Модели на основе деревьев решений

Рассмотрим задачу регрессии. Для этого используем датасет о стоимости зданий. Для предобработки можете использовать методы, изученные в прошлом году. Также для упрощения работы предоставляю вам ноутбук с EDA.

1. Выделите обучающую и отложенную выборку. Рассмотрим индивидуальные деревья (4 балла)
    a. Выполните 5-кратную перекрестную проверку, чтобы определить, какой будет наилучшая максимальная глубина для одного дерева регрессии, используя все признаки обучающего набора.
    b. Визуализируйте предсказания со средними отклонениями +/- 1 стандартное отклонение во всех наборах перекрестной проверки.
    c. Постройте зависимость точности (+/- 2 стандартных отклонения) от числа деревьев
    d. Визуализируйте предсказания на отложенной выборке. Можно ли получить интервальные предсказания и для нее?
2. Перейдите к беггингу. С помощью кросс-валидации постройте и обучите несколько отдельных деревьев (4 балла).
    a. Подведите итоги работы каждого из отдельных деревьев (как численно, так и визуально) с помощью R^2. Как они работают в среднем?
    b. Объедините деревья в один прогноз с помощью беггинга и оцените его с помощью R^2. Улучшились ли результаты? Оцените, как будет меняться точность в зависимости от глубины. Используйте не только среднее предсказание, но и дисперсию.
    c. Постройте графики зависимости bias-variance в зависимости от числа деревьев (до 500 деревьев)
3. Повторите предыдущие шаги, но для случайного леса (4 балла). Можете использовать любую из опций из oob_score и кросс-валидации. Объясните свой выбор.
    a. Также оцените важность признаков. Какие 5 признаков наиболее важны?
    b. Как изменится результат при использовании только 3/5/8 важнейших признаков?
    c. * покажите разницу между использованием oob_score и кросс-валидации с помощью графиков/таблиц (2 балла со *)
4. (*) Обучите случайный лес с помощью “mse”. Покажите разницу в точности на тестовом наборе. На каких примерах разница наиболее заметна? (2 балла)
5. Повторите шаги из пункта 3 для AdaBoost. Также постройте зависимость точности от глубины дерева. (4 балла)
6. (*) какова связь между остатками и градиентом? Покажите в свободной форме, на примере (можно использовать формулы + графики/иллюстрации, набор данных взять любой, синтетический или реальный). 5 баллов
7. (*) Повторите шаги из пункта 3 для градиентного бустинга. Проведите анализ зависимости точности от параметров (то, что оценивать, выберите сами. Покажите, как еще можно оценивать такие модели. (5 баллов)

8. Соревнование! Набор данных тот же, что и для линейной регрессии, баллы будут даваться тоже соответственно по четвертям. . Можете использовать любые модели По баллам 2-4-6-8.