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

Ансамбли алгоритмов. Композиции алгоритмов для решения задачи регрессии.

Практикум на ЭВМ

Выполнил:
Кадченко И. Е.

2022

Содержание

Постановка задачи

В задании предлагается познакомиться с композициями алгоритмов для решения задачи регрессии. Поэтапно требуется реализовать алгоритмы случайного леса и градиентного бустинга, провести предобработку исходных данных и разделить исходную выборку, а также проанализировать зависимости данных алгоритмов от входящий параметров. Данный отчёт отражает проделанную мной работу по описанному заданию.

Данные содержат 21613 объекта по задаче определения стоимости продажи недвижимости в Houses Sales in King County, US. Для обучения берётся около 70% выброки, для тестовой выборки - остальные 30%, из которых 20% используются для валидационной выборки, на которой будет производиться подбор гиперпараметров.

Предобработка данных

Сперва выделим целевую переменную из данных, а затем удалим её из признакового пространства. Заметим, что все признаки представлены числами, за исключением 'даты'. Этот признак имеет строковый тип, поэтому переведём его в число. На дату продажи или покупки недвижжимости не так сильно оказывает влияние день, как месяц или год, которые содержат важную информацию о сезонности, поэтому создадим новый признак - номер месяца (первый месяц, май 2014-ого, кодируем 5, последний, май 2015-ого, кодируем 17). Удалим исходный строковый признак, а также идентификатор, чтобы снизить переобучение.

По части данных, выведенной на экран, можно сделать предположение, что признаки 'waterfront' или 'view' могут оказаться нулевыми, поэтому проверим их значения. Признаки оказались преимущественно нулевыми, но, при этом, отличные от нуля значения присутствуют. Поэтому оставим эти признаки без изменений.

Зависимость случайного леса от параметров

Изучим поведение случайного леса по изменению RMSE и времени работы алгоритма на отложенной выборке. Будем варьировать количество деревьев в ансамбле (n_estimators), размерность подвыборки признаков (max_features) и максимальную глубину для одного дерева (max_depth).

Обособленно от остальных параметров, посмотрим на зависимость от количества деревьев в ансамбле, рассматривая диапазон от 1 до 1000.

Увелечение числа деревьев, когда их ещё немного (до 50), даёт существенный прирост в качестве: значение RMSE снижается с более чем 230000 до порядка 140000, как видно из Рис.1. Далее увелечение числа деревьев также снижает значение функции ошибки, но уже не так сильно, а после 300 деревьев RMSE не меняется. Время работы алгоритма имеет квадратичную зависимость от количества деревьев. Далее будем брать для экспериментов 300 деревьев в случайном лесе.

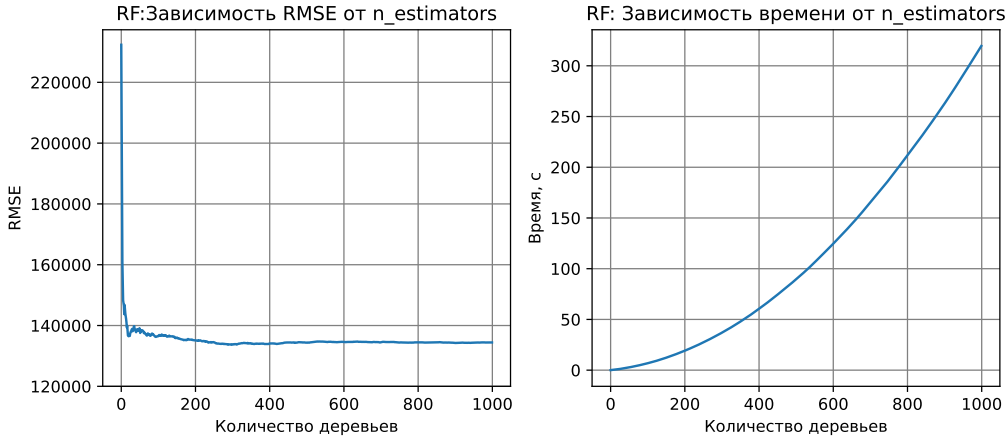

Рис. 1: Зависимость RF от количества деревьев

Параметры глубины одного дерева и размера подвыборки признаков будем перебирать вместе. Глубину дерева будем брать из множества $[2, 3, 7, 11, 15]$, отдельо рассмотрим случай, когда глубина дерева не ограничена. Размеры подвыборки признаков рассмотрим следующие $[3, 7, 10, 15, 18]$. Отранжируем по RMSE результаты и посмотрим на десять лучших комбинаций (Рис.2).

	RMSE	TIME
max_features=10, max_depth=None	131541.18872911623	14.555967092514038
max_features=7, max_depth=None	133017.46286844055	10.902682781219482
max_features=15, max_depth=None	133199.10776989913	20.77192711830139
max_features=10, max_depth=15	133257.94808476404	12.258431911468506
max_features=15, max_depth=15	134114.67229977806	17.80390191078186
max_features=7, max_depth=11	134789.47478148033	7.148885011672974
max_features=18, max_depth=None	134867.42945800585	24.40219521522522
max_features=18, max_depth=15	135569.86066125968	21.16316795349121
max_features=7, max_depth=15	135697.85709649956	8.958775043487549
max_features=10, max_depth=11	135785.63221764643	9.819741010665894

Рис. 2: Таблица зависимости RMSE и времени работы от max_features и max_depth для RF

Можно заметить, что оказалось важным рассматривать данные параметры вместе, потому что, как видно из таблицы, по одному из параметров вывод сделать не получится. Наименьшую ошибку показала модель с $max\_features = 10$ и $max\_depth = None$, имея при этом удовлетворительное время работы, поэтому будем считать эти параметры оптимальными. В целом на данном датасете лучшим образом показывает себя глубина дерева 15 или None и размер подвыборки признаков не меньше 7. С увелечением как и глубины одного дерева, так и его размера подвыборки признаков алгоритм начинает работать медленнее.

Зависимость градиентного бустинга от параметров

На вариации тех же параметров, а также шага обучения (learning_rate) посмотрим, как меняется RMSE и время работы градиентного бустинга.

Сначала установим зависимость от количества деревьев в ансамбле (Рис. 3).

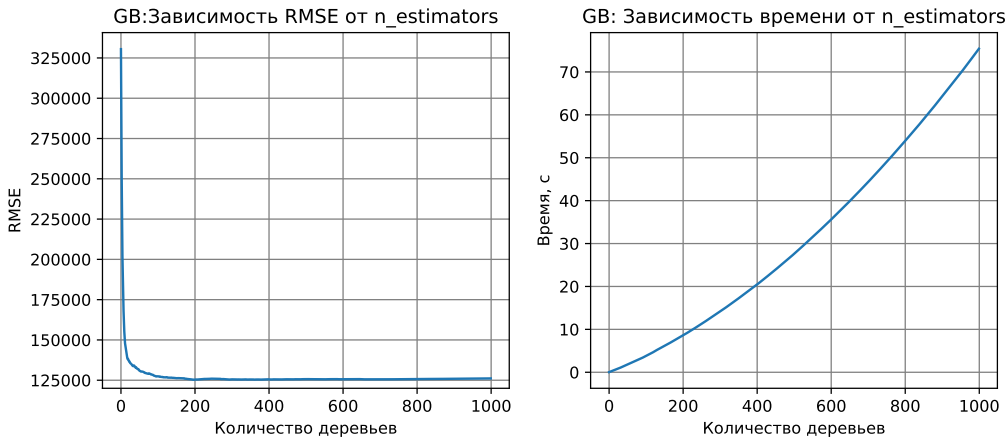

Рис. 3: Зависимость GB от количества деревьев

Значение функции ошибки выходит на плато, когда число деревьев в ансамбле достигает 200, и заметно уменьшается при стремлении к этому значению слева. Время работы алгоритма, как и в случае со случайным лесом имеет квадратичную зависимость. В градиентном бустинге будем использовать 200 деревьев.

На тех же значениях параметров и аналогичным способом посмотрим на зависимость градиентного бустинга от глубины дерева и размера подвыборки признаков. Лучшие 10 комбинаций отражены на Рис. 4.

	RMSE	TIME
max_features=7, max_depth=3	123977.76447886691	1.5509040355682373
max_features=10, max_depth=7	124036.03934408945	4.760010004043579
max_features=15, max_depth=3	124188.82637033	3.047178030014038
max_features=3, max_depth=3	125102.83606039252	0.748784065246582
max_features=10, max_depth=3	126518.56294661028	2.100318193435669
max_features=18, max_depth=3	127986.24404264538	3.6689300537109375
max_features=18, max_depth=7	128233.33161828294	8.284440040588379
max_features=15, max_depth=7	128951.53281738558	7.019269704818726
max_features=7, max_depth=2	130576.46287712992	1.0785627365112305
max_features=7, max_depth=7	132152.3694107631	3.3881900310516357

Рис. 4: Таблица зависимости RMSE и времени работы от max_features и max_depth для GB

Лучший RMSE показала модель с $max\_features = 7$ и $max\_depth = 3$, время работы которой отнюдь не значительное. Эти параметры и возьмэм за оптимальные для градиентного бустинга. Модели с низкой глубиной дерева показывают лучшие RMSE и время работы, что и характерно для градиентного бустинга - деревья ансамбля должны быть простыми, неглубокими.

Подберём шаг обучения для градиентного бустинга. Рассмотрим значения из множества $[0.05, 0.1, 0.2, 0.4, 0.7, 0.9]$, отранжируем результаты по RMSE.

	RMSE	TIME
learning_rate = 0.2	121791.58952930209	1.52522611618042
learning_rate = 0.1	127460.65068512045	1.5337412357330322
learning_rate = 0.05	128327.90056912291	1.538916826248169
learning_rate = 0.4	165088.1151573469	1.5113799571990967
learning_rate = 0.7	258784.4324618186	1.5583109855651855
learning_rate = 0.9	588749.2610114976	1.5222129821777344

Рис. 5: Таблица зависимости RMSE и времени работы от learning_rate для GB

Минимум функции ошибки с выше подобраными параметрами достигается при шаге обучения приблизительно равном 0.2. При заметном уменьшении или, наоборот, увеличении шага алгоритм заметно теряет в качестве предсказания, переобучаясь при большом шаге и недообучаясь при маленьком. Как и можно было предположить, время работы алгоритма почти не изменяется при варьировании шага обучения.

Сравнение случайного леса и градиентного бустинга

Теперь посмотрим на одном графике какой из ансамблей показывает меньшее время и большую точность на оптимальных параметрах в зависимости от количества деревьев в ансамбле.

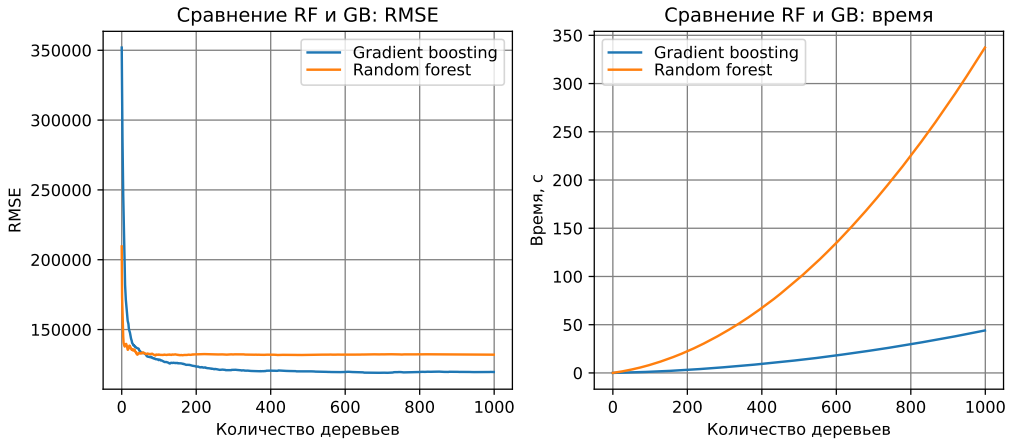

Рис. 6: Сравнение RF и GB по времени работы и RMSE

Для случайного леса требуется гораздо меньше деревьев, чтобы величина ошибки вышла на плато. При этом, плато для градиентного бустинга имеет

меньшую ошибку. Это видно из левого графика Рис.6. Оба алгоритма, как и было подмечено ранее, имеют квадратичную зависимость от количества деревьев, но случайный лес выполняется в несколько раз дольше.

Заключение

В ходе выполнения задания были построены собственные реализации случайного леса и градиентного бустинга с использованием библиотечных базовых деревьев решений. На данных **House Sales in King County, USA** были проведены исследования по зависимости времени и метрики RMSE для ансамблей от максимальной глубины, количества деревьев и подвыборки признакового пространства, шага обучения для градиентного бустинга. Предварительно данные были предобработаны. Также было проведено сравнение алгоритмов на оптимальных параметрах.