Московский государственный университет
имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математических методов прогнозирования

Отчет по заданию №3

«Ансамбли алгоритмов. Веб-сервер. Композиции алгоритмов для решения задачи регрессии.»

Практикум 317 группы

Выполнил:
студент 317 группы
Абрамов В. А.

Москва
2022

Содержание

Постановка задачи **2**

Эксперименты **2**

Предобработка данных 2

Исследование поведения алгоритма случайный лес 2

Исследование поведения алгоритма градиентный бустинг 4

Выводы **7**

Постановка задачи

Необходимо ознакомиться с методами ансамблирования.

Для ознакомления с ансамблевыми методами требуется написать собственную реализацию алгоритмов **случайный лес** и **градиентный бустинг** на языке **Python**.

Затем необходимо провести эксперименты, используя датасет данных о продажах недвижимости **House Sales in King County, USA**.

Эксперименты

Предобработка данных

Проведём предобработку имеющихся данных.

1. В данных отсутствуют пропуски.
2. В данных 21 признак: 20 численных и признак “date” типа object. Переведем признак “date” в тип datetime и извлечем из него день, месяц и год. Добавим их в датасет, удалим признак “date”.
3. Признак “id” следует удалить, так как из-за него модель может переобучиться.
4. Признак “price” – наш таргет. Создадим из него вектор таргета и удалим из обучающей выборки.
5. Переведем обучающую выборку и вектор таргета в numpy ndarray.
6. Разделим данные на обучение и контроль в соотношении 7:3.

Исследование поведения алгоритма случайный лес

Изучим зависимость RMSE на отложенной выборке и время работы алгоритма сначала только от количества деревьев, а потом от следующих гиперпараметров:

- количество деревьев в ансамбле: 10, 50, 100, 300, 500;
- размерность подвыборки признаков для одного дерева: 0.1, 0.3, 0.5, 0.7, 1.0;
- максимальная глубина дерева: 1, 4, 7, 10, 13, 16, неограниченная.

Рассмотрим RMSE на отложенной выборке и время обучения в зависимости от количества деревьев в ансамбле при неограниченной максимальной глубине и доле размерности признакового пространства равной $\frac{1}{3}$:

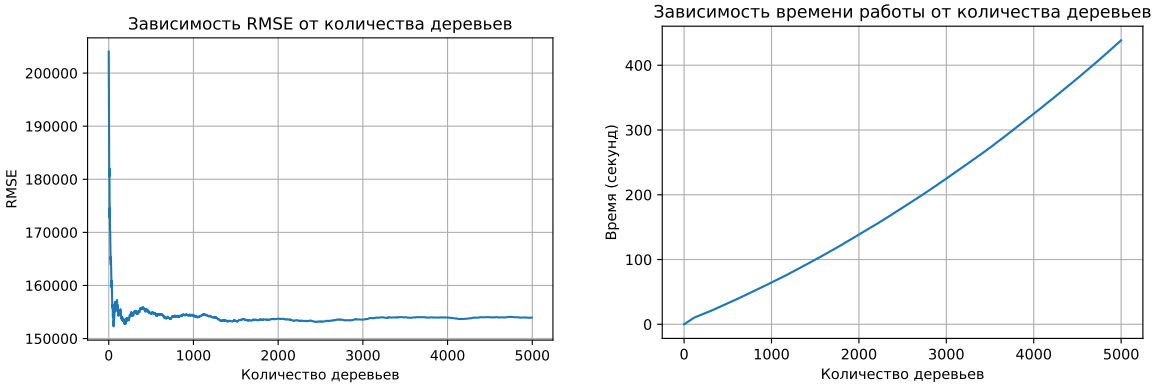


Таблица 1: RMSE выводится для количества дереьев ≥ 5

Из графика для RMSE видим, что эффект переобучения не наблюдается при таком большом количестве деревьев. Это объясняется тем, что голосование алгоритмов не влияет на bias в BVD разложении, а влияние на variance уменьшается при скоррелированности базовых алгоритмов. Время обучения растёт нелинейно, но близко к линейному.

Теперь рассмотрим зависимость RMSE на отложенной выборке и времени обучения от размерности подвыборки признаков для одного дерева при неограниченной максимальной глубине:

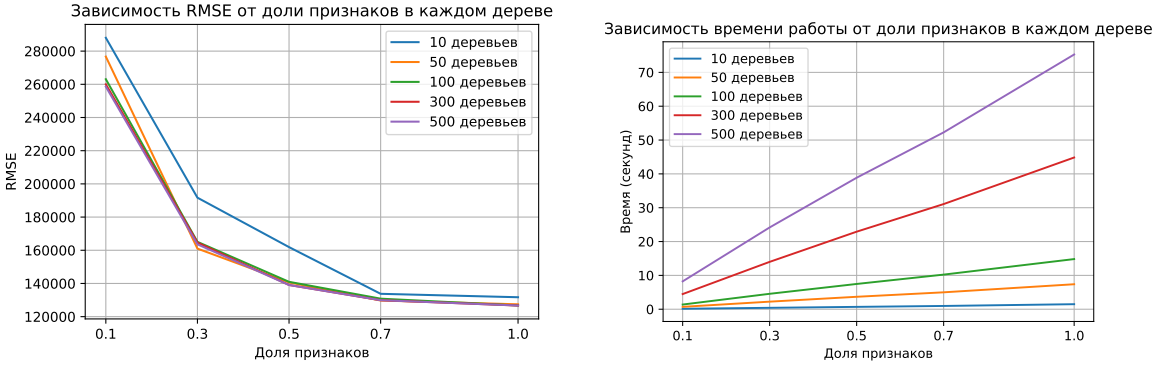
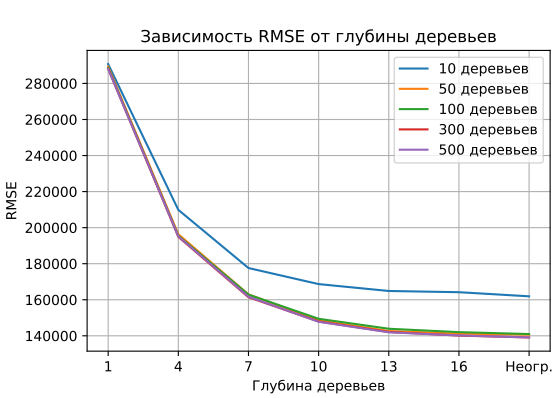


Таблица 2: RMSE и время работы для различных долей размерности признаков.

Видим, что при отсутствии ограничений на глубину деревьев увеличение размерности признакового пространства ведет к уменьшению RMSE, но происходит это ценой времени обучения. При этом время растет линейно, а качество улучшается нелинейно. В связи с этим оптимальным значением доли признаков я бы назвал 0.5.

Теперь рассмотрим RMSE на отложенной выборке и время обучения в зависимости от глубины одного дерева при доле размерности признакового пространства 0.5:



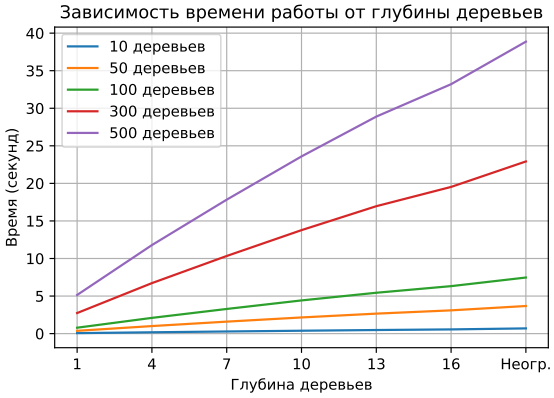


Таблица 3: RMSE и время работы для различной глубины деревьев.

Предыдущая ситуация повторилась: при увеличении глубины деревьев получаем улучшение качества, но время растет линейно, а RMSE уменьшается нелинейно. В связи с этим оптимальным значением глубины я бы выбрал 13.

Из экспериментов видны следующие закономерности для случайного леса:

- отсутствие переобучения при большом числе базовых алгоритмов;
- увеличение таких гиперпараметров, как глубина дерева и размерность признакового пространства также не приводит к переобучению;
- использование слишком больших значений гиперпараметров неоптимально, так как это сильно замедляет процесс обучения и дает небольшой прирост качества.

Исследование поведения алгоритма градиентный бустинг

Изучим зависимость RMSE на отложенной выборке и время работы алгоритма сначала только от количества деревьев, а потом от следующих гиперпараметров:

- количество деревьев в ансамбле: 50, 100, 300, 500;
- размерность подвыборки признаков для одного дерева: 0.1, 0.3, 0.5, 0.7, 1.0;
- максимальная глубина дерева: 1, 4, 7, 10, 13, 16, неограниченная.
- темп обучения: 10^{-3}, 10^{-2}, 10^{-1}, 1

Рассмотрим RMSE на отложенной выборке и время обучения в зависимости от количества деревьев в ансамбле при максимальной глубине равной 5, доле размерности признакового пространства равной $\frac{1}{3}$ и темпе обучения равном 10^{-1}:

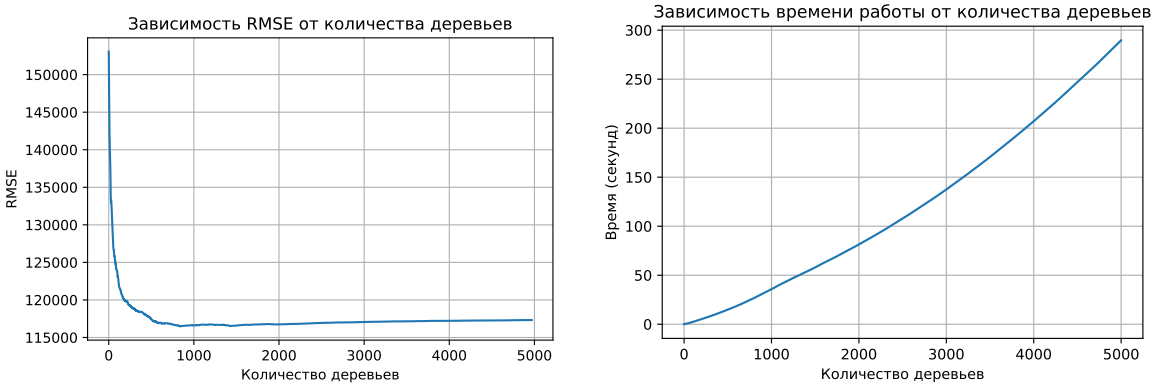


Таблица 4: RMSE выводится для количества дереьев ≥ 30

Из графика для RMSE видим, что эффект переобучения присутствует, но в слабой форме. Время обучения так же, как и у случайного леса, растёт нелинейно, но близко к линейному.

Теперь рассмотрим RMSE на отложенной выборке и время обучения в зависимости от размерности подвыборки признаков для одного дерева при максимальной глубине 4 и темпе обучения равном 10^{-1}:

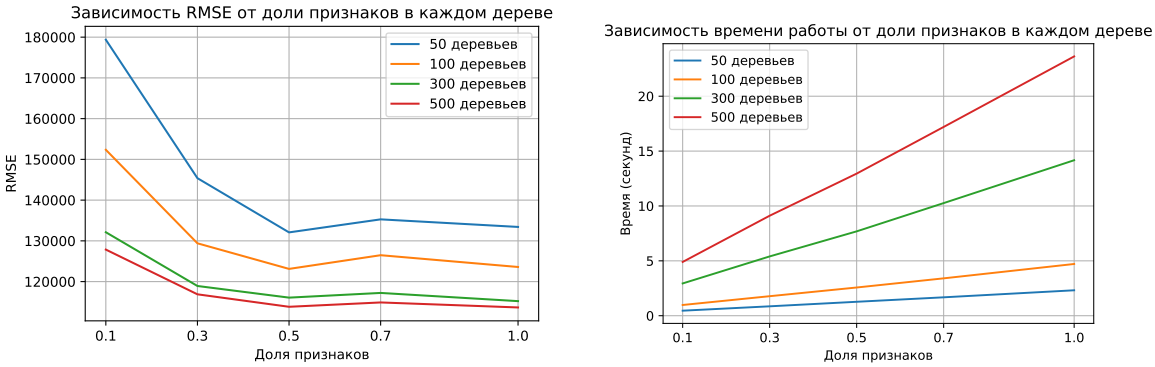


Таблица 5: RMSE и время работы для различных долей размерности признаков.

Видим немонотонную зависимость RMSE от доли признаков, оптимальным значением является 0.5. Время работы алгоритма линейно зависит от этого гиперпараметра.

Теперь рассмотрим зависимость RMSE на отложенной выборке и времени обучения от глубины одного дерева при темпе обучения равном 10^{-1} и доле размерности признакового пространства 0.5:

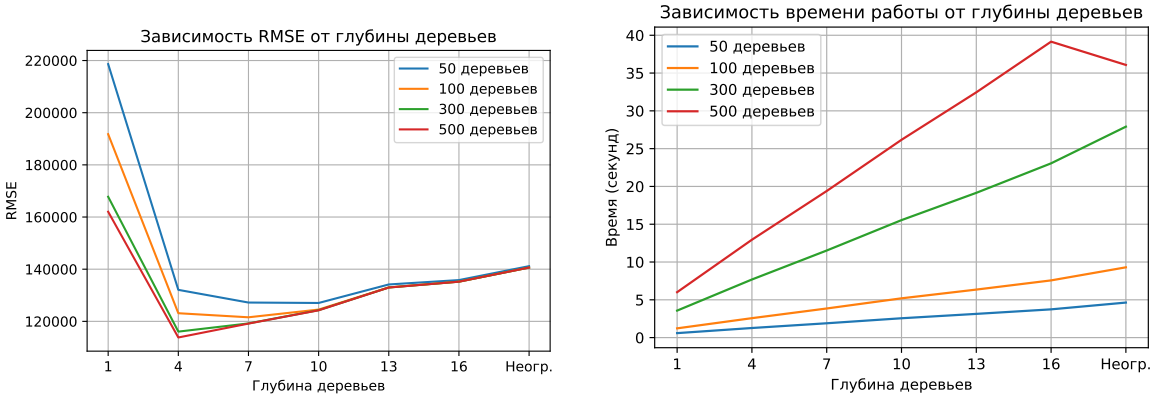


Таблица 6: RMSE и время работы для различной глубины деревьев.

Снова получаем немонотонную зависимость RMSE от значений гиперпараметра: оптимальной глубиной для числа деревьев, большего 300, является 4. При меньшем числе деревьев оптимальными являются большие значения. Зависимость от времени линейная за исключением снижения времени обучения алгоритма при количестве базовых алгоритмов равном 500. Это может быть связано с особенностью работы решающих деревьев из библиотеки **sklearn** – например, ранней остановкой.

Рассмотрим RMSE на отложенной выборке и время обучения в зависимости от темпа обучения при максимальной глубине 4 и доле размерности признакового пространства 0.5:

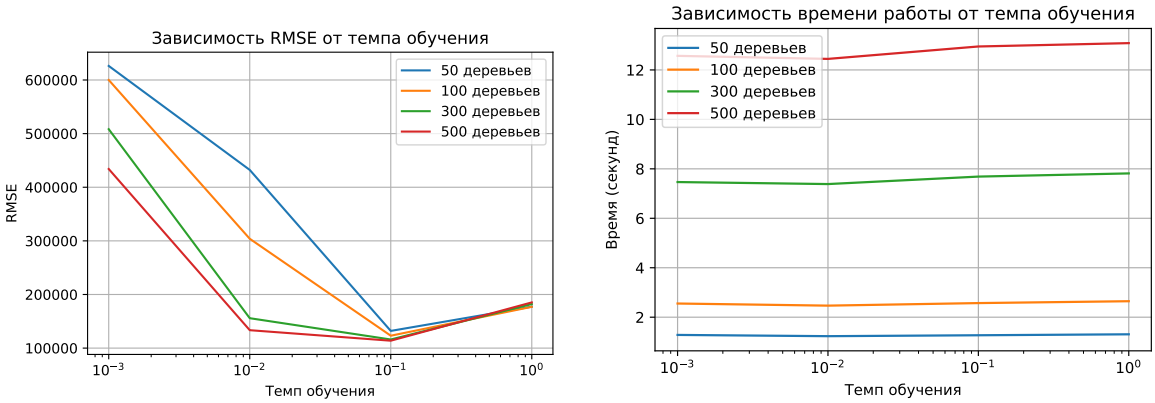


Таблица 7: RMSE и время работы для различных темпов обучения.

Слишком маленький темп обучения негативно влияет на качество, так как ансамбль “не успевает” обучиться, однако значения темпа обучения на один-два порядка меньшие 1 положительно влияют на итоговое качество. Малейшие изменения времени обучения вероятнее всего являются выбросами и объясняются особенностями операционной системы.

Из экспериментов видны следующие закономерности для градиентного бустинга:

- увеличение количества базовых алгоритмов приводит к переобучению;
- RMSE зависит от гиперпараметров немонотонно, поэтому есть оптимальные значения гиперпараметров;
- темп обучения меньше 1 положительно влияет на RMSE, если отличается не более чем на 2 порядка.

Выводы

Ансамбли алгоритмов позволяют получить приемлемую точность с использованием простых базовых алгоритмов, таких как дерево решений.

Случайный лес проще градиентного бустинга, его обучение проще организовать параллельными вычислениями и он не переобучается при большом количестве деревьев, но этот метод ансамблирования имеет гораздо меньшую точность.

Градиентный бустинг имеет отличную точность и обучается быстрее (при однопоточных вычислениях), но легко переобучается и немонотонно зависит от гиперпараметров, перебор которых может занять достаточно большое время.