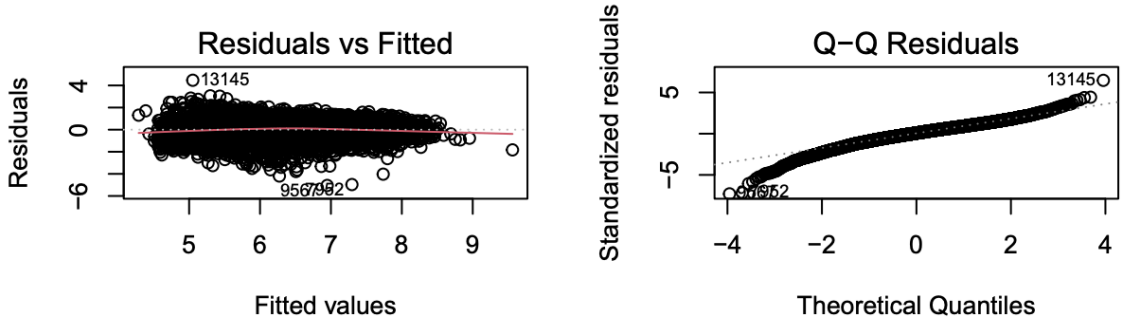
MyAnimeList Anime Scoring Executive Analysis

Aaron Yu

MyAnimeList (MAL) is the largest public database of Japanese animations. The data we used was gathered directly from the website using a crawler and compiled by Matěj Račinský. The dataset contains information on 302,675 unique users, and 14,478 unique shows. The particular data we are interested in is the show data, under AnimeList.csv, which contains 14,478 shows, and 30 columns, including ratings, genre, and other information. In particular our goal is to predict the score, which is on a scale of 1-10 from information about the show to gauge how well received the show would be. After cleaning the data we ended up using a linear regression model with a power transformation that predicts the score from a total of 34 predictors, consisting of the number of episodes, the number of people who scored the show, the type of animation, the reference source, the rating, and which of the five most popular genres that the show fit into (Comedy, Action, Fantasy, Adventure, and Drama).

We started by cleaning the data, we wanted meaningful results, so we accomplished this by first removing shows that had only been rated by less than 10 users, these were often extremely niche projects that were added to MAL and would have skewed rating data to be lower and higher, many of these also were essentially unscored and had a score of 0 which also contributed to outliers. We also removed any unreleased shows as those also had an effective score of 0. After that we started by selecting two quantitative and three categorical predictors for our baseline model. We then testing our linearity assumptions using a Residuals vs Fitted and Q-Q Residuals plot. We can see that we have a bit of a fan shape in the residuals plot and the tail for the Q-Q plot is lower, this resulted in us decided to use a



Box-Cox test to determine that a power-transformation with $\lambda = 2.2$ would be appropriate. After this we decided to fine tune our data by incorporating genre data. This was tricky since each show had its own set of genres that they fit into, what we ended up doing was compiling genre counts over all the shows, then chose to create new categorical variables for the top five genres indicating whether a show fit into one of these categories. We then fit our new model and used stepwise model selection with AIC in order to decide what the best model would be, we ended up with a model keeping all our original variables and incorporating all five genre variables.

Our data had several interesting results, given how effective our model was for predicting the scores of shows, we can use the model structure to understand what comprises a successful and well received show. Some important results are the following, TV shows and Movies performed the best, with Music animations performing the worst. Text that was adapted from Web Manga/Manga/Novels performed the best, indicating that having an existing fanbase is important for the success of the show. We also saw that more family friendly shows performed better with PG, PG-13, and G animations doing the best, and adult animations doing the worst. We also learned that Adventure and Drama type shows performed well. These are all important data points for studios when considering what shows they want to greenlight and produce.

Finally, it's important to understand the limitations of our data. MyAnimeList is an inherently western-leaning dataset, and because of the nature of the platform it consists of users who are already predisposed to the genre, and more hardcore fans, which may result in lower perceived quality than how the general public would actually react to a show. We were also unable to accommodate as much genre information as we would have liked due to the nature of the dataset. Another limitation was time data, we were unable to penalize/account for when shows were aired, older shows naturally have larger fanbases which can significantly affect overall score and viewership, making it more difficult to determine what current trends make popular shows.

Sources

Matěj Račinský, “MyAnimeList Dataset.” Kaggle, 2018, doi: 10.34740/KAGGLE/DSV/45582.