Работа с датасетом MovieLens

https://grouplens.org/datasets/movielens/

На выбор два датасета (разница в размере):

- 1M оценок https://grouplens.org/datasets/movielens/1m/
  Здесь можно объединить все три датасета, и при желании еще сделать дополнительное EDA с учетом характеристик пользователя. Есть опасность, что будет долго считаться. Можно сделать подвыборку.

- 100K оценок (Small https://grouplens.org/datasets/movielens/latest/)
  В этом датасете берем movies.csv, ratings.csv. Нет информации о пользователе.

Пожелания к оформлению:

В EDA добавляйте комментарии только там, где есть конкретный вопрос, на который не отвечает таблица/визуализация, которую вы водите, но оставляйте подзаголовки к каждому пункту.

В пункте создания модели оставляйте подзаголовки и кратко комментируйте свои действия/выводы.

Провести EDA

1. Оценить количество фильмов, пользователей и оценок
2. Оценить распределения (предлагается построить распределения количества оценок по фильмам и пользователям и увидеть т.н. "длинные хвосты" в распределениях)
3. Посмотреть на смещение оценок, увидеть сколько пользователей занижают оценки и завышают оценки (посмотреть разницу со средним значением, например).
4. Оценить средние оценки по фильмам и по пользователям
5. Выбрать критерий, показывающий, что фильм нравится всем, или не нравится никому, и вывести топ фильмов из этих списков. Знаете ли вы эти фильмы, согласны ли с оценками?
6. Выбрать лучшие фильмы.

Создание рекомендательной системы

1. Построить рекомендательную систему с помощью библиотеки surprise (или другой, по желанию). Применить подбор параметров модели на сетке, оценить качество полученной модели по выбранной вами метрике.

2. Создать нового пользователя (имитируем регистрацию нового пользователя в системе). Получить рекомендации для нового пользователя. Совпадают ли они с лучшими фильмами?
3. Поставьте оценку какому-нибудь фильму (который вы знаете и можете поставить оценку) и получите рекомендации, насколько они качественные по вашим ощущениям? Сделайте это для фильма с негативной оценкой и для фильма с позитивной оценкой.
4. Добавляйте оценки и посмотрите, как изменяются рекомендации фильмов. Соответствуют ли они вашим предпочтениям? Есть ли недостатки у системы? Сколько понадобилось оценок для того, чтобы рекомендации стали более или менее релевантными? (в этом пункте добавляйте оценки в том числе из интересного вам жанра для пункта 11).

Задания со звездочкой

1. Построить рекомендательную систему по подвыборке для одного или нескольких жанров. Сделать простой интерфейс (user input в Jupyter более чем достаточно), который спрашивает пользователя, хочет ли он получить рекомендацию в конкретном жанре или общую, выдать результат по всей выборке или жанровой подвыборке на основе выбора пользователя.
2. Сделать самостоятельное исследование и выяснить, как добавить нового пользователя\фильм в систему без переобучения всей модели, описать архитектуру такого решения