Movie Recommendation Model

Data Science Project
-Endsem Edition

Mohd Irfan Raza

Mayank Kumar

Abhishek Kushwaha

Modeling Approaches

Model 1: Content-Based Filtering

- Utilizes movie genres and tags.
- Data Preparation:
  - Merged and cleaned datasets (movies, ratings, tags).
  - One-hot encoded genres and tags.
- Model Structure:
  - Feedforward Neural Network.
  - Layers: Input (genres + tags), 128 neurons, 64 neurons, output (rating).
  - Activation: ReLU.

Model 2: Collaborative Filtering

- Utilizes user and movie embeddings
- Data Preparation:
  - Factorized user and movie IDs.
  - Merged user and movie embeddings with genres and tags.
- Model Structure:
  - Embedding Layers: 500-dimensional vectors for users and movies.
  - Fully Connected Layers: 128 neurons, 64 neurons, output (rating).
  - Activation: ReLU.

Model Training and Evaluation

Model 1: Content-Based Filtering

Training Details:

- Data Split: 80% training, 20% testing.
- Batch Size: 64.
- Optimizer: Adam, learning rate 0.001.
- Loss Function: MSE.
- Training Epochs: 35

Running Time:

- Per Epoch: ~0.6 minute.
- Total Training Time: ~21 minutes.

Training Accuracy:

- Loss: 0.0972, Accuracy: 53.93%.

Testing Accuracy:

- RMSE: 1.0912
- Accuracy (±0.5 tolerance): 36.99%.

Model 2: Collaborative Filtering

Training Details:

- Data Split: 80% training, 20% testing.
- Batch Size: 64.
- Optimizer: Adam, learning rate 0.001.
- Loss Function: MSE.
- Training Epochs: 10

Running Time:

- Per Epoch: ~1 minutes.
- Total Training Time: ~10 minutes.

Training Accuracy:

- Loss: 0.0743, Accuracy: 93.20%.

Testing Accuracy:

- RMSE: 1.0536
- Accuracy (±0.5 tolerance): 39.74%.

Model Recommendations for Specific User

```
predict_for_user_id(1, model_ratings, movies_ratings_tags_cleaned)
```

```
Top-10 Recommended movies for User 1:
1. Braveheart (1995) (Predicted Rating: 6.08)
2. Sirens (1994) (Predicted Rating: 5.84)
3. Vampire in Venice (Nosferatu a Venezia) (Nosferatu in Venice) (1986) (Predicted Rating: 5.84)
4. Pink Flamingos (1972) (Predicted Rating: 5.72)
5. Gossip (2000) (Predicted Rating: 5.56)
6. Jennifer 8 (1992) (Predicted Rating: 5.53)
7. White Christmas (1954) (Predicted Rating: 5.52)
8. Baraka (1992) (Predicted Rating: 5.46)
9. Breaker! Breaker! (1977) (Predicted Rating: 5.43)
10. Monster's Ball (2001) (Predicted Rating: 5.40)
```

Randomized Scaling Technique Used

Random Sampling

Prevents Overfitting

- Random sampling creates subsets of the data, enabling the model to **generalize better across different data points** and reducing the risk of overfitting.

Speeding Up Experimentation

- Allows us to test our models on a smaller set of data, **giving us faster feedback** to guide our experimentations.

Reducing Computational Cost and Time

- Helps us to quickly test our model, tune parameters and evaluate performance without waiting for long training times.

Ensuring Representativness

- By selecting random samples, we reduce the chance of introducing any bias that might occur when manually selecting subsets of the data and see if the sample is representative of the dataset.

Model Training and Evaluation For Small Dataset

Model 1: Content-Based Filtering

Training Details:

- Data Split: 80% training, 20% testing.
- Batch Size: 64.
- Optimizer: Adam, learning rate 0.001.
- Loss Function: MSE.
- Training Epochs: 35.

Running Time:

- Per Epoch: ~0.5 minute.
- Total Training Time: ~17.5 minutes.

Training Accuracy:

- Final Epoch (35/35): Loss: 0.1032, Accuracy: 55.33%.

Testing Accuracy:

- RMSE: 1.1986
- Accuracy (±0.5 tolerance): 35.00%.

Model 2: Collaborative Filtering

Training Details:

- Data Split: 80% training, 20% testing.
- Batch Size: 64.
- Optimizer: Adam, learning rate 0.001.
- Loss Function: MSE.
- Training Epochs: 10

Running Time:

- Per Epoch: ~1 minutes.
- Total Training Time: ~10 minutes.

Training Accuracy:

- Final Epoch (10/10): Loss: 0.0283, Accuracy: 98.79%.

Testing Accuracy:

- RMSE: 1.1231
- Accuracy (±0.5 tolerance): 33.71%.

Random Sampling Recommendations for Specific User

```
predict_for_user_id(1, model_scaled_ratings, movies_ratings_tags_sample)
```

```
Top-10 Recommended movies for User 1:
1. Johnny Mnemonic (1995) (Predicted Rating: 6.35)
2. Naked (1993) (Predicted Rating: 6.32)
3. Truth About Cats & Dogs, The (1996) (Predicted Rating: 6.29)
4. Pharaoh's Army (1995) (Predicted Rating: 6.05)
5. That Old Feeling (1997) (Predicted Rating: 6.03)
6. MURDER and murder (1996) (Predicted Rating: 5.99)
7. Air Force One (1997) (Predicted Rating: 5.99)
8. Year of the Horse (1997) (Predicted Rating: 5.98)
9. No Looking Back (1998) (Predicted Rating: 5.96)
10. Halloween 4: The Return of Michael Myers (1988) (Predicted Rating: 5.91)
```

Actual Dataset Vs Small Dataset

Training Performance

- **Content-Based Filtering (Model 1)** shows moderate training accuracy, decreasing a bit with randomized sampling with higher testing RMSE (1.1986).

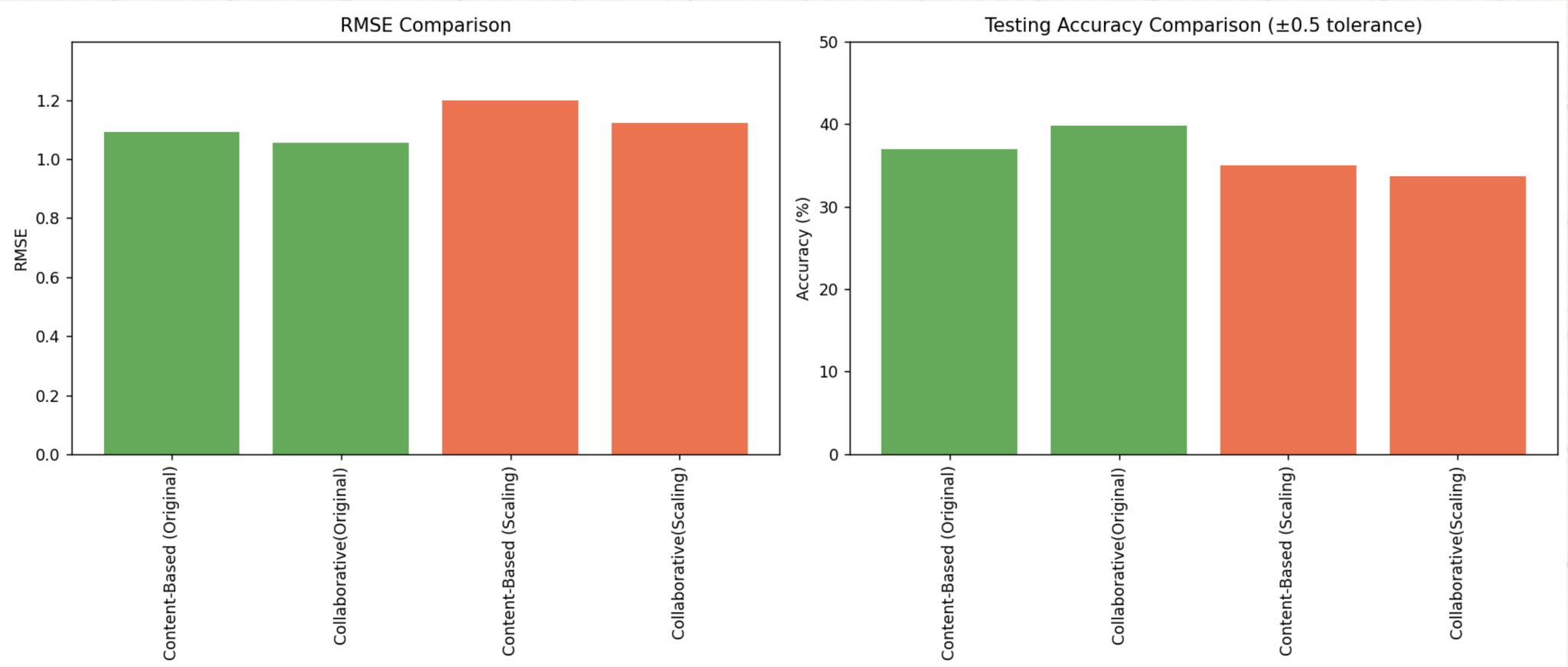
- **Collaborative Filtering (Model 2)** performs significantly better in terms of training accuracy (up to 98.79%) but shows a slight decline in testing performance when compared to the original.

Testing Accuracy

- **Content-Based Filtering (Model 1)** tends to have slightly higher RMSE in both sampling scenarios, indicating less predictive power for unseen data.

- **Collaborative Filtering (Model 2)** shows relatively consistent performance, with minor degradation under randomized sampling.

Graphical Representation



Conclusion

- This data science project aimed to develop a movie recommendation system by comparing two prevalent approaches: Content-Based Filtering and Collaborative Filtering. The evaluation metrics included training accuracy, testing Root Mean Square Error (RMSE), and accuracy within a ±0.5 tolerance.

- Collaborative Filtering proved to be a powerful approach due to its ability to learn user preferences and uncover latent patterns, while Content-Based Filtering excels in cold-start scenarios where user data is limited.

- Overall, recommendation systems must balance accuracy, efficiency, and scalability to serve diverse and evolving user needs effectively.

Future Work

- Integrate additional data, such as user demographics or temporal behaviour.

- Include additional metrics, like Mean Absolute Error (MAE) or precision/recall, to better assess recommendation quality.

- Use dynamic learning to adapt recommendations based on real-time user interactions.


THANK
YOU