Predicting MyAnimeList Score of Modern Anime based on their Features

Razan Ahsan Rifandi (1009562108)
30 April 2025

1 Introduction

1.1 Background

Anime is a genre of animated works with styles that has their roots originating from Japan. The word itself originated from the same Japanese term, derived from a shortening of the English word 'animation'. Historically, it is used to specifically refer to animation produced in Japan. In recent times, it more commonly refers to the distinct group of animation styles itself disregarding the work's country of origin, although a vast majority of them still originates from Japan. Anime itself has had its fair share of history, with the style evolving over time to reflect advancements in animation technology and a growing audience. A noticeable shift is around the year 2012 where HD quality and a defining art style more reminiscent of today's anime is starting to be adopted.

Anime is typically released via television broadcasting in seasonal cycles. Some may also be broadcasted through theaters instead as movies outside the seasonal cycles. Many anime are adaptations of existing source material, such as manga (Japanese comics), light novels, or video games, while others are original creations. The production of anime is often handled by specialized animation studios (e.g. MAPPA, Kyoto Animation, etc.), with some being associated with their production quality.

Measuring how "good" an anime is inherently subjective, especially considering the difference in themes such as stories and animation style and how well they resonate with a viewer's preferences. For the purposes of this research, I am interested in the overall community rating for an anime from the website MyAnimeList (MAL). It is one of the largest online databases and communities for anime alongside AniList and Kitsu.io. Each anime entry on MAL includes detailed information such as its type (TV series, movie, etc.), source material, genres, studios, and demographic ratings (e.g., G, PG-13, R). Users can rate anime on a scale of 1 to 10. Additionally, MAL tracks the number of "members" (users who have added the anime to their list) and "favorites" (users who have marked the anime as a favorite). These metrics could provide insights on the correlations of an anime's features and its rating, which is the main motivation of this research.

1.2 Research Question

For the purposes of this research, I will be defining anime as television series or theatrical releases that are listed as entries on MAL. Aside from those, MAL also lists OVAs (Original Video Animations), commercials, previews, and even music as entries. As they are not what would generally be categorized as animes, they will not be included in this research. Furthermore, this research will specifically focus on anime released over the 10 year period from 2015-2024, post the shift in quality around 2012s, which I will be referring to as "Modern Anime". With that in mind, the research question for this study is: **"How can we predict the MAL score of modern anime based on type, source material, age rating, popularity metrics, animation studios, and genres?"**. To answer this research question, I will be exploring several machine learning models such as Multiple Linear Regression, Regularized Regression, Support Vector Regression, Random Forest, and XGBoost. After training, the best models will be interpreted in order to find the most influential features of determining MAL score.

2 Methods

The analysis is conducted using the programming language R, using relevant libraries. Subsequent parts (such as EDA and model training) will be discussed with the context of R and relevant libraries.

2.1 Data Collection, Cleaning, and Wrangling

The data for this research was collected using Jikan, an unofficial and open-source API for MAL. The data was collected for all anime released between 2015 and 2024 accessed by season for each year (Winter, Spring, Summer, Fall). Since the API uses pagination, the data collection iterates through the pages of each season to retrieve all entries. The response is in JSON format and includes nested structures for some fields such as studios genres. In order to convert it into a proper R dataframe, the nested values were flattened by extracting the names and combining them into comma-separated strings. After collecting the data for each season, the individual datasets were merged into a single dataframe for analysis. The columns found in the dataset are described in Table 1 below. The data is then further processed for analysis as follows:

1. Removing duplicates, MAL may list the same anime entry in multiple seasons (e.g. if the anime was still airing during data collection). Duplicates were removed using the unique `mal_id` column which is then dropped.
2. Filtering relevant entry types, The dataset was filtered to include only entries that align with our definition of anime: animated shows with television or theatrical broadcasts. This is represented with the value `TV` and `Movie` in the `type` column.
3. Removing currently airing shows, as their scores and popularity metrics may not be stable. They were removed using the `airing` column which is then dropped.

Table 1: Description of Variables in the Anime Dataset

Variable	Description
`mal_id`	Unique MyAnimeList identifier
`title_english`	English title (or romanized Japanese title if no English title exists)
`airing`	Boolean indicating if anime was airing during API call.
`type`	Format of the anime (TV, Movie, OVA, etc.)
`source`	Source material (manga, light novel, original, etc.)
`year`	Release year of the anime
`rating`	Demographic rating (G, PG-13, R, etc.)
`score`	Average user rating on MyAnimeList (1-10 scale)
`members`	Number of users who added the anime to their list
`favorites`	Number of users who marked the anime as favorite
`studios`	Animation studio(s) responsible for production
`genres`	Associated genres (Action, Romance, Fantasy, etc.)

4. Reformatting the columns, `studios` and `genres` columns were reformatted so that empty strings (a byproduct of the data collection process) and `Unknown` genre were replaced with NA values. The NA values were subsequently observed in order to process them.
5. Removing the NA values, the columns with the most NA values are studios and score, which is the response variable. Since there are a lot of overlap in the rows with NA values for these columns, I decided to remove all NA values for convenience of data usage in the future.

After cleaning, the final dataset consists of 2,550 entries and 10 columns (with the columns `mal_id` and `airing` being removed). The raw code for calling the API, collecting the data, and cleaning the data will be provided in the .Rmd file, but as the entries could update in the future, the generated clean dataset will be provided in a .csv file for reproducibility.

2.2 Exploratory Data Analysis

EDA was conducted to examine the variable distributions, distribution of scores across categories, and potential relationships between predictors as follows:

- Count of ‘type‘, ‘source‘, ‘rating‘ (Summary Table)
- Anime distribution by Year (Barplot)
- Distribution of ‘score‘, ‘members‘, and ‘favorites‘ (Summary Table and Histograms)
- ‘scores‘ vs. ‘type‘, ‘source‘, and ‘year‘ (Boxplot)
- ‘scores‘ vs. ‘studios‘ and ‘genres‘ (Summary Table)
- Pair scatterplot of ‘scores‘, ‘type‘, and ‘rating‘

The visualizations are available in the project’s EDA page.

The EDA showed that an anime’s score on MAL seems to vary by type, source material, age rating, animation studios, and genres. Additionally, popularity metrics such as members

and favorites show a strong linear correlation with scores, indicating that more popular anime are likely to receive higher ratings. Additionally, There are more anime created pre-2019 than post-2019, suggesting a potential decline in the number of new anime releases in recent years (possibly due to the COVID-19 pandemic). The score distribution does not vary significantly across years, which justifies not using it as a predictor for score.

2.3 Modeling and Evaluation

Before modeling, the data was further preprocessed in order to enable model training. To address multicollinearity, the `favorites` was dropped, retaining `members` as the only numerical predictor. I justified this since I believe `members` is more reliable metric as users need to add an anime to their list in order to rate it, but do not always mark it as a favorite. Based on EDA findings, a log transformation was applied to `members` in order to address skewness. For categorical columns, I used one-hot encoding for `type`, `source`, and `rating` and multi-hot encoding for `genres` and `studios`. Non-predictor columns such as `title_english` and `year` were also dropped. The final encoded dataset was comprised of 428 features.

The relatively limited entries and rare categorical values suggests that training and model selection needs to be done carefully. For example, a rare value (e.g. the studio `Pierrot Films` having only 1 entry associated to it) might not appear in the training dataset, resulting in an invariant column which might cause problems for training models such as Regression Trees. To mitigate these issues, 10-fold cross validation is employed when training the models to efficiently make use of the small dataset and ensure generalization. Each model with hyperparameters are also tuned using grid search. As training can be quite computationally expensive, parallel methods using 4 cores are used to speed up computation.

To compare model performance, these metrics were used:

1. R^2, representing how much variability were captured in the model.
2. $RMSE$ or Root Mean Squared Error. This metric is used to penalize large errors.
3. MAE or Mean Average Error. It is similar to $RMSE$ in that it also measures prediction error, but it directly calculates the average errors instead of first squaring the errors.

To choose the best model according to our research question, $RMSE$ will be used for hyperparameter tuning, which is the most widely used metric to minimize performance error. Furthermore, $RMSE$ is preferred over R^2 in order to mitigate the possibility of the dataset not being well-behaved enough, which is relevant to our research question as anime scoring could inherently have a lot of variance.

The following models were considered:

- **Multiple Linear Regression**
  This model assumes a direct linear relationship between predictors and scores and thus is used as an easily interpretable baseline model. As the dataset mostly consists of categorical variables (`members` being the only numerical variable), the model is directly fit to the dataset without any further variable transformation to provide baseline metrics.
- **Regularized Regression (ElasticNet)**

ElasticNet is a linear regression model that combines L1 (Lasso) and L2 (Ridge) regularization to avoid overfitting and improve prediction accuracy and interpretability. Lasso regularization helps in feature selection by shrinking some coefficients to zero, while Ridge regularization reduces overfitting by penalizing large coefficients. By blending both penalties, ElasticNet is particularly useful when dealing with datasets that have multicollinearity or when the number of predictors is large relative to the number of observations. The hyperparameters to be tuned are `alpha` (the balance between L1 and L2 penalties, with 0 being pure ridge and 1 being pure Lasso) and `lambda` (overall strength of regularization).

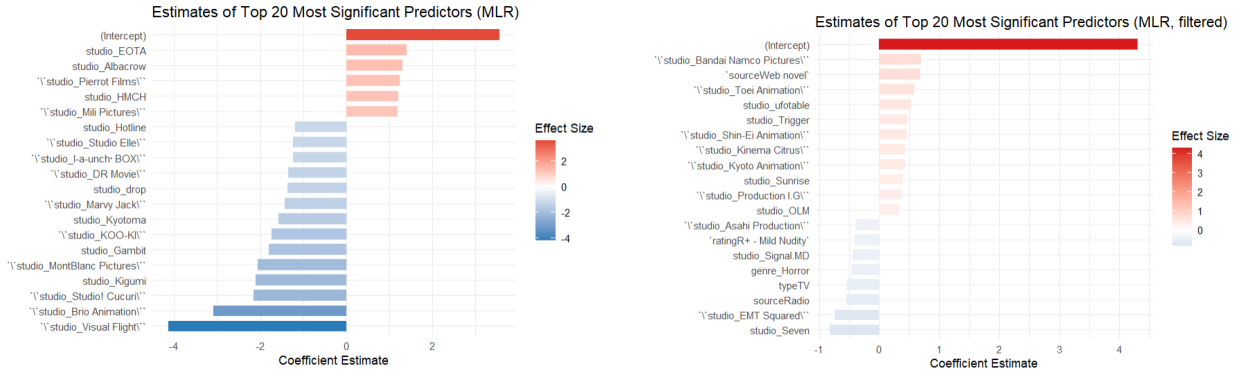
- **Random Forest**
  This model is an ensemble learning method that constructs multiple decision trees during training and averages their predictions to reduce variance and improve accuracy. This approach makes the model robust against overfitting while capturing complex, non-linear relationships in the data like regular Regression Trees. To train the model, the `ranger` package was used instead of the `randomForest` library for better computational efficiency and support for modern tuning parameters. The hyperparameters to be tuned are `mtry` (number of randomly selected features considered at each split) and `min.node.size` (minimum observations in terminal node).
- **XGBoost**
  XGBoost is a scalable end-to-end tree boosting system. It is a highly-optimized framework that combines gradient boosting (constructing trees sequentially with each tree correcting the errors of the previous one, minimizing $RMSE$ using gradient descent), regularization, and hardware optimizations. The `xgboost` library is used to train the model. The hyperparameters tuned are `nrounds` (how many sequential trees or boosting iterations is built), `eta` (learning rate), `max_depth` (maximum tree depth), `colsample_bytree` (proportion of features considered at each split), `min_child_weight` (minimum weight needed in child nodes), and `subsample` (fraction of samples per tree).

3 Results

The four previous models were trained sequentially. Initial fitting using Multiple Linear Regression was done using all predictors. The most significant predictors were either the intercept or studios with few associated entries (Figure 1), which concerns us with overfitting. The model was then retrained on filtered dataset including only encoded values with $\geq$12 associated entries. The refit model still identified several studios as significant predictors (Figure 1), which suggests studio information may genuinely influence MAL scores.

Regularized regression was then employed which maintained linear assumptions while penalizing model complexity. The variable importance plot (Figure Y) showed studios remained among the most influential predictors after regularization, further justifying including them to our models.

Finally, two non-linear tree-based models were trained: Random Forest and XGBoost. The final model performance metrics for all predictive models can be found in (Table 2) below. Additionally, the hyperparameters value considered and final model value can be



(a) Initial MLR (b) Filtered MLR

Figure 1: MLR coefficient estimates of top 20 most significant variables (A) before and (B) after filtering rare studio entries.

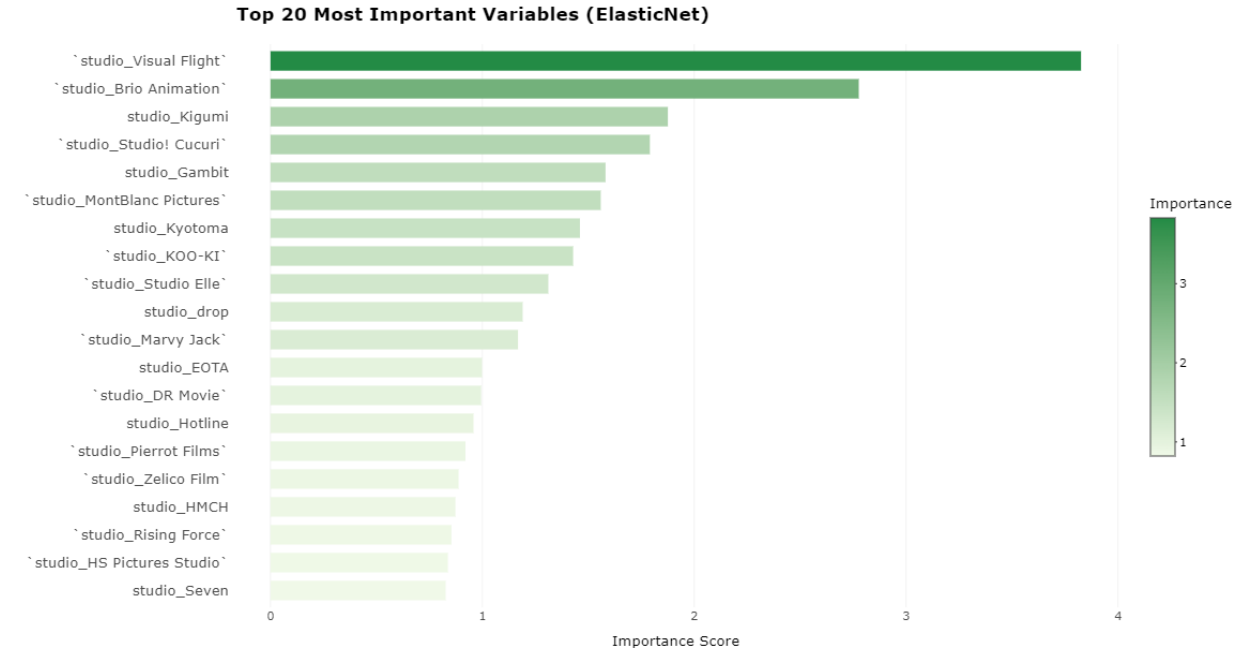


Figure 2: Variable Importance Plot of Regularized Regression model

found in (Table 3).

Table 2: Performance Comparison of Built Models

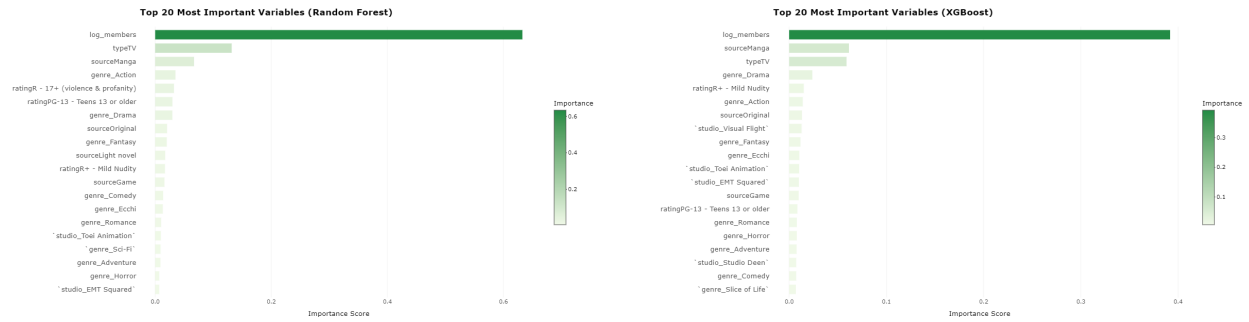
Model	**RMSE**	**R^2**	**MAE**
MLR	0.59815	0.5464	0.46136
MLR (Filtered)	0.60834	0.52662	0.46969
Regularized Regression	0.5832	0.56461	0.44931
Random Forest	0.5619	0.60261	0.41889
XGBoost	0.55792	0.6025	0.42065

Table 3: Hyperparameter Tuning Details for Predictive Models

Model	**Hyperparameter**	**Values Considered**	**Final Value**
Regularized Regression	alpha	0, 0.1, ..., 1	0.5
	lambda	10^{-3} to 10^{-1} (100 values)	0.012
Random Forest	mtry	5, 10, 20, 50, 100	100
	min.node.size	1, 3, 5	1
XGBoost	nrounds	100, 200	200
	max_depth	3, 6, 9	9
	eta	0.01, 0.1, 0.3	0.1
	colsample_bytree	0.6, 0.8	0.6
	min_child_weight	1, 3	1
	subsample	0.8, 1	1

From the performance metrics, we can see that the tree-based models (Random Forest and XGBoost) showed similar performance on 10-fold CV. The top 20 variable importance plot (Figure 3) for both models agrees on the importance of these variables: `log_members`, `typeTV`, `sourceManga`, `genre_Action`, `ratingPG-13`, `genre_Drama`, `sourceOriginal`, `genre_Fantasy`, `sourceLight novel`, `ratingR+`, `sourceGame`, `genre_Comedy`, `genre_Ecchi`, `genre_Romance`, `studio_Toei Animation`, `genre_Adventure`, `genre_Horror`, and `studio_EMT Squared`.

On both models, `log_members` is the most important variable by a wide margin, which is expected as `log_members` directly measures popularity and more popular works are often associated with higher rating. Furthermore, entries with high members (amount of rating) would imply that the ratings should be more stable and predictable for our model given that more people took part to rate it. `typeTV` is also an important predictor in both model, as our EDA and previous linear models does indicate that movies are typically higher rated than TV series. We can also see certain ratings such as PG-13 and R+ which are associated with more mature shows being important predictors. Lastly, the appearance of certain sources (`sourceManga`, `sourceOriginal`, `sourceLight novel`, `sourceGame`), genres (`genre_Action`, `genre_Drama`, `genre_Fantasy`, `genre_Comedy`, etc.), and studios (`studio_Toei Animation`,



(a) Random Forest (b) XGBoost

Figure 3: Variable Importance Plot of (A) Random Forest and (B) XGBoost

`studio_EMT Squared`) indicate that the values in these features imply associations with the score (such as positive, negative, and non-linear/polarizing).

Finally, the Actual vs Predicted plot using the Random Forest model (Figure 4) shows the variation in the dataset were captured well. The plot shows that outlier values with low scores tend to be overestimated by the model, while high scores outliers are rarer more predictable relative to low scores.

4 Conclusions and Summary

The study evaluated multiple predictive models to analyze the factors influencing MAL (MyAnimeList) scores. The first model considered was Multiple Linear Regression (MLR), which initially raised concerns about overfitting due to the significance of predictors with few entries. After filtering rare studio entries, the refit MLR model still identified several studios as significant predictors, suggesting a genuine influence of studio information on MAL scores.

To address model complexity, we employed Regularized Regression, which maintained studios among the most influential predictors, further justifying their inclusion. Finally, we trained two non-linear tree-based models—Random Forest and XGBoost—which demonstrated superior performance compared to linear models, as evidenced by lower $RMSE$ and higher R^2 valued.

The conclusion to our research question was that **popularity, type, source material, age rating, animation studios, and genres seems to have an effect on MAL score of modern animes**. The variable `log_members`, which measures popularity, was the most influential predictor, indicating that an anime's popularity strongly correlates with its score. Movies are also associated with higher rating than TV series as well as more mature series. Specific sources (e.g., manga, original, light novel) and genres (e.g., Action, Drama, Fantasy) were consistently important, which suggests inherent associations with MAL score. Lastly, studios appeared as notable predictors, justifying our assumption that studios (and consequently their work quality) plays a role in determining MAL score.

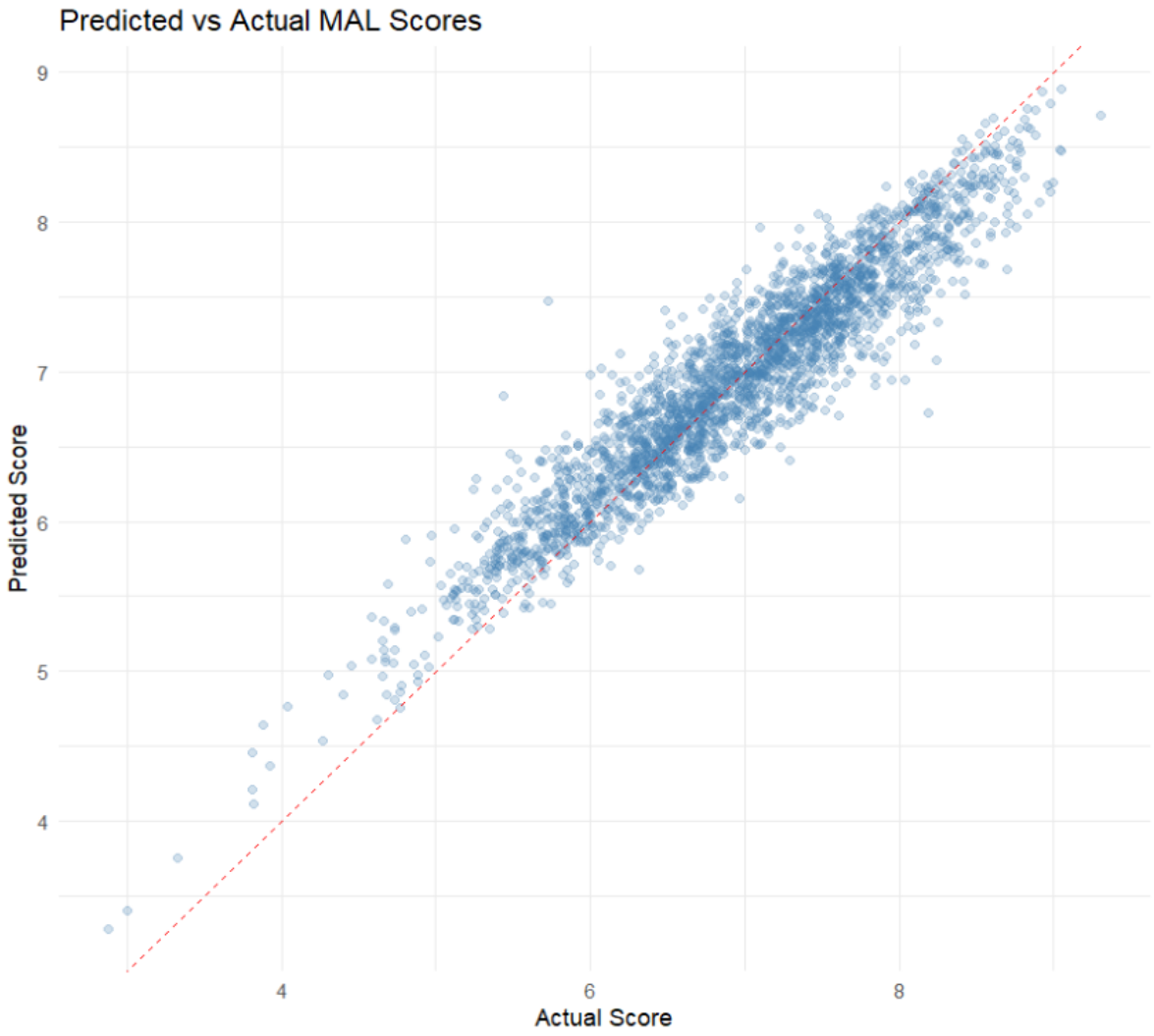


Figure 4: Actual vs Predicted plot using full dataset (Random Forest)