Исследование методов построения ансамблей в задаче предсказания цен на дома

Богачев Владимир Александрович

19 декабря 2023 г.

1 Введение

Многие изученные ранее алгоритмы моделировали простые зависимости между признаковым описанием объекта и его целевой переменной. Методы, основанные на построении ансамблей алгоритмов, могут позволить более качественную и более устойчивую модель для решения описанных выше задач. В данной работе будет рассмотрена задача предсказания цен на дома. Для решения данной задачи будут использованы такие методы построения ансамблей, как градиентный бустинг и случайный лес.

2 Общее описание методов построения ансамблей

Рассмотрим следующую задачу регрессии: по данному семейству алгоритмов $\mathcal{F}$ по выборке $(x_i, y_i)_{i=1}^l \in (X \times Y)^l$ и функции потерь $\mathcal{L}$ необходимо построить стратегию $\mu : (X \times Y)^l \to \mathcal{F}$, удовлетворяющую следующему правилу: $\mu(x, y) = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \mathcal{L}(f(x), y)$. Предположим, что выборка $(x_i, y_i)_{i=1}^l$ порождается совместным распределением (X, Y). В силу данного предположения возможно оценить средний эмпирический риск для стратегии μ:

$$L(\mu) = \underbrace{\mathbb{E}_{x,y}\left[y - \mathbb{E}(y|x)\right]^2}_{\text{шум}} + \underbrace{\mathbb{E}_x\left[\mathbb{E}_X(\mu(X) - \mathbb{E}(y|x)\right]^2}_{\text{смещение (bias)}} + \underbrace{\mathbb{E}_x\,\mathbb{E}_X\left[\mu(X) - \mathbb{E}_X(\mu(X))^2\right]^2}_{\text{разброс(variance)}} \tag{1}$$

Для рассмотренных ранее алгоритмов имеет место `bias-variance tradeoff`. Стратегии, действующие в бедное параметрическое семейство, будут склонны к недообучению, другими словами, смещение данных алгоритмов будет велико, а смещение мало. Стратегии, действующие в богатое параметрическое семейство, наоборот, склонны к переобучению, другими словами, смещение данных алгоритмов будет мало, а разброс велик.

При построении ансамблей появляется возможность одновременно оптимизировать и смещение и разброс. Поэтому, можно предположить что при наличии сложной зависимости между признаковыми описаниями объектов x_i и целевыми переменными y_i методы, основанные на построении ансамблей будут давать более «качественные» алгоритмы.

3 Методы оценки качества полученных алгоритмов

В данной работе решается задача регрессии с квадратичной функцией потерь $\mathcal{L}(f, x, y) = \frac{1}{l}\sum_{i=1}^l (f(x_i) - y_i)^2$. Поэтому естественно оценивать качество модели по `MSE`.

`MSE` зависит от дисперсии целевой переменной, поэтому данная метрика качества может оказаться сложно интерпретируемой. Поэтому можно рассмотреть метрику $R^2(f, x, y) = 1 - \frac{\sum_{i=1}^l (f(x_i) - y_i)^2}{\sigma^2}$, где $\sigma^2 = \sum_{i=1}^l (y_i - \sum_{i=1}^l y_i^2)$

4 Построение ансамблей

В данной работе будут рассмотрены две стратегии построения ансамблей:

Бэггинг — независимое построение n алгоритмов, нацеленных на решение одной задачи. При построении необходимо минимизировать корреляцию между ошибками алгоритмов

Бустинг — построение последовательности алгоритмов, в который каждый следующий алгоритм нацелен на исправление ошибок суперпозиции предыдущих алгоритмов.

4.1 Описание случайного леса

Случайный лес — разновидность бэггинга. Данный алгоритм строит суперпозицию решающих деревьев. Каждое дерево обучается по подвыборке, полученной методом бутстрапа из исходной обучающей выборки. При обучении каждой вершины дерева рассматриваются признаки, взятые из случайного подмножества признакового описания.

Для дальнейшего анализа введем следующие обозначения:

- $\mathcal{B}$ — класс базовых алгоритмов (решающие деревья)
- $\mathcal{A} = L(\mathcal{B})$ — класс линейных комбинаций базовых алгоритмов
- $\hat{b} : (X \times Y)^l \to \mathcal{B}$ — стратегия построения базового алгоритма по выборке
- $\hat{a} : (X \times Y)^l \to \mathcal{A}$

Пусть базовый алгоритм $\hat{b}_i(X, Y)$ — дерево, обученное по подвыборке $(\widetilde{X}, \widetilde{Y})$, полученной из (X, Y) методом бутсрапа. Причем номер i однозначно определяет преобразование $(X, Y) \to (\widetilde{X}, \widetilde{Y})$. Тогда можно получить соотношение на средние алгоритмы, построенные по данным стратегиям:

$$\mathbb{E}_{(X,Y)}\hat{b}_1(X, Y) = \ldots = \mathbb{E}_{(X,Y)}\hat{b}_n(X, Y) = \mathbb{E}_{(X,Y)}\hat{b}(X, Y)$$

Тогда можно получить выражение для смещения алгоритма $\mathtt{a}_n$:

$$\mathtt{bias}(a_n) = \mathbb{E}_x \left[\mathbb{E}_X \left(\frac{1}{n} \sum_{i=1}^{n} b_i(X)(x) - \mathbb{E}(y|x) \right) \right]^2 = \mathbb{E}_x \left[\mathbb{E}_{(X,Y)}\hat{b}(X, Y)(x) - \mathbb{E}(y|x) \right]^2 \tag{2}$$

Из (2) можно сделать вывод, что смещение случайного леса не зависит от количества базовых алгоритмов в ансамбле. Данная компонента зависит исключительно от силы класса $\mathcal{B}$.

Получим аналогичное соотношение для разброса:

$$\begin{aligned}
&\mathtt{variance}(a_n) = \mathbb{E}_x \mathbb{E}_X \left[\frac{1}{n} \sum_{i=1}^{n} b_i(X, Y)(x) - \mathbb{E}_{(X,Y)} \left(\frac{1}{n} \sum_{i=1}^{n} b_i(X, Y)(x) \right) \right]^2 = \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E}_x \mathbb{E}_X \left(b_i(X, Y)(x) - \mathbb{E}_{(X,Y)} b_i(X, Y)(x) \right)^2 + \\
&+ \frac{1}{n^2} \sum_{i \neq j} \left(b_i(X, Y)(x) - \mathbb{E}_{(X,Y)} b_i(X, Y)(x) \right) \left(b_i(X, Y)(x) - \mathbb{E}_{(X,Y)} b_i(X, Y)(x) \right) = \\
&= \frac{1}{n} \mathtt{variance}(b) + \widetilde{cov}(b_1, \ldots, b_n)
\end{aligned} \tag{3}$$

Из формулы (3) можно сделать вывод, что смещение случайного леса раскладывается на $\frac{1}{n}\mathtt{variance}(b)$ и «ковариацию» предсказаний базовых алгоритмов. При увеличении $n \to \infty$ будет $\frac{1}{n}\mathtt{variance}(b) \to 0$. Следовательно, необходимо минимизировать «ковариацию» предсказаний базовых алгоритмов.

4.2 Градиентный бустинг

Градиентный бустинг — разновидность бустинга. Данный алгоритм строит композицию решающих деревьев по следующему итерационному алгоритму:

0. Пусть построена композиция a_{n-1} по выборке (x, y).
1. Вычисляется градиент ошибки алгоритма по выборке $S_n = \nabla_x \mathcal{L}(a_{n-1}, x, y)$
2. Строится базовый алгоритм $\widetilde{b}_n = \underset{b \in \mathcal{B}}{\operatorname{argmin}} (b(x) - S_n)^2$
3. Строится базовый алгоритм $b_n = \gamma_n \alpha_n \widetilde{b}_n$, где γ_n — шаг обучения, а α_n — решение одномерной задачи оптимизации $\alpha_n = \underset{\alpha \in \mathbb{R}}{\operatorname{argmin}} (\alpha_n \widetilde{b}_n - S_n)^2$
4. Строится композиция $a_n = a_{n-1} + b_n$

Данный алгоритм минимизирует смещение. В работе [1] было показано, что данная стратегия не склонна к переобучению. Таким образом, градиентный бустинг является, пожалуй, одним из самых сильных алгоритмов машинного обучения.

5 Выбор оптимальных параметров

Для дальнейшего исследования полученных алгоритмов необходимо подобрать оптимальные гиперпараметры, а также изучить зависимость качества полученных моделей от выбора гиперпараметров.

5.1 Выбор оптимальных параметров для случайного леса

Будем рассматривать следующие гиперпараметры случайного леса:

- Количество базовых моделей
- Объем случайной выборки признаков, рассматриваемых при обучении вершины дерева
- Максимальная глубина решающего дерева

Исходя из теоретических ошибок смещения (2) и разброса (3) случайного леса, можно сделать предположение, что ошибка алгоритма будет не убывать при увеличении количества базовых алгоритмов.

Увеличение сложности базового алгоритма приводит к уменьшению среднего смещения, а следовательно и разброса. Глубокие деревья становятся склонны к переобучению, поэтому разброс каждого базового алгоритма увеличивается. С увеличением разброса можно бороться увеличением количества базовых алгоритмов. Поэтому при увеличении максимальной глубины дерева ошибка ансамбля, состоящего из большого количества моделей, будет не убывать

При увеличении объема случайной выборки признаков, рассматриваемых при обучении вершины дерева, будет увеличиваться сложность дерева. Следовательно, смещение будет не убывать. Однако с ростом данного гиперпараметра будет уменьшаться «разнообразие» моделей, то есть будет увеличиваться коэффициент корреляции между предсказаниями моделей. Следовательно, будет увеличиваться разброс ансамбля. Таким образом, возможно выдвинуть лишь гипотезу об унимодальности зависимости качества модели от объема выборки признаков, рассматриваемых при обучении вершины дерева. В [2] утверждается, что оптимальный объем выборки признаков для задачи регрессии соответствует $\frac{1}{3}$ от общего числа признаков.

5.1.1 Постановка вычислительного эксперимента

Для проведения вычислительного эксперимента будет использован следующий датасет: House Sales in King County, USA. Будем изучать зависимость качества модели от выбора гиперпараметров в следующем порядке:

1. Зависимость качества модели от количества базовых алгоритмов в ансамбле, при неограниченной максимальной глубине дерева и объеме выборки признаков $= \frac{1}{3}$ от количества всех признаков
2. Зависимость качества модели от максимальной глубины дерева при количестве базовых алгоритмов, соответствующем оптимальному значению и объеме выборки признаков $= \frac{1}{3}$ от количества всех признаков признак
3. Зависимость качества модели от объема выборки признаков при выборе оптимальных гиперпараметров

5.1.2 Анализ результатов вычислительного эксперимента

Зависимость качества модели от количества базовых алгоритмов изображена на рис. 1. Наблюдается близкая к монотонной зависимость, значения метрик стабилизируются после примерно 250 базовых алгоритмов в ансамбле. Следовательно, гипотеза о характере данной зависимости подтверждена.

Зависимость качества модели от числа признаков, рассматриваемых при обучении вершины изображена на рис. 2. Данная зависимость близка к монотонной, следовательно, гипотеза о характере данной зависимости не подтверждена.

Зависимость качества модели от максимальной глубины решающего дерева изображена на рис. 3. Данная зависимость является монотонной, наилучшее качество достигается при неограниченной глубине. Следовательно, гипотеза о характере данной зависимости подтверждена.

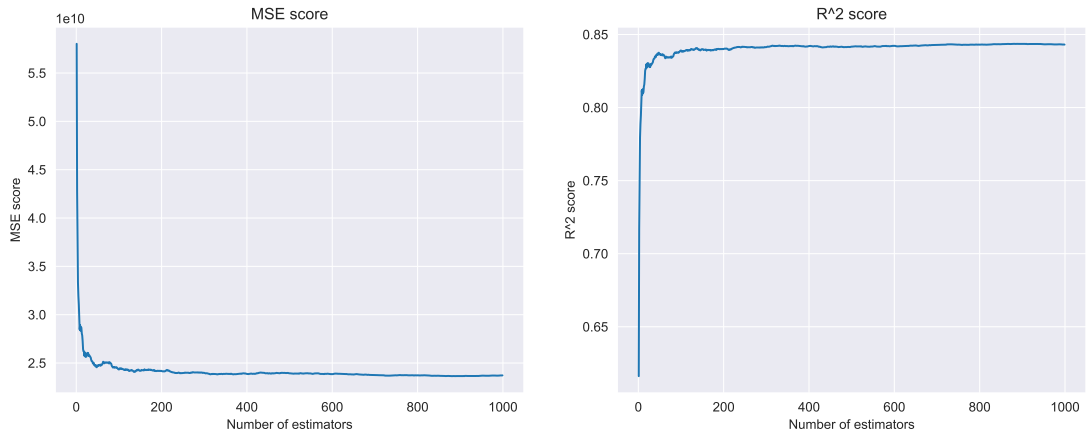

Рис. 1: Зависимость качества модели от количества базовых алгоритмов

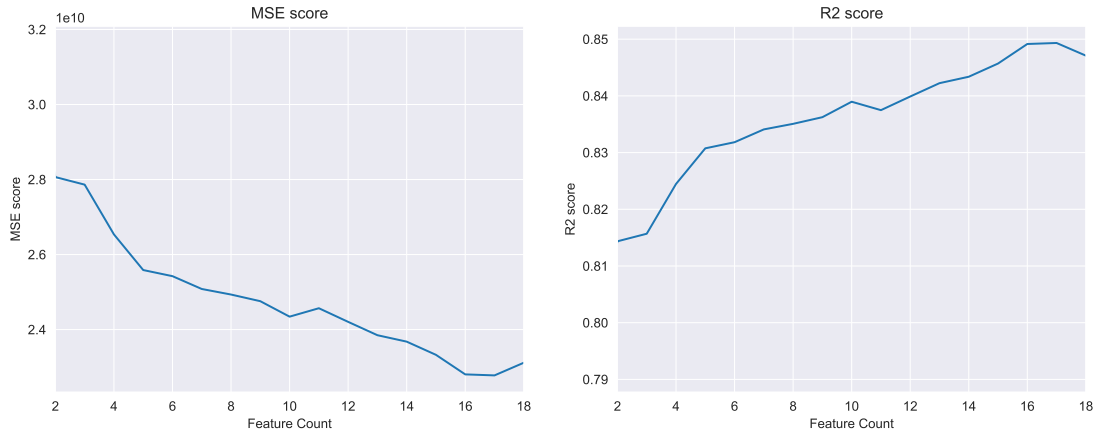

Рис. 2: Зависимость качества модели от объема выборки признаков, рассматриваемых при обучении вершины

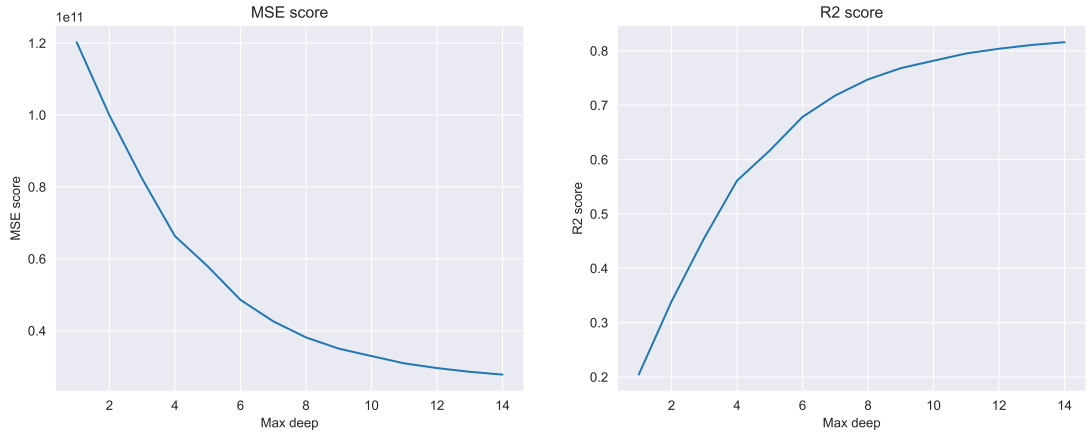

Рис. 3: Зависимость качества модели от максимальной глубины решающего дерева

5.2 Выбор оптимальных параметров для градиентного бустинга

Будем рассматривать следующие гиперпараметры градиентного бустинга:

- Количество базовых моделей
- Объем случайной выборки признаков, рассматриваемых при обучении вершины дерева
- Максимальная глубина решающего дерева
- Шаг обучения

Так как каждый следующий базовый алгоритм нацелен на исправление ошибок ансамбля, то при увеличении количества базовых алгоритмов в ансамбле смещение модели будет убывать. В работе [1] было показано, что градиентный бустинг не склонен к переобучению при увеличении числа базовых алгоритмов. Следовательно, качество ансамбля должно возрастать при увеличении числа базовых алгоритмов.

При использовании в качестве $\mathcal{B}$ богатого множества, базовый алгоритм может слишком сильно переобучиться, получив избыточное влияние в ансамбле. Возможны случаи, когда первый алгоритм «выучивает» всю обучающую выборку, в таком случае все последующие базовые алгоритмы обучаются на предсказание тождественного нуля. Следовательно, нужно строить ансамбль из слабых базовых алгоритмов. Поэтому, максимальная глубина решающего дерева может оказать сильное влияние на качество итогового алгоритма.

5.2.1 Постановка вычислительного эксперимента

Для проведения вычислительного эксперимента будет использован следующий датасет: House Sales in King County, USA. Будем изучать зависимость качества модели от выбора гиперпараметров в следующем порядке:

1. Зависимость качества модели от количества базовых алгоритмов в ансамбле, при неограниченной максимальной глубине дерева, объеме выборки признаков $= \frac{1}{3}$ от количества всех признаков и при шаге обучения 0.05
2. Зависимость качества модели от максимальной глубины дерева при количестве базовых алгоритмов, соответствующем оптимальному значению, объеме выборки признаков $= \frac{1}{3}$ от количества всех признаков и при шаге обучения 0.05
3. Зависимость качества модели от объема выборки признаков при оптимальном выборе описанных выше гиперпараметров и при шаге обучения 0.05

5.2.2 Анализ результатов вычислительного эксперимента

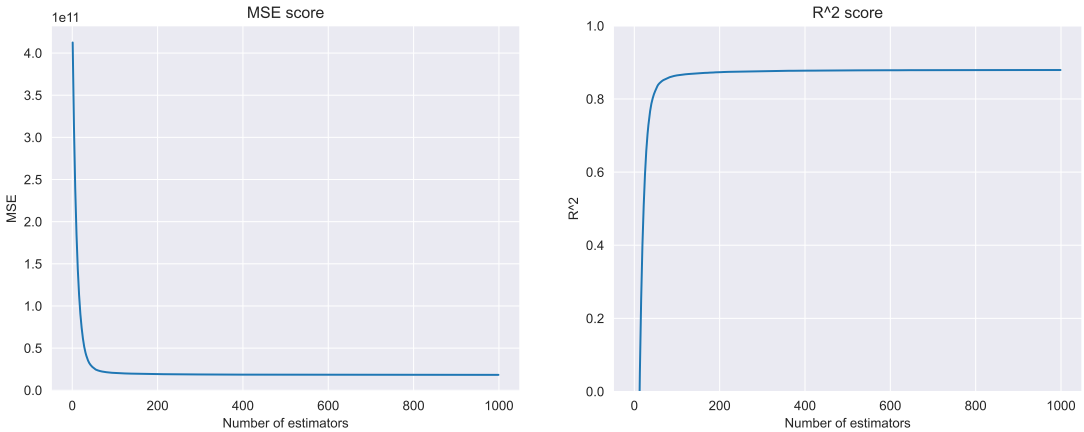

Рис. 4: Зависимость качества модели от количества базовых алгоритмов

Зависимость качества модели от количества базовых алгоритмов изображена на рис. 4. Наблюдается близкая к монотонной зависимость, значения метрик стабилизируются после примерно 300 базовых алгоритмов в ансамбле. Следовательно, гипотеза о характере данной зависимости подтверждена.

Зависимость качества модели от числа признаков, рассматриваемых при обучении вершины изображена на рис. 2. При объеме выборки ≥ 8 качество стабилизируется. Следовательно, данные ограничения на класс $\mathcal{B}$ не привносят существенных изменений.

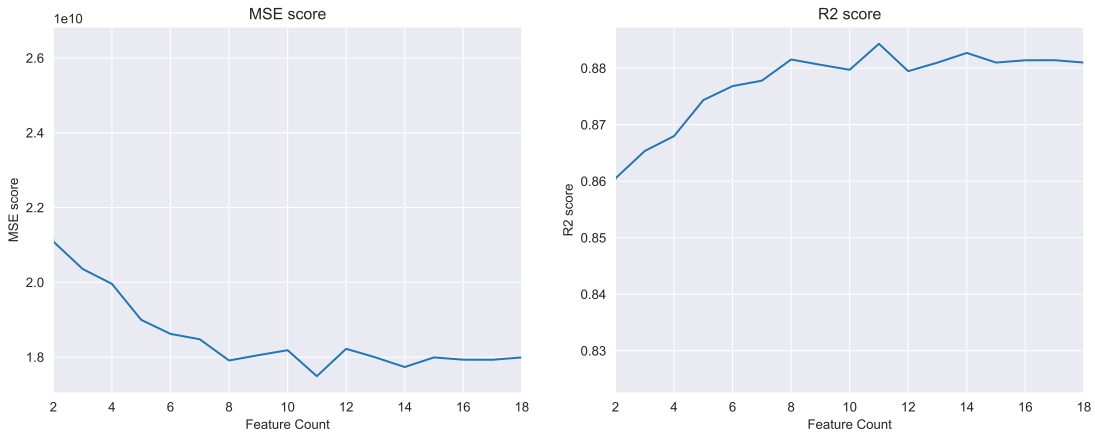

Рис. 5: Зависимость качества модели от объема выборки признаков, рассматриваемых при обучении вершины

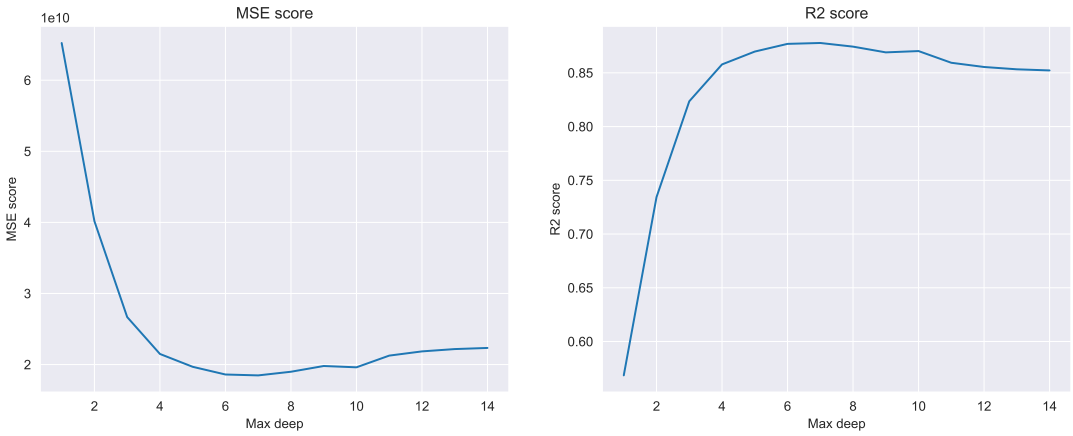

Рис. 6: Зависимость качества модели от максимальной глубины решающего дерева

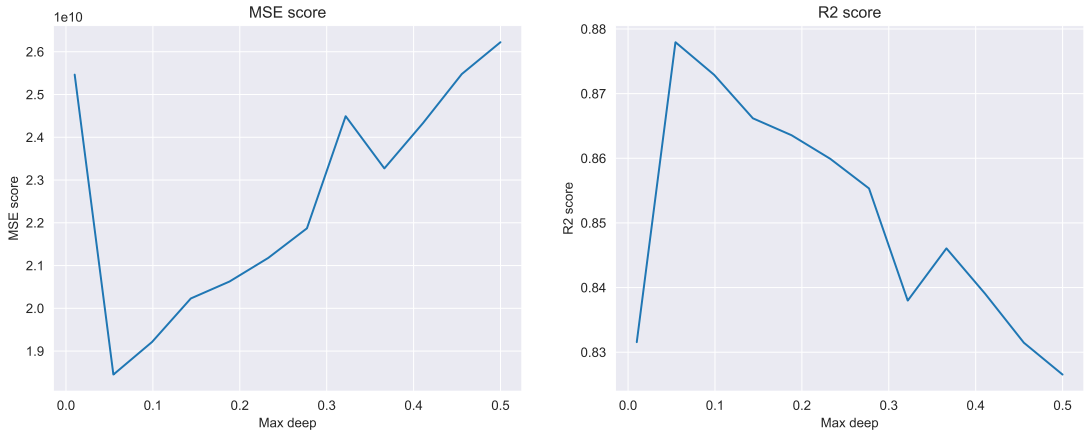

Рис. 7: Зависимость качества модели от шага обучения

Зависимость качества модели от максимальной глубины решающего дерева изображена на рис. 6. Данная зависимость близка к унимодальной. Наилучшее качество достигается при глубине 7.

Зависимость качества модели от шага обучения изображена на рис. 7. Оптимальное значение шага нахо-

Модель	RMSE	R^2
Случайный лес	154095	0.84
Градиентный бустинг	137418	0.88
Ridge	212120	0.70
SVM	230030	0.65

Таблица 1: Caption

Модель	Смещение	Разброс	Ошибка
Случайный лес	0.42	0.60	0.44
Градиентный бустинг	0.31	1.0	0.33
Ridge	0.87	0.09	0.88
SVM	1.0	0.05	1.0

Таблица 2: Разложение ошибки на смещение и разброс

дится в окрестности точки 0.05.

6 Сравнение алгоритмов

В данном разделе будет произведено сравнение ансамблей, построенных в предыдущих разделах, с изученными ранее моделями Ridge и SVM регрессии.

Для оценки качества моделей будут использованы не только метрики R^2, MSE, но и анализ разложения среднего эмпирического риска на шум, смещение и разброс.

В таблице 1 представлены значения метрик качества для исследуемых моделей. В таблице 2 представлено разложение ошибок исследуемых моделей на смещение и разброс. Столбцы 2 нормированы на максимальное значение.

Наилучшее качество достигается при использовании градиентного бустинга. Данная модель обеспечивает наименьшее смещение среди всех исследуемых моделей. Разброс данной модели максимальный среди исследуемых моделей. На втором месте находится случайный лес. Смещение данной модели выше, чем у градиентного бустинга, разброс меньше.

Модели, не основанные на построении ансамблей (Ridge, SVM), имеют значительно более высокое смещение, разброс данных моделей мал. Из этих данных можно сделать вывод, что данные линейные модели плохо моделируют зависимость между признаковым описанием объекта и целевой переменной.

В таблице 3 показано, что основной вклад в ошибку вносит смещение. Это свидетельствует о высокой сложности зависимости между признаковым описанием объекта и целевой переменной. Превосходство моделей, основанных на построении ансамблей, над моделями, не использующими ансамбли, не опровергает выдвинутую гипотезу об эффективности моделирования сложных зависимостей с помощью ансамблей.

7 Выводы

В данной работе были изучены такие методы построения ансамблей, как случайный лес и градиентный бустинг. Было показано, что в задаче предсказания цен на дома зависимость целевой переменной от признакового описания сложная. Было показано, что с данной задачей лучше справляются методы, основанные на построении ансамблей. Также было показано, что в данной задаче градиентный бустинг, несмотря на высокую сложность семейства $\mathcal{A}$, переобучается незначительно.

Список литературы

[1] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[2] К. В. Воронцов. Линейные ансамбли.
https://github.com/MSU-ML-COURSE/ML-COURSE-22-23/blob/main/slides/msu21-compos1.pdf.

A Ненормированная таблица разложения ошибки на смещение и разброс

Модель	Смещение	Разброс	Ошибка
Случайный лес	19479336077	769668197	20553121443
Градиентный бустинг	14238006506	1274072736	15542559460
`Ridge`	40542568762	117740838	40624449347
`SVM`	46365351191	68777376	46253483394

Таблица 3: Разложение ошибки на смещение и разброс