Отчёт по практическому заданию № 3 "Ансамбли алгоритмов. Веб-сервер. Композиции алгоритмов для решения задачи регрессии"

Практикум 317 группы

Суржанский Александр

Декабрь 2023

Содержание

1 Введение

Данное практическое задание посвящено исследованию ансамблевых алгоритмов на примере моделей Random Forest (Случайный лес) и Gradient Boosting (Градиентный бустинг). Целью исследования является рассмотрение зависимости точности алгоритмов и времени их работы от параметров метода. Ставится задача регрессии - предсказания стоимости недвижимости в зависимости от признаков.

2 Предобработка данных

Дана задача регрессии - необходимо предсказать стоимость недвижимости по её признакам на основе набора данных House Sales in King County, USA. Он содержит информацию о 21613 объектах, характеризуемых 21 признаком.

Удалим из датасета столбец `id` (идентификатор) из-за отсутствия информативности. Данные содержат 3 календарных признака - `date` - дата публикации объявления, `yr_built` - год постройки дома и `yr_renovated` - год последнего проведённого ремонта. Последний столбец может содержать нулевые значения - это означает, что ремонт не проводился. В этом случае будем считать годом проведения ремонта год постройки дома и заменим данные нулевые значения на соответствующие значения из `yr_built`. Преобразуем данные признаки в количество лет, прошедших от данного события (завершение строительства или ремонт) до даты публикации объявления и запишем их в столбцы `yr_built` и `yr_renovated` соответственно. Признак `date` заменим на номер дня в году, чтобы было возможно отслеживать сезонные тренды стоимости продажи.

Разделим данные на обучающую и тестовую выборку в отношении 80% и 20% соответсвенно. Построим матрицу корреляций для признаков из обучающей выборки (Рис. 8 из Приложения). В датасете содержатся 2 пары признаков, сильно коррелирующих друг с другом. Удалим один такой признак из каждой пары - столбцы `sqft_above`, `yr_built`.

3 Эксперименты

3.1 Анализ поведения алгоритма случайный лес

Исследуем поведение алгоритма случайный лес (Random Forest). Изучим зависимость RMSE на отложенной выборке и времени работы алгоритма от следующих факторов:

- количество деревьев в ансамбле;
- размерность подвыборки признаков для одного дерева;
- максимальная глубина дерева.

Рассмотрим, как количество базовых алгоритмов влияет на качество модели и время её работы (Рис. 1). Минимум RMSE достигается на 50 деревьях. Значение ошибки колеблется, однако изменяется не сильно. Время обучения случайного леса зависит от количества алгоритмов линейно.

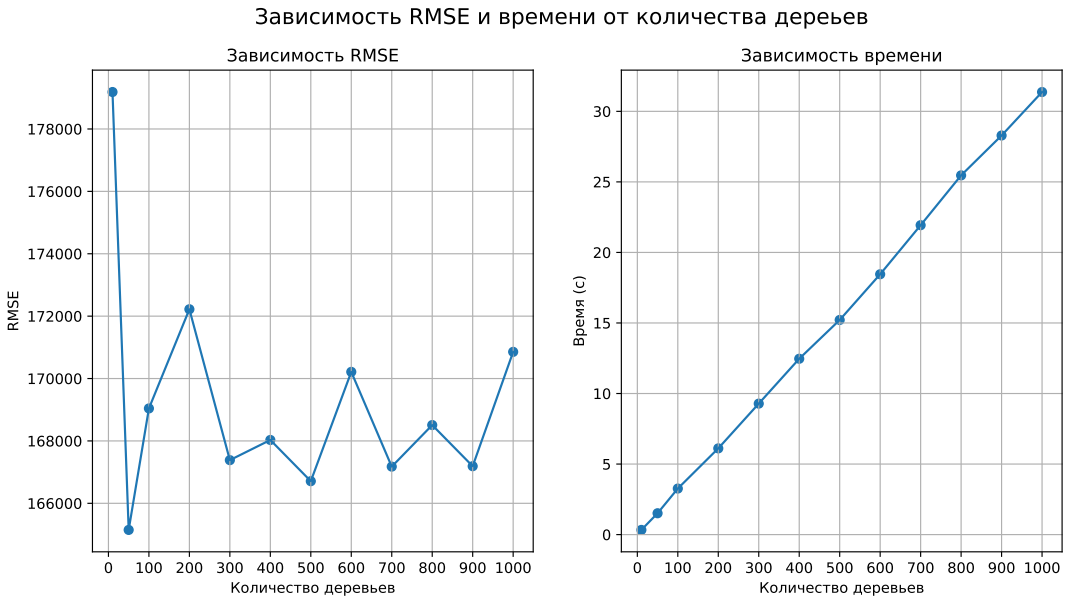

Рис. 1: Зависимость качества и времени работы случайного леса от количества деревьев (max_depth = 20)

Далее рассмотрим, как влияет на качество и время размерность подвыборки признаков (Рис. 2). Можно наблюдать, что оптимальнее всего обучать алгоритм на 80% - 90% признаков. Это объясняется тем, что удаляя часть признаков, деревья становятся более разнообразными, но при этом их остаётся достаточно для того, чтобы делать точные прогнозы. Для времени обучения - также линейная зависимость.

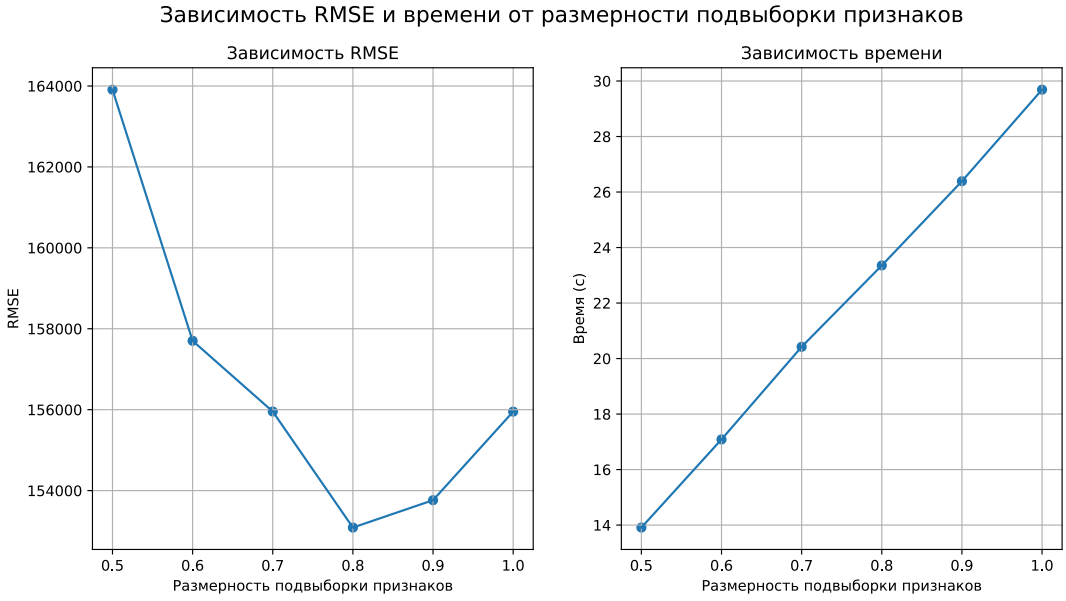

Рис. 2: Зависимость качества и времени работы случайного леса от размерности подвыборки признаков (n_estimators = 500, max_depth = 10)

Рассмотрим, как максимальная глубина дерева влияет на RMSE и время работы

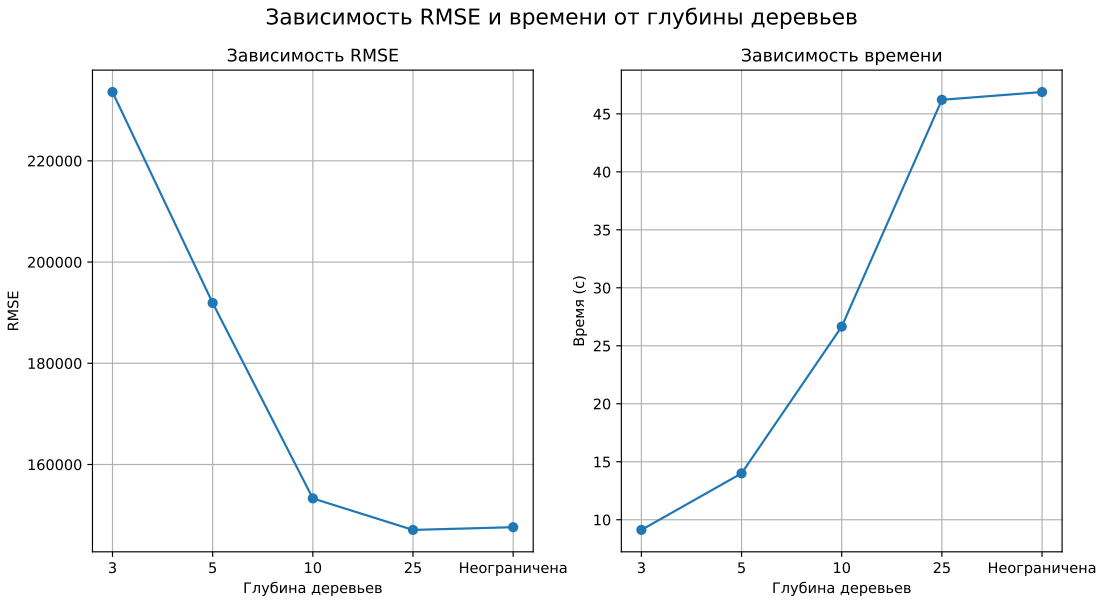

Рис. 3: Зависимость качества и времени работы случайного леса от максимальной глубины дерева (n_estimators = 500, feature_subsample_size = 0.9)

алгоритма (Рис. 3). При увеличении глубины дерева ошибка уменьшается и остаётся примерно одинаковой, начиная со значения 25. Время работы алгоритма линейно зависит от глубины вплоть до этого же значения, далее - растёт медленно. Можно прийти к выводу, что оптимальнее всего для данной модели использовать более разнообразные и сложные деревья, которые будут иметь маленькое смещение и большой разброс.

3.2 Анализ поведения алгоритма градиентный бустинг

Исследуем поведение алгоритма градиентный бустинг (Gradient Boosting). Изучим зависимость RMSE на отложенной выборке и времени работы алгоритма от следующих факторов:

- количество деревьев в ансамбле;
- размерность подвыборки признаков для одного дерева;
- максимальная глубина дерева;
- выбранный `learning_rate`.

Рассмотрим, как количество базовых алгоритмов влияет на качество модели и время её работы (Рис. 4). В отличие от случайного леса увеличение количества деревьев почти всегда уменьшает значение RMSE. Данное наблюдение можно объяснить тем, что градиентный бустинг стремится уменьшить имеющуюся ошибку при каждом следующем построении базового алгоритма. Для времени также наблюдается линейная зависимость.

Далее рассмотрим, как влияет на качество и время размерность подвыборки признаков (Рис. 5). Для данной модели можно наблюдать, что наилучшее качество достигается

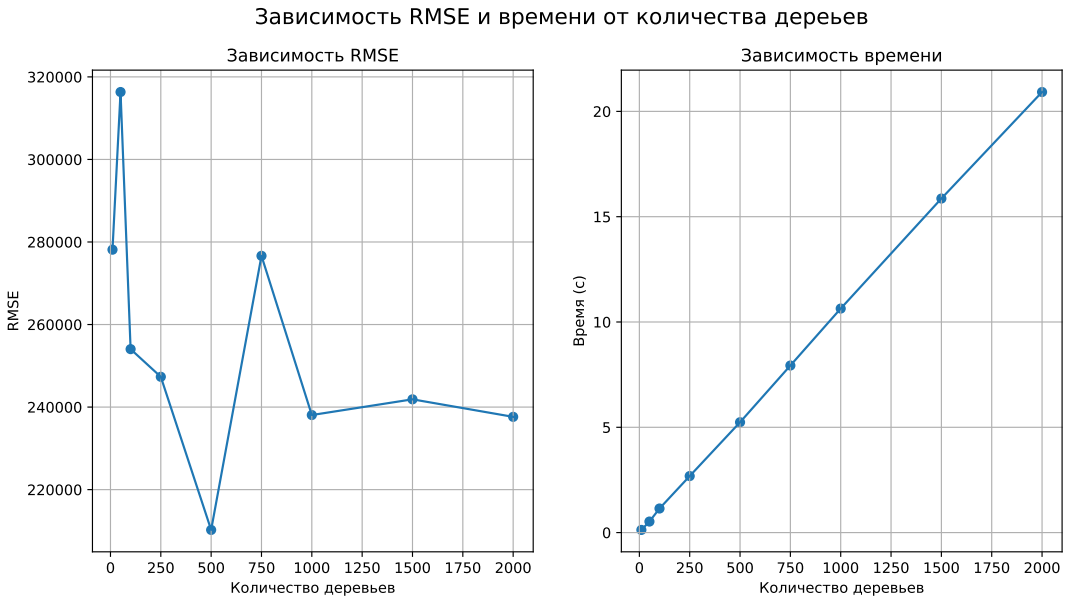

Рис. 4: Зависимость качества и времени работы градиентного бустинга от количества деревьев (max_depth=5)

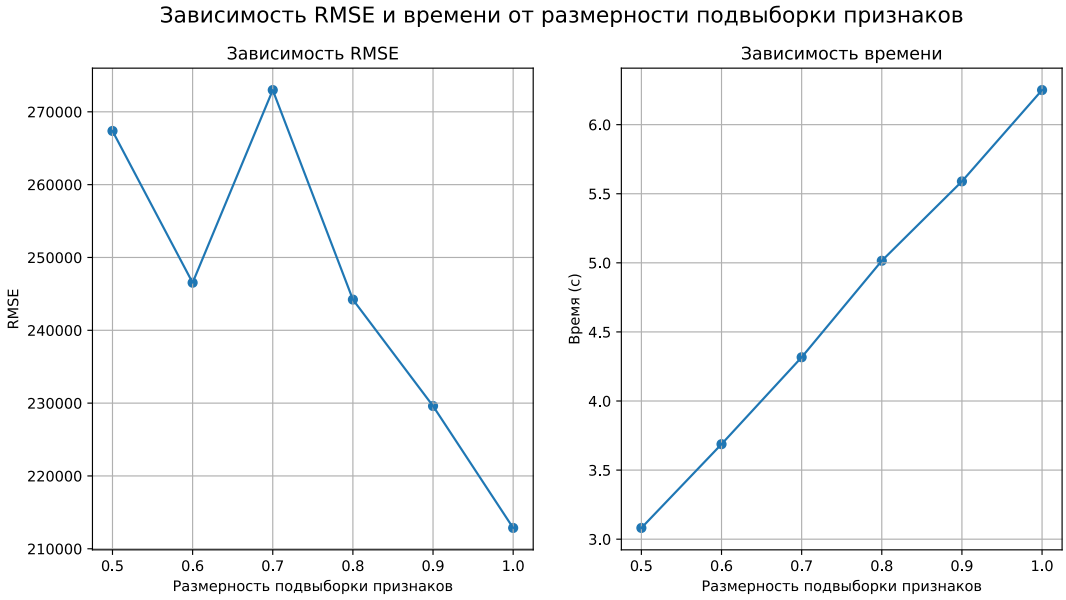

Рис. 5: Зависимость качества и времени работы градиентного бустинга от размерности подвыборки признаков (n_estimators = 100, max_depth = 10)

тогда, когда градиентный бустинг обучается на большем числе признаков. Для времени - линейная зависимость.

Рассмотрим, как максимальная глубина дерева влияет на RMSE и время работы алгоритма (Рис. 6). В отличие от случайного леса минимум ошибки достигается при маленьких значениях глубины. В градиентном бустинге оптимально использовать простые деревья, чтобы избежать переобучения. Для времени обучения алгоритма наблюдается линейная зависимость.

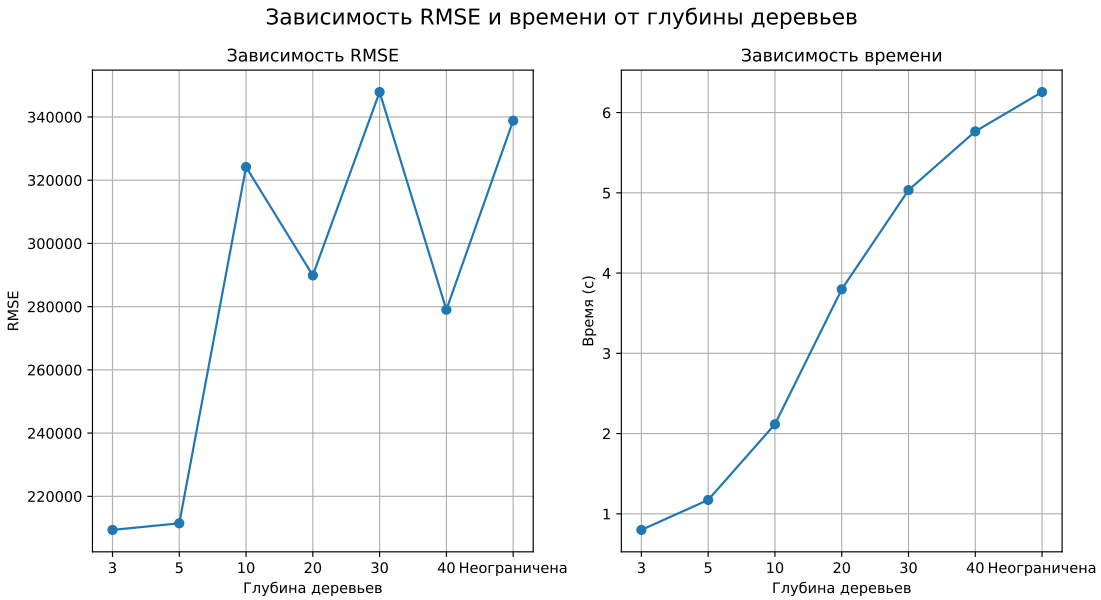

Рис. 6: Зависимость качества и времени работы градиентного бустинга от максимальной глубины дерева (n_estimators = 100)

Рассмотрим, как темп обучения влияет на ошибку алгоритма и время его работы (Рис. 7). По данным графикам невозможно наблюдать очевидных закономерностей. Время работы алгоритма практически не зависит от значения learning_rate и изменяется слабо.

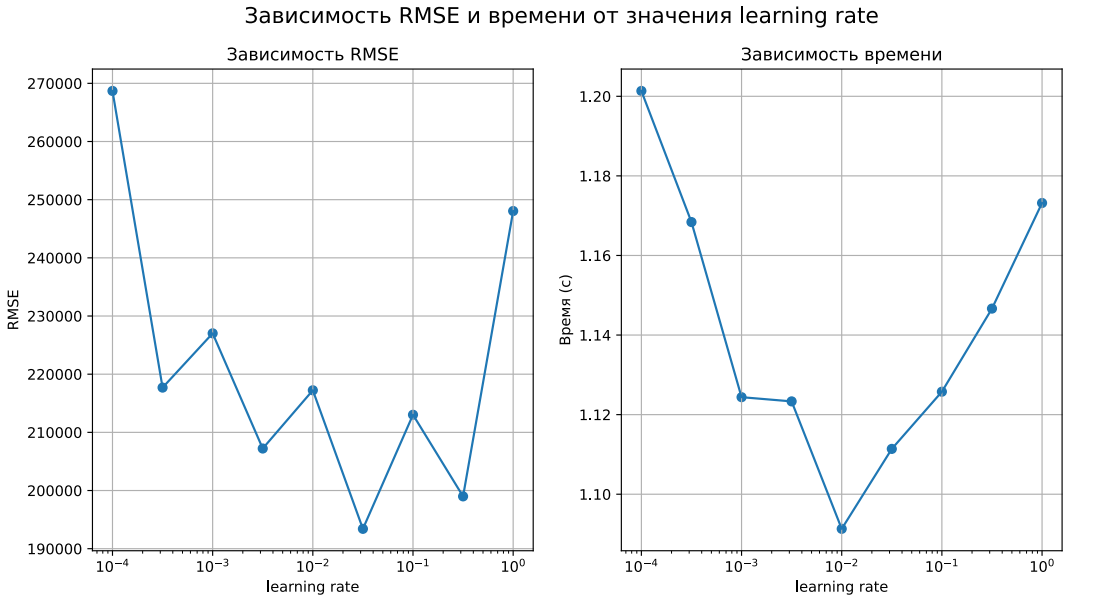

Рис. 7: Зависимость качества и времени работы градиентного бустинга от learning_rate (n_estimators = 100, max_depth=5)

4 Заключение

В данном практическом задании были изучены и реализованы ансамблевые алгоритмы регрессии на примере случайного леса и градиентного спуска. На основе датасета о продажах недвижимости были проведены эксперименты по измерению времени и качества работы метода в зависимости от параметров.

Приложение

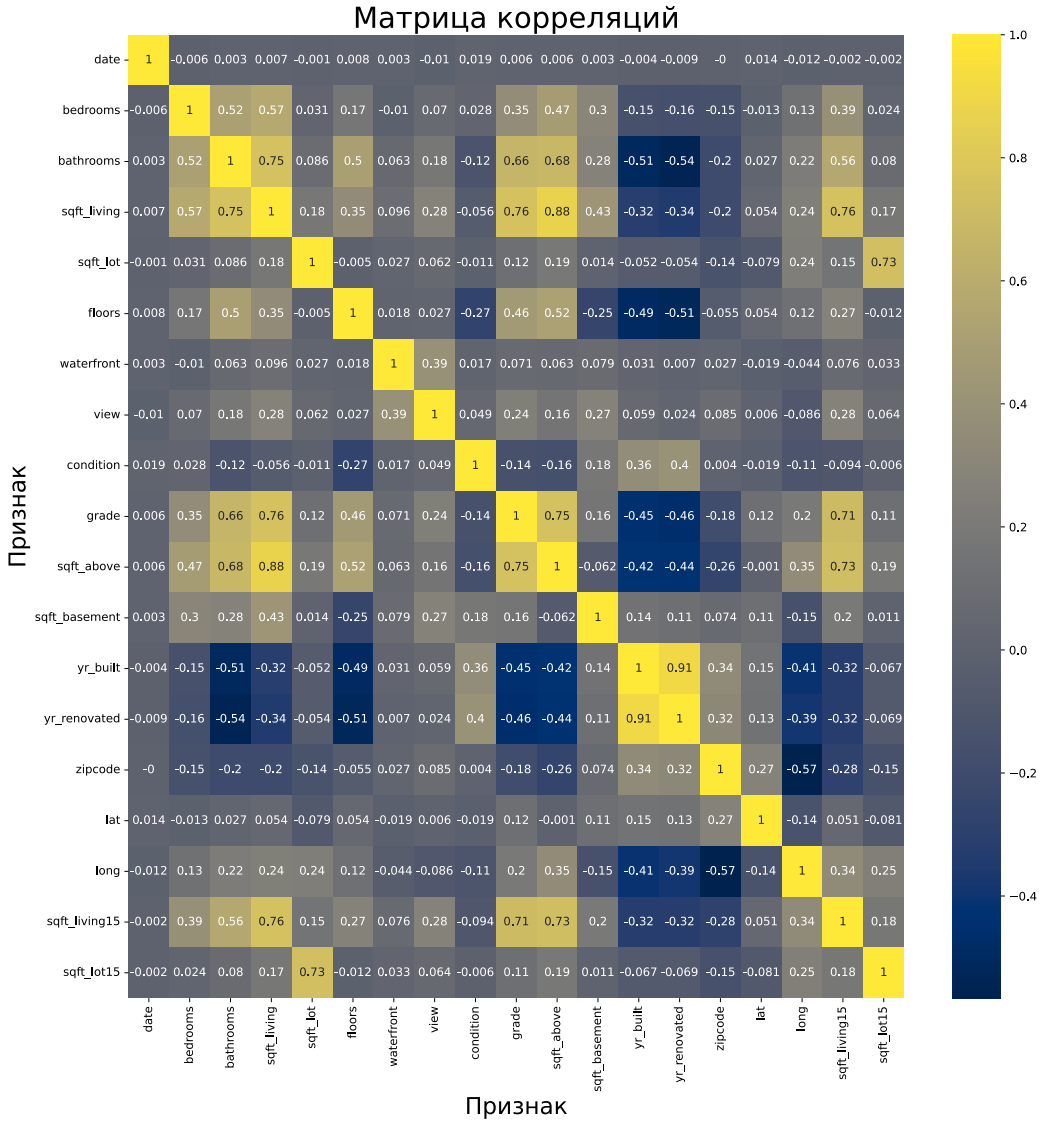

Рис. 8: Корреляционная матрица признаков для обучающей выборки