Anime Recommendation System

CSPB4502 Spring 2022

Matthew Fournier
Matthew.Fournier@colorado.edu

Christie Hui
chhu9206@colorado.edu

PROBLEM STATEMENT

The aim of this project is to create an anime recommendation system for users based on existing user data obtained from MyAnimeList.net, which is essentially a cataloging site which allows users to keep lists of which anime they've seen/plan to see, view user reviews of anime titles, and rate shows based on a scale from 1-10[1]. The motivation behind choosing this topic is due to a shared common interest in anime as well as prior familiarity with MyAnimeList site functionality. The knowledge and results gained from the analysis of this data set may be beneficial for future personal decisions regarding choosing an anime to watch.

Based on certain features from the data set – such as anime titles, genres, and ratings, to name a few – we aim to discover and answer interesting questions such as:

- Which genres (of anime) contribute to ambiguous users' watching decisions?
- How much do features such as rating, episode count, and anime type influence a user's viewing choices?
- Is the popularity of an anime or similarity to other users' profiles/viewing preferences more significant in impacting a user's preferences?

In addition to the questions listed above, we may also discover or explore other potentially interesting questions that may arise during the EDA portion or data mining/analysis phase of the project. Our project will aim to answer the questions above at minimum.

LITERATURE SURVEY

Prior work on this topic can be seen on popular anime streaming platforms such as Crunchyroll[2], which has an existing recommendation system in place which suggests various titles to a user based on the users' previous watch history on the site. The MyAnimeList site also provides a list of recommendation to users who have added anime entries to their profile, and these recommendations differ from those of Crunchyroll's since they also incorporate other users' unique feedback to support the recommendation[3].

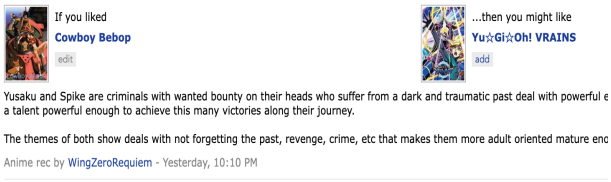

Figure 1: An example of recommendation with unique user commentary provided by MyAnimeList.net.

DATA SET

The data set that our project will be primarily using is titled Anime Recommendations Database, and it was

[1] Summary taken from https://myanimelist.net/about.php
[2] https://www.crunchyroll.com
[3] https://myanimelist.net/recommendations.php?s=recentrecs&t=anime

obtained from Kaggle.com. The data set can be accessed via this URL: https://www.kaggle.com/CooperUnion/anime-recommendations-database. The data set is comprised of two separate CSV files, Anime.csv and Rating.csv, and it contains information on user preference data from 73,516 users on 12,294 different anime titles[4]. The Anime.csv file contains various features (columns) describing anime titles, such as anime_id (the unique identifier assigned to an anime title by MyAnimeList), name, genre, type, number of episodes, rating, and members (the amount of users on the site who have added a particular anime to their profile). The Rating.csv file contains the features user_id (a randomly generated user ID), anime_id (the unique identifier assigned to an anime that the particular user has rated), and rating (the rating out of 1-10 that the particular user has assigned).

PROPOSED WORK

The data set obtained from Kaggle.com is already quite clean. We plan on performing an initial Exploratory Data Analysis on the data before the data pre-processing phase by creating various data visualizations. In the Rating.csv file, the "rating" column (which specifies the rating out of 1-10 that the user has assigned to a particular anime) contains the numeric value -1 if the user has added the anime to their profile but has not designed a rating, whether intentionally or not. To account for these values, we plan on either dropping these values entirely or substituting them with another constant value during the data pre-processing phase. In addition, it is important to note that in the Anime.csv file, there are several titles which belong to the same anime franchise (for example, it contains an entry for the first season of the anime "Haikyuu!!" as well as an entry for a special movie episode of "Haikyuu!!"). To account for these occurrences, we also plan to combine the entries (based on having the same title) which belong to the same anime franchise in the data pre-processing step, since these entries typically consist of 1 episode (as opposed to the typical number of 12 to 24 for most anime series) and could possibly skew the distribution of the ratings for a particular anime(s).

Our project will also aim to develop and train a proper classification model for the data set. A feature that we plan to incorporate that will differ from the prior work done that was mentioned in the Literature Survey portion is to allow for a two-way recommendation system in which user input can also be provided in addition to the existing user data in the data set. The output will be a result of shows that pass a percentage threshold of matching for one of a few models that will test different types of similarity. This can include multiple ways of abstracting the data (studio, art style, type of media). We will also be checking multiple types of models either clustering, decision trees, random forest, k nearest, linear regression, naïve bayes, or k nearest.

TOOLS

The tools which we plan on using for this project include:

- Jupyter Notebook (for formatting and code)
- Pandas, numpy, sklearn, scikitlearn (for calculations, EDA, data processing, and data analysis)
- Matplotlib, seaborn (for visualizations)

EVALUATION METRICS

The evaluation metrics which we plan to use to evaluate our classification model currently include, but are not limited to, are: cross validation, precision, accuracy, F1 score, recall, and R-squared value. The

[4] From the description of the dataset found at https://www.kaggle.com/CooperUnion/anime-recommendations-database

metrics will be used to help us decide which model is effective and drive the best recommendation. The problem that we have really identified with inferior recommendations is due to not having a high enough recall and precision. Recall tells us how often we are correctly identifying a good recommendation while precision will tell us how the model performs over the whole data set. The f1 score will give us a representation of recall and precision together. Ideally, we want to maximize all the scores with focus on the precision. The focus on precision allows us to make the claim of a "better" recommendation system.

MILESTONES

At the time of this proposal, we plan to have at least completed the Exploratory Data Analysis and data-preprocessing steps by the end of Week 10 (March 18). In addition, we hope to be able to finish a majority of the coding portion and model training by the end of Week 11 (April 1).

Milestones Completed

We have completed the Exploratory Data Analysis and Data Preprocessing steps, and we have begun work on the coding portion of the project. Due to some scope increase of the assignment this has been pushed back. This is majorly due to the increased amount of data exploration needed.

Milestones To Do

At the time of writing the project Progress Report (Project Milestone Part 3), we are currently still working on the rest of the coding portion of the project. The coding portion will mainly consist of the recommendation system algorithms we plan on implementing, which includes the machine learning/AI models and analysis to support and solve the goal of creating tailored anime series suggestions based on provided user input and the existing user data from our data sets. As we continue to develop our algorithms and models, we aim to flesh out our project's final report. Lastly, after completing the development of the project code and completing the final project writeup, we will work on the project presentation per the final requirements outlined in the project guide. We are certain that with the time remaining, and given the current status of our project timeline, we can complete the rest of the currently in-progress and future planned milestones.

RESULTS SO FAR

The approach that was decided upon for the Exploratory Data Analysis portion of the project was to look for interesting features in the data that we could pick for future in depth exploration. The main bulk of the EDA consists of examining the data points from the Anime and Rating data sets, checking for anomalies or unwanted aspects in the data, data preprocessing, and data abstraction.

Our first step consisted of checking both how much and what type of information we could potentially utilize in our development of the recommendation system. As previously mentioned, the Rating.csv file contains the fields user_id, anime_id and rating, each of which are the data type int64. This data set consists of 7,813,737 rows, where effectively each row is equivalent to one unique review. When diving further into the data in the Rating.csv file, we noticed that it contained -1 values, which signifies that the user for that particular row included a specific anime on their list, but did not assign a rating to their listing (see Figure 2 below). If there is a negligible amount of these values/missing ratings, we might not do any other work besides preprocessing out these values. We will attempt to explain or utilize these values in our suggestion algorithm.

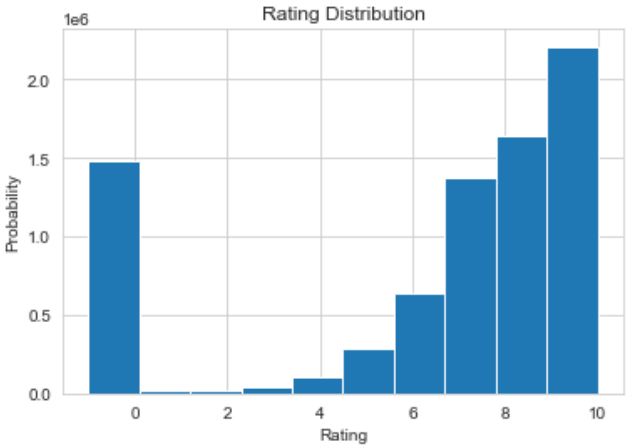

Figure 2: Unprocessed data distribution for the user ratings in the Rating.csv file.

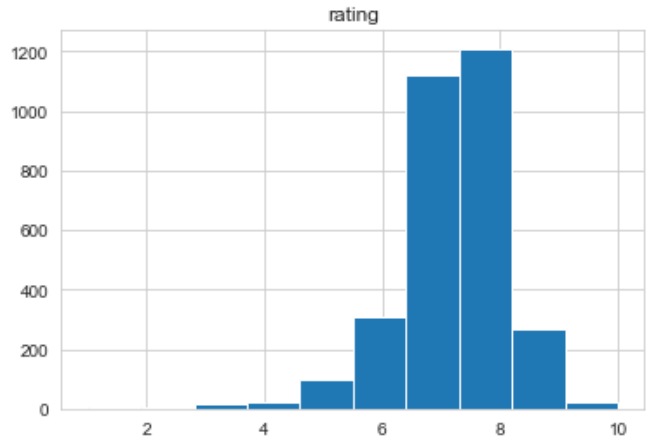

Figure 3: The data after pre-processing out the null values, grouping the anime by anime_id and removing non-tv show types.

Our next task was to ensure that the anime metadata from the Anime.csv file was merged with the user data from the Ratings.csv file. Both data sets were merged into one Pandas data frame on the attribute anime_id, since the data in the Ratings.csv file contains only the corresponding anime_id and not the associated anime title for that particular anime. In addition, we also made sure to drop the "rating" attribute from the Anime.csv data set because this was at an anime_id level abstraction and we stepped down to a user rating level. We then adjusted the merged data frame to only include rows with the value of "TV" in the show type attribute, as we are only attempting to develop an anime series recommendation system and are not focusing on including recommendations for movies, OVAs, or special episodes.

The Anime.csv file data set has more fields than the Rating.csv file data set, as it primarily contains the bulk of the metadata for the anime titles. We have taken some preliminary guesses as to which features will be the most important, with them being genre and episodes (at least for this data set). There is some exploration later that will dive further into the potential significance/influence of these attributes; we made sure to get a picture of the data at a wholistic view in order to add some perspective to our analysis.

Shown in Figure 3 is the data distribution for the user ratings in the Rating.csv file after preprocessing out the -1 values in the Pandas dataframe after grouping the data points by the anime_id attribute. As seen above, there is quite a change when we aggregate the data to this level, with a shift from the higher end of the rating scale to a heavier concentration around the 7-8 rating. This is likely due to polarized ratings on both ends balancing out with middle of the road ratings. There is also an extremely low number of shows that are at the bottom end of the spectrum which means that we will have to make sure that our models do not skew to the positive side of the scale. This is an important observation because models more often than not will take the road of least resistance when predicting an outcome. Based on the distribution of the data, it seems that the distribution will be adequate to make a meaningful prediction. Moreover, by including several outliers from the data, the developed model will be able to make stronger predictions if there really is a degree of correlation.

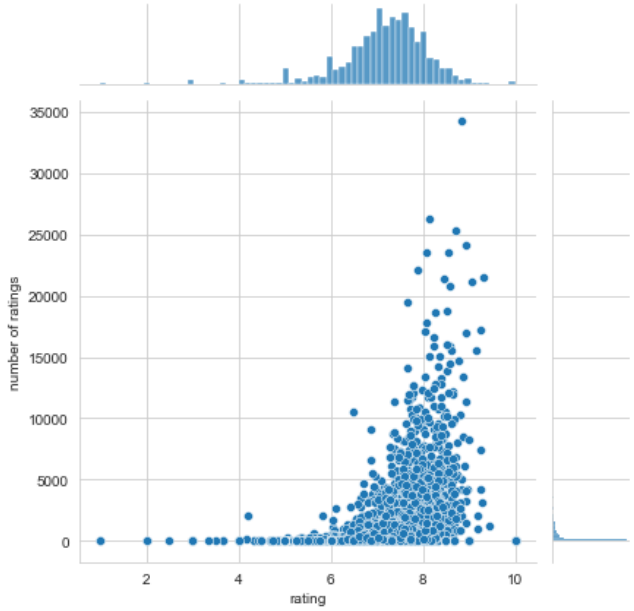

Figure 4: Scatter plot and histogram of the data abstracted to a user rating level.

features. Because this data does not really provide many features it is important that we can provide as many levels of abstraction as possible. This will be explored further in the next page, but it includes one of our personal predictions on what feature we think is likely to be more influential than the others, namely genre.

The graph shown in Figure 5 can help shed some more light into more abstractions of the data. In order to create this visualization, an algorithm was developed that abstracts the data to the genre level. This was initially difficult to achieve because of the way the data was originally presented in the data set. The genre attribute values were formatted in a string list as values that were sepereated by commas and a space. Since many of the data points have several different values under the genre attribute in the Anime.csv data set, we also ran a loop that would capture the combined ratings for each of the genres

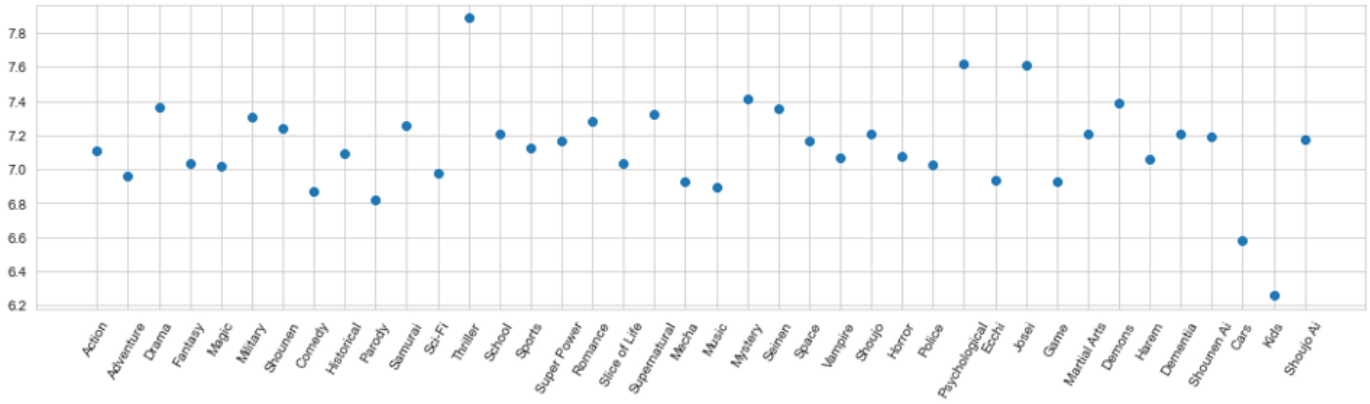

Figure 5: The graph shows the average rating for each genre represented in the Anime.csv data set.

From the scatterplot in Figure 4, we can observe that the more users there are that give a rating for an anime title, there is an increased likelihood that it receives an overall higher rating. This is interesting because it is not necessarily intuitive information that we can abstract from this chart, which is really the whole goal of the project. It is likely that the model will pick up on this trend when it is making predictions and will probably be one of the best

for each individual data point. We also kept an accumulator counter variable for each of the genres as a whole by using a Python dictionary data structure. After all of the loops were complete, the totaled rating for each of the categories along with the total count for each category were used to come up with an average rating, which is shown above in Figure 5. The plot in Figure 5 allows one to observe that there is arguably some level of variance in terms of average rating within each of the genres. While it initially

seems that the average values for the genres are not too far apart at first glance, if observed more closely, one can see that there is a decent amount of variance, albeit not as apparent. For example, the data point for the average rating for anime titles under the “Kids” genre is significantly lower than the average rating values for the other genres; in contrast, the data point for the average rating for anime titles under the “Thriller” genre is noticeably higher than its other counterparts. Based on these observations, we presume that this will likely be the source of some of the negative predictions especially when a user has previously rated a certain genre as low (somewhere along the lines of a rating of 5 or below). It is quite likely that if the rating is consistent with the average values as shown in Figure 5 for anime titles under a particular genre, the algorithm will not suggest many, or any other shows with that same genre.

We plan to include further analysis and visualizations as we continue to round out our project, especially when we begin implementing the clustering algorithm and other machine learning aspects of the project. We have put in a lot of work in the pipeline to decide what features are going to be predictably the most important. The exploratory research performed thus far will help increase awareness of the ways in which the data can be processed/transformed to uncover new aspects, thus aiding in the quest to develop the most optimal anime recommendation system based on user input.

REFERENCES

[1] https://myanimelist.net/about.php
[2] https://www.crunchyroll.com
[3]https://myanimelist.net/recommendations.php?s=recentrecs&t=anime
[4] https://www.kaggle.com/CooperUnion/anime-recommendations-database