Захаров Владимир Олегович

В качестве тестового задания был предложен датасет с отзывами к фильмам. Необходимо было построить модели для классификации и прогнозирования рейтинга. Были обучены несколько типов моделей, а именно:

а) для классификации:

1) логистическая регрессия
2) нейронные сети с 2 и 3 полносвязными слоями, слоями BatchNormalization и нелинейностью ReLU
3) случайный лес
4) градиентный бустинг

б) для прогнозирования рейтинга:

1) линейная регрессия
2) случайный лес
3) градиентный бустинг

В обоих случаях лучше всего выступили линейные модели.
Точность при классификации на тестовой выборке: 0.9026 (accuracy_score)

При прогнозировании рейтинга была применена следующая концепция: поскольку нейтральных отзывов в выборках нет, то рейтинг положительных отзывов был смещен на - 2 при обучении. А после прогнозирования все разворачивалось в обратную сторону, метки > 4 были смещены на + 2. Такой способ дал наилучший результат.

На тренировочной выборке при прогнозировании рейтинга имеем ошибку 1.9968, то есть в среднем мы ошибаемся на 2 позиции рейтинга.