Задача: формирование обучающей выборки, проектирование валидации.
Итог: сформирована обучающая, тестовая и валидационная выборки, построен пайплайн валидации.

Валидация

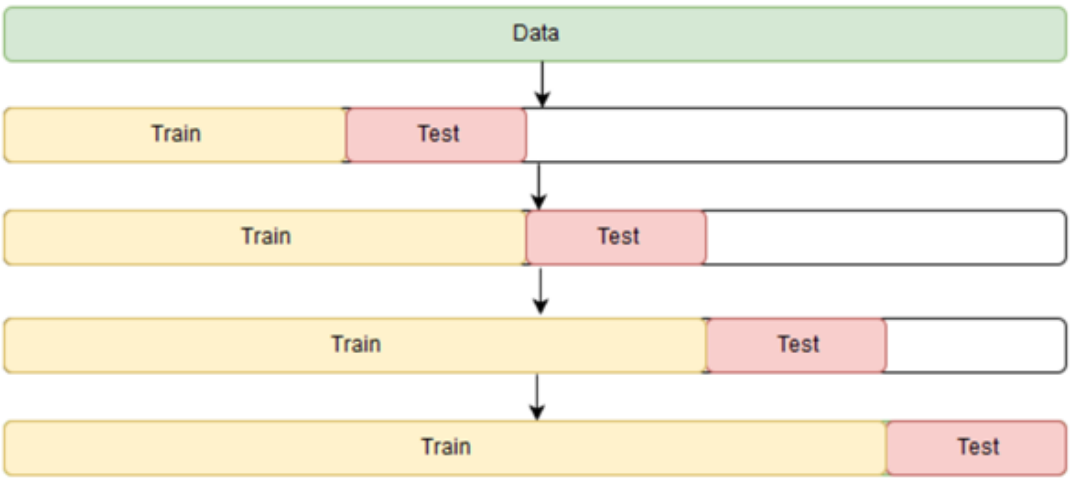
Для отбора кандидатов будут использованы 2 модели:

- ➔ Матричная факторизация.
- ➔ Item2item модель с метрикой близости BM-25. Формула показателя для случая рекомендаций выводилась на Этапе L2;
- ➔ Логистическая матричная факторизация;
- ➔ Bayessian Personalized Ranking.

Прочие модели: https://github.com/benfred/implicit/blob/main/examples/lastfm.py#L47.

Базовым алгоритмом решения задачи будет матричная факторизация. Его результат буду пытаться превзойти подходом, реализуемым в 2 этапа: отбор кандидатов и ранжирование.

Данные представляют собой временной ряд, поэтому делать стандартную K-Fold кросс-валидацию нельзя, иначе может потеряться последовательность взаимодействий. Можно зафиксировать временное окно (2 недели) и последовательно сдвигать его. Такой подход называется кросс-валидацией с кумулятивным сплитом, скользящее временное окно. На части train-выборки, которая туда не вошла, считаем значение метрики. Повторяем процедуру, пока не дойдём до конца исследуемого промежутка. Усредняем значение метрики (простое среднее). Полученный результат будет отражать качество модели с конкретным набором параметров.



Метрики качества

Метрики ошибок отражают, насколько значение прогноза близко к реальному. Применяются к небинарным данным. Это MSE, RMSE, MAE. Численное значение метрик ошибок не несёт смысл о качестве рекомендаций.

В работе метрики, подходящие для оценки качества рекомендательной системы, названы метрика точности классификации модели (classification accuracy metrics). Они применяются в случаях, когда нам не важно прогнозируемое значение само по себе. Это Precision, Recall, ROC-AUC.

Полезные ссылки:
https://www.jmlr.org/papers/volume12/duchi11a/duchi11a.pdf
https://arxiv.org/ftp/arxiv/papers/1205/1205.2618.pdf
https://github.com/benfred/implicit/blob/main/implicit/bpr.py
https://arxiv.org/ftp/arxiv/papers/1205/1205.2618.pdf
https://github.com/AmazingDD/daisyRec
http://ethen8181.github.io/machine-learning/recsys/4_bpr.htm
https://github.com/shah314/BPR