ANIME RECOMMENDATION SYSTEM BASED ON USER RATINGS.

Member:
21127327 - Nguyễn Trần Trung Kiên
21127329 - Châu Tấn Kiệt
21127170 - Nguyễn Thế Thiện

Instructors:
Thầy Bùi Duy Đăng
Cô Nguyễn Ngọc Thảo

Problem

- **Problem: An anime recommendation for users based on user ratings.**
- **Description: Analyze and process user ratings on anime titles using Big Data and Machine Learning tools, in order to recommend animes the user has yet to watch, based on that user's rating history.**
- **Input: Existing user ID in the rating dataset.**
- **Output: A ranking list for recommended animes fitting the user's taste.**

Dataset

Anime Dataset 2023: A Comprehensive Collection of Anime Information

Author: Sajid (Kaggle)

MyAnimeList

Dataset

This score is calculated by Kaggle.

Completeness · 100%

- ✓ Subtitle
- ✓ Tag
- ✓ Description
- ✓ Cover Image

Credibility · 100%

- ✓ Source/Provenance
- ✓ Public Notebook
- ✓ Update Frequency

Compatibility · 100%

- ✓ License
- ✓ File Format
- ✓ File Description
- ✓ Column Description

---------- **"anime-dataset-2023.csv"** ----------

- `anime_id` : Unique ID for each anime.
- `Name` : The name of the anime in its original language.
- `English name` : The English name of the anime.
- `Other name` : Native name or title of the anime(can be in Japanese, Chinese or Korean).
- `Score` : The score or rating given to the anime.
- `Genres` : The genres of the anime, separated by commas.
- `Synopsis` : A brief description or summary of the anime's plot.

Activity Overview

Views

68.9K

3216 in the last 30 days

Downloads

11.9K

564 in the last 30 days

Dataset

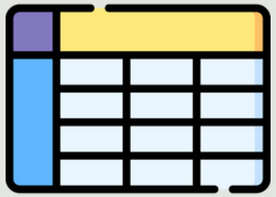
- final-animedataset.csv (4.55 GB)
- anime-filtered.csv (9.72 MB)
- user-filtered.csv (1.55 GB)
- users-details-2023.csv (73.93 MB)
- anime-dataset-2023.csv (15.92 MB)
- users-score-2023.csv (1.16 GB)

2018 - 2020

2023

Dataset

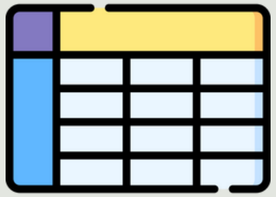
User



anime-details-2023.csv

731,290 rows/users x 24 cols.

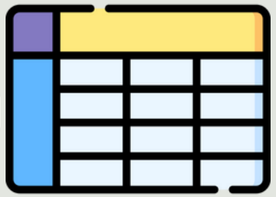
Anime



users-details-2023.csv

24,905 rows/animes x 16 cols.

Rating

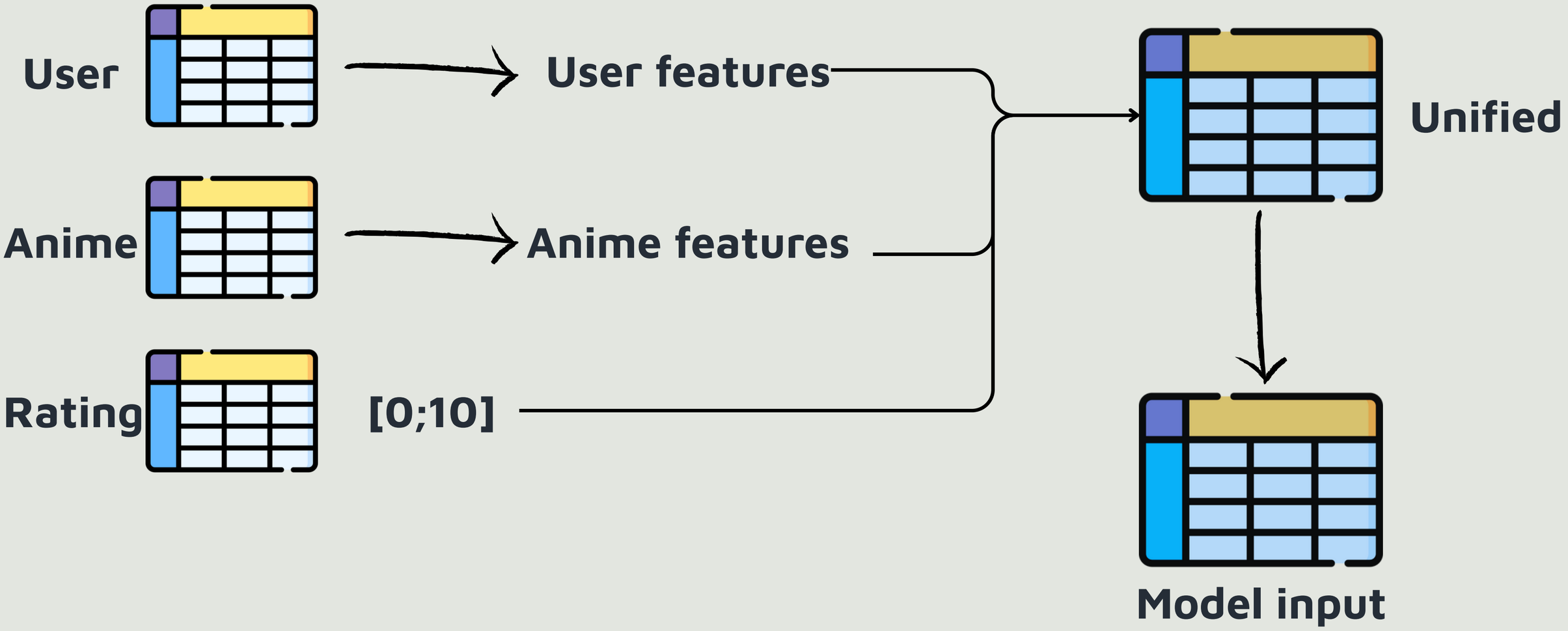


users-score-2023.csv

270,033 unique users, 16,611 unique animes.
24,325,191 rows/ratings x 5 cols.
99.48% sparsity.

Total joined:
> 1 B entries.
> 14 GB data.

Data Preprocess



Data Preprocess

User 

User features

User personal information

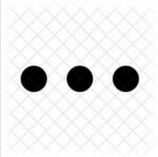
ID

Username

Gender

Birthday

Location

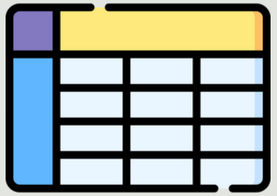


```python
user_df.select('Location').where('Location is not NULL').show(truncate=False)
```

```
+-----------------------------------------------+
|Location                                       |
+-----------------------------------------------+
|California                                     |
|Oslo, Norway                                   |
|Melbourne, Australia                           |
|Bergen, Norway                                 |
|Canada                                         |
|Land of Rain and Fjords                        |
|31f288172a11dea9f2781a6d87e0a200               |
|Calgary, AB                                    |
|Paris, France                                  |
|Seattle, Washington                            |
|Canada                                         |
|Latvia                                         |
|good ol' Europe                                |
|London,England                                 |
|Lule&aring;, Sweden                            |
|Ontario, Canada                                |
|UAE                                            |
|Locked up in Shuuka basement working on cards :p|
|Tampere, Finland                               |
|Finland, Pori                                  |
+-----------------------------------------------+
```

Data Preprocess

User



User features

Watching routine

Days Watched

Mean Score

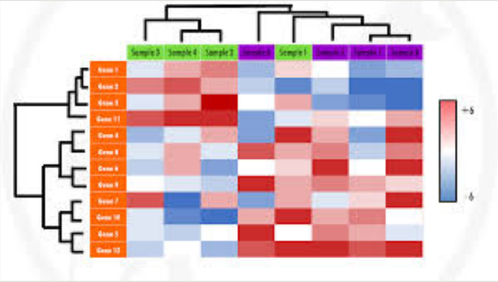
Watching

On Hold

Dropped

Plan to Watch

●●●



PCA

Data Preprocess

Anime

Anime features

ID	Name	English name	Other name	Image URL
	ぼっち・ざ・ろっく!	Bocchi The Rock!	Botchi Za Rokku!	

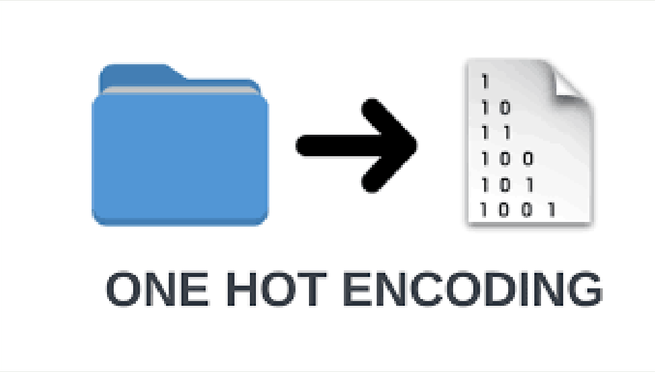
Data Preprocess

Anime

Anime features

List-based / Classifying information

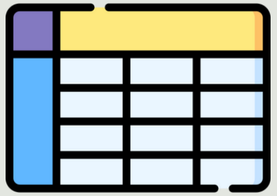
Episodes	Aired	Type	Genres	Producers	Studios	Licensors	Source



PCA

Data Preprocess

Anime



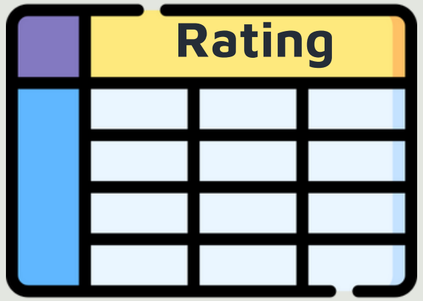
Anime features

Removed

Synopsis

Scored by

Score

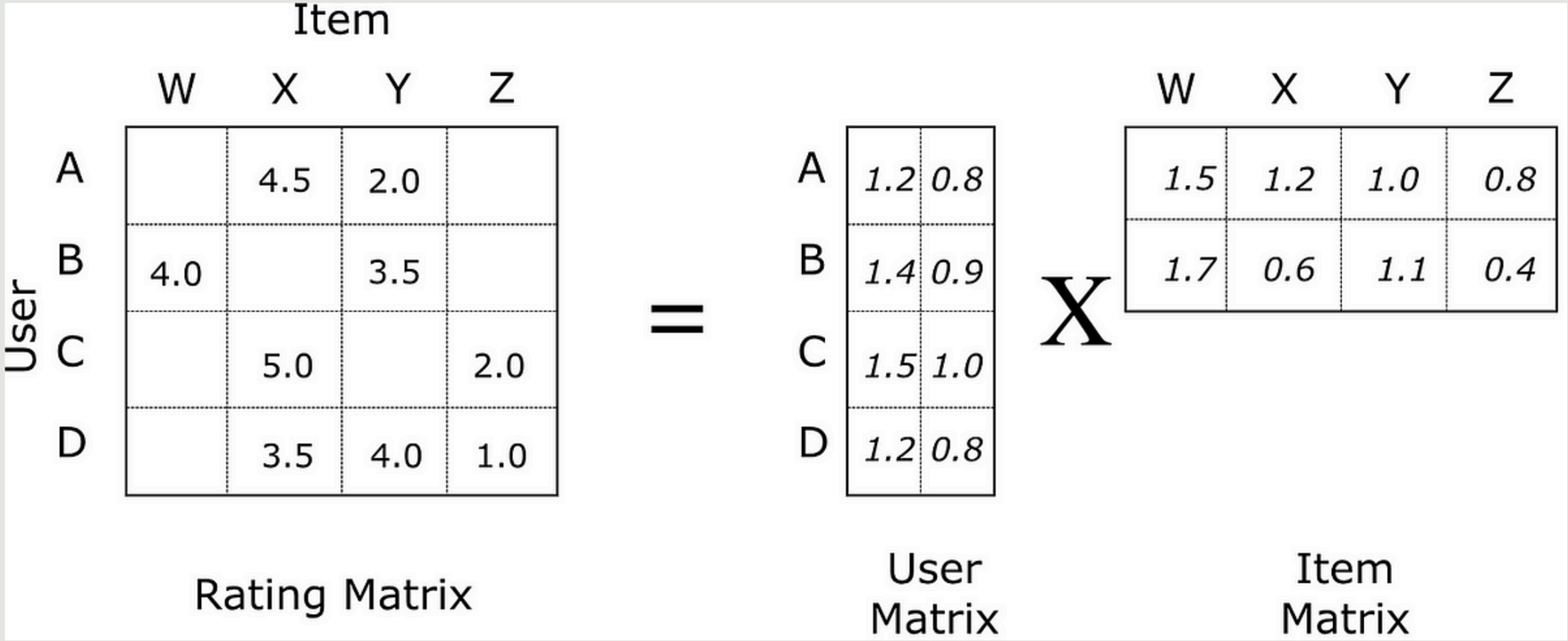


Rank

Popularity

MyAnimeList

ALS Model


Item
W X Y Z
User
A
B
C
D
4.5 2.0
4.0 3.5
5.0 2.0
3.5 4.0 1.0
=
A 1.2 0.8
B 1.4 0.9
C 1.5 1.0
D 1.2 0.8
X
W X Y Z
1.5 1.2 1.0 0.8
1.7 0.6 1.1 0.4
Rating Matrix
User Matrix
Item Matrix

ALS Model

Supposing R is the user-item rating matrix — shape (m×n) where:

m = number of users

n = number of items

ALS tries to approximate:

R≈U×P^T where:

- U is the user factors matrix — shape (m×k)
- P is the item factors matrix — shape (n×k)
- k is the number of latent factors

Each row of U is the latent vector for a user (user preferences).

Each row of P is the latent vector for an item (item features).

RMSE Metrics

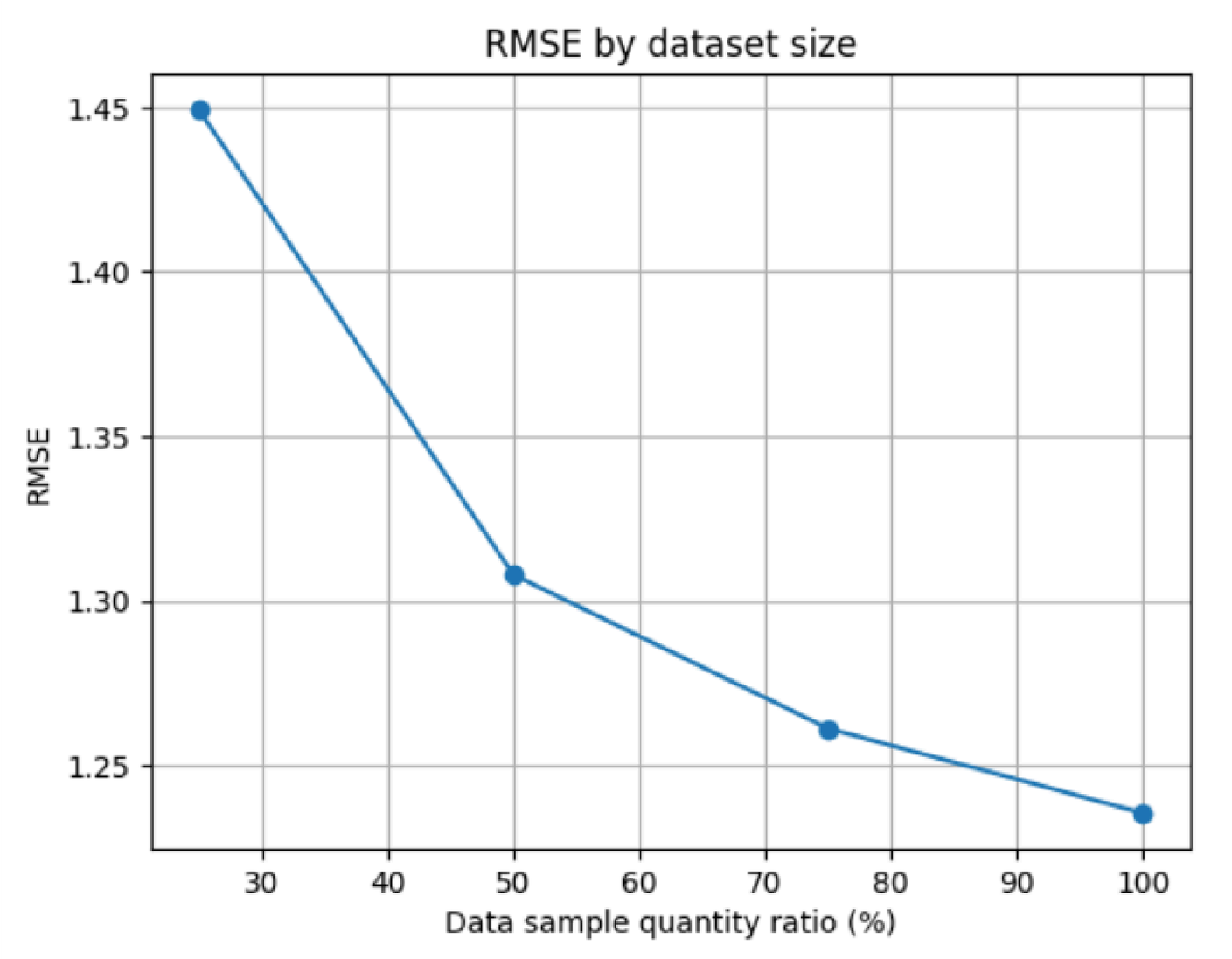
RMSE indicates how many units each model's prediction deviates from the actual average value; the smaller the RMSE, the better the model.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_{\mathrm{true},i} - y_{\mathrm{pred},i})^2}$$

- **N is the number of test samples.**
- **y_pred_i = Predicted value by the ALS model (can be outside 1–10).**
- **y_true_i = Actual value from the test data (always between 1–10).**

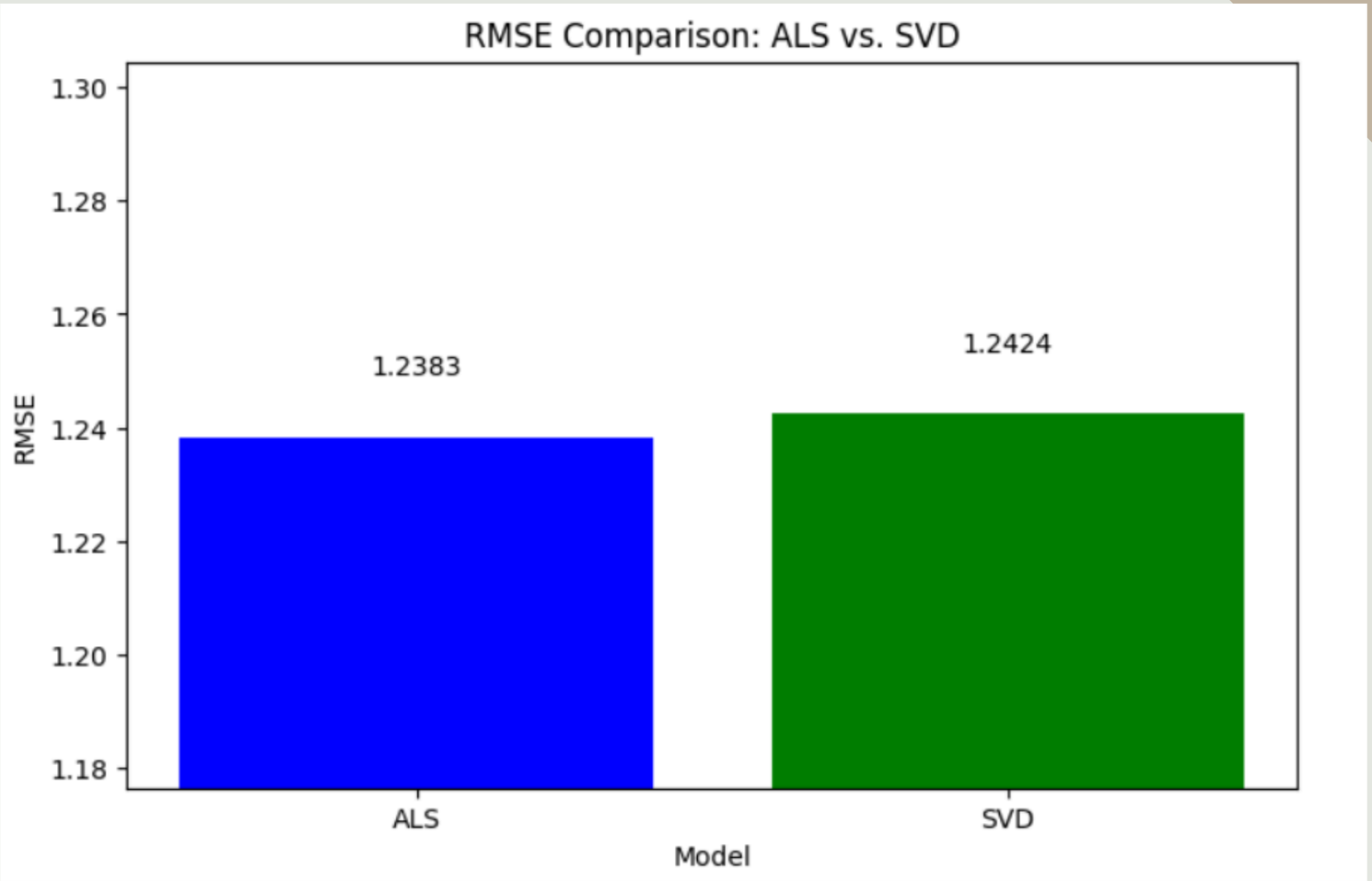
ALS Model

- The y-axis is the RMSE score achieved by the model
- The x-axis is the size of the dataset used (25%, 50%, 75%, and 100% of the original dataset size)
- As the size of the dataset increases, the RMSE decreases, meaning the model becomes more accurate
- This demonstrates the suitability of ALS for large datasets



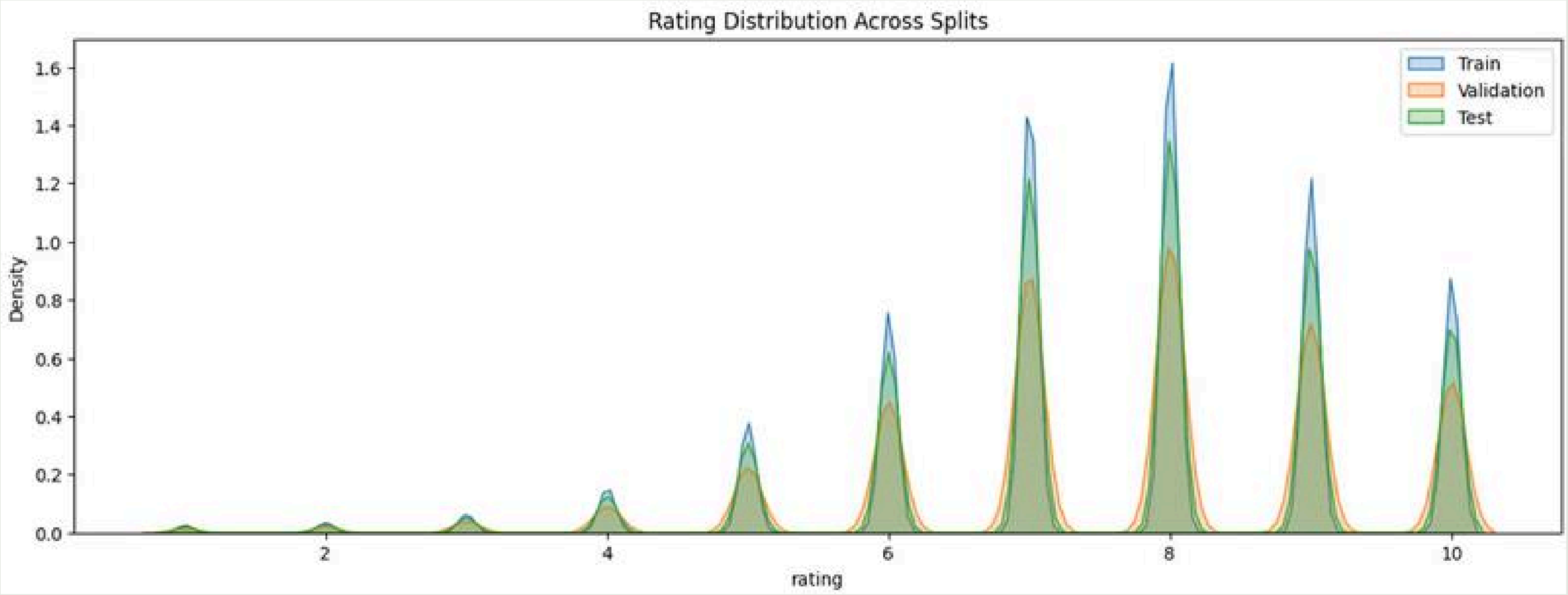
Compare RMSE with SVD

- The y-axis is the RMSE score achieved by the model
- The x-axis is 2 models commonly used for Recommendation tasks (ALS and SVD)
- ALS gives better results than SVD with lower RMSE
- Based on the above results, ALS was chosen as the model for this project instead of SVD.



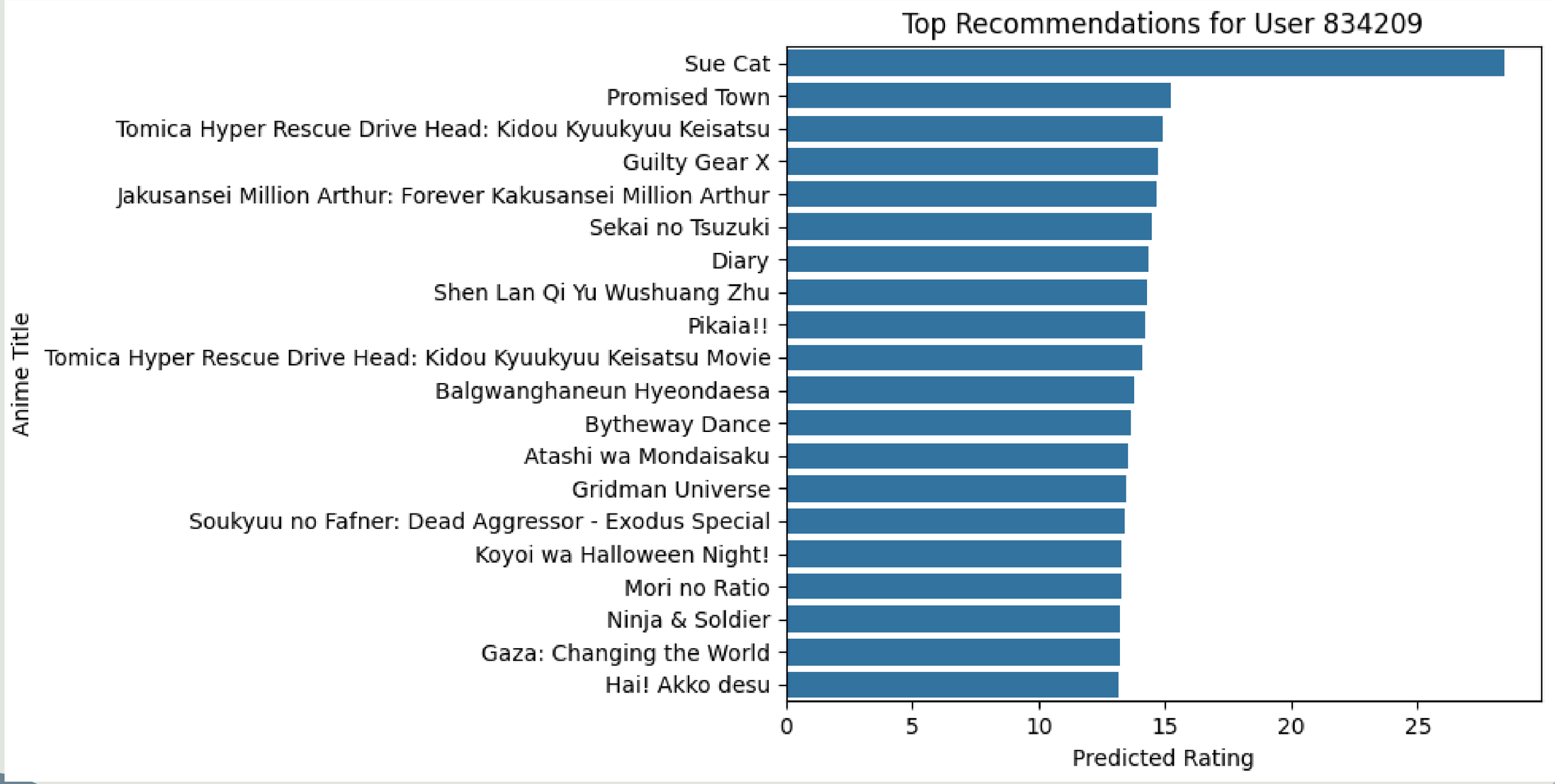
Training Methods

The preprocessed dataset is split into 3 parts: Train, Validation and Test with the ratio of 70:5:25

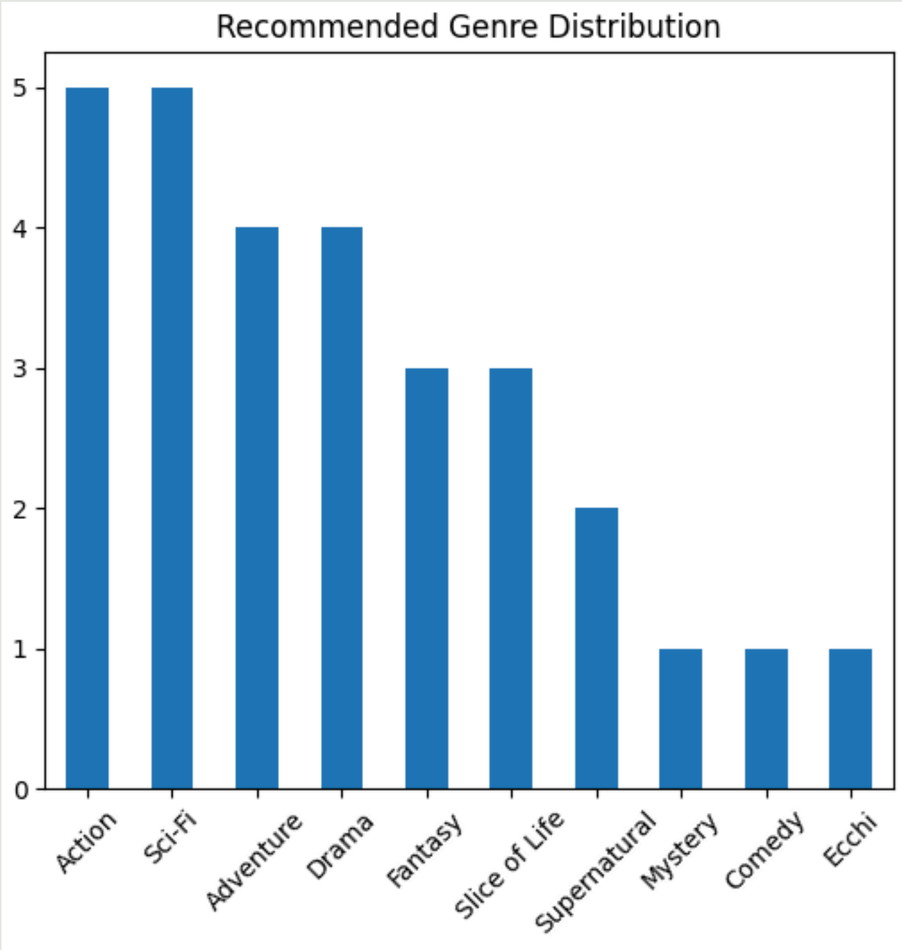


Since the Train, Validation, and Test curves overlap well with each other, there is no major skew, and the distribution is balanced

Recommendation Results



Recommendation Results



Insights about Recommended Genre stribution:1. Dominant genres

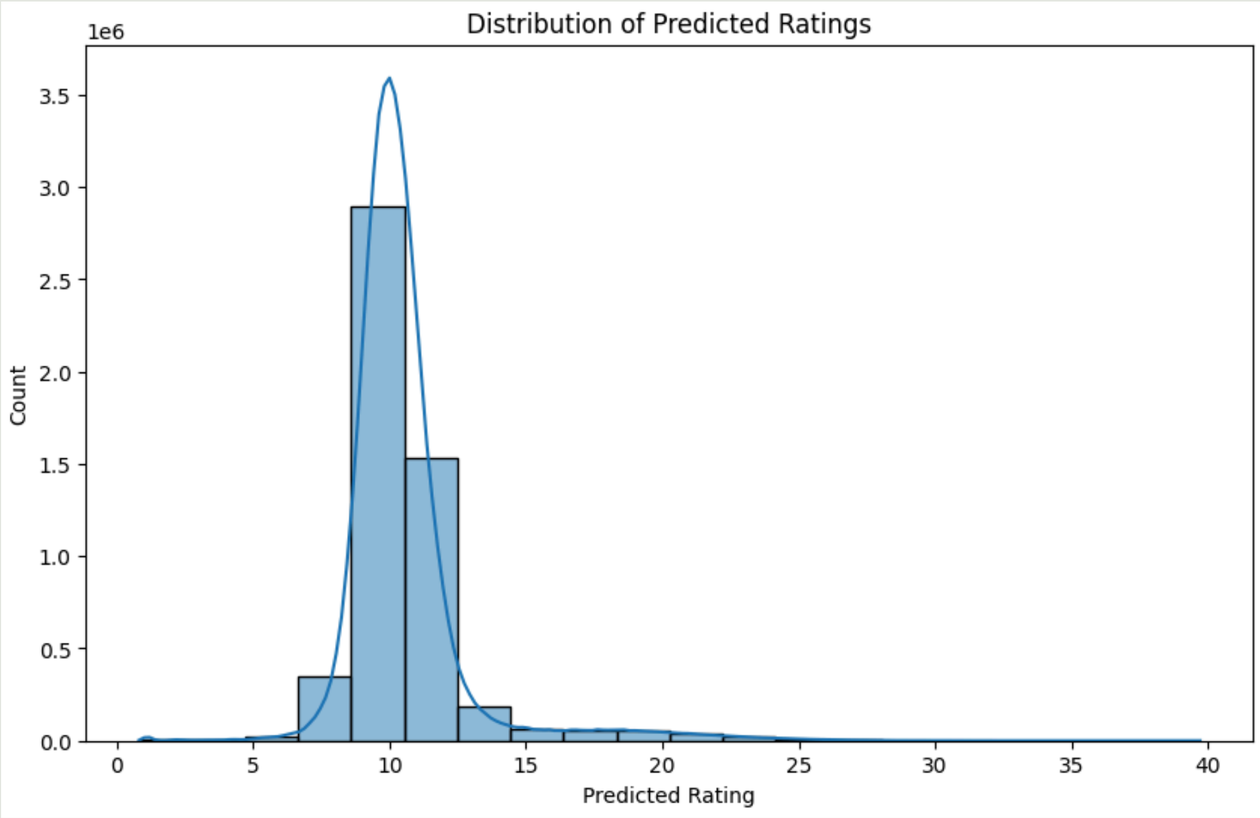
- **Action and Sci-Fi are the most recommended genres (5 animes each).**
- **Followed by Adventure, Drama,and Fantasy (4 animes each).**

2. Least recommended genres:

- **Comedy, Mystery, and Ecchi have the fewest recommendations (1 anime each).**

This suggests the user's preferences are skewed towards action-packed, futuristic, and emotionally engaging genres.

Recommendation Results



Insights about Predicted Rating Distribution:

- **Bell-shaped, right-skewed distribution: The bulk of predicted ratings fall between 8 and 12, with a sharp peak around 10.**
- **Long tail: A smaller number of ratings extend all the way up to ~38, but these are rare.**
- **Most common range: Between 9 and 11, suggesting the model tends to predict ratings in this range for the majority of items.**

Implications

- **The recommendation model is conservative, predicting high ratings only when it has high confidence.**
- **The long right tail shows that outlier items with very high predicted appeal are present but very few.**

Thank You

For your attention