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

Практикум 317 группы, ММП ВМК МГУ.

Морозов Иван Дмитриевич

Декабрь 2023

Содержание

1 Введение

Данное практическое задание посвящено исследованию свойств ансамблей и композиций алгоритмов в машинном обучении на примере случайного леса и градиентного бустинга в задаче предсказания стоимости квартиры.

Цели исследования:

- Написание реализаций вышеупомянутых алгоритмов.
- Изучение зависимости ошибки **RMSE** в зависимости от параметров алгоритмов.

2 Эксперимент №1. Предобработка данных.

Исходный датасет содержит информацию о параметрах квартир и их ценах. Целевой переменной является цена **price**. На графике (рис. 1) отображается корреляционная зависимость между целевой переменной и признаками. Видно, что всех меньше коррелирует с ценой **id**, это ожидаемо, поскольку идентификатор мало связан с ценой. Также очень слабая корреляция между длиной **long** и ценой, однако, на других датасетах эта зависимость может быть выше. С почтовым индексом **zipcode** корреляция немного выше, однако он мало даёт информации о цене квартиры, поэтому было принято решение удалить этот признак, как и идентификатор. **date** была переведена в специальный формат. Корреляция рассматривалась только по числовым признакам. Пропуски были заполнены средним значением. Выборка была разделена на обучающую и тестовую в соотношении 8 к 2. Данные переведены в `numpy.ndarray`.

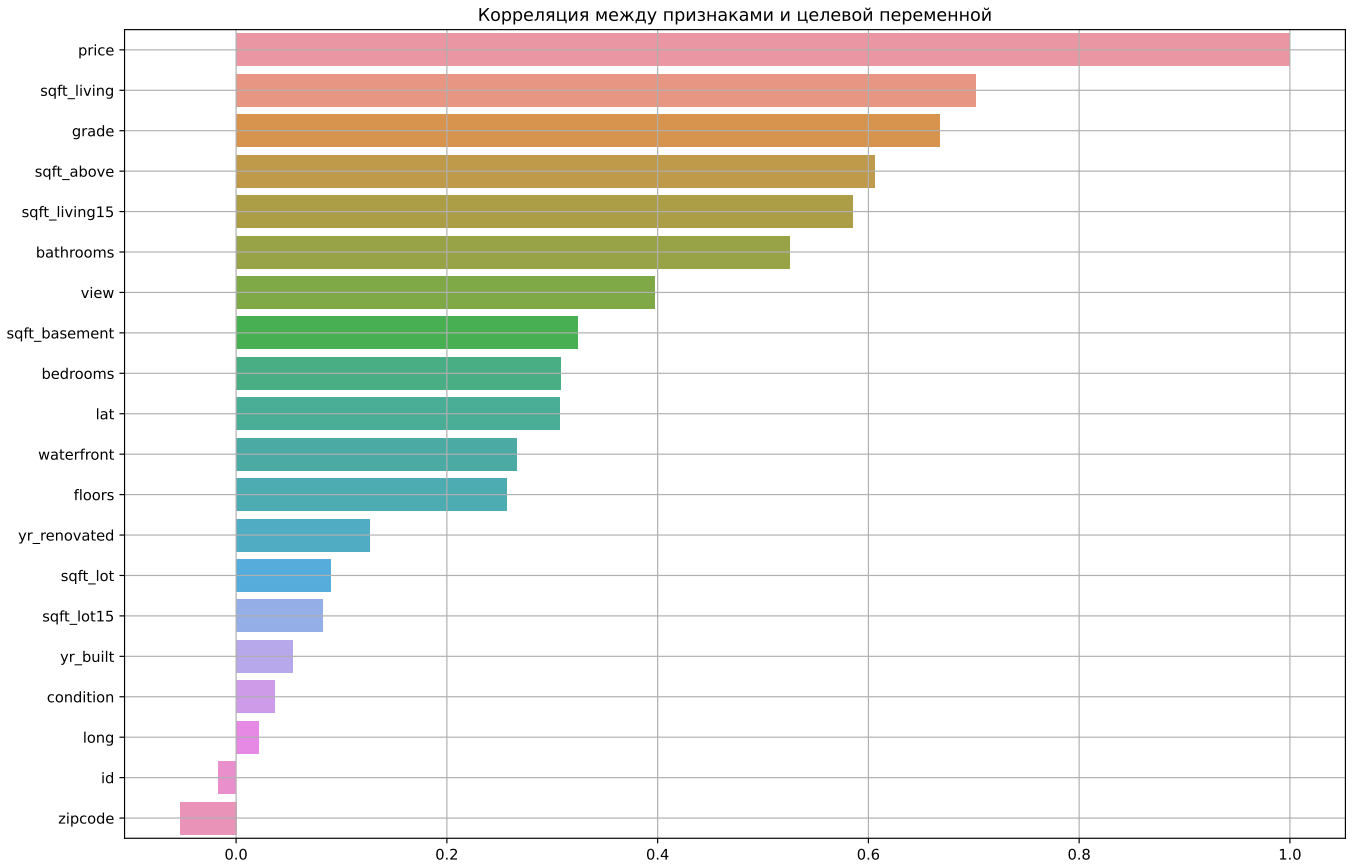

Рис. 1: Корреляция

3 Эксперимент №2. Исследование поведения алгоритма "Случайный лес".

В эксперименте изучается зависимость **RMSE** для алгоритма "Случайный лес"на тестовой выборке от количества деревьев в ансамбле, размерности подвыборки признаков для одного дерева, максимальной глубины дерева (она может быть и неограничена). Также замерялось время обучения и предсказания. Результаты приведены в таблице на рис. 2. Использовались следующие обозначения: **N_estimators** - количество деревьев в ансамбле, **Max_depth** - максимальная глубина дерева, значение **nan** соответствует отсутствию ограничения, **Feature_size** - количество признаков, **Training_time** - время обучения и предсказания (в мск).

3.1 Наблюдения и выводы

- При увеличении количества деревьев не происходит переобучения, что является достоинством алгоритма "Случайный лес".
- Наиболее малая ошибка достигается в случае, когда берётся большое число признаков. Это связано с тем, что признаков в датасете не очень много, каждый из них полезен.
- Время обучения увеличивается пропорционально числу деревьев, это ожидаемо.
- Наиболее низкие значения **RMSE** при прочих равных достигались при отсутствии ограничения на глубину дерева. Это связано с тем, что признаков мало.
- При прочих равных дальнейшее увеличение числа деревьев не приводит к уменьшению ошибки.

Зависимость RMSE для RandomForestMSE от параметров

N_estimators	Max_depth	Feature_size	Training_time	RMSE
10.0	7.0	2.0	1613.7	224852.2
10.0	15.0	2.0	1613.9	176596.9
10.0	30.0	2.0	1614.2	156481.3
10.0	nan	2.0	1614.6	160792.1
10.0	7.0	6.0	1614.7	205501.3
10.0	15.0	6.0	1615.0	146334.2
10.0	30.0	6.0	1615.4	142644.2
10.0	nan	6.0	1615.7	138632.6
10.0	7.0	12.0	1615.9	166859.1
10.0	15.0	12.0	1616.2	126577.9
10.0	30.0	12.0	1616.6	134685.2
10.0	nan	12.0	1617.1	132246.2
10.0	7.0	16.0	1617.3	159013.7
10.0	15.0	16.0	1617.6	126574.8
10.0	30.0	16.0	1618.1	125801.6
10.0	nan	16.0	1618.6	127682.6
50.0	7.0	2.0	1619.7	219631.8
50.0	15.0	2.0	1621.5	164725.6
50.0	30.0	2.0	1625.5	141826.9
50.0	nan	2.0	1630.1	142193.9
50.0	7.0	6.0	1631.4	196186.5
50.0	15.0	6.0	1633.9	140790.5
50.0	30.0	6.0	1638.7	135990.0
50.0	nan	6.0	1643.3	135228.1
50.0	7.0	12.0	1645.0	161896.1
50.0	15.0	12.0	1648.1	126474.4
50.0	30.0	12.0	1653.2	123211.5
50.0	nan	12.0	1658.4	122613.6
50.0	7.0	16.0	1660.0	157463.3
50.0	15.0	16.0	1663.4	124234.7
50.0	30.0	16.0	1668.9	120605.2
50.0	nan	16.0	1674.4	120123.1
100.0	7.0	2.0	1678.2	218978.6
100.0	15.0	2.0	1684.7	164886.5
100.0	30.0	2.0	1698.9	139538.3
100.0	nan	2.0	1715.4	140563.6
100.0	7.0	6.0	1719.4	198025.8
100.0	15.0	6.0	1728.0	143330.9
100.0	30.0	6.0	1744.0	132191.6
100.0	nan	6.0	1760.5	134756.8
100.0	7.0	12.0	1765.4	163707.9
100.0	15.0	12.0	1776.2	126137.3
100.0	30.0	12.0	1793.0	122657.8
100.0	nan	12.0	1810.1	121957.2
100.0	7.0	16.0	1815.1	153593.6
100.0	15.0	16.0	1826.3	122978.9
100.0	30.0	16.0	1844.0	120432.1
100.0	nan	16.0	1861.8	119467.1

Рис. 2: Зависимость случайного леса от параметров

4 Эксперимент №3. Исследование поведения алгоритма "Градиентный бустинг".

В эксперименте изучается зависимость **RMSE** для алгоритма "Градиентный бустинг"на тестовой выборке от количества деревьев в ансамбле, размерности подвыборки признаков для одного дерева, максимальной глубины дерева (она может быть и неограничена) и темпа обучения - **learning_rate**. Также замерялось время обучения и предсказания. Результаты приведены в таблице на рис. 3. Использовались следующие обозначения: **N_estimators** - количество деревьев в ансамбле, **Max_depth** - максимальная глубина дерева, значение **nan** соответствует отсутствию ограничения, **Feature_size** - количество признаков, **Learning_rate** - темп обучения, **Training_time** - время обучения и предсказания (в сек).

4.1 Наблюдения и выводы

- При прочих равных (например, если взять максимальную глубину 10, количество признаков 3, темп обучения 0.1) при меньшем количестве деревьев величина ошибки меньше, что говорит о том, что "Градиентный бустинг"склонен к переобучению, пусть и не очень большому.

- Алгоритм даёт меньшую ошибку, если брать не все признаки (например, если взять количество

Зависимость RMSE для GradientBoostingMSE от параметров

N_estimators	Max_depth	Feature_size	Learning_rate	Training_time	RMSE
10.0	10.0	3.0	0.1	0.1	631577.8
10.0	10.0	3.0	0.8	0.1	631600.8
10.0	50.0	3.0	0.1	0.4	631645.4
10.0	50.0	3.0	0.8	0.4	631666.9
10.0	nan	3.0	0.1	0.4	631661.3
10.0	nan	3.0	0.8	0.4	631673.8
10.0	10.0	16.0	0.1	0.2	631669.5
10.0	10.0	16.0	0.8	0.2	631671.1
10.0	50.0	16.0	0.1	0.6	631679.8
10.0	50.0	16.0	0.8	0.6	631680.1
10.0	nan	16.0	0.1	0.6	631680.3
10.0	nan	16.0	0.8	0.6	631681.2
50.0	10.0	3.0	0.1	3.0	631680.8
50.0	10.0	3.0	0.8	2.7	631681.0
50.0	50.0	3.0	0.1	6.6	631681.9
50.0	50.0	3.0	0.8	6.2	631681.8
50.0	nan	3.0	0.1	6.4	631681.9
50.0	nan	3.0	0.8	6.0	631681.9
50.0	10.0	16.0	0.1	3.7	631681.9
50.0	10.0	16.0	0.8	3.2	631681.9
50.0	50.0	16.0	0.1	7.7	631681.9
50.0	50.0	16.0	0.8	5.6	631681.9
50.0	nan	16.0	0.1	5.6	631681.9
50.0	nan	16.0	0.8	5.0	631681.9
200.0	10.0	3.0	0.1	41.5	631681.9
200.0	10.0	3.0	0.8	35.9	631681.9
200.0	50.0	3.0	0.1	34.2	631681.9
200.0	50.0	3.0	0.8	33.9	631681.9
200.0	nan	3.0	0.1	33.7	631681.9
200.0	nan	3.0	0.8	33.2	631681.9
200.0	10.0	16.0	0.1	44.0	631681.9
200.0	10.0	16.0	0.8	36.8	631681.9
200.0	50.0	16.0	0.1	33.9	631681.9
200.0	50.0	16.0	0.8	33.9	631681.9
200.0	nan	16.0	0.1	33.5	631681.9
200.0	nan	16.0	0.8	34.0	631681.9
500.0	10.0	3.0	0.1	215.9	631681.9
500.0	10.0	3.0	0.8	227.8	631681.9
500.0	50.0	3.0	0.1	232.1	631681.9
500.0	50.0	3.0	0.8	227.5	631681.9
500.0	nan	3.0	0.1	225.2	631681.9
500.0	nan	3.0	0.8	219.7	631681.9
500.0	10.0	16.0	0.1	219.7	631681.9
500.0	10.0	16.0	0.8	221.9	631681.9
500.0	50.0	16.0	0.1	220.5	631681.9
500.0	50.0	16.0	0.8	224.1	631681.9
500.0	nan	16.0	0.1	223.7	631681.9
500.0	nan	16.0	0.8	219.9	631681.9

Рис. 3: Зависимость градиентного бустинга от параметров

деревьев 10, максимальную глубину 10, темп обучения 0.1 и сравнить результат для 3 и 16 признаков).

- Время обучения увеличивается пропорционально числу деревьев, это ожидаемо.
- Наиболее низкие значения **RMSE** при прочих равных достигались при ограничении глубины деревьев, это связано с тем, что для градиентного бустинга подходят простые базовые модели (достаточно посмотреть на строки с количеством деревьев 10, количеством признаков 16, темпом обучения 0.1, видно, что для максимальной глубины 10 ошибка меньше, чем для 50).
- С ростом количества деревьев величина ошибки выходит на плато.
- Темп обучения, меньший единицы, является полезным способом улучшить качество алгоритма (достаточно сравнить 1 и 2 строку таблицы).
- Стоит заметить, что сетка перебора параметров была выбрана не очень удачно, что привело к тому, что получилось много близких значений **RMSE**.

5 Заключение

Отчёт посвящён ансамблевым алгоритмам. В результате проведения экспериментов и исследования особенностей поведения алгоритмов в зависимости от параметров были сделаны следующие ключевые выводы:

- Алгоритм "Случайный лес"устойчив к переобучению.
- Для случайного леса полезно брать как можно больше признаков для обучения и не ограничивать глубину деревьев.
- Алгоритм "Градиентный бустинг"использует параметр **Learning_rate**, это следует брать небольшим.
- Базовые модели градиентного бустинга должны быть простыми.

Список литературы

[1] Евгений Соколов, Курс «Машинное обучение», 2022, Лекции 8 - 11, https://youtube.com/playlist?list=PLEqoHzpnmTfChItexxg2ZfxCsm-8QPsdS&si=CkPTF9ZojuqASApx