Практическое задание №2. Применение линейных моделей для определения сентимента новости.

Петренко Дарья, 317 группа

24 ноября 2018 г.

Содержание

Введение

В ходе выполнения задания были написаны собственные реализации методов случайный лес и градиентный бустинг, для работы с деревьями использовался класс $Decision\_Tree\_Regressor$ из библиотеки $scikit\_learn$.

Эксперименты проводились на датасете для предсказания цен на недвижимость. Кждый объект представляет из себя вектор из значений 20-ти признаков, описывающих характеристики единицы недвижимости, а целевая переменная - число, обозначающее ее стоимость.

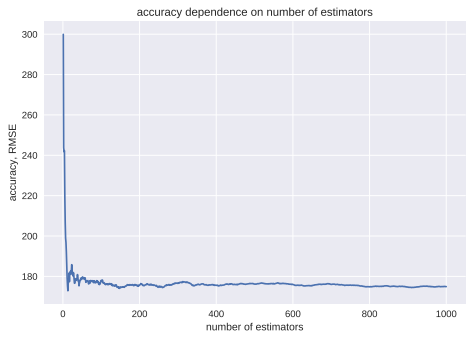
Предобработка данных

Была произведена предобработка данных - удалено поле $'index'$, а из полной даты, указанной в поле $'date'$, извлечена только информация о месяце и закодирована при помощи one hot encoding как категориальный признак. Количество признаков после предобработки составляет 28.

Исследование поведения алгоритма случайный лес

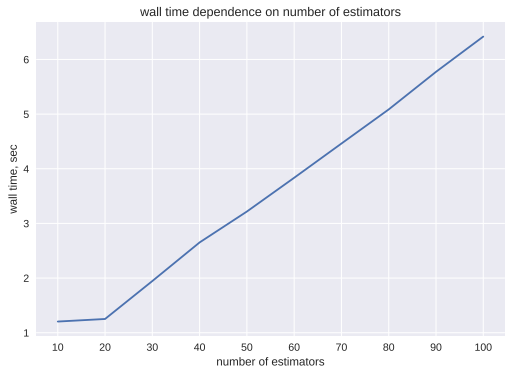
Исследование зависимости от количества деревьев

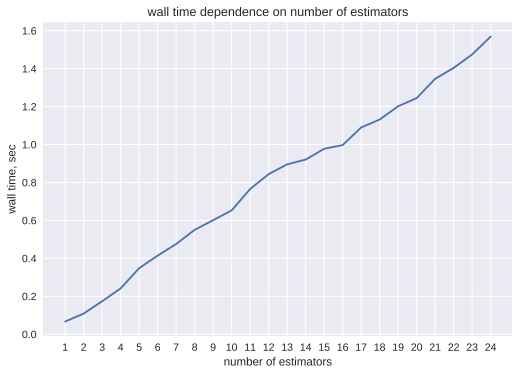
Для упрощения исследования зависимости точности от количества деревьяв функция predict была модифицирована: на вход она принимает дополнительный параметр - вектор значений целевой переменной, и на каждом шаге алгоритма (при добавлении к сумме предсказания следующего дерева из списка) измеряется точность предсказания при данном числе деревьев. Построим график полученных величин.



Как и ожидалось, с учетом теоретичесикх знаний о работе метода, он устойчив к переобучению, с ростом числа деревьев полученная точность перестает меняться и сходится к некоторой асимптоте.

Так как от количества деревьев зависит не только время работы метода predict, но и время работы fit, то для исследования зависимости времени работы от числа итераций проведем независимые запуски для разных значений параметра $n\_estimators$. Полученный график изображен ниже. Можно видеть, что зависимость линейная.

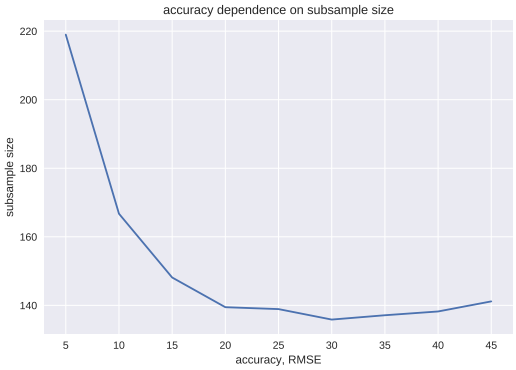


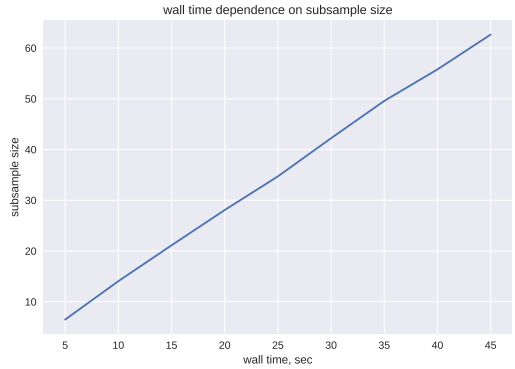


Может показаться, что при малых значениях параметра линейности не наблюдается, но для опровержения этого факта был построен более детальный график зависимости (справа).

Исследование зависимости от размера подвыборки признаков

Для проведения эксперимента возьмем значение параметра $n\_estimators$ равным 200. По результатам предыдущего эксперимента при этом значении достигается достаточно высокая точность, но время работы относительно небольшое.

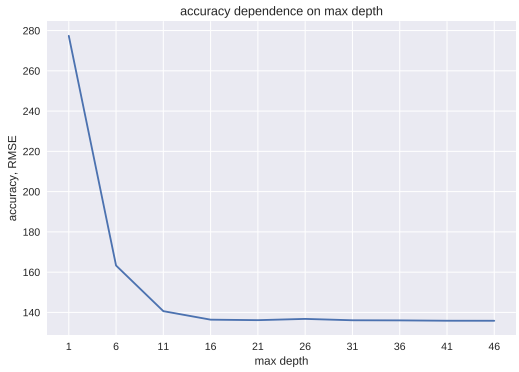


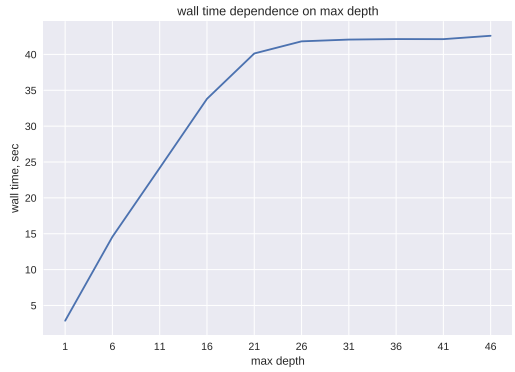


Зависимость от времени линейная. До достижения минимума при значении параметра, равном 30, качество стремительно растет при увеличении размера подвыборки, после начинает медленно ухудшаться.

Исследование зависимости от максимальной глубины дерева

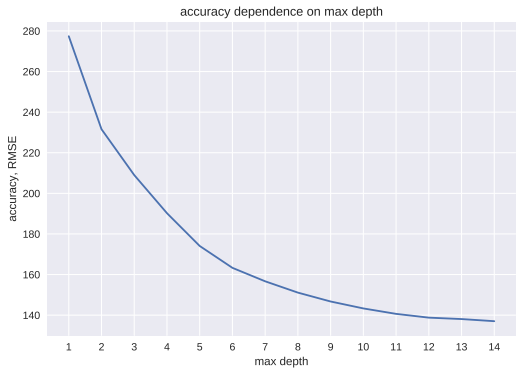
В первой части эксперимента была рассмотрена зависимость в большом диапазоне с большим шагом, чтобы посмотреть на общую тенденцию.

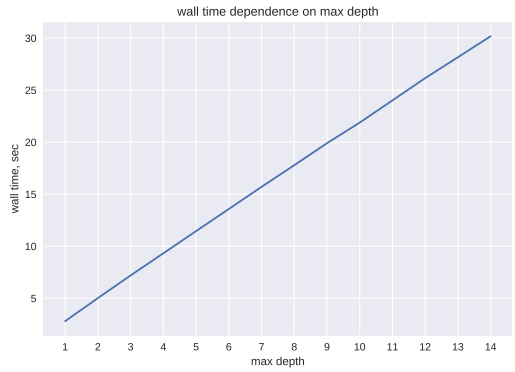




Можно заметить, что при строгих ограничениях (при маленьких значениях максимальной глубины дерева) точность стремительно растет, а время работы резко увеличивается при увеличении максимального значения. При достижении параметром значения, близкого к 15, точность и время работы перестают меняться.

Рассмотрим поведение при маленькх значениях параметра более детально.
Точность увеличивается как логарифм от максимальной глубины. Это кажется логичным,



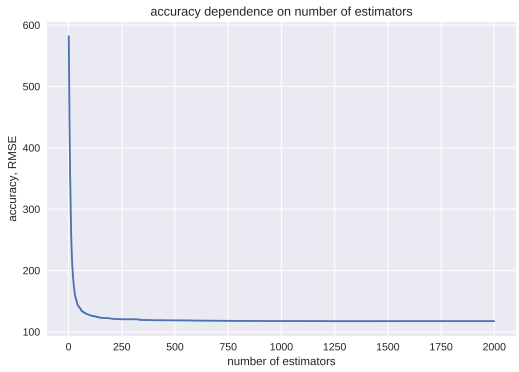


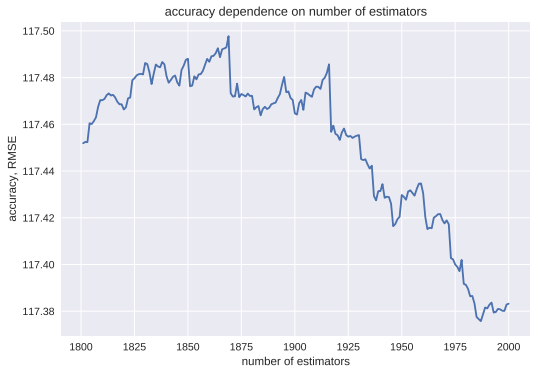
тк при небольших значениях максмальная глубина приблизительно равна числу разбиений пространства, совершаемых при построении дерева. Время работы растет линейно.
В последней част эксперимента были измерены точность и качество модели без ограничения на глубину. Точность составила 135.85, время - 42.65 секунд. Это примерно описывает асимптоту, к которой стремится модель при увеличении максимальной глубины (видно на первый графиках).

Иследование поведения алгоритма градиентный бустинг

Исследование зависимости от количества деревьев

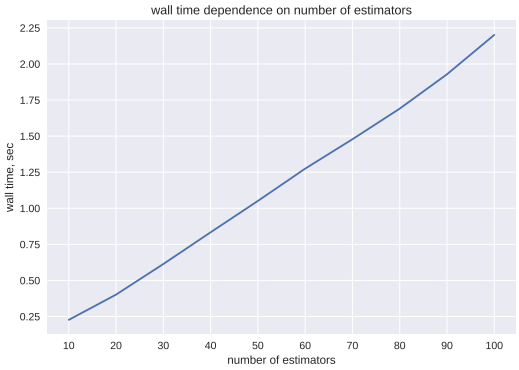
Аналогично исследванию алгоритма случайный лес, была произведена модификация метода predict: на вход подавался дополнительный параметр - вектор целевый переменных, и точность подсчитывалась после добавления каждого следующего дерева из созданного списка. Полученный график зависимости точности от числа деревьев представлен ниже.





Точность плавно увеличивается с ростом количества деревьев. Как известно из теории, при достижении большого количества деревьев точность может начать постепенно снижаться, тк метод склонен к переобучению. Для проверки этого факта был выведен график последних 200 итераций (справа). Предположение подтвердилось: на последних итерациях точность постепенно снижается.

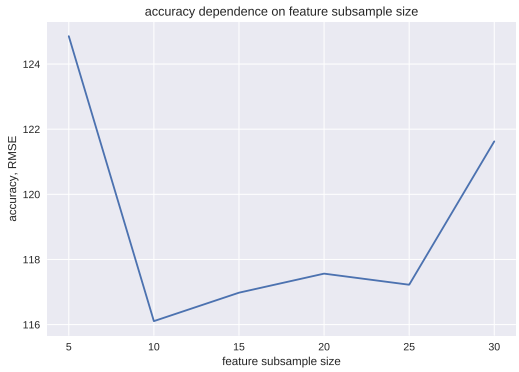
Чтобы установить зависимость времени работы метода от числа итераций, были преведены отдельные запуски программы для различны значений параметра $n\_estimators$, так как число деревьев влияет на время работы сразу нескольких вызываемых отдельно методов класса. Полученный график представлен ниже.

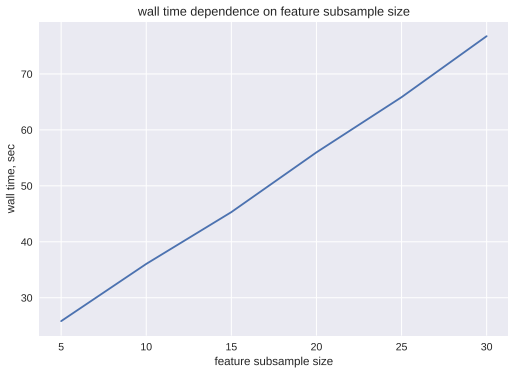


Наблюдаемая зависимость близка к линейной.

Исследование зависимости от размера подвыборки признаков

Для проведения эксперимента в качестве значения параметра $n\_estimators$ выберем оптимальное по результатам предыдущего эксперимента, равное 1000. Ниже представлены графики полученных зависимостей.

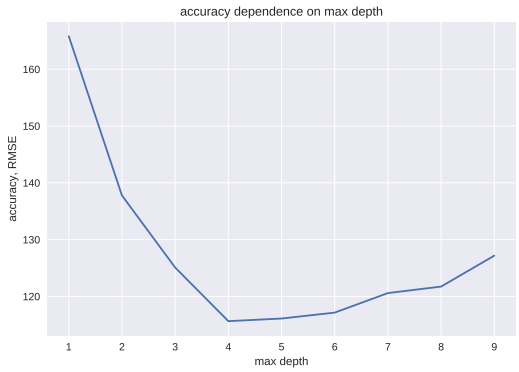




Лучшая точность достигается при значении параметра, равном 10. При этом график имеет сложную форму: при небольших точность значениях резко растет, затем меняется незначительно, а после значения параметра, равного 25, начинает стремительно падать.
Зависимость от времени линейная .

Исследование зависимости от максимальной глубины дерева

Эксперимент проводился при найденных оптимальных параметрах$n_estimators = 1000$, $feature\_subsample\_size = 10$. Полученные графики приведены ниже.



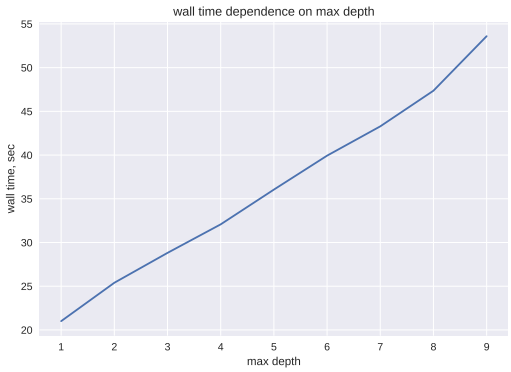
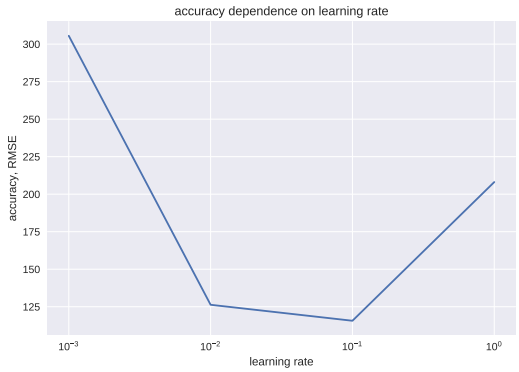
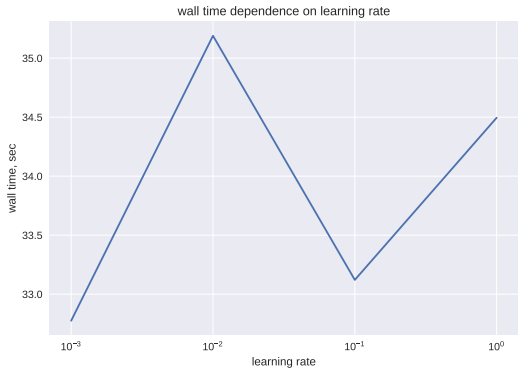


График зависимости точности от глубины дерева имеет ярко выраженную точку минимума при значении параметра равном 4. Зависимость от времени близка к линейной.

Исследование зависимости от learning rate

Параметры при проведении эксперимента - оптимальные, найденные из предыдущих экспериментов.





На графике видна ярко выраженная точка минимума, совпадающая с дефолтным значением. Зависимость времени от параметра нелинейна, график имеет множество скачков.