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

Практикум 317 группы, ММП ВМК МГУ.

Марьясов Максим Михайлович.

Декабрь 2024.

Содержание

1 Введение **1**

2 Эксперименты **1**

2.1 Эксперимент №1: Предобработка данных 2

2.1.1 Изучение таргета 2

2.1.2 Общее описание признаков и их предобработка 2

2.2 Эксперимент №2: Исследование поведения случайного леса 4

2.2.1 Исследование в зависимости от количества деревьев 5

2.2.2 Исследование в зависимости от размерности подвыборки признаков для вершины дерева 5

2.2.3 Исследование в зависимости от максимальной глубины деревьев 5

2.2.4 Исследование при неограниченной глубине деревьев 6

2.3 Эксперимент №3: Исследование поведения градиентного бустинга 6

2.3.1 Исследование в зависимости от количества деревьев 7

2.3.2 Исследование в зависимости от размерности подвыборки признаков 7

2.3.3 Исследование в зависимости от максимальной глубины деревьев 8

2.3.4 Исследование при неограниченной глубине деревьев 9

2.3.5 Исследование в зависимости от скорости обучения ансамбля 9

3 Заключение **10**

4 Библиография **10**

1 Введение

Данное задание направлено на изучение ансамблей алгоритмов, таких как случайный лес и градиентный бустинг, на примере задачи регрессии, заключающейся в предсказывании цены недвижимости из датасета **House Sales in King County, USA**. Цель исследования - реализовать модели RandomForestMSE (случайный лес) и GradientBoostingMSE(градиентный бустинг) с целью изучить поведение моделей в зависимости от гиперпараметров в задаче регрессии.

2 Эксперименты

Эксперименты в рамках данного задания включают в себя:

- **Предобработка данных**
- **Исследование поведения случайного леса в зависимости от гиперпараметров**
- **Исследование поведения градиентного бустинга в зависимости от гиперпараметров**

2.1 Эксперимент №1: Предобработка данных

Данный эксперимент направлен на изучение датасета и предобработки его для дальнейшего использования в обучении и валидации. Качественная предобработка дает возможность качественный этап обучения модели.

2.1.1 Изучение таргета

На Рис.1 наблюдается, что относительно реальных значений распределения стоимости недвижимости не является нормальным, но если прологарифмировать каждое значение, то распределение становится нормальным.

В следующих экспериментах мы будем использовать RMLSE в качестве критерия оценки.

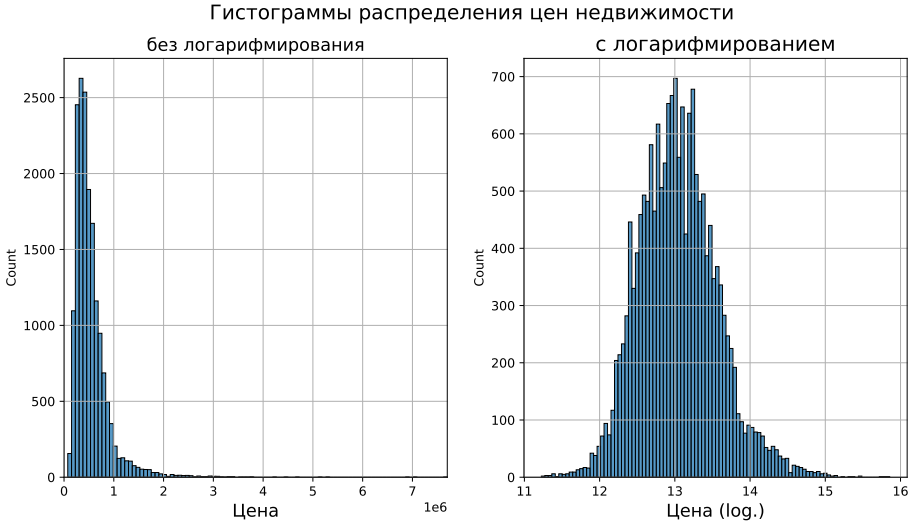

Рис. 1: График значения таргета без логарифмирования и с логарифмированием

2.1.2 Общее описание признаков и их предобработка

В датасете представлены множество признаков и некоторые из них требуют предобработки для повышения информативности и вклада признака в обучение модели.

Визуализируем корреляционную таблицу, интерпретировав каждый признак, как численный (если это можно). Результаты представлены на Рис. 2.

Была заметна сильная корреляция признаков, отвечающих за размеры недвижимости (sqft_leaving, sqft_lot, sqft_above, sqft_leaving15, sqft_lot15). Был удален признак sqft_above (размер жилой площади без подвала) (так как он является разницей между признаками sqft_leaving и sqft_basement (размер жилой площади и размер подвала, соответственно). Также были преобразованы признаки sqft_leaving15 и sqft_lot15 в признаки dev_leaving15, dev_lot15 (как разница между средней площадью ближайших 15 объектов недвижимости и площадью рассматриваемого объекта). Это было сделано для того, чтобы сделать признак более информативным - он становится отклонением площади от ближайших объектов. Результат данной предобработки в качестве корреляционной таблицы представлен на Рис. 3.

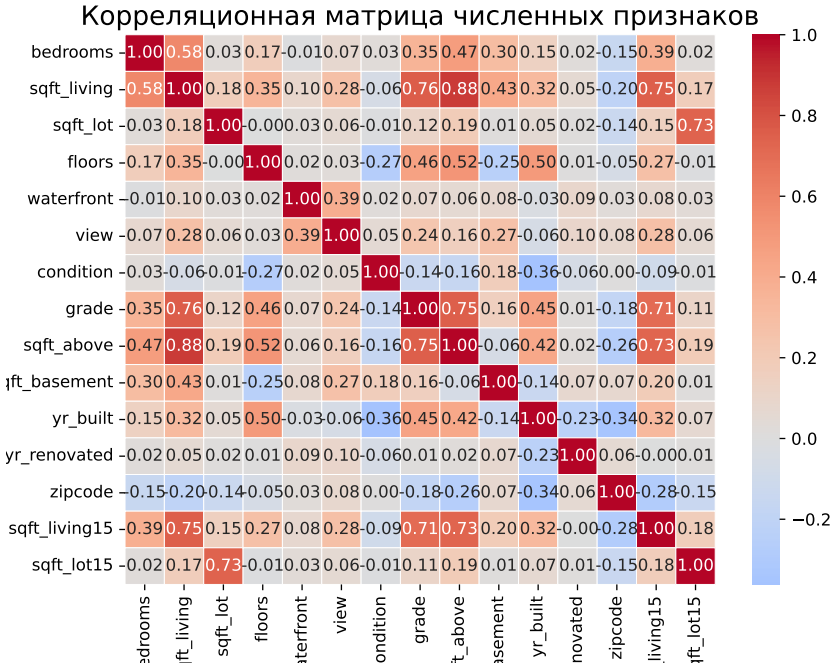

Рис. 2: Корреляционная матрица до предобработки

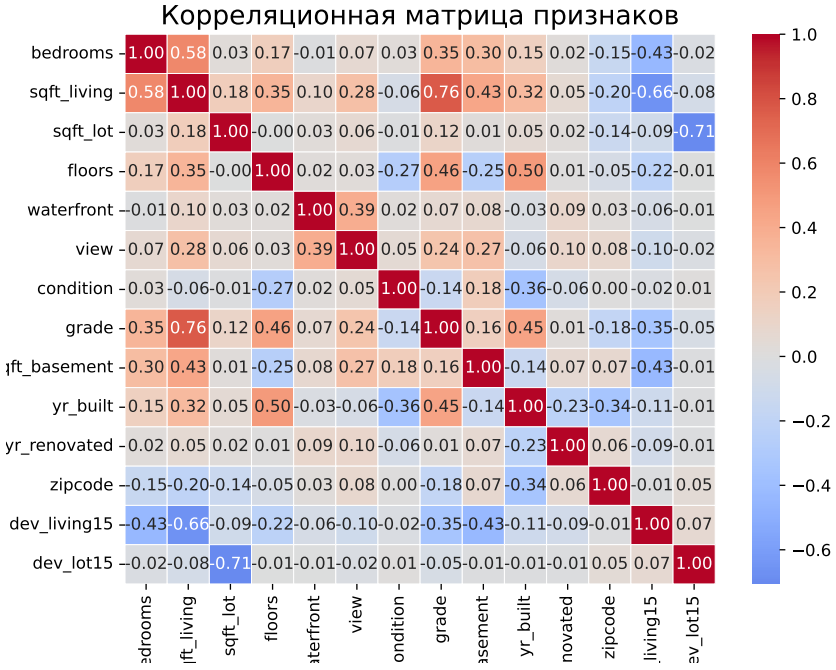

Рис. 3: Корреляционная матрица после предобработки площади

Далее была совершена предобработка признаков с датами, такие как date, yr_built и yr_renovated, для их использования в обучении модели. По таблице корреляции можно заметить, что год постройки, взятый в числовом виде коррелирует с классом жилья. Было сделано следующее:

- yr_built и yr_renovated были преобразованы в значения (2015 - значение). Это было сделано для повышения информативности признака - он становится возрастом жилья (с постройки и с

реновации, соответственно).

- date была преобразована в категориальное значение от 0 до 19. Даты объектов обучающей выборки делились на 20 множеств с приблизительно равным размером, где множество представляли отрезок на прямой из дат.

  Результат данной предобработки дат представлен в качестве корреляционной таблицы на Рис. 4.

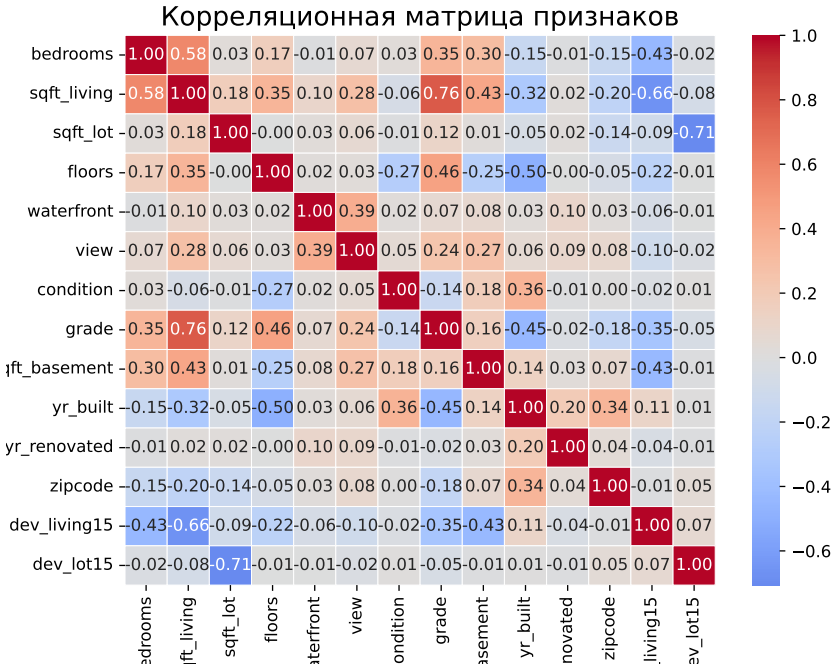

Рис. 4: Корреляционная матрица после предобработки дат

Далее была совершена предобработка двух колонок lat и long (широта и долгота) с целью извлечь из них информацию о местоположении недвижимости. Карта была разделена на прямоугольник и были выбраны 50 прямоугольников с самой большой частотой недвижимости. Заметим, что среди признаков уже есть zipcode, отвечающий за почтовый адрес, но во время исследования было получено, что эти два признака вместе повышают качество модели.

В предобработке признаков использовались OneHotEncoder, StandartScaler из библиотеки scikit-learn и MeanTargetEncoder, реализованный самостоятельно. Это было сделано для того, чтобы некоторые признаки кодировать значением цены с шумом, если это важно. Это сделано в предположении того, что, к примеру, в одном квартале стоимость жилья будет приблизительно равна среднему по кварталу.

2.2 Эксперимент №2: Исследование поведения случайного леса

Данный эксперимент направлен на исследование поведения случайного леса (RandomForestMSE) в зависимости от различных гиперпараметрах, таких как: max_depth - максимальная глубина деревьев в лесе, max_features - размерность подвыборки признаков для вершины деревьев при их обучении и n_estimators - количество деревьев в лесе.

2.2.1 Исследование в зависимости от количества деревьев

В результате данной части эксперимента было изучено поведение случайного леса в зависимости от количества деревьев (от 1 до 10000). Результаты приведены на Рис. 5.

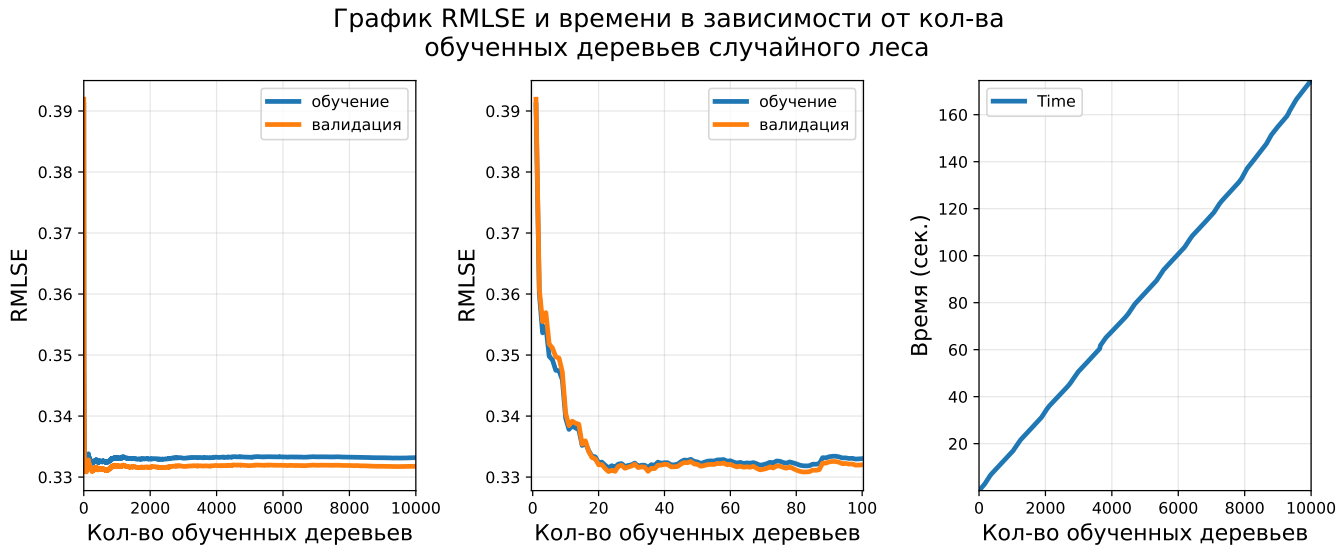

Рис. 5: Графики исследования случайного леса в зависимости от количества деревьев

Данный результат показывает, что случайный лес, как средний ответ по всем деревьям на больших количествах деревьев приходит к определенному среднему. Первые 80 деревьев дают выше качество на валидации и обучении, не доходя еще до равновесного среднего.

В дальнейшем, при исследовании случайного леса будет использован параметр patience равный 2000 при обучении модели, чтобы остановить модель когда она начнет показывать качество ниже, чем в 2000 итерациях до этого. Также на графике времени видно, что зависимость от количества деревьев линейна.

2.2.2 Исследование в зависимости от размерности подвыборки признаков для вершины дерева

В результате данной части эксперимента было изучено поведение случайного леса в зависимости от размерности подвыборки признаков для вершины дерева. Результаты приведены на Рис. 6.

Данный результат показывает, что качество на валидации становится выше при увеличении количества признаков в подвыборке. В тоже время коэффициент наклона в графике зависимости времени увеличивается, потому что увеличиваются затраты по времени на выбор "оптимального"признака для построения одного дерева.

2.2.3 Исследование в зависимости от максимальной глубины деревьев

В результате данной части эксперимента было изучено поведение случайного леса в зависимости от максимальной глубины деревьев. Результаты приведены на Рис. 7.

Данный результат показывает, что качество на валидации становится выше при увеличении глубины деревьев. В тоже время коэффициент наклона в графике зависимости времени увеличивается, потому что увеличиваются затраты по времени на построение большего в глубину дерева для его построения.

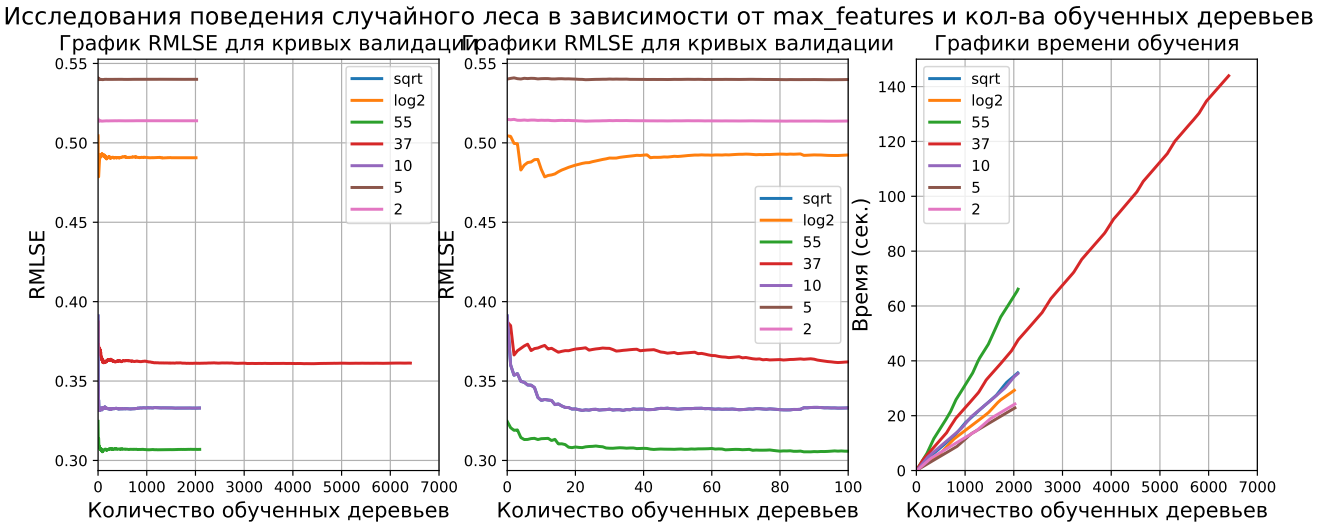

Рис. 6: Графики исследования случайного леса в зависимости от размерности подвыборки признаков для вершины дерева

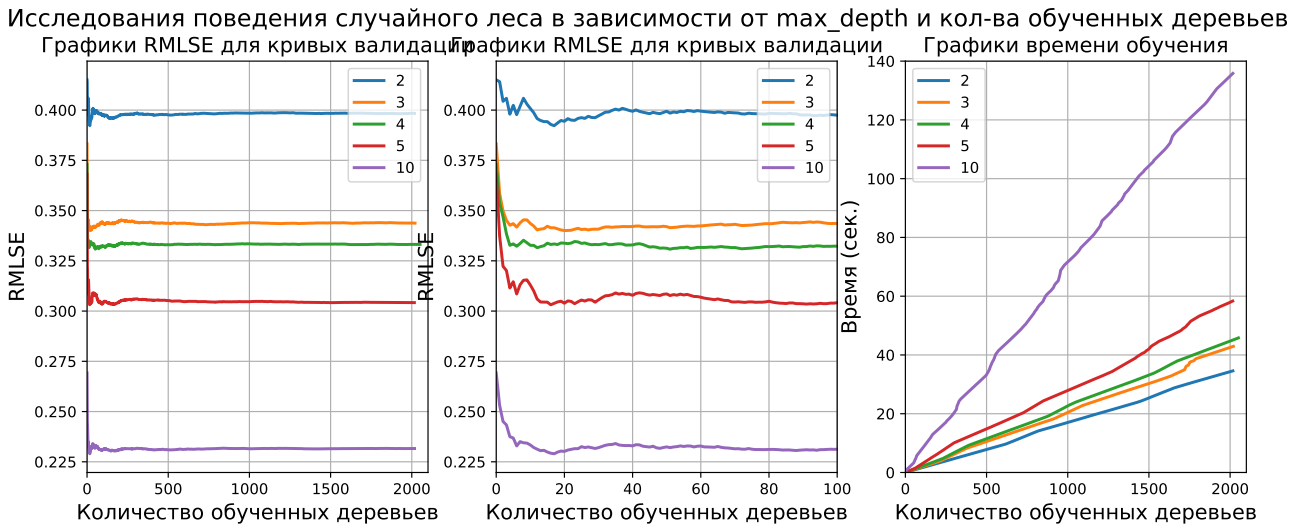

Рис. 7: Графики исследования случайного леса в зависимости от максимальной глубины деревьев

2.2.4 Исследование при неограниченной глубине деревьев

В результате данной части эксперимента было изучено поведение случайного леса при неограниченной глубине деревьев. Результаты приведеы на Рис. 8.

Данный результат показывает, что качество валидации становится заметно выше, но время потраченное на обучение заметно выше относительно сравнения с случайным лесом с ограниченной глубиной деревьев.

2.3 Эксперимент №3: Исследование поведения градиентного бустинга

Данный эксперимент направлен на исследование поведения градиентного бустинга (GradientBoostingMSE) в зависимости от различных гиперпараметров, таких как: n_estimators - количество деревьев в ансамбле, max_depth - максимальная глубина деревьев, и learning_rate - скорость обучения.

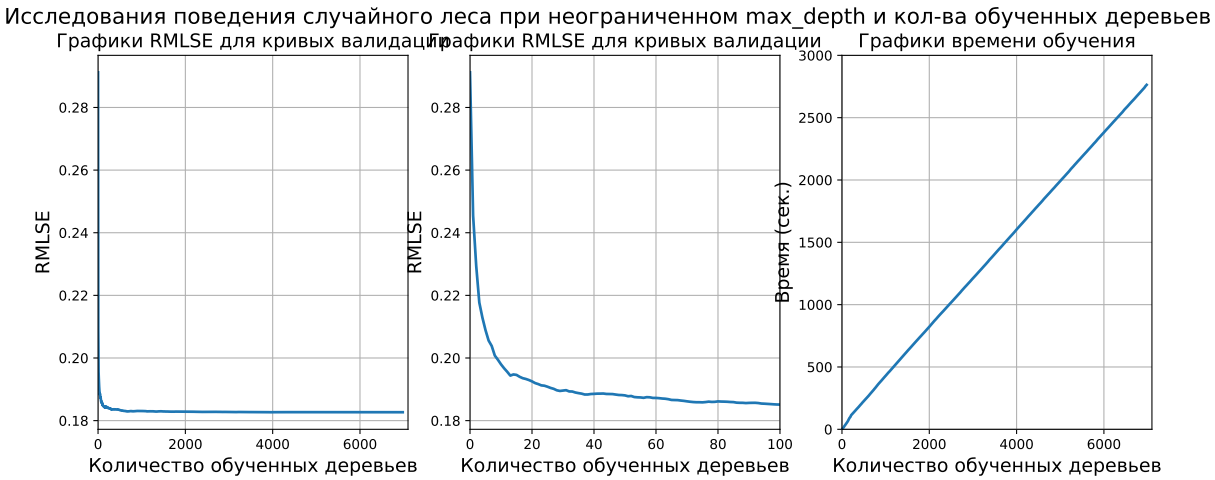

Рис. 8: Графики исследования случайного леса при неограниченной глубине деревьев

2.3.1 Исследование в зависимости от количества деревьев

В результате данной части эксперимента было изучено поведение градиентного бустинга в зависимости от количества деревьев. Результаты приведены на Рис. 9.

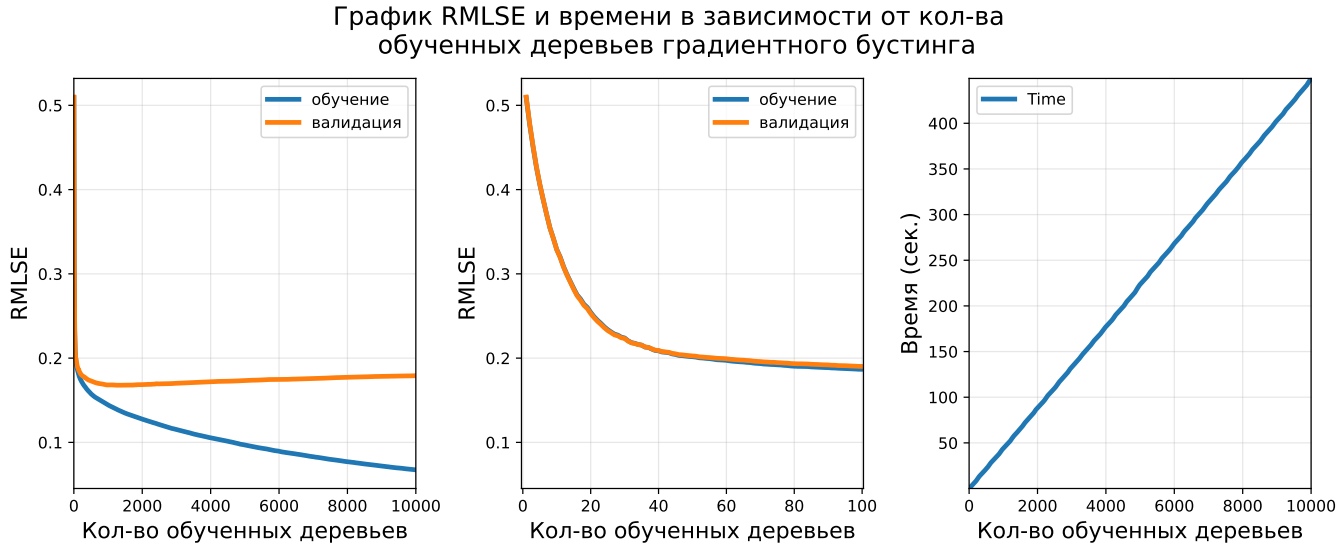

Рис. 9: Графики исследования градиентного бустинга в зависимости от количества деревьев

В отличие от случайного леса момент переобучения сильно заметен на графиках. Мы наблюдаем, что градиентный бустинг максимально старается минимизировать критерий качества обучения, что приводит к плохому качеству на валидации. График времени также линейный, как и у случайного леса.

Далее, будет использоваться параметр patience = 500 при обучении градиентного бустинга, чтобы предотвратить его переобучение.

2.3.2 Исследование в зависимости от размерности подвыборки признаков

В результате данной части эксперимента было изучено поведение градиентного бустинга в зависимости от размерности подвыборки признаков для каждой вершины дерева. Результаты приведены на Рис. 10.

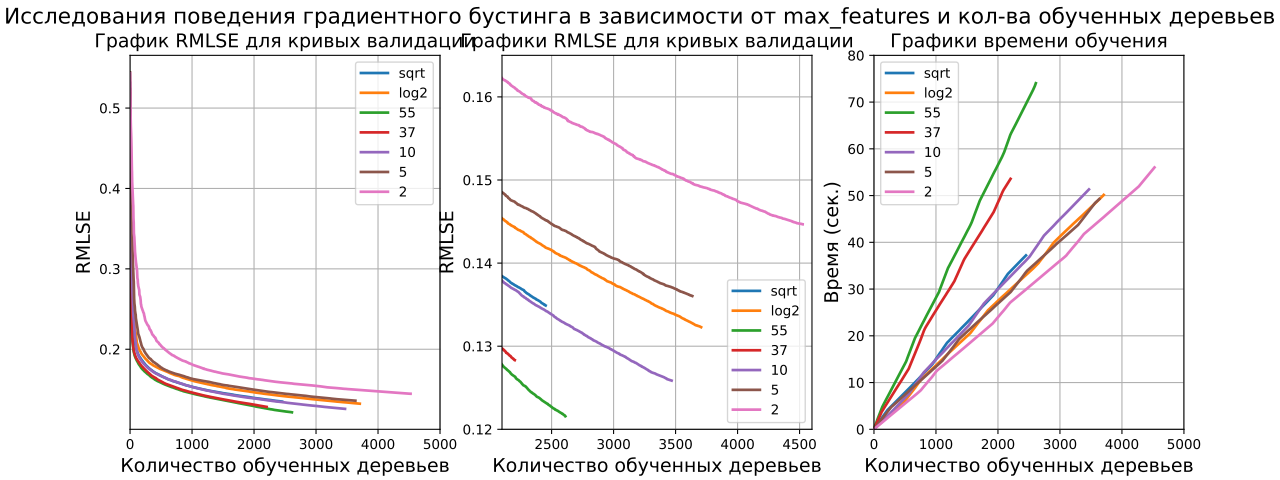

Рис. 10: Графики исследования градиентного бустинга в зависимости от размерности подвыборки признаков

На графике можно наблюдать, что как и в случайном лесе, большее число признаков повышает качество, но увеличивает затраты по времени за итерацию, но не за полное обучение, так как увеличивается скорость сходимости. Это объясняется тем, что алгоритму дано большее количество для выбора "оптимального"для вершины, ускоряя сходимость.

2.3.3 Исследование в зависимости от максимальной глубины деревьев

В результате данной части эксперимента было изучено поведение градиентного бустинга в зависимости от максимальной глубины деревьев. Результаты приведены на Рис. 11.

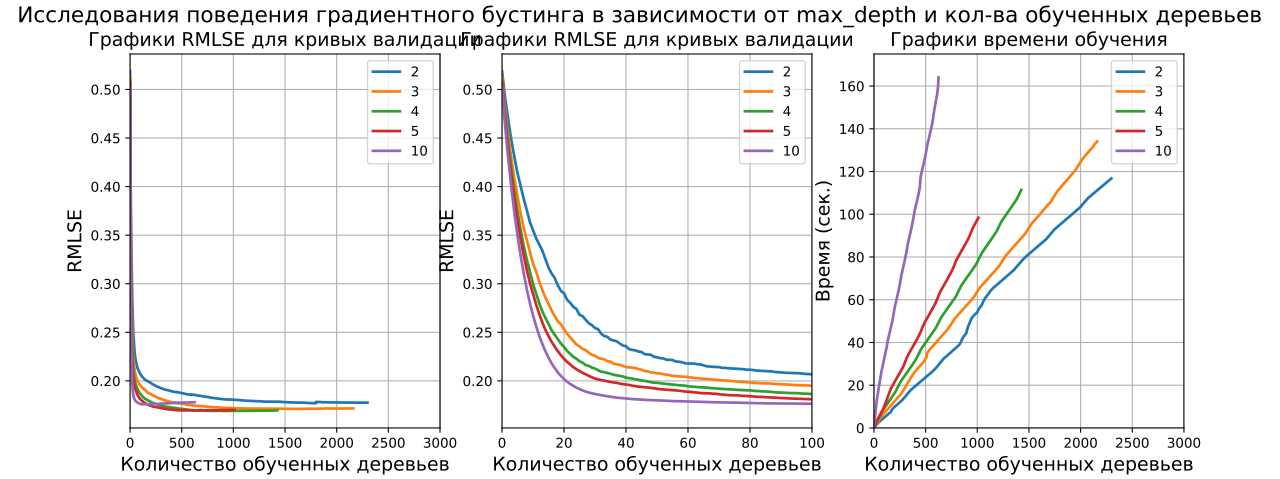

Рис. 11: Графики исследования градиентного бустинга в зависимости от максимальной глубины деревьев

На графике можно наблюдать, что при увеличении глубины скорость сходимости увеличивается, затраты по времени тоже, но также, в основном, увеличивается качество (до глубины равной 5 так). Эти выводы применимы к деревьям до глубины 5 включительно, для более больших значений глубины требуются отдельный анализ.

2.3.4 Исследование при неограниченной глубине деревьев

В результате данной части эксперимента было изучено поведение градиентного бустинга при неограниченной глубине деревьев. Результаты приведены на Рис. 12.

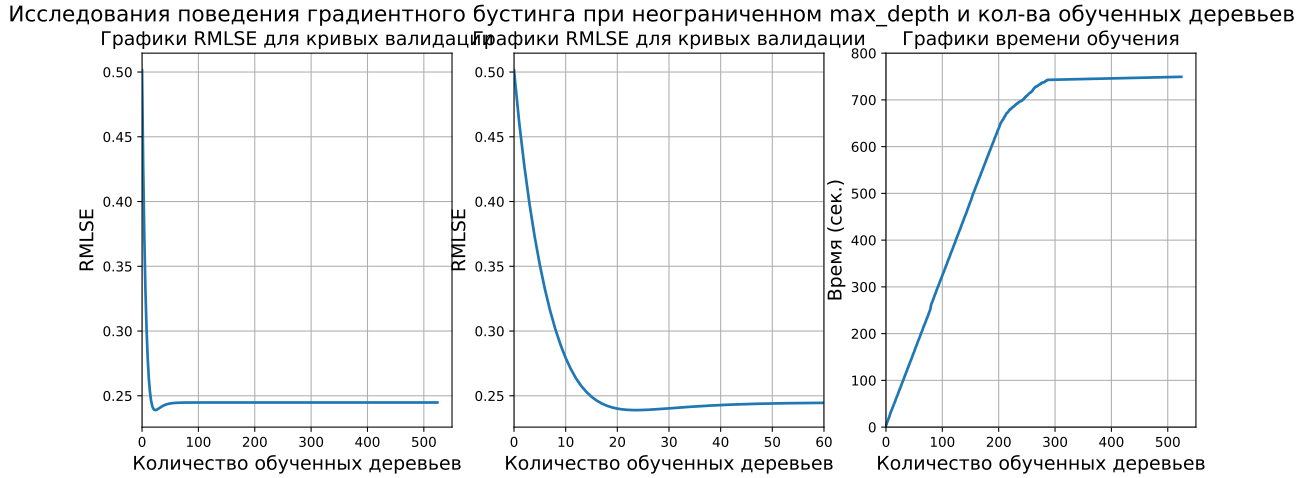

Рис. 12: Графики исследования градиентного бустинга при неограниченной глубине деревьев

На графике можно заметить, что качество градиентного бустинга заметно хуже, чем его качество с деревьями с ограниченной глубиной. Затраты на обучение тоже вырастают, но для такого алгоритма требуется меньшее число деревьев.

2.3.5 Исследование в зависимости от скорости обучения ансамбля

В результате данной части эксперимента было изучено поведение градиетного бустинга в зависимости от скорости обучения ансамбля. Результаты приведены на Рис. 13.

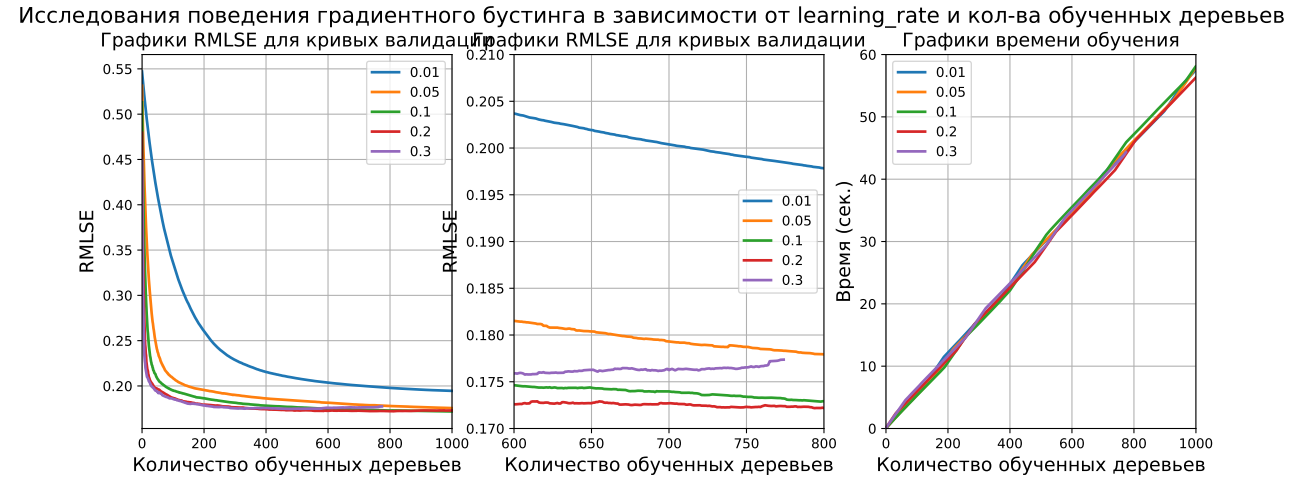

Рис. 13: Графики исследования градиентного бустинга в зависимости от скорости обучения ансамбля

На графиках заметно следующее: скорость обучения не влияет на время обучения одного дерева. Хорошее качество показывают коэффициенты 0.1 и 0.2, 0.3 введет себя не предсказуемо, а 0.01 и 0.05 имеют меньшее качество (в силу меньшей скорости обучения).

3 Заключение

Результатом данной работы являются навыки работы со случайным лесом и градиентным бустингом в задачах регрессии, изучение их гиперпараметров. В рамках данного задания были написаны две реализации, предобработан датасет. Мы наблюдаем, что данные модели требуют качественные данные, хоть они и устойчивы к выбросам.

4 Библиография

1. **Лекция ММРО по ансамблям**
   https://github.com/mmp-mmro-team/mmp_mmro_fall_2024/blob/main/seminars/Seminar_12_structure_of_decision_tree.pdf

2. **Воронцов К. В. Градиентный бустинг**
   https://github.com/MSU-ML-COURSE/ML-COURSE-24-25/blob/main/slides/2_stream/msu24-compos2.pdf

3. **Воронцов К. В. Линейные ансамбли**
   https://github.com/MSU-ML-COURSE/ML-COURSE-24-25/blob/main/slides/2_stream/msu24-compos1.pdf