1. фильтруете по порогам Q1+-1.5IQR -> точно ли это выбросы? пробовали сравнивать качество без фильтров / с фильтрами? если в выборке есть репетиторы с большим опытом и высокой ценой - почему нет, это может быть вполне реальный случай, не выброс, обратите внимание на физический смысл признаков и не стоит фильтровать просто потому, что объекты вылезают за "усы ящика", должны быть более веские основания)

2. "Change all outliers to average lesson_price for quallification group" - ручные корректировки теста не лучшая идея, более чем достаточно просто отмасштабировать данные, но по порогам что-то хардкодить не стоит. прирост качества на 1% мог быть вызван тем, что у вас на трейне тоже ограничения по порогам было

3. "Drop highly correlated feature" - тоже не стоит. вообще коррелирующие признаки большого вреда не наносят, они просто прироста не дают, но когда признаков и так мало, выкидывать целый признак не нужно

Удаляем сильно скоррелированные, если они по природе схожи и особенно если удаление не даёт просадку по качеству.
Для отбора признаков можно использовать L1, если линейная регрессия.
В нелинейных моделях есть feature_importance_

Можно рассчитывать статистики, показывающие связь между признаками и таргетом: Хи квадрат, взаимная информация. Есть permutation importance, есть information value.. Много разных техник есть на отбор признаков, в следующих курсах вы будете их рассматривать тоже

4. стандартизацию неверно сделали: вы расчитали на трейне

means = np.mean(train, axis=0)
stds = np.std(train, axis=0)

и преобразовали трейн - ок. дальше вы масштабируете тест

means = np.mean(test, axis=0)
stds = np.std(test, axis=0)

но так нельзя, вы должны мастабировать тест к тому же масштабу, что трейн, поэтому тест надо преобразовать, используя те же means и stds, что расчитали на трейне. это скорее всего даст ощутимый прирост в качестве

5. округляете ответы float_format='%.1f' разве это требуется? вы таким образом можете отрезать себе немного качества

6. параметры модели не настроены - на 4 итерации уже плато по качеству, уменьшайте шаг и увеличивайте кол-во итераций, модель недообучилась

7. целевая метрика соревнования R2, почему не выводите ее, а только MSE смотрите? ваши сабмиты должны иметь ожидаемое качество - какой R2 на отложенной выборке, примерно такой вы должны видеть и при сабмите на кэгле, это будет значить, что вы делаете всё правильно и валидация модели адекватная

8. «создание новых переменных несет в себе смысл" - генерировать на основании имеющихся может помочь только в случае линейных моделей, потому что зависимость может быть, например, квадратичная между признаком и таргетом, поэтому создать новые признаки, как квадраты имеющихся, иногда помогает вырасти в точности. в случае нелинейных моделей (аля бустинг на деревьях) генерация признаков из имеющихся особой пользы не приносит обычно, лучше добавлять новые данные в датасет, но в нашей задаче других данных нет, поэтому это не поможет :)