MOVIE RECOMMENDER CAPSTONE PROJECT

Cristina Andronescu

PROJECT GOAL

- Noticed my missing references to various cinematic landmarks
- Improve user familiarity with movies of the past decades by getting viewing suggestions that correspond to personal tastes

AGENDA

- Recommendation systems
- Infrastructure setup
- Initial EDA and data preparation
- Building the recommender model
    - training the model
    - getting recommendations
- Next steps

RECOMMENDER SYSTEMS

- **Working definition**: family of methods that enable filtering a set of available items that a user could choose to provide recommendations for items that the user has not experienced

- Types: Content based – uses item information to infer preferences for the current user

  Collaborative filtering – uses explicit or implicit user preferences to determine a prediction for the current user

  Hybrid

CONTENT BASED VS. COLLABORATIVE FILTERING

- Content based:

  advantages: can make recommendations as soon as there is item info

  disadvantages: limited to recommend content of the same category the user is using

- Collaborative filtering:

  advantages: no need for item content; subtle, can recommend items out of user's category of experience (serendipity)

  disadvantages: user preference sparsity may make it hard to find users with similar tastes; difficult to make predictions for users whose preferences are not consistently similar

RECOMMENDER FEATURES

- Uses explicit collaborative filtering at its core
- Can provide recommendations for a user in the data set
- Can provide rating for a movie a particular user has not yet rated
- Can provide recommendations for a new user – focus here

Results features

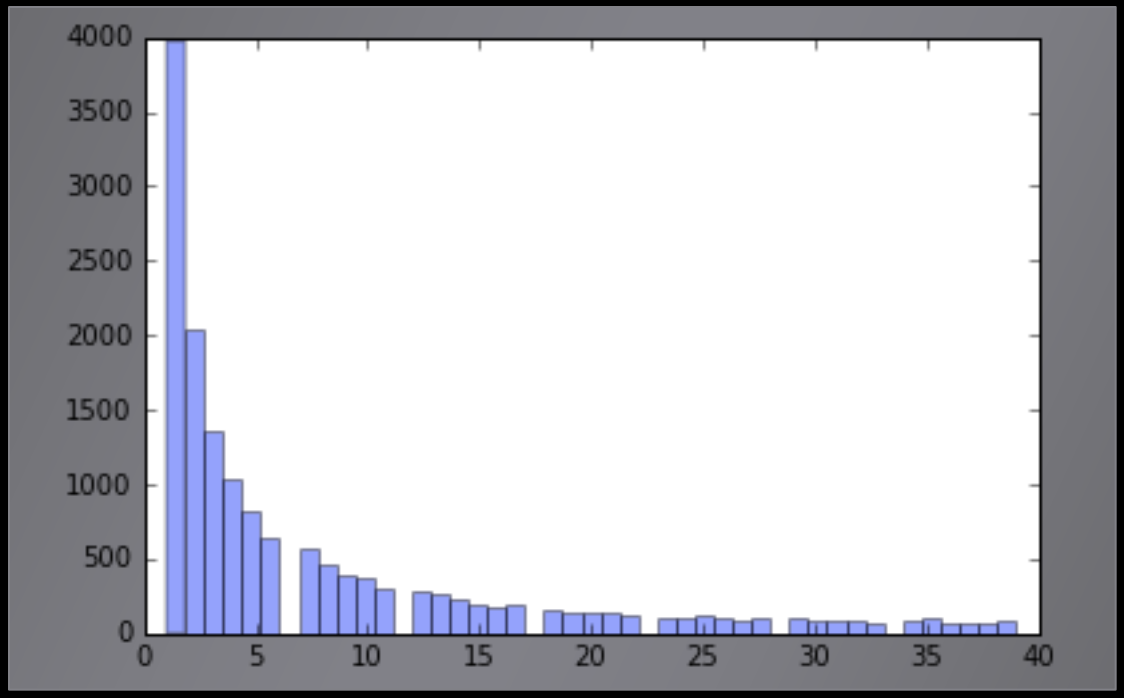
- Sort by genre
- Sort by recency of ratings
- Sort by minimum number of ratings – focus here

INFRASTRUCTURE SETUP

- Created professional account on Databricks
- Linked account to AWS
- Created AWS bucket
- Loaded data into the AWS bucket
- Created dedicated workspace files on Databricks

INITIAL EDA AND DATA PREPARATION

- Used the research recommended MovieLens dataset containing 20 000 000 ratings, for 27000 movies, by 138493 viewers
- Each user has a minimum of 20 ratings, sparsity on user side not a problem
- However



INITIAL EDA AND DATA PREPARATION

- To avoid making recommendations based on movies with only 1 rating, while at the same time maintain generality, I removed movies with fewer than 5 ratings.

- Resulting data has 19 984 024 ratings, for 18 346 movies and 13 8493 viewers

Number of ratings per user			
Min.	Median	Mean	Max.
16	68	144.29	8540

TRAINING THE MODEL

- Split data into training and testing subsets, then spit training into testing, training and validation sections.

- Tuned across a range of regularization parameters, number of iterations and ranks

- Selected model with regularization parameter= 0.5, rank= 30 and 50 iterations as best model

EVALUATING MODEL PERFORMANCE

- Model produced a RMSE of 0.79663
- Also tested performance against a naïve model which predicted the global rating average for each movie
- The best model produced an improvement of 24. 32% in RMSE over the naïve model

NEW USER RECOMMENDATIONS

- Tran new model adding new user ratings:

```
# Line format is (userID, movieID, rating)
new_user_ratings = [
     (0,10,8), # GoldenEye (1995)
     (0,253,9), # Interview with the Vampire: The Vampire Chronicles (1994)
     (0,296,4), # Pulp Fiction (1994)
     (0,19,2), # Ace Ventura: When Nature Calls (1995)
     (0,837,7), # Matilda (1996)
     (0,355,3), # Flintstones, The (1994)
     (0,2571,10), # Matrix, The (1999)
     (0,329,6), # Star Trek: Generations (1994)
     (0,858,5) , # Godfather, The (1972)
     (0,912,4) # Casablanca (1942)
    ]
```

NEW USER RECOMMENDATIONS

- Top 20 recommended movies after filtering out any movies that had been already rated by the new user

```
TOP recommended movies (with more than 25 reviews):
(u'"Lord of the Rings: The Fellowship of the Ring', 9.28661624160383, 37553)
(u'"Lord of the Rings: The Return of the King', 9.212673691281186, 31577)
(u'"Lord of the Rings: The Two Towers', 9.141703216766633, 33947)
(u'Voices from the List (2004)', 8.904142331416093, 30)
(u'So Close (Chik Yeung Tin Sai) (2002)', 8.638311066965732, 123)
(u'Alice (2009)', 8.62846539829182, 115)
(u'Labyrinth (1986)', 8.463849867400107, 8305)
(u'Inception (2010)', 8.266318534637225, 14023)
(u'Loose Change 9/11: An American Coup (2009)', 8.259518624185514, 27)
(u'"Slipper and the Rose', 8.231262159814197, 138)
(u'X-Men (2000)', 8.230755468430957, 26846)
(u'"Fifth Element', 8.202739982046014, 27660)
(u'X2: X-Men United (2003)', 8.200225861950628, 15573)
(u'Asterix and Cleopatra (Ast\xe9rix et Cl\xe9op\xe2tre) (1968)', 8.179881130976183, 39)
(u'V for Vendetta (2006)', 8.15200796205134, 14356)
(u'Minority Report (2002)', 8.151272284291975, 23642)
(u'Gattaca (1997)', 8.133948743456944, 18573)
(u'Children of Dune (2003)', 8.122554343086057, 829)
(u'"Hobbit', 8.113271456997923, 266)
(u'About Time (2013)', 8.087285797851928, 728)
```

THANK YOU!