Отчет о практическом задании «Ансамбли алгоритмов»

Практикум 317 группы, ММП ВМК МГУ

Борисов Иван Максимович

17.12.2023

Содержание

Аннотация

Данное практическое задание посвящено изучению ансамблевых методов машинного обучения и композиций алгоритмов, таких как случайный лес и градиентный бустинг, в задаче регрессии — предсказания стоимости квартиры. Главными целями исследования являются: написание методов случайного леса и градиентного бустинга и их применение в задаче предсказания стоимости квартиры; изучение поведения алгоритма случайного леса в зависимости от количества деревьев, размерности подвыборки признаков и максимальной глубины дерева; исследование поведения алгоритма градиентного бустинга в зависимости от количества деревьев, размерности подвыборки признаков, максимальной глубины дерева и длины шага обучения.

1 Эксперименты

1.1 Данные

Посмотрим на данные в таблице [1.1]:
Все данные имеют численный тип, значит предобработку категориальных признаков делать не нужно. Также в данных нет пропусков. Удалим из обучающей выборки поля *id, date, yr_built, yr_renovated, zipcode, lat, long*, так как большие значения данных коллонок не обязательно соответствуют большим значеням цены. Итого имеем 13 признаков.
Производить нормализацию/стандартизацию данных нет необходимости, так как модели построены на деревьях, в вершинах которых стоят предикаты вида $x_i < t$, где порог t подбирается по значеням признака x_i.
Разделим выборку на обучение и валидацию в отношении $8 : 2$ с $random\_state = 777$. Далее будем рассматривать качество модели и функцию потерь только на валидационной выборке.

	Column	Non-Null Count	Dtype
0	id	21613 non-null	int64
1	date	21613 non-null	object
2	bedrooms	21613 non-null	int64
3	bathrooms	21613 non-null	float64
4	sqft living	21613 non-null	int64
5	sqft lot	21613 non-null	int64
6	floors	21613 non-null	float64
7	waterfront	21613 non-null	int64
8	view	21613 non-null	int64
9	condition	21613 non-null	int64
10	grade	21613 non-null	int64
11	sqft above	21613 non-null	int64
12	sqft basement	21613 non-null	int64
13	yr built	21613 non-null	int64
14	yr renovated	21613 non-null	int64
15	zipcode	21613 non-null	int64
16	lat	21613 non-null	float64
17	long	21613 non-null	float64
18	sqft living15	21613 non-null	int64
19	sqft lot15	21613 non-null	int64

Таблица 1: Данные

1.2 Случайный лес

Рассмотрим зависимости лосс-функции $RMSE$ и метрики качества $MAPE = \frac{1}{n}\sum_{i=1}^{n} |\frac{a_N(x_i)-y_i}{y_i}|$ (где $a_N(x_i)$ — предсказание ансамля из N базовых алгоритмов на объекте x_i, y_i — верный ответ), а также время обучения, от следующих факторов:

- Количество деревьев в ансамбле
- Размерность подвыборки признакового постранства для базового алгоритма
- Максимальная глубина дерева

Ниже на приведенных графиках можно видеть, что наилучшее качество достигается при следующих параметрах: максимальное число деревьев (то есть алгоритм не переобучается), 9 или 13 признаков и отсутствие ограничений по глубине дерева (depth=None) или при достаточно больших значениях (depth=50). При этом время обучения существенно зависит от максимального числа базовых алгоритмов, но ближе к значению в 50 выходит на один уровень, то есть для данной задачи деревья глубины 50 приблизительно соответствуют деревьям, построенным без ограничений по глубине. Также уменьшение признакового пространства с 13 до 9 дает выигрыш по времени обучения в среднем на 3-4 секунды при большой глубине деревьев.

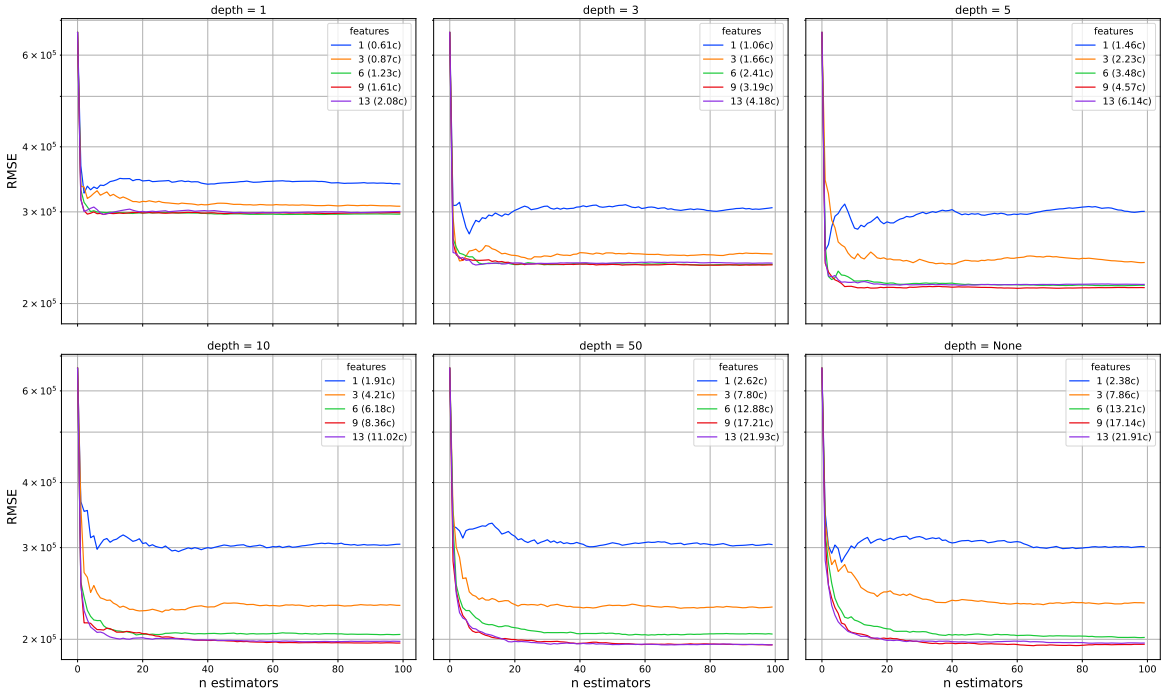

Рис. 1: Значение функции потерь от количества базовых алгоритмов, максимальной глубины дерева и числа признаков.

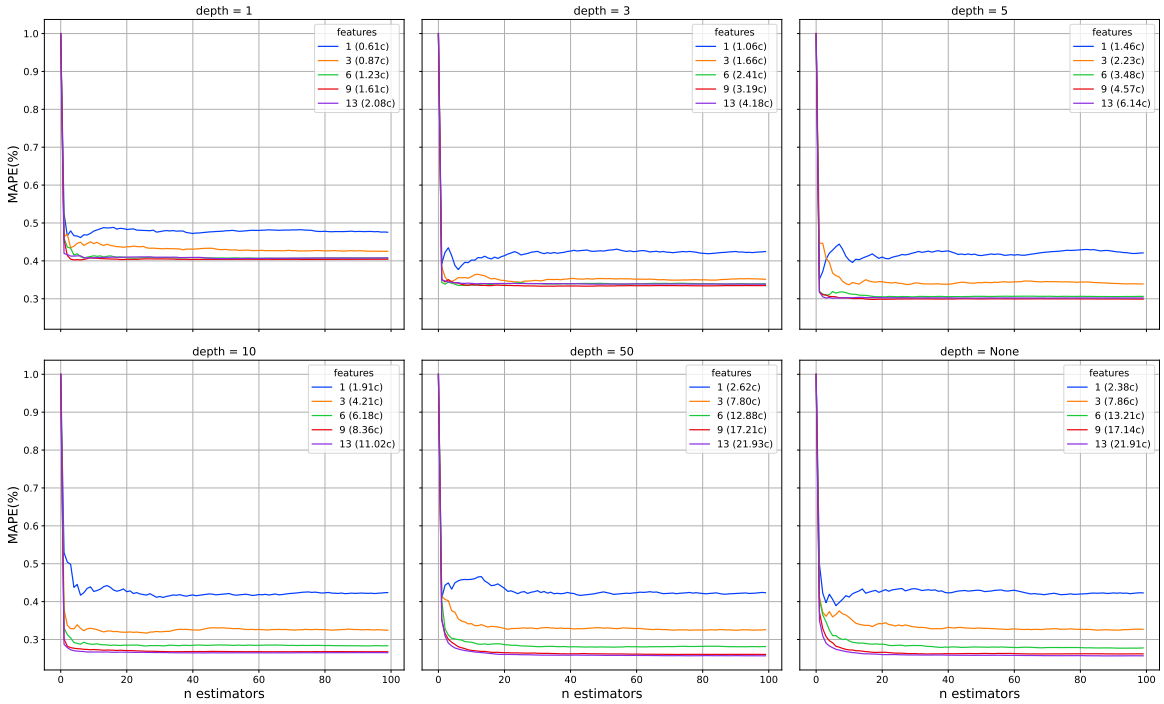

Рис. 2: Значение $MAPE$ от количества базовых алгоритмов, максимальной глубины дерева и числа признаков.

1.3 Градиентный бустинг

Рассмотрим зависимости лосс-функции $RMSE$ и метрики качества $MAPE = \frac{1}{n}\sum_{i=1}^{n}|\frac{a_N(x_i)-y_i}{y_i}|$ (где $a_N(x_i)$ — предсказание ансамля из N базовых алгоритмов на объекте x_i, y_i — верный ответ), а также время обучения, от следующих факторов:

- Количество деревьев в ансамбле
- Размерность подвыборки признакового пространства
- Длины шага обучения

Для начала зафиксрируем $learning\_rate = 0.1$ и переберем оставшиеся параметры.

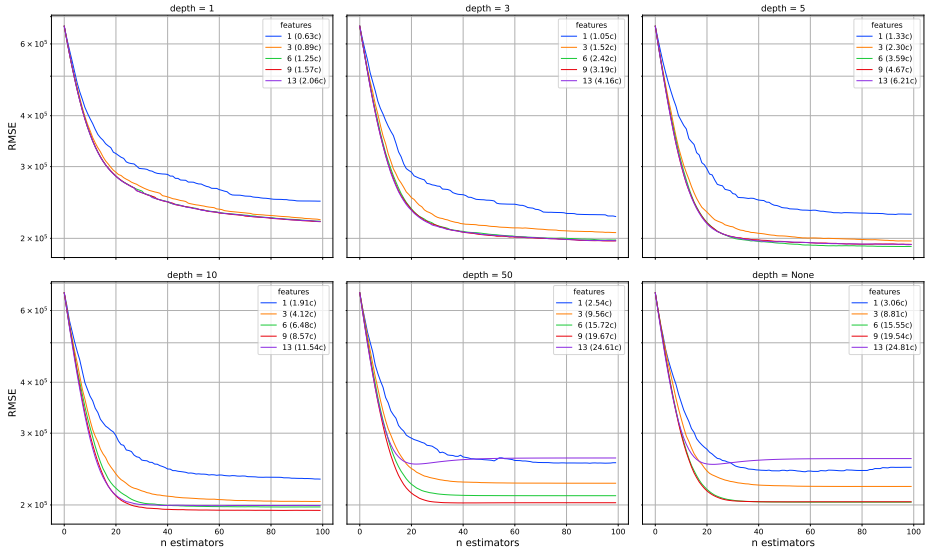

Рис. 3: Значение функции потерь от количества базовых алгоритмов, максимальной глубины дерева и числа признаков.

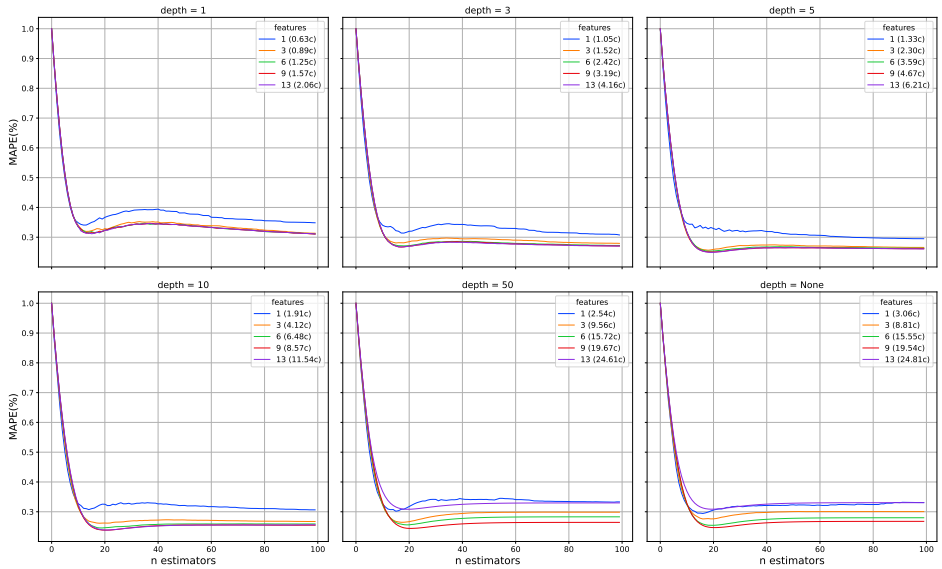

Рис. 4: Значение $MAPE$ от количества базовых алгоритмов, максимальной глубины дерева и числа признаков.

В отличие от случайного леса для бустинга оптимальной глубиной является $depth = 10$, при этом увеличение глубины приводит только к ухудшению качества модели. Число признаков, которое необходимо рассматривать - 9, качество модели также ухудшается, если брать все признаковое пространство. Что удивительно, для "глубоких"деревьев брать все признаковое пространство невыгодно даже по сравнению с 6 и 3 признаками, а местами даже и с 1 признаком. Также стоит отметить, что по графику качества модели можно судить о переобучении ансамбля, то есть оптимальным числом количества базовых алгоритмов в данной задаче является 20 штук. Время обучения для градиентного бустинга имеет те же тенденции, что и для случайного леса, при этом дополнительно можно сказать, что при одинаковых значениях параметров время обучения обоих видов алгоритмов примерно одинаковое.

Теперь зафиксируем максимальную глубину деревьев равную 10, и пусть каждое дерево обучается на случайных 9 признаках. Рассмотрим зависимость функции потерь от длины шага обучения.

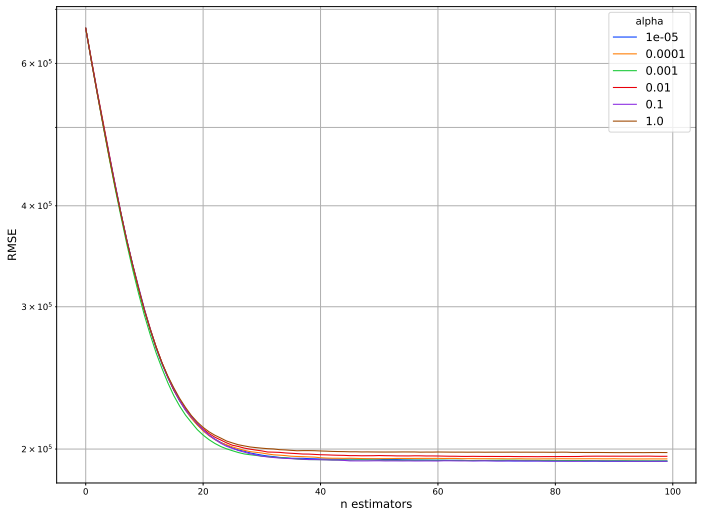

Рис. 5: Значение функции потерь от длины шага обучения

Как можно видеть, в данном случае разница между различными значениями несущественна. Оптимальным значением является 0.1, неоптимальным значением является 1.

2 Заключение

В ходе экспериментов были изучены случайный лес и градиентный бустинг в задаче прогнозирования стоимости квартир. В рамках исследования были написаны собственные реализации методов случайного леса и градиентного бустинга.

Для случайного леса проведено исследование влияния различных параметров на его эффективность. Установлено, что в данной задаче оптимальным является использование большей части признаков в подвыборке, а также отсутствие ограничений на глубину дерева.

Градиентный бустинг также был исследован на зависимости от параметров. Установлено, что увеличение количества базовых алгоритмов не приводит к повышению качества, противоположно случаю со случайным лесом. Также важно отметить, что для градиентного бустинга необходимо использовать не сильно переобученные базовые модели, то есть стоит ограничивать глубину базовых алгоритмов небольшими значениями - в данном случае 10.