Building a Movie Recommender System: Evaluation of ML Algorithms Using the Python Surprise Library

By Charles Luswata

(Professional Certificate in Machine Learning and Artificial Intelligence, Berkeley Haas)

**Problem Statement: ** This project focused on developing and assessing a recommender system for predicting movie ratings based on historical data. The objective was to determine the most effective algorithm by minimizing the root mean squared error (RMSE) through cross-validation. The algorithms evaluated included:

* **KNNBasic**: ** A fundamental k-nearest neighbors collaborative filtering algorithm.

* **SVD**: ** Singular Value Decomposition, a matrix factorization technique.

* **NMF**: ** Non-negative Matrix Factorization, a matrix factorization method with non-negativity constraints.

* **SlopeOne**: ** An item-based algorithm that computes predictions based on pairwise item similarities.

* **CoClustering**: ** A co-clustering algorithm that groups both users and items into clusters to make recommendations.

Methodology

Data Understanding

* ** **Data Source:**** The ` MovieLens` dataset from `https://grouplens.org/datasets/movielens/ ` which includes movie user ratings, tags and movie information (N = 100,836).

* ** **Features:**** The dataset comprises `userId`, `movieId`, `rating`, `tag`, `title`, `genres`, and `year` .

Data Preparation and Feature Engineering

The user ratings, tags, and movie information datasets were merged, with checks for duplicates and missing data. The final dataset was filtered to include complete ratings (1, 2, 3, 4, 5) from the year 1990 onwards.

```
      Unique Ratings: [1.0, 2.0, 3.0, 4.0, 5.0]
      Number of Unique Ratings: 5
      Number of Users: 609
      Number of Items: 5770
      Number of Ratings: 52054
```

Model Training and Evaluation

* **Hyperparameter Tuning:** We tested `n_factors` values of 2, 3, 5, 10, 20, 30, 40, 50 for the SVD algorithm.

* **Model Selection:**Evaluated five different algorithms from the Surprise library; KNNBasic, SVD, NMF, SlopeOne, and CoClustering.

* **Model Evaluation:** ** Applied 5-fold cross-validation to compute the RMSE for each algorithm. The selected model's performance was analyzed for movie rating predictions.

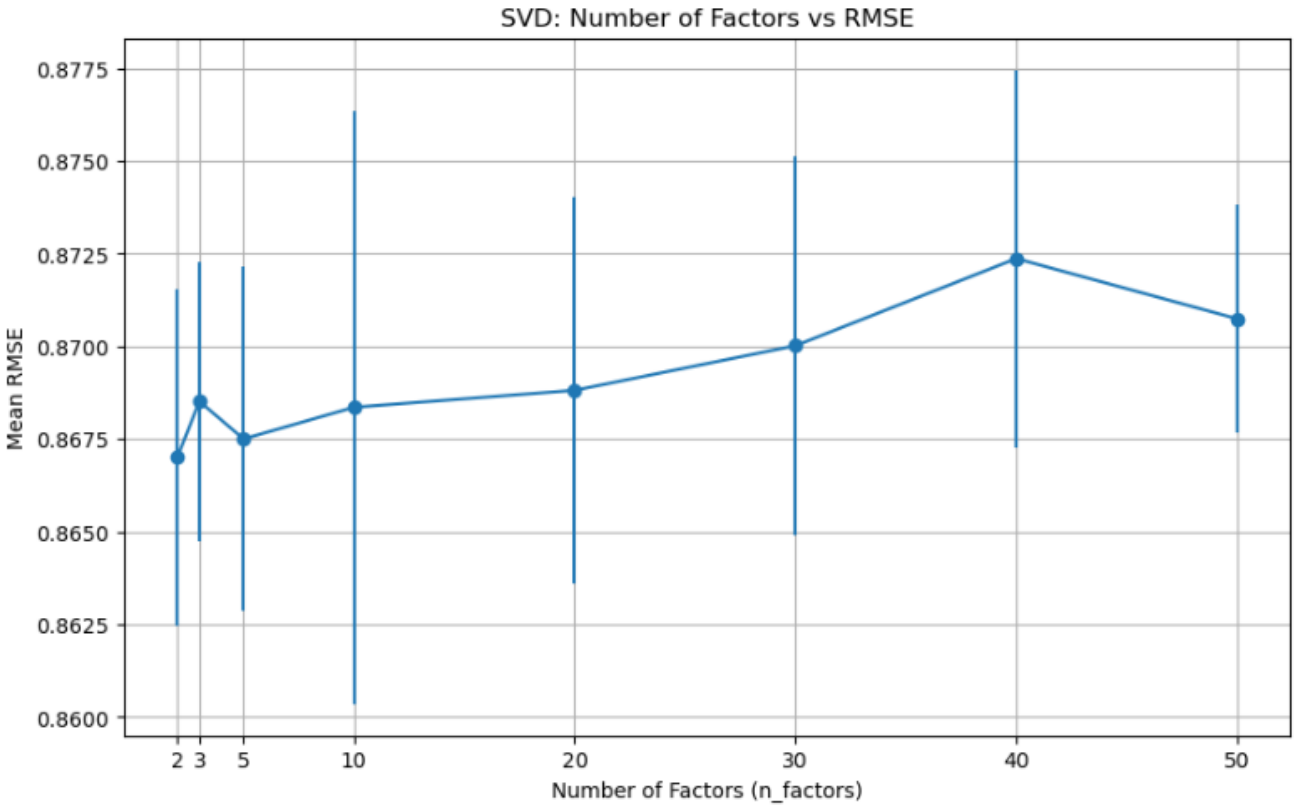
* **Results: **

```
Evaluated RMSE of algorithm KNNBasic on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    0.9494  0.9498  0.9653  0.9651  0.9563  0.9572  0.0070
Fit time          0.02    0.04    0.03    0.03    0.03    0.03    0.00
Test time         0.21    0.24    0.22    0.28    0.25    0.24    0.03

Evaluated RMSE of algorithm SVD on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    0.8693  0.8816  0.8825  0.8698  0.8684  0.8743  0.0063
Fit time          0.24    0.27    0.23    0.23    0.24    0.24    0.01
Test time         0.03    0.03    0.02    0.08    0.03    0.04    0.02

Evaluated RMSE of algorithm NMF on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    0.9376  0.9289  0.9457  0.9398  0.9262  0.9356  0.0072
Fit time          0.55    0.48    0.52    0.52    0.47    0.51    0.03
Test time         0.02    0.02    0.02    0.02    0.02    0.02    0.00

Evaluated RMSE of algorithm SlopeOne on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    0.9027  0.8979  0.9025  0.9157  0.9013  0.9040  0.0061
Fit time          0.37    0.35    0.32    0.35    0.34    0.35    0.01
Test time         0.55    0.57    0.46    0.52    0.48    0.52    0.04

Evaluated RMSE of algorithm CoClustering on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    0.9134  0.9191  0.9261  0.9462  0.9240  0.9258  0.0111
Fit time          0.49    0.49    0.49    0.51    0.47    0.49    0.01
Test time         0.02    0.02    0.03    0.02    0.02    0.02    0.00
```

Algorithm	mean_rmse	std_rmse
KNNBasic	0.957158	0.006997
SVD	0.874325	0.00633
NMF	0.935633	0.007169
SlopeOne	0.904001	0.006117
CoClustering	0.925756	0.011093

* **Interpretation: ** The **SVD** (Singular Value Decomposition) achieved the lowest mean RMSE, indicating it performs the best among the evaluated algorithms for predicting movie ratings in this dataset. This suggests SVD provides the most accurate predictions for this recommendation task.

Hyperparameter tuning for SVD




```
Best Parameter:
Number of Factors: 2.0
Mean RMSE: 0.8670
Standard Deviation of RMSE: 0.0045
```

*** **Fitting SVD Model****

The SVD model was trained on the whole dataset and evaluated using 5-fold cross-validation. In this method, the dataset is divided into 5 subsets (folds), and the model is trained on 4 folds and tested on the remaining fold. This process is repeated 5 times, each time with a different fold as the test set.

The RMSE values for each fold represent the model's prediction errors compared to the actual values. Lower RMSE values indicate better performance.

```
Fitting SVD model on the entire dataset...
Evaluating RMSE of algorithm SVD on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    0.8771  0.8711  0.8611  0.8660  0.8624  0.8675  0.0059
Fit time          0.07    0.07    0.07    0.07    0.07    0.07    0.00
Test time         0.03    0.02    0.02    0.02    0.02    0.02    0.00

SVD Cross-validation Results:
{'test_rmse': array([0.87709319, 0.87113359, 0.86105576, 0.86596794, 0.8624414 ]), 'fit_time': (0.07270669937133789, 0.06877017021179199, 0.0707633495330
8105, 0.07183957099914551, 0.06982278823852539), 'test_time': (0.028901338577270508, 0.024037837982177734, 0.023973941802978516, 0.022976398468017578, 0.
023864269256591797)}
```

** **Interpretation:** **

The average RMSE across all folds was 0.8675, indicating that the model's predictions are approximately 0.8675 units away from the true values on average. The standard deviation of 0.0059 shows that RMSE values are consistently low across different folds, with minimal variation. The average test time was 0.02 seconds, reflecting stable and efficient performance.

**** Random sample of 20 predictions****

userId	movie_Title	true_rating	predicted_rating	details
177	Spider-Man 3 (2007)	1.0	2.804	{'was_impossible': False}
394	Addams Family Values (1993)	2.0	2.561	{'was_impossible': False}
89	Tuxedo, The (2002)	5.0	3.282	{'was_impossible': False}
603	The Devil's Advocate (1997)	3.0	3.163	{'was_impossible': False}
105	Ocean's Eleven (2001)	4.0	4.289	{'was_impossible': False}
103	Gulliver's Travels (2010)	3.0	3.501	{'was_impossible': False}
610	Big Fish (2003)	4.0	4.016	{'was_impossible': False}
387	Three Kings (1999)	3.0	3.204	{'was_impossible': False}
380	Hitchhiker's Guide to the Galaxy, The (2005)	4.0	3.579	{'was_impossible': False}
42	Sweet November (2001)	3.0	3.398	{'was_impossible': False}
276	That Thing You Do! (1996)	5.0	4.147	{'was_impossible': False}
104	Nancy Drew (2007)	3.0	3.431	{'was_impossible': False}
279	World's End, The (2013)	2.0	3.141	{'was_impossible': False}
239	X2: X-Men United (2003)	4.0	4.324	{'was_impossible': False}
387	House of Wax (2005)	2.0	2.864	{'was_impossible': False}
534	Hangover, The (2009)	4.0	3.847	{'was_impossible': False}
603	Afterglow (1997)	4.0	3.592	{'was_impossible': False}
385	Shadow, The (1994)	3.0	2.891	{'was_impossible': False}
317	Lucky Number Slevin (2006)	2.0	3.689	{'was_impossible': False}
304	Hercules (1997)	3.0	3.756	{'was_impossible': False}

**** Paired T-test statistic comparing true ratings to predicted ratings****

* **T-statistic:** -1.1464

* **P-value:** 0.2516

* **Interpretation:** No significant difference was found between true and predicted ratings.

**** Predictions for new 'unseen' data**, including average and median ratings from the analysis dataset.**

userId	movie_Title	predicted_rating	average_rating	median_rating
239	Spider-Man 3 (2007)	3.54	2.72	3.0
105	Hangover, The (2009)	3.99	3.69	4.0
279	The Devil's Advocate (1997)	3.19	3.38	3.0
106	X2: X-Men United (2003)	4.56	3.94	4.0
89	Lucky Number Slevin (2006)	4.17	3.81	4.0
603	Ocean's Eleven (2001)	3.97	3.88	4.0
317	World's End, The (2013)	3.10	3.36	4.0
105	Hercules (1997)	4.24	3.57	4.0

**** Conclusion ****

The SVD model demonstrated stable and consistent performance with a low RMSE, indicating high predictive accuracy. The model's computational efficiency was also reliable, with consistent fit and test times.

**** Future Work****

Further research could include:

- Hyperparameter optimization for each algorithm to enhance performance.
- Exploration of additional algorithms or hybrid methods combining multiple approaches.
- Utilization of other datasets or integration of metadata (e.g., movie genres, tags) for improved recommendations.

**** References****

Surprise Library Documentation: http://surpriselib.com/

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