Assignment 2 - Anime Dataset

1.Dataset and Exploratory Analysis:

The dataset[9] was originally scraped from MyAnimeList [8] and consists of three files. The three files are animes.csv, reviews.csv and profiles.csv. Exploratory analysis was conducted on each dataset file to understand the dataset and its features.

The file animes.csv contains 19311 data points with 12 features namely: uid, title, synopsis, genre, aired, episodes, members, popularity, ranked, score, img_url and link. This file contains some duplicate values after whose removal leads to 16216 data points. After removing duplicate rows, we deal with NaN values in the score and episodes category. The feature 'scores' contains 341 NaN values which are replaced with a global average score calculated from the whole dataset. The feature 'episodes' contains 492 NaN values which are replaced with global average episode number calculated from the whole dataset. We also extract the total number of genres mentioned in the 'genre' feature of the dataset. The total number of genres mentioned in the dataset are 43. We also find the distribution of genres over the whole dataset to understand animes of which genres are frequent.

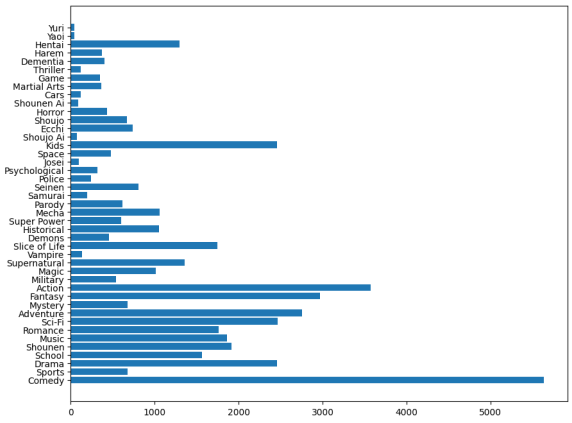

Fig 1. Distribution of genres over the whole dataset

We found the top five most frequent genres: comedy, action, fantasy, mystery and drama. We also find the distribution of genres over the years to understand the trend of genres over the period of roughly 90 years. Through this, we can understand the trend of genres over the years and what genres have become more popular over the recent years. We can see anime with genres of comedy , action, fantasy and especially slice of life have become more frequent over the recent years.

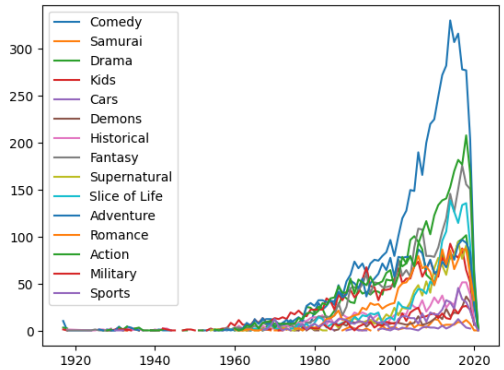

Fig 2. Distribution of Genres over the Years

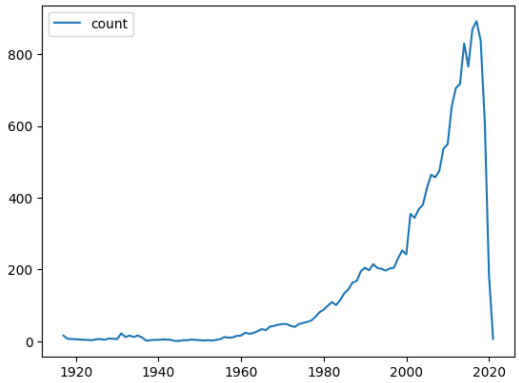

Fig. 3 New Anime Count Per Year

We also plot the distribution of the number of anime being aired every year over the period of 90 years. From 1998 to 2019, we see a spike in the number of anime being aired every year and after 2019, we see a drop in the number of 'new' anime being aired, which can be correlated to the pandemic effect. We also plotted the scatterplots of score v. rank and score v. popularity to understand the distribution of scores for the features of 'rank' and 'popularity'.

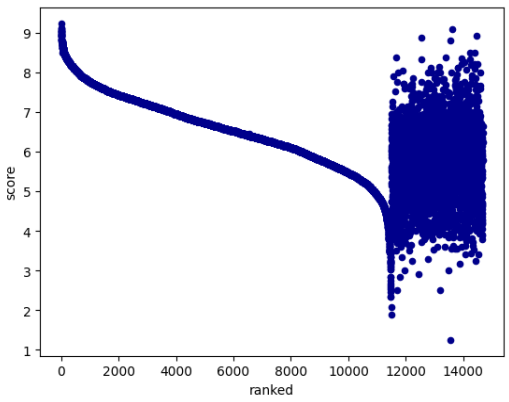

Fig 4. Scatter plot of Score v/s Rank

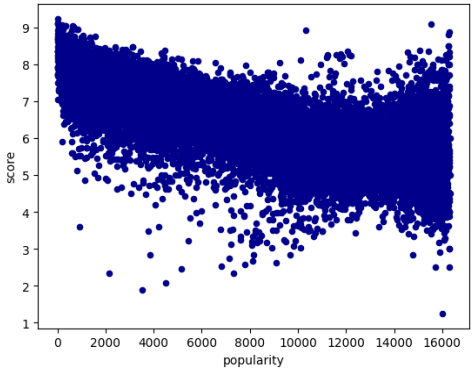

Fig 5. Scatter plot of Score v/s Popularity

In Fig. 4 , we can see that for a certain range of ranks, we can see some correlation with the score, but as the ranks increase beyond a certain value, we find many outliers. In Fig. 5, we can see there is no definite correlation between popularity and score. We can extrapolate that animes which have not been there for a long period of time may not be popular but they can have a high score.

The file reviews.csv contains 192112 reviews with 7 features namely: uid, profile, anime_uid, text, score, scores and link. The feature 'scores' contains 6 fields which are Overall, Animation, Character, Sound, Enjoyment and Story. For every anime reviewed in the dataset, we calculated the average of each of the fields in the 'scores' category for each anime and number of reviews per anime. We try to understand the ranks given to the animes in the 'animes.csv' can be correlated with the average score for each anime with the number of reviews. We find that the average score does not correlate with the rank as per the dataset and the number of reviews need to be taken into account along with the average score to estimate rank.

The file profiles.csv contains 81727 data points with 5 features namely: profile, gender, birthday, favorites_anime and link. This file contains some duplicate values after whose removal leads to 47885 data points. We use the gender feature of the dataset to understand the distribution of the users' gender. We find the majority of the user base is male. Gender distribution data is not conclusive as for many users, it may not have been scrapped properly. In the dataset, we have been given a list of favorite anime lists per user. We leverage this to find which animes occur frequently in every user's list and we rank the animes based on how many times the anime has appeared in all the users' favorite list. We compared the estimated ranks with the ranks provided in the animes.csv file for each anime. We do see some correlation and hence determine the baseline prediction case to predict the score an user will give to the anime.

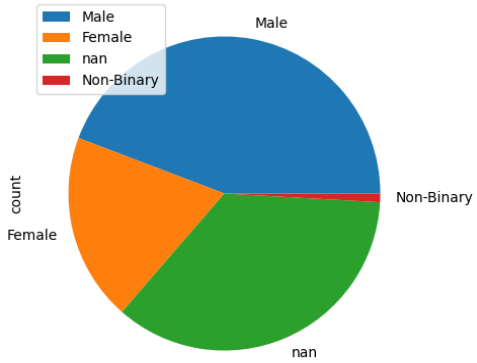

Fig 6. Gender Distribution Of User Base

	count	Title	Rank Calculated Using Favourite Count	Rank As Per Dataset
5114	4915	Fullmetal Alchemist: Brotherhood	1	1
9253	4624	Steins;Gate	2	2
1535	3360	Death Note	3	52
11061	3149	Hunter x Hunter (2011)	4	3
4181	2792	Clannad: After Story	5	12
1575	2779	Code Geass: Hangyaku no Lelouch	6	31

Fig 7. Correlation between Anime Rank Calculated Using Favorite Count & Rank As Per Dataset

2. Predictive Task:

We have performed 2 main predictive tasks for this dataset.

	profile	gender	birthday	favorites_anime	link
0	DesolatePsyche	Male	Oct 2, 1994	['33352', '25013', '5530', '33674', '1482', '2...	https://myanimelist.net/profile/DesolatePsyche
1	baekbeans	Female	Nov 10, 2000	['11061', '31964', '853', '20583', '918', '925...	https://myanimelist.net/profile/baekbeans
2	skrn	NaN	NaN	['918', '2904', '11741', '17074', '23273', '32...	https://myanimelist.net/profile/skrn
3	edgewalker00	Male	Sep 5	['5680', '849', '2904', '3588', '37349']	https://myanimelist.net/profile/edgewalker00
4	aManOfCulture99	Male	Oct 30, 1999	['4181', '7791', '9617', '5680', '2167', '4382...	https://myanimelist.net/profile/aManOfCulture99
5	eneri	NaN	NaN	['5114', '4898', '2904', '1575', '1482']	https://myanimelist.net/profile/eneri
6	Waffle_Empress	NaN	May 29, 1996	['338', '322', '440', '199', '28223', '12815',...	https://myanimelist.net/profile/Waffle_Empress
7	NIGGER_BONER	Male	Jan 1, 1985	['11061', '30', '6594', '28701', '10087', '674...	https://myanimelist.net/profile/NIGGER_BONER
8	jchang	Male	Jul 29, 1992	['846', '2904', '5114', '2924', '72']	https://myanimelist.net/profile/jchang

Fig 8. Dataset Head for Task 1

For the first task we have been given the information regarding what anime a particular user likes. We have reformatted this data into anime user pairs and converted the data into a prediction task of whether a given user will like the anime or not. For training and checking the accuracy of the dataset we have created negative examples of user anime pairs as per the domain knowledge similar to assignment 1 would-play task. Essentially given an anime-user pair, the model generated

in the end should be capable of predicting if the user will like that anime or not. This is a binary classification task. A suitable baseline for this predictive task is the popularity of the anime i.e we will not take into account the users preferences and simply predict that they will like the anime if it is popular. This baseline gave an accuracy score of 74.22%.

	uid	profile	anime_uid	text	score	scores	link
0	255938	DesolatePsyche	34096	\n \n \n \n ...	8	{'Overall': '8', 'Story': '8', 'Animation': '8...	https://myanimelist.net/reviews.php?id=255938
1	259117	baekbeans	34599	\n \n \n \n ...	10	{'Overall': '10', 'Story': '10', 'Animation': ...	https://myanimelist.net/reviews.php?id=259117
2	253664	skm	28891	\n \n \n \n ...	7	{'Overall': '7', 'Story': '7', 'Animation': '9...	https://myanimelist.net/reviews.php?id=253664
3	8254	edgewalker00	2904	\n \n \n \n ...	9	{'Overall': '9', 'Story': '9', 'Animation': '9...	https://myanimelist.net/reviews.php?id=8254
4	291149	aManOfCulture99	4181	\n \n \n \n ...	10	{'Overall': '10', 'Story': '10', 'Animation': ...	https://myanimelist.net/reviews.php?id=291149
5	10046	eneri	2904	\n \n \n \n ...	10	{'Overall': '10', 'Story': '9', 'Animation': '...	https://myanimelist.net/reviews.php?id=10046
6	247454	Waffle_Empress	16664	\n \n \n \n ...	6	{'Overall': '6', 'Story': '6', 'Animation': '9...	https://myanimelist.net/reviews.php?id=247454
7	140903	NIGGER_BONER	2904	\n \n \n \n ...	8	{'Overall': '8', 'Story': '7', 'Animation': '9...	https://myanimelist.net/reviews.php?id=140903
8	23791	jchang	2904	\n \n \n \n ...	10	{'Overall': '10', 'Story': '10', 'Animation': ...	https://myanimelist.net/reviews.php?id=23791

Fig 9. Dataset Head for Task 2

For the second task we are given reviews of anime given by particular users. This data doesn't require much reformatting and after extraction from the dataset we have data containing anime user pairs and a score out of 10 that the user has given to that anime. We transform these data of anime-user pairs into dict of sets namely, itemsPerUser and usersPerItem for applying regressive models discussed in the Class. It's a regression task of predicting the score that a particular user will give to an Anime. For checking the accuracy of the model created we have split the data into 80% training and 20% test set. For the baseline of this task we chose to predict the users average score if the user has been seen in the training data or the average of scores in the training data if the user wasn't seen previously. This simple baseline gave an MSE of 3.7915.

3. Selection and Design of Model:

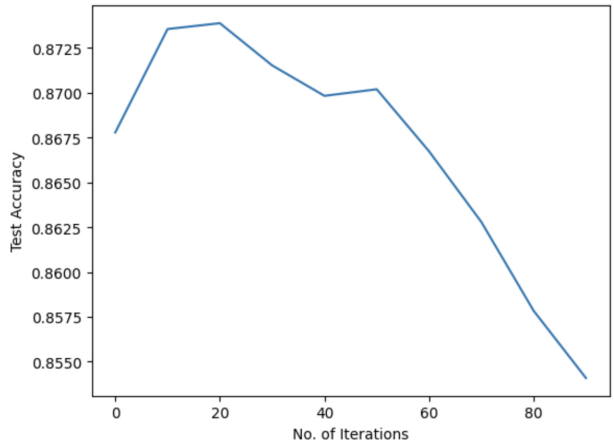

Fig 10. BPR model Test accuracy dropping after 50 iterations(Overfitting) for Task 1

Predictive Task 1 - The final model that worked for this classification task was a Bayesian Personalized Ranking Model. The BPR model makes use of an ItemsPerUser dict of sets as well as a set of all items to generate negative samples for training purposes. The model performance after hyper parameter tuning was significantly better than the baselines and gave an accuracy of 87.29% after 30 iterations. The model was overfitting after 50 iterations as the training accuracy was becoming better but the test accuracy was getting worse as can be seen in fig 10. Although this was a Classification problem, typical classical models like logistic regression were not feasible since the data was in the form of user-item pairs unlike the numerical features that can be used in a logistic model.

Predictive Task 2 - The final model that gave the best results was the fastFM FMregression model. The model completely outperformed the baselines MSE of 3.7915 by giving an MSE of 1.908 on the Test set. In order to apply this model to the data the data had to be converted to a 2D array of size(dataset_size, no. of Users + no. of Anime) where each row is only set 1 in position for the corresponding anime and user. This model takes into account both the user features, anime features as well as their combined features in order to make the prediction and is an example of collaborative based filtering. We also tried the model for gradient descent with regularization. We tried different values of lambda and no. of iterations but the model always gave an MSE greater than the baselines as can be seen in fig 11. This can be because the model doesn't take into account the combined features of the user and anime.

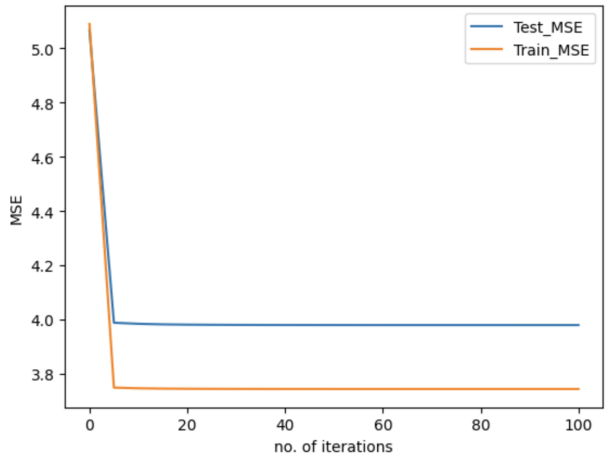

Fig 11. Gradient Descent with regularization, test and train MSE Vs no. of iterations for Task 2

4. Literature Review:

The datasets were originally scapred from the openly accessible website MyAnimeList [8] and acquired from here [9]. MyAnimeList is a social networking and cataloging website for anime and manga that is sometimes shortened as MAL. It offers a comprehensive database of manga and anime and facilitates the search for individuals with similar interests. MAL features a thriving community that is enthusiastic about anime and manga, and it stores information on over 17.5K of them. It is the go-to resource for anime information for otakus worldwide, with numerous reviews and rankings for anime. Three distinct datasets are included in the contents. The first dataset contains a list of animes with their title, genre, rank, popularity, score, aired date and number of episodes. The second dataset contains profile information of users such as their username, gender, birth date and favorite anime list. The third dataset contains reviews from users about animes which includes the review text and distribution of scores in different categories. Other similar datasets can be found on Kaggle [1-3]. The dataset [2] contains information about the preference list of 73516 users on 12294 anime and their respective ratings. The dataset [3] contains over 130k comments scraped from the openly accessible website MyAnimeList [8] since 2006 and their respective ratings.

One of the papers [4] reviewed used natural language processing and random oversampling methods on the dataset [3] to extract sentiments and reduce the class imbalance. Word2vec model was used to obtain vector representations from the sentiment data extracted to create an embedding layer. The embedding layer served as input for the CNN model to train a predictive model. The model achieved an accuracy of 99.85%. The acquired findings demonstrated that the suggested convolutional neural network performs better than a number of methods, including k-nearest neighbors, random forests, and decision trees. Another paper [5] used machine learning techniques like content-based filtering, collaborative-based filtering and popularity based filtering on Kaggle dataset [1] to obtain top matching recommendations and predicted ratings. This paper demonstrated traditional recommendation techniques that can reduce model complexity and provide top anime recommendations for particular users. In another study [6], a model was proposed to predict the ratings of Bollywood movies. The data used for the analysis were collected from various sources. Seven different machine learning algorithms were trained on the dataset collected and ANN performed the best among them in the prediction task. The paper [7] devised an unsupervised based machine learning approach for movie recommendation tasks. Several algorithms such as K-Means algorithm, birch algorithm, agglomerative clustering algorithm, spectral clustering algorithm, etc were trained on the MovieLens dataset . The purpose of this research was to optimize K and use different evaluation methods to compare the clustering performance of the different models trained. The study showed that the birch algorithm performed better than the other models.

5. Conclusion and Results:

Both the prediction tasks that we have performed can be used very well for recommender systems. In the first task using the BPR for a given user we can find the scores for anime that the user hasn't watched and recommend the ones with the maximum score as the model suggests there is a high chance of the user liking these anime. Using the second model of FMregression we can predict the ratings that a user might give to an anime that they haven't watched and can recommend the ones that will be rated highly by the user as per the model. This is an example of how different classification and regression tasks can be used in order to make user specific recommendations.

Both models that gave better results than the baselines make use of the features animePerUser and usersPerAnime which are majorly used features in collaborative filtering based recommendation systems. The models that worked take into account the combined features of the anime and user rather than considering just the features of anime or just the features of anime and user separately. Thus we can claim that taking into account the combined features of users and the items will more often give a model with better predictions than taking their features separately and fitting the model.

6. References

[1]https://www.kaggle.com/datasets/azathoth42/myanimelist
[2]https://www.kaggle.com/datasets/CooperUnion/anime-recommendations-database

[3]https://www.kaggle.com/datasets/natlee/myanimelist-comment-dataset
[4] S. M. AlSulaim and A. M. Qamar, "Prediction of Anime Series' Success using Sentiment Analysis and Deep Learning," 2021 International Conference of Women in Data Science at Taif University (WiDSTaif), Taif, Saudi Arabia, 2021, pp. 1-6, doi: 10.1109/WiDSTaif52235.2021.9430244.
[5] https://dx.doi.org/10.2139/ssrn.4121831
[6] A. Kanitkar, "Bollywood movie success prediction using machine learning algorithms", 2018 3rd International Conference on Circuits Control Communication and Computing (I4C), pp. 1-4, 2018.
[7] Cintia Ganesha Putri, D.; Leu, J.-S.; Seda, P. Design of an Unsupervised Machine Learning-Based Movie Recommender System. Symmetry 2020, 12, 185.
[8] https://myanimelist.net/
[9]https://www.kaggle.com/datasets/marlesson/myanimelist-dataset-animes-profiles-reviews