Задание 1.

Ответ. Да, такая константа существует при линейной модели. Const = 0.3

Обоснование.
1) Допустим, использовалась модель линейной регрессии
Вычислим MSE по исходным данным задачи, где:
для 80 случаев: y_ист – y_предск = 0,5
для 20 случаев: y_ист – y_предск = -0,3

MSE = 1/100 * ((80*(0,5)^2 + 20*(-0,3)^2) = 1/100 * (20 + 1,8) = 0,218

2) 80% значений находится выше графика линейной регрессии (y_ист – y_предск = 0,5 > 0), значит, для уменьшения MSE нужно сдвигать прямую выше, то есть ближе к 80% значений.

Сдвинем предсказанные значения вверх на 0,1. Тогда для 80 значений y_ист – y_предск = 0,4; для 20 значений: y_ист – y_предск = -0,4; а MSE:
MSE = 1/100 * (80*(0,4)^2 + 20*(-0,4)^2) = 0,16

Сдвинем значения еще вверх на 0,1 + 0,1 = 0,2. Тогда для 80 значений y_ист – y_предск = 0,3; для 20 значений: y_ист – y_предск = -0,5; а MSE:
MSE = 1/100 * (80*(0,3)^2 + 20*(-0,5)^2) = 0,122

Попробуем сдвинуть еще: на 0,3. Тогда для 80 значений y_ист – y_предск = 0,2; для 20 значений: y_ист – y_предск = -0,6; а MSE:
MSE = 1/100 * (80*(0,2)^2 + 20*(-0,6)^2) = **0,104**

Двигаем еще выше: на 0,4. Тогда для 80 значений y_ист – y_предск = 0,1; для 20 значений: y_ист – y_предск = -0,7; а MSE:
MSE = 1/100 * (80*(0,1)^2 + 20*(-0,7)^2) = 0,106

3) Видим, MSE стало ухудшаться. Проверим последний случай:

Наконец, попробуем провести линию предсказаний ровно по точкам, в которых истинное значение отклоняется от предсказанного на 0,5. Иначе говоря, поднимем прямую на 0,5. Тогда для 80 значений y_ист – y_предск = 0; для 20 значений: y_ист – y_предск = -0,8; а MSE:
MSE = 1/100 * (80*0^2 + 20*(-0,8)^2) = 0,128

Действительно, выше сдвигать уже смысла нет

Таким образом, наименьший MSE мы получили при прибавлении к ответам модели константы 0,3

Задание 2.

Ответ: отрицательные значения может возвращать градиентный бустинг.

Объяснение: принцип работы градиентного бустинга предполагает, что мы пошагово идем по направлению к лучшему предсказанию, отталкиваясь от предыдущих результатов и функции ошибок. Поэтому на каких-то шагах модель может «слишком сильно шагнуть влево», то есть уйти в минус, если предыдущий результат дал значение больше, чем целевая переменная, а сама целевая переменная в этой точке небольшая. Для предсказаний модель будет учитывать все значения, которые получались на разных шагах, поэтому минусовые значения могут попасть в предсказания модели.

Случайный лес работает иным образом. Он представляет собой совокупность решающих деревьев, а каждое дерево занимается классификацией заданных фичей в том диапазоне, в котором они даны. Поэтому случайный лес не сможет «выйти» за область допустимых значений и будет выдавать предсказания только в рамках этого интервала.

Задание 3.

Ответ и пояснение. Такая ситуация могла произойти, если присутствует не только гетероскедастичность, но и автокорреляция остатков, то есть корреляция между ошибками, например, соседними. В таком случае R-квадрат, который является отношением между дисперсией предсказаний и дисперсией целевого признака

$R^2 = \sum (y_{пред} - y_{сред})^2 \ / \ \sum (y_{истинное} - y_{сред})^2$

при применении теста Уайта останется неизменным. Числитель и знаменатель R-квадрат будут изменяться пропорционально, так как остатки регрессии (или их квадраты) связаны между собой.