Oliver Green, Jiayue Zhao, Qianyong Hu

DS 3000 – Dataset

Topic Idea #1: Anime Popularity Prediction

1. Problem Statement

Anime studios have to pick between hundreds of manga and light novels when deciding what to adapt for their next show. If they choose poorly, they can end up wasting a lot of time, money and effort for nothing. According to blogger ghostlightning anime can cost, "about 50 million yen for a late-night timeslot across 5-7 stations for a 52 episode series." So picking the right material to adapt is a decision that a lot of money rides on.

2. Significance of the Problem

Studios can use their own discretion and experience when choosing what source material they believe will become successful anime. They can look at the popularity of the source material itself, if it fits that studio's target demographic, or the state of the market, like if a lot of similar anime have already been released and people might be tired of that genre. (This isn't the most scholarly source but the information is good: https://anime.stackexchange.com/questions/11446/what-is-required-for-a-manga-to-get-an-anime-adaptation) That process, however, can be time consuming and prone to human error. It also focuses on mostly external factors. It is more difficult to judge a potential anime on its own merits and characteristics. So an ML algorithm that lets studios figure out if an anime will be popular based solely on the idea itself would be a useful tool for preliminary screening.

Anime lovers can also use this tool to find their next favourite anime. Our system could be a great platform for the smaller studios to get noticed.

3. Potential Datasets

- ✔ https://myanimelist.net/ is a website that catalogs detailed information about almost every single anime that comes out. Users can then submit reviews and scores that contribute to a global ranking list.

- ✔ This Kaggle dataset contains information about several thousand anime from MyAnimeList (MAL), including their ranking on a global list and their user rating.

- ✔ This unofficial API can be used to scrape MAL for whatever specific data is needed for the project.

Oliver Green, Jiayue Zhao, Qianyong Hu

anime.csv

Out-of-the-box datasets won't be sufficient:

We could use this dataset to build our ML model around and then once we figure out which features are most important we can scrape MAL ourselves and update this dataset with the most recent values. We can also add more rows for the anime that have aired after this dataset was constructed. Alternatively, we could use this whole dataset for training and construct a separate testing dataset based off of anime that have come out after this dataset was constructed.

Describe your variables below (add more rows if necessary):

Variable name in file	**Description (what the variable represents/means)[1]**	**Feature/ Outcome[2]**
MAl_ID	MyANimelist ID of the anime	Feature
Name	full name of the anime	
Score	average score of the anime given from all users in MyAnimelist database.	Outcome
Genres	comma separated list of genres for this anime.	Feature
English name	full name in English of the anime.	
Japanese name	full name in Japanese of the anime.	
Type	TV,Movie,OVA,etc	
Episodes	number of chapters	
Aired	Broadcast date	
premiered	season premiere	
Producers	comma separated list of producers	
Licensors	comma separated list of licensors	
Studios	comma separated list of studios	
Source	Manga, Light novel, Book, etc.	
Duration	duration of the anime per episode	
Rating	age rate	
Ranked	position based in the score	Outcome
Popularity	position based in the the number of users who have added the anime to their list.	Outcome
Member	number of community members that are in this anime's "group".	Feature
Favourite	number of users who have the anime as "favorites".	
Watching	number of users who are watching the anime.	
Completed	number of users who have complete the anime	
On-Hold	number of users who have the anime on Hold.	
Dropped	number of users who have dropped the anime	
Score-10	number of users who scored 10.	Outcome
...	...	Outcome
Score-1	number of users who scored 1.	Outcome

[1] Refer to dataset descriptions in sklearn

[2] In the Feature/Outcome column, indicate whether the variable is a feature or outcome variable. You need to have at least one outcome variable and nine feature variables.

There are 13 outcome variables and 22 feature variables in this dataset. The outcome variables all relate to the user score of the anime. This includes the average score that gets displayed on the anime's page as well as a breakdown of how many people rated the anime for each possible score (1-10 discrete). Additionally MAL includes two global ranking lists for all anime. The first is based off of the average score of the anime and is recorded in this dataset under the "ranked" variable. The second list is based on how many users have added the anime to their personal anime lists and is called here "popularity." Using these variables we could

calculate a combined metric based off of an anime's average position in the score based and popularity based rankings as well as through other metrics like the distribution of scores. Two anime may have the same average score but one could have a wider distribution of score meaning that it is a more polarizing show and this could be used in the combined metric as well. The feature variables are everything else in the dataset like the name of the show, the genre, number of episodes, type of source material, ect.