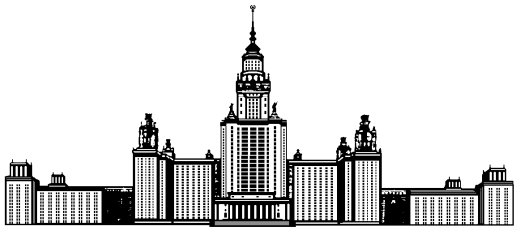
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики
Кафедра Математических Методов Прогнозирования

Отчет по третьему заданию курса "Практикум на ЭВМ"

"Ансамбли алгоритмов. Веб-сервер. Композиции алгоритмов для решения задачи регрессии."

Выполнила:
студентка 3 курса 317 группы
Кривуля Полина Юрьевна

Москва, 2022

Содержание

1 Введение

Случайный лес – это ансамбль равноправных решающих деревьев, не передающих друг другу никакой информации. В градиентном бустинге, в отличие от случайного леса, каждый следующий алгоритм получает от предыдущего информацию о допущенных ошибках.

В рамках данного задания выполнена ручная реализация методов случайного леса и градиентного бустинга. Выполнены эксперименты с датасетом данных о продажах недвижимости House Sales in King County, USA, позволяющие исследовать зависимость времени работы и качества RMSE методов от параметров. Реализован веб-сервер, позволяющий использовать данные модели с выбранными параметрами и смотреть на поведение алгоритма на обучающей, валидационной и тестовой выборках.

2 Предобработка имеющихся данных

В исходных данных о продажах недвижимости House Sales in King County, USA 21613 объекта, 21 признак, включая таргет (столбец "price"). Пропущенных значений нет. Типы данных: float, int, кроме колонки с датой. У последней изначально тип object. Выполнено преобразование даты в тип datetime, из даты выделен год, месяц, день недели, день месяца. На графике 1 видно, что средняя цена на недвижимость зависит от этих параметров. Завимость от года не показана, так как там всего два значения, но средняя цена также меняется. Сам столбец с изначальной датой удален.

Рис. 1: Зависимость средней цены на недвижимость от даты

Выполнен подсчет числа уникальных значений (таблица 2) и так как почти все id уникальны, было принято решение удалить данный столбец. Более подробная информация о числе уникальных значений и типах данных каждого столбца представлена в таблице. Изначально смутило, что в некоторых столбцах, которые, казалось бы, должны быть целочисленными, тип данных – float (например, "bathrooms"и "floors"). Однако ознакомление со значениями столбцов и с данными соревнования показало, что так и задумано.

Стандартизация данных не использовалась, так как методы устойчивы к немасштабированным признакам.

Категориальные признаки выбирались как те, которые нельзя упорядочить, к которым нельзя применить операции сравнения. Как категориальные рассматривались изначально "waterfront", "view", "condition", "grade", "zipcode". Однако после первого же эксперимента (пункт 3) в качестве категориальных были оставлены "waterfront" и "zipcode" как показавшие лучший результат. К категориальным признакам поочередно применено One-Hot кодирование и кодирование с помощью счетчиков со сглаживаниям (TargetEncoder) и выполнено сравнение результатов первого эксперимента, после чего было оставлено кодирование с помощью счетчиков.

3 Случайный лес

В данных экспериментах исследуется зависимость качества RMSE на обучающей и отложенной выборке и времени работы алгоритма случайный лес в зависимости от количества деревьев в

ансамбле, размерности подвыборки признаков для одного дерева и максимальной глубины дерева. Параметры, зависимость от которых не исследуется в экспериментах, остаются по умолчанию, если не сказано другое.

- Количество деревьев в ансамбле

  В данном эксперименте также проведено сравнение поведения алгоритма в зависимости от вида используемого кодирования. Как видно на графике 2, значение RMSE в начале монотонно убывает, а затем выходит на асимптоту. Время работы растет линейно, так как метод является ансамблевым. наименьшее RMSE на валидационной выборке показывает способ кодирования с помощью счетчиков. Также такой способ кодирования позволяет модели работать в разы быстрее, чем кодирование с помощью One-Hot Encoding, так как размерность признакового пространства не увеличивается, что тоже является существенным плюсом. Далее оставим кодирование с помощью счетчиков со сглаживанием.

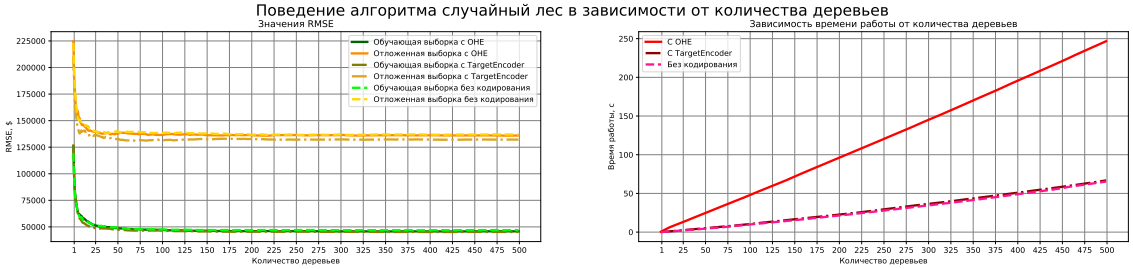


Рис. 2: Зависимость поведения алгоритма случайный лес от числа деревьев при различных способах кодирования

Минимальные достигнутые значения RMSE в зависимости от используемого кодирования представлены в таблице 1.

Способ кодирования	Минимальное RMSE
OHE	135797
TargetEncoder	131126
Без кодирования	136534

Таблица 1: Зависимость RMSE от используемого способа кодирования

Минимальное значение RMSE при TargetEncoder достигается на 82 деревьях.

- Размерность подвыборки признаков для одного дерева

  На графике 3 представлена зависимости поведения алгоритма от размерности подвыборки признаков для 1 дерева и для 82 деревьев (показавшее себя оптимальным в предыдущем пункте значение). Количество признаков перебрано от 1 до 22 включительно, то есть, всевозможные варианты. В данном случае RMSE убывает (но не монотонно) с ростом числа признаков. Время работы увеличивается практически линейно (для случая 1 дерева время тоже растет, но не так заметно). На отложенной выборке для 82 деревьев наименьшее значение RMSE равно 128645 и оно достигнуто при 17 признаках.

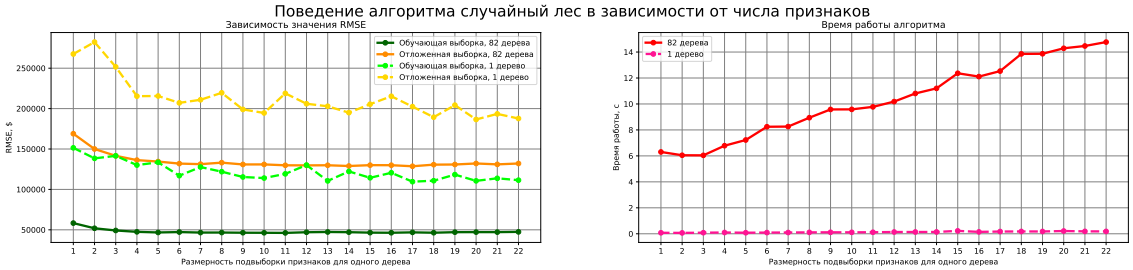


Рис. 3: Зависимость поведения алгоритма случайный лес от числа признаков при различных количествах деревьев

- Максимальная глубина дерева

  Ясно, что чем меньше глубина, тем быстрее строится и работает случайный лес. При увеличении глубины возрастает качество на обучении (т.к. каждое дерево лучше настраивается на свою подвыборку), но и на контроле оно, как правило, увеличивается [1]. В данном эксперименте проводится зависимость поведения алгоритма от максимальной глубины дерева. Значения перебираются от 1 до 34 включительно, отдельно рассматривается случай неограниченной глубины. При этом число деревьев берется равным 82, а количество признаков 17, так как именно такие значения параметров показали себя как лучшие. На графике 4 представлены полученные результаты.

  Значения ошибки уменьшаются с ростом числа признаков, а затем выходит на асимптоту. До глубины, равной 20, время работы растет линейно, а после 20 меняется в небольших пределах. На отложенной выборке наименьшее значение RMSE равно 127829 и оно достигнуто при глубине, равной 24.

Рис. 4: Зависимость поведения алгоритма случайный лес от максимальной глубины дерева

4 Градиентный бустинг

В данных экспериментах исследуется зависимость качества RMSE на обучающей и отложенной выборке и времени работы алгоритма случайный лес в зависимости от количества деревьев в ансамбле, размерности подвыборки признаков для одного дерева, максимальной глубины дерева и темпа обучения. Сразу используется кодирование с помощью счетчиков со сглаживанием. Параметры, зависимость от которых не исследуется в экспериментах, остаются по умолчанию, если не сказано другое.

- Количество деревьев в ансамбле

  Полученная зависимость поведения алгоритма от числа деревьев представлена на графике 5. Ошибка на отложенной выборке в начале монотонно убывает, а затем выходит на асимптоту. Ошибка на обучающей выборке убывает к нулю. Время работы растет линейно. Лучшее значение RMSE на валидационной выборке равно 113899 (что уже меньше, чем при всех выбранных лучших параметрах случайного леса) и достигается при количестве деревьев, равном 967.

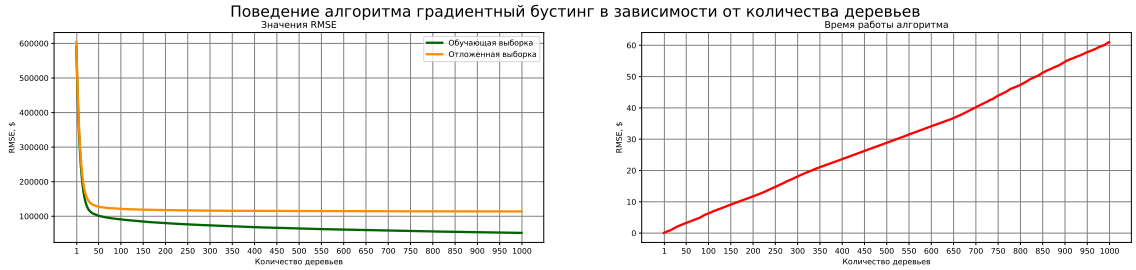


Рис. 5: Зависимость поведения алгоритма градиентный бустинг от числа деревьев

- Размерность подвыборки признаков для одного дерева

  На графике 6 представлены полученные результаты зависимости поведения алгоритма от размерности признакового пространства. Число деревьев равно 967. В отличие от случайного леса, здесь сильнее выражена нестабильность на валидации. Время работы растет линейно. Наименьшее значение RMSE на отложенной выборке равно 112095 и достигнуто при 21 признаке.

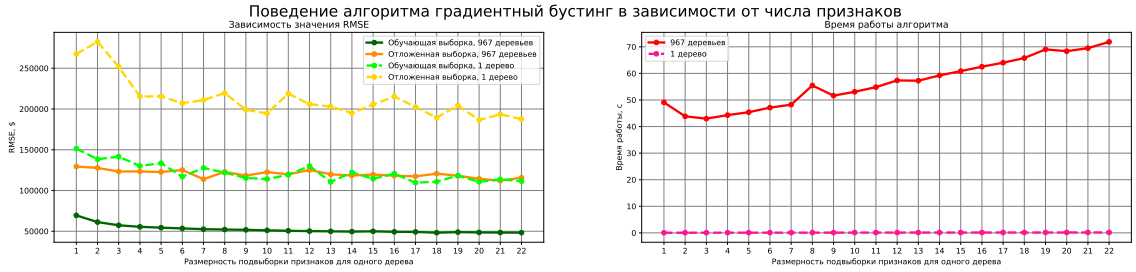


Рис. 6: Зависимость поведения алгоритма градиентный бустинг от числа признаков при различных количествах деревьев

- Максимальная глубина дерева

  В данном эксперименте число деревьев равно 967, число признаков 21. Перебирается глубина деревьев от 1 до 19 включительно, отдельно рассматривается случай, когда глубина не ограничена. Причина уменьшения числа рассматриваемых значений связана с тем, что как правило, бустинг показывает высокое качество над неглубокими деревьями [2], а также с большим количеством деревьев и, соответственно, увеличением времени работы. На графике 7 наглядно видно, что при больших значениях глубины модель перебучается. Время обучения растет линейно. Наименьшее значение RMSE равно 112095 – оно такое же, как и в предыдущем эксперименте, так как достигается при значении глубины по умолчанию (это значение 5).

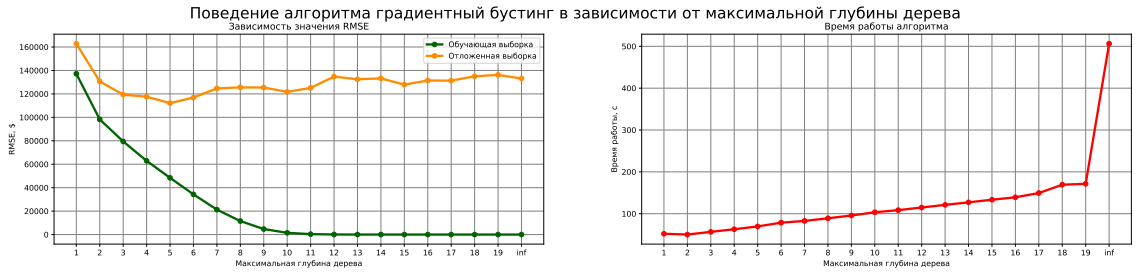


Рис. 7: Зависимость поведения алгоритма градиентный бустинг от максимальной глубины дерева

- Выбранный темп обучения

В данном эксперименте число деревьев равно 967, число признаков 2, максимальная глубина дерева 5. На графике 8 видно, что с ростом темпа обучения RMSE на обучающей выборке монотонно убывает, а на отложенной выборке растет, то есть, модель переобучается. Явной зависимости времени работы от темпа обучения не наблюдается. Наименьшее значение RMSE на валидации равно 112095 и достигается при темпе обучения 0.1, что является значением по умолчанию.

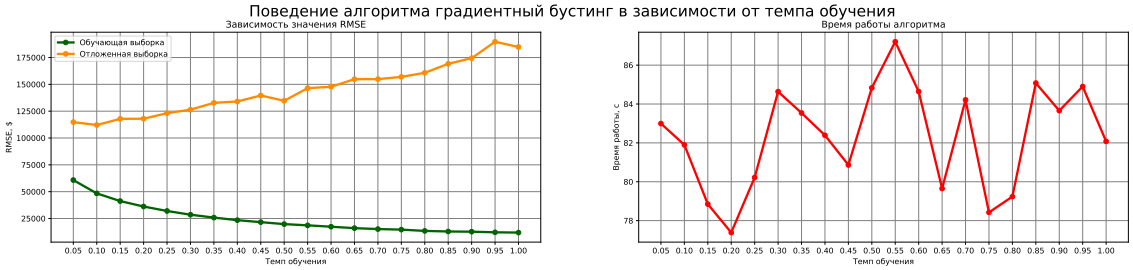


Рис. 8: Зависимость поведения алгоритма градиентный бустинг от темпа обучения

5 Выводы

Исследована зависимость времени работы и качества RMSE случайного леса и градиентного бустинга от параметров: число деревьев, размерность подвыборки признаков для одного дерева, максимальная глубина дерева, темп обучения (для градиентного бустинга). Градиентный бустинг показал качество лучше на валидационной выборке, чем случайный лес. Однако, в случае градиентного бустинга регулярно был риск переобучения, в отличие от случайного леса. Для случайного леса подбор параметров помог достичь более заметного изменения качества.

Приложение

Название столбца	Тип данных	Число уникальных значений
id	int	21436
price	float	4028
bedrooms	int	13
bathrooms	float	30
sqft_living	int	1038
sqft_lot	int	9782
floors	float	6
waterfront	int	5
view	int	5
condition	int	5
grade	int	12
sqft_above	int	946
sqft_basement	int	306
yr_built	int	116
yr_renovated	int	70
zipcode	int	70
lat	float	5034
long	float	752
sqft_living15	int	777
sqft_lot15	int	8689

Таблица 2: Информация о столбцах датасета

Список используемой литературы

[1] Александр Дьяконов «Анализ малых данных»: ГЛАВА XX. Ансамбли алгоритмов

[2] Александр Дьяконов «Введение в анализ данных и машинное обучение»: ГРАДИЕНТНЫЙ БУСТИНГ