SOEN 471 - Final Presentation Anime Recommender System

El Mamoun OUAHID ALAOUI, Subhanshu ROWAL,
Iphigenia Jenna PAPPAS, Omar ELKASSOUANI

Research Questions

- Problem:
    - From niche to popular
    - Huge catalogue, which to pick?
    - Daunting for newcomers and avid anime enjoyers
- How can we recommend anime that users would enjoy?
- How can we create and improve an anime recommender system?

Dataset (1)

Purpose: analyze user profiles and suggest an anime matching their interests

Used 2 Datasets from Kaggle:

1) **anime.csv** : csv file that contains a list of animes with numerous features.
2) **ratings.csv** is a csv file that contains user ids and their ratings for different animes

Goal: Use this data to create a **cross referencing table** that will **map users** who have the same taste of anime **based on their ratings** and **suggest a list of animes** that can fit their profile

Dataset (2) - Anime Dataset

We can see that our anime data set is composed of 12 294 rows and contains 7 columns.

- anime_id: id to identify the anime
- name: title of the anime
- genre: Genre of the anime
- type: tye of the anime
- episodes: number of episodes of the anime
- rating: average rating out of 10 for this anime, provided by the website provider.
- members: number of community members that follow the particular anime

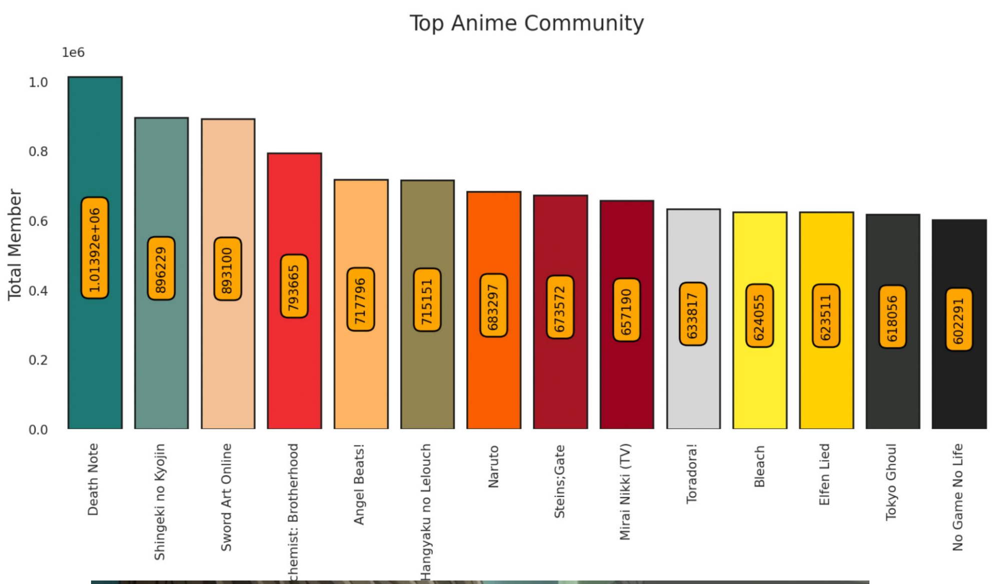
	anime_id	name	genre	type	episodes	rating	members
0	32281	Kimi no Na wa.	Drama, Romance, School, Supernatural	Movie	1	9.37	200630
1	5114	Fullmetal Alchemist: Brotherhood	Action, Adventure, Drama, Fantasy, Magic, Mili...	TV	64	9.26	793665
2	28977	Gintama°	Action, Comedy, Historical, Parody, Samurai, S...	TV	51	9.25	114262
3	9253	Steins;Gate	Sci-Fi, Thriller	TV	24	9.17	673572
4	9969	Gintama'	Action, Comedy, Historical, Parody, Samurai, S...	TV	51	9.16	151266
5	32935	Haikyuu!!: Karasuno Koukou VS Shiratorizawa Ga...	Comedy, Drama, School, Shounen, Sports	TV	10	9.15	93351
6	11061	Hunter x Hunter (2011)	Action, Adventure, Shounen, Super Power	TV	148	9.13	425855
7	820	Ginga Eiyuu Densetsu	Drama, Military, Sci-Fi, Space	OVA	110	9.11	80679
8	15335	Gintama Movie: Kanketsu-hen - Yorozuya yo Eien...	Action, Comedy, Historical, Parody, Samurai, S...	Movie	1	9.10	72534
9	15417	Gintama': Enchousen	Action, Comedy, Historical, Parody, Samurai, S...	TV	13	9.11	81109

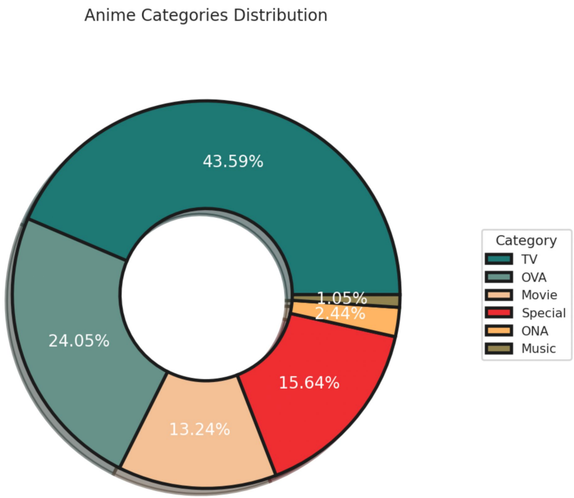
Dataset (3) - User Dataset

We can see that our user data set is composed of 3 columns.

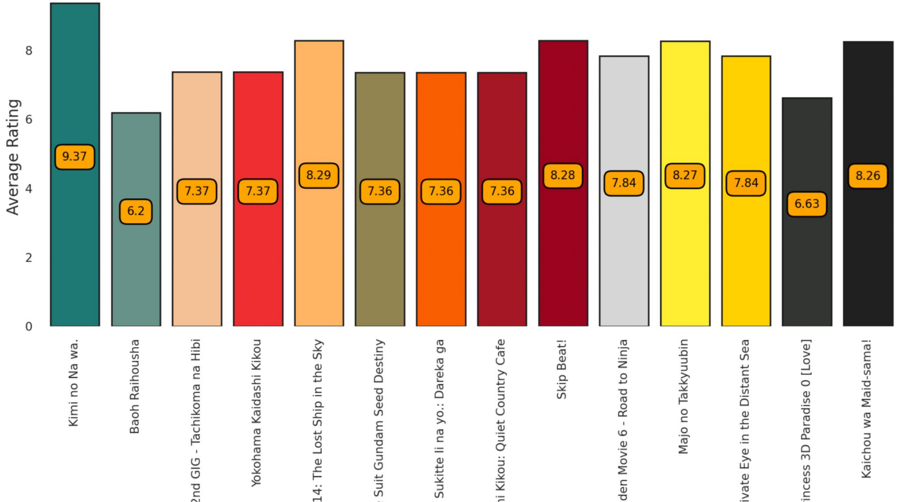
- user_id : non identifiable randomly generated user id.
- anime_id : the anime that this user has rated.
- rating : rating out of 10 this user has assigned (-1 if the user watched it but didn't assign a rating).

	user_id	anime_id	rating
0	1	20	-1
1	1	24	-1
2	1	79	-1
3	1	226	-1
4	1	241	-1


Top Anime Community
1e6
1.0
0.8
0.6
0.4
0.2
0.0
Total Member
1.01392e+06
896229
893100
793665
717796
715151
683297
673572
657190
633817
624055
623511
618056
602291
Death Note
Shingeki no Kyojin
Sword Art Online
chemist: Brotherhood
Angel Beats!
Hangyaku no Lelouch
Naruto
Steins;Gate
Mirai Nikki (TV)
Toradora!
Bleach
Elfen Lied
Tokyo Ghoul
No Game No Life


Anime Categories Distribution
43.59%
24.05%
13.24%
15.64%
2.44%
1.05%
Category
TV
OVA
Movie
Special
ONA
Music


DEATH NOTE


Average Rating
8
6
4
2
0
9.37
6.2
7.37
7.37
8.29
7.36
7.36
7.36
8.28
7.84
8.27
7.84
6.63
8.26
Kimi no Na wa.
Baoh Raihousha
2nd GIG - Tachikoma na Hibi
Yokohama Kaidashi Kikou
14: The Lost Ship in the Sky
Suit Gundam Seed Destiny
Sukitte Ii na yo.: Dareka ga
i Kikou: Quiet Country Cafe
Skip Beat!
den Movie 6 - Road to Ninja
Majo no Takkyuubin
ivate Eye in the Distant Sea
incess 3D Paradise 0 [Love]
Kaichou wa Maid-sama!

Model Design (1) - Preprocessing

Further filtering is done to ensure that our models can predict better. We subjected the dataset to 2 major constraints:

- filter out users who have rated less than 200 animes
- filter out animes that have received less than 100 ratings

Then, we constructed a pivot table and replaced all NaN values by 0 as Matrix Factorization works best when NaN are set to 0

Model Design (2) - SVD

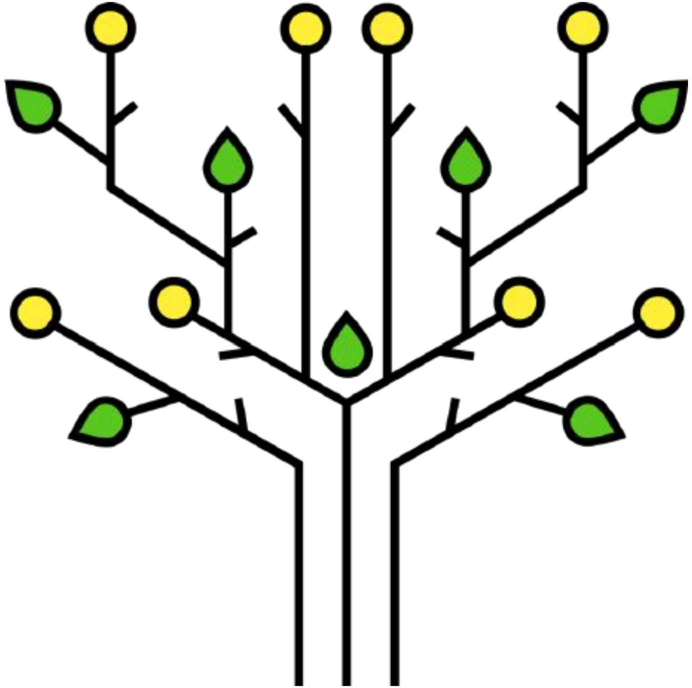
- Implemented an SVD model and set the number of components to 100.
- Number of components is a hyperparameter that can be changed. Results change according to that parameter.
- We then created a correlation matrix using the results from SVD transformation.
- Recommendations were made based on the top 1 anime.
- Filtered the animes by correlation to the top 1. Tries showed an upper bound correlation of 1.0 and a lower bound correlation of 0.878 provide the best results.

```
list(anime_names[(corr_popular_anime < 1.0) & (corr_popular_anime > 0.878)])
```

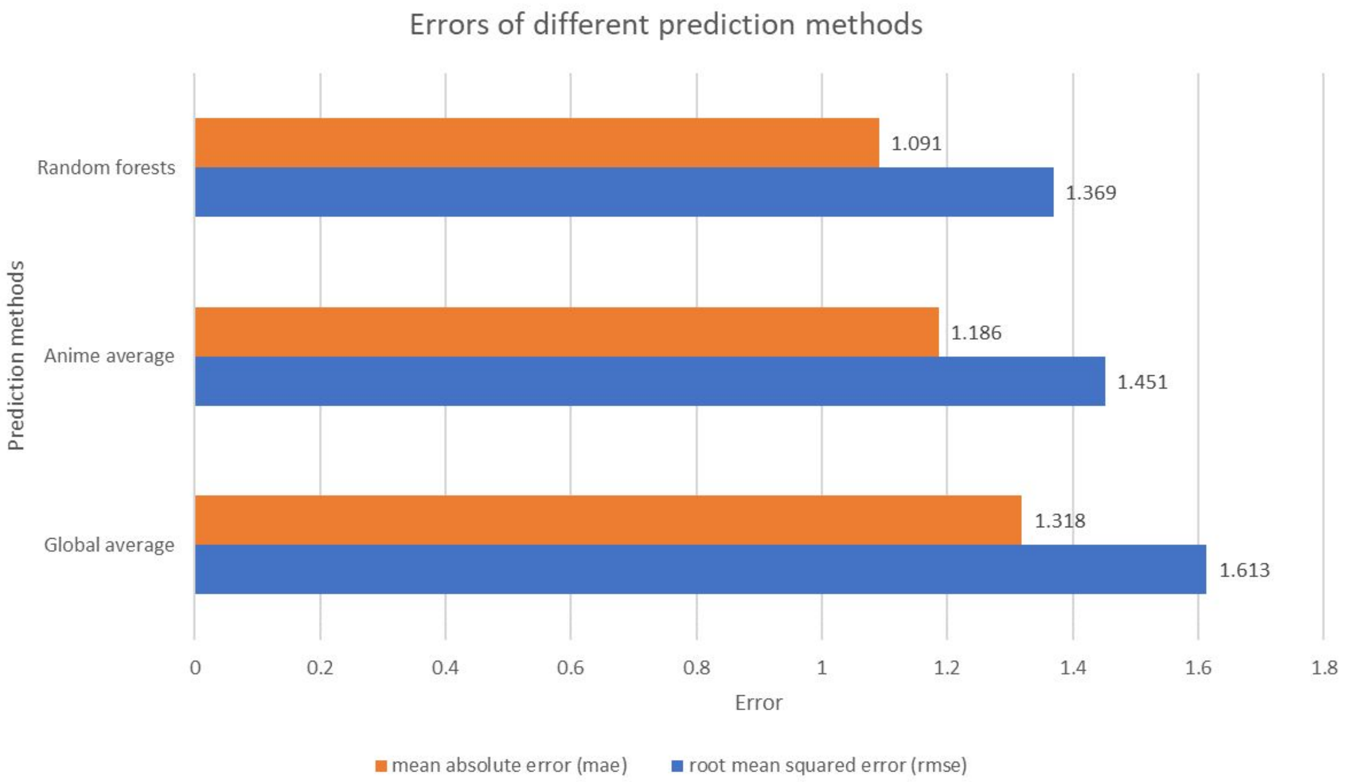
```
['Angel Beats!',
 'Code Geass: Hangyaku no Lelouch',
 'Code Geass: Hangyaku no Lelouch R2',
 'Elfen Lied',
 'Fullmetal Alchemist: Brotherhood',
 'Shingeki no Kyojin']
```

Model Design (3) - Random Forest

- Supervised Learning
- User-user Recommender
- Collaborative Filtering Approach
- Random Forest Algorithm
  - Manually Tuned Hyperparameters
- Regression model that predicts Ratings
- Results?



Results



- RMSE and MAE: most popular regression measurements metrics
- Used global average (7.287) and anime averages as baseline
- Lower is better

Conclusion and Discussion

- Retrospective on research questions.
- Satisfied with results?
  - Almost satisfied. The results are encouraging but we are interested in how we can optimize it and provide more precise and accurate results while ensuring a low running time.
- Issues?
  - Long computing times (ex: RandomSearchCV)
- What can be improved?
  - Further analysis on the effect of hyperparameter tweaking is considered for the TruncatedSVD.
  - User biases
  - Compare with content based recommender

Questions?