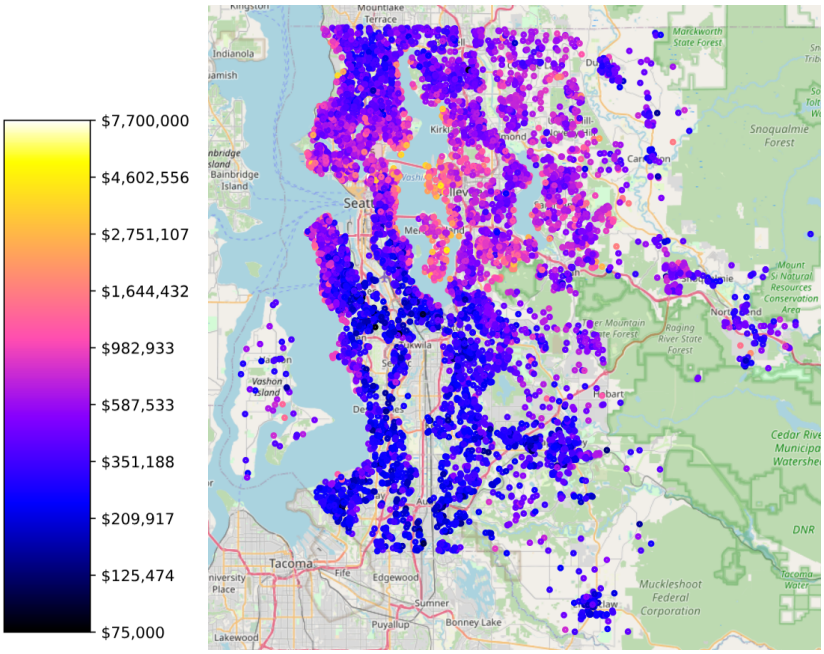

Рис. 1: Зависимость цены недвижимости от расположения на карте.

Итого была получена следующая корреляционная матрица:

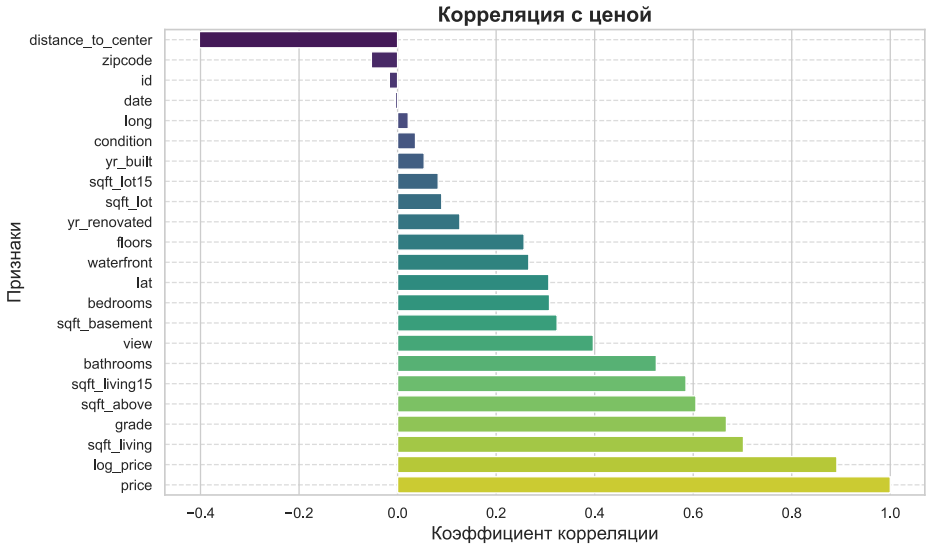

Рис. 2: Корреляция признаков с таргетом

Во избежание сильного увеличения признакового описания объекта (например, `zipcode` содержит 70 уникальных значений), все категориальные признаки были закодированы средним значением таргета.

2.2 Случайный лес

В этом пункте были проведены эксперименты с моделью случайного леса для определения оптимальных гиперпараметров жадным алгоритмом.

Количество деревьев в ансамбле

Были рассмотрены разные количества деревьев в ансамбле из множества $\{1, 5, 10, 20, 50, 100, 200\}$.

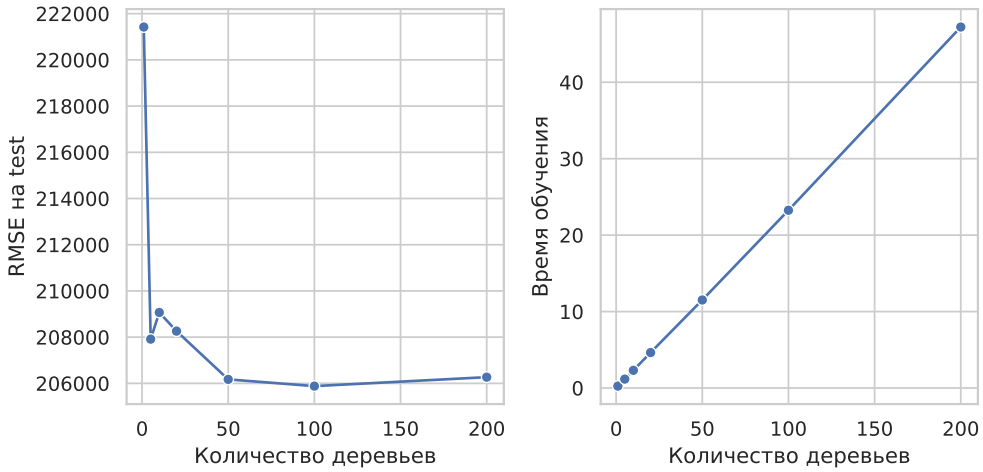

Рис. 3: Зависимость RMSE и времени от количества деревьев в RandomForest

Наиболее оптимальным количеством оказалось 100. Видно, что до этого значение RMSE на тесте уменьшается, а после увеличивается, что говорит о переобучении. Время, очевидно, будет иметь линейную зависимость.

Размерность подвыборки признаков для одной вершины дерева

Были рассмотрены разные значения размерности подвыборки. По горизонтальной оси отмечены числа, соответствующие доле используемых признаков от всех.

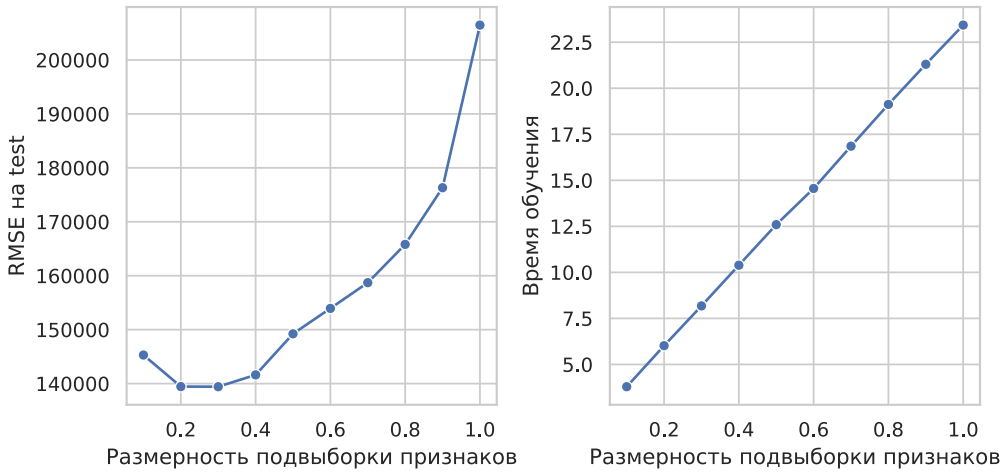

Рис. 4: Зависимость RMSE и времени от размерности подвыборки признаков для одной вершины дерева в RandomForest

Наиболее оптимальным оказалось значение 0.2, то есть в 20% от всех признаков.

Максимальная глубина дерева

Кроме того, были рассмотрены разные максимальные значения глубины.

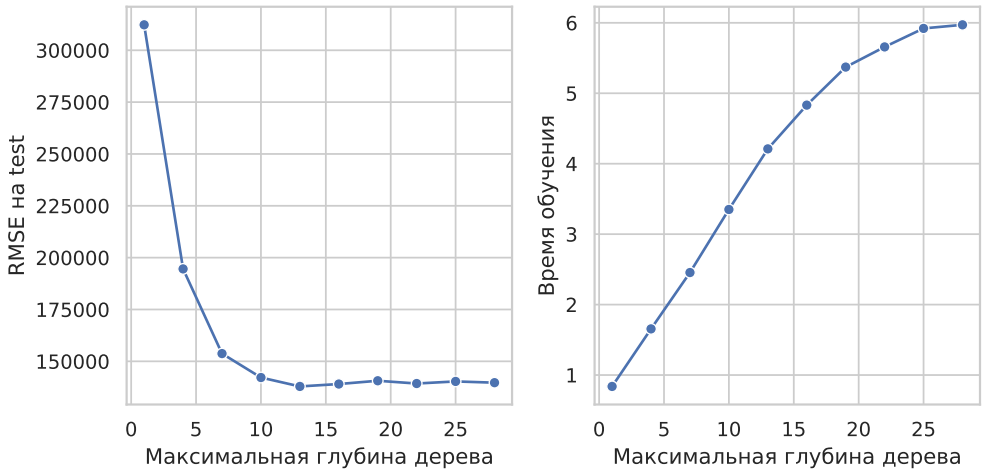

Рис. 5: Зависимость RMSE и времени от максимальной глубины дерева в RandomForest

Наиболее оптимальным значением получилось 13, при таком значении RMSE=137889.66. При неограниченной глубине деревьев RMSE = 140399.78.

2.3 Градиентный бустинг

Аналогичные эксперименты были проведены и с градиентным бустингом.

Количество деревьев в ансамбле

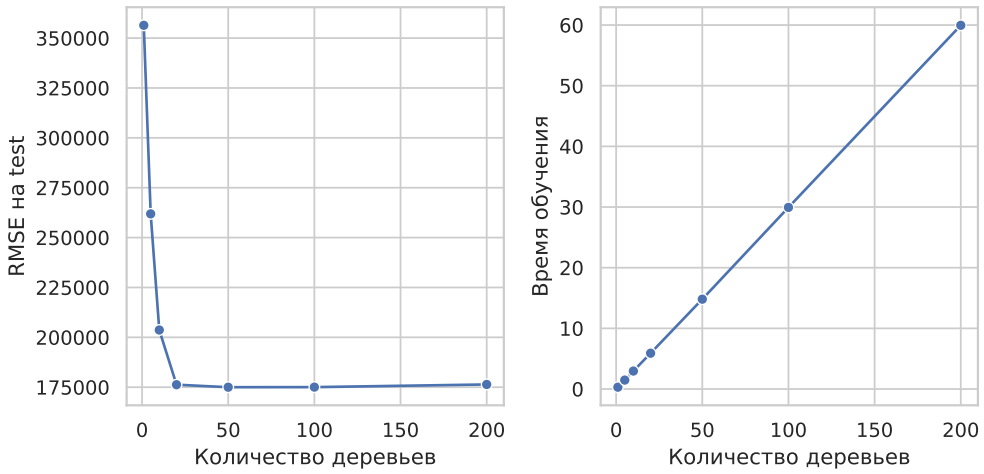

Рис. 6: Зависимость RMSE и времени от количества деревьев в GradientBoosting

Наиболее оптимальное количество деревьев в ансамбле – 50.

Размерность подвыборки признаков для одной вершины дерева

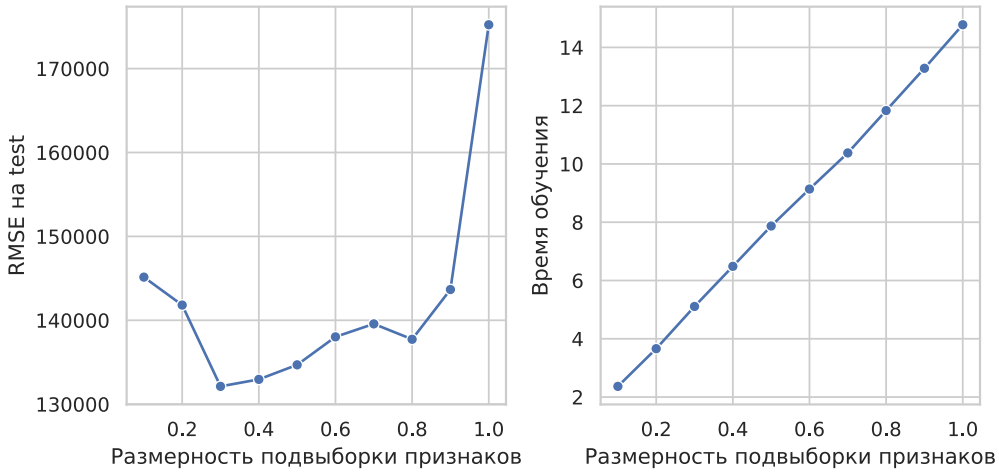

Рис. 7: Зависимость RMSE и времени от размерности подвыборки признаков для одной вершины дерева в GradientBoosting

Наиболее оптимальное – 0.3, то есть в каждой новой вершине дерева рассматривается только 30% от всех признаков.

Максимальная глубина дерева

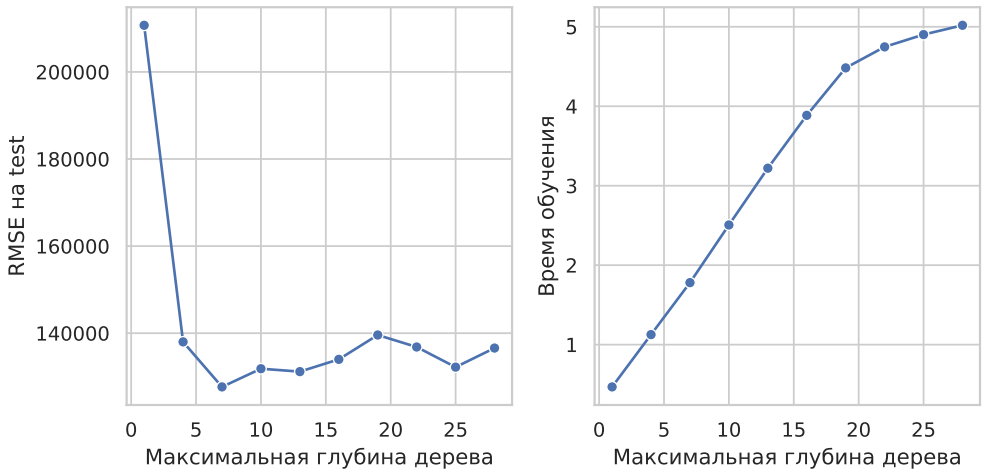

Рис. 8: Зависимость RMSE и времени от максимальной глубины дерева в GradientBoosting

Наиболее оптимальное – 7, при нём RMSE = 127659.01. При более глубоких деревьях наблюдается переобучение, в частности, при неограниченной глубине дерева RMSE=134833.11.

Выбранный `learning_rate`

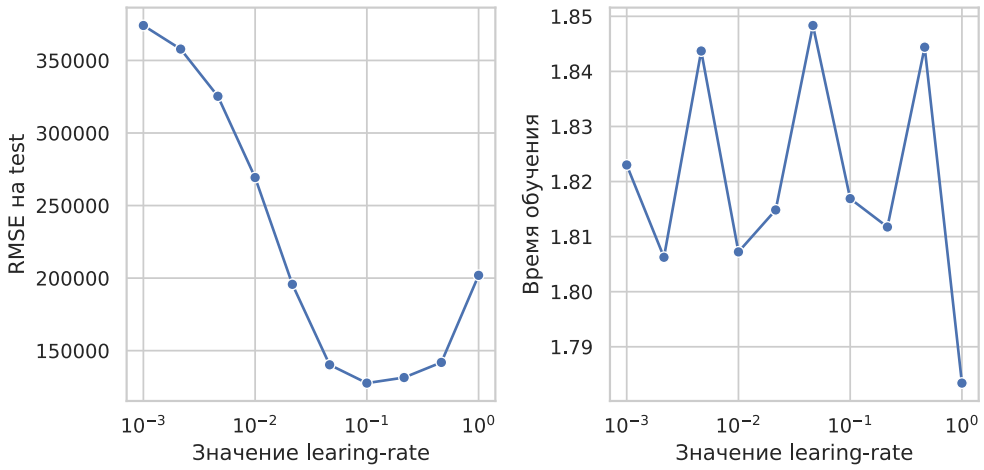

Рис. 9: Зависимость RMSE и времени от learining rate в GradientBoosting

Наиболее оптимальное – 0.1. При таком значении RMSE= 127649.33.

3 Выводы

Оба метода — случайный лес и градиентный бустинг — показали свою эффективность в решении задачи регрессии. Однако градиентный бустинг оказался более точным, с RMSE 127649.33 по сравнению с 137889.66 у случайного леса.

Алгоритм случайныого леса более интерпретируем, так как каждое дерево в ансамбле можно анализировать отдельно. Градиентный бустинг, напротив, менее интерпретируем, так как деревья строятся последовательно и зависят друг от друга

Кроме того, градиентный бустинг требует значительно большего времени для обучения по сравнению с параллельной реализацией случайного леса. Это связано с тем, что каждое дерево в градиентном бустинге строится последовательно, тогда как в случайном лесу деревья могут строиться параллельно, независимо друг от друга.