Final Paper

STOR 520 Group 2

December 07, 2023

Introduction

In today's diverse anime industry, a deep understanding and analysis of the emotional trends in anime works are crucial for creators, distributors, and even anime enthusiasts. Anime is not just a form of entertainment, but also a key channel for cultural and emotional expression. By exploring the emotional elements in anime, we can better understand its appeal and impact on audiences. Our project, through a detailed analysis of a vast anime database, mainly focuses on shoujo (girls') and shounen (boys') anime, aiming to answer the following two core questions:

1. How can sentiment analysis techniques be used to understand the emotional trends in shoujo and shounen anime over the years?

This research focuses on whether anime maintains emotional stability and investigates whether there are significant emotional differences between different types of anime to reveal the evolution and characteristics of emotional expression in anime targeted at different audiences. What types of themes and genre combinations do girls and boys prefer? For example, storylines paired with romantic subtexts. The aim of this exploration is to understand the emotional atmosphere such combinations can create to cater to the tastes of adolescent audiences.

2. When examining the scores of anime targeting teenage audiences, which models are used to ensure the better predictions?

Our study employs Linear Regression, Random Forest Regression, and XGBoost Regression models to predict missing scores in our dataset. We compare these predictions with IMDb ratings and audience reviews to validate the effectiveness of our models in predicting viewer preferences.

Through in-depth research on these two questions, our goal is to reveal the diversity and complexity of emotional expression in shoujo and shounen anime, providing richer insights for the creation and promotion of anime. Whether you are interested in the emotional aspects of anime works or want to understand the uniqueness of emotional expression in anime for girls and boys audiences, this project will offer valuable information and fresh perspectives.

Data

In our project, we utilized a dataset named `Anime Database 2022` available on the Kaggle platform. However, this dataset was not collected by Kaggle itself, but rather through web scraping techniques from the MyAnimeList.net, one of the most popular anime websites that operate as a user-generated database, allowing registered users to create and maintain lists of anime and manga they have watched or read. This data scraping was conducted on September 20, 2022, and it gathered comprehensive details of 21,460 anime entries. Our primary focus was on specific variables related to the sentiment analysis of shoujo (girls') and shounen (boys') anime. We obtained it by filtering the **Demographics** column. The following data table consists the entire 2,701 entries of such anime. Each observation (or data row) represents an individual anime work and includes variables such as **Title**, **Favorites**, **Ratings**, and **Members**. Furthermore, for our subsequent analysis of audience preferences in shoujo and shounen anime, we also considered other variables like anime **Themes** and **Genres**. The following table provides an overview of the most important variables in our dataset:

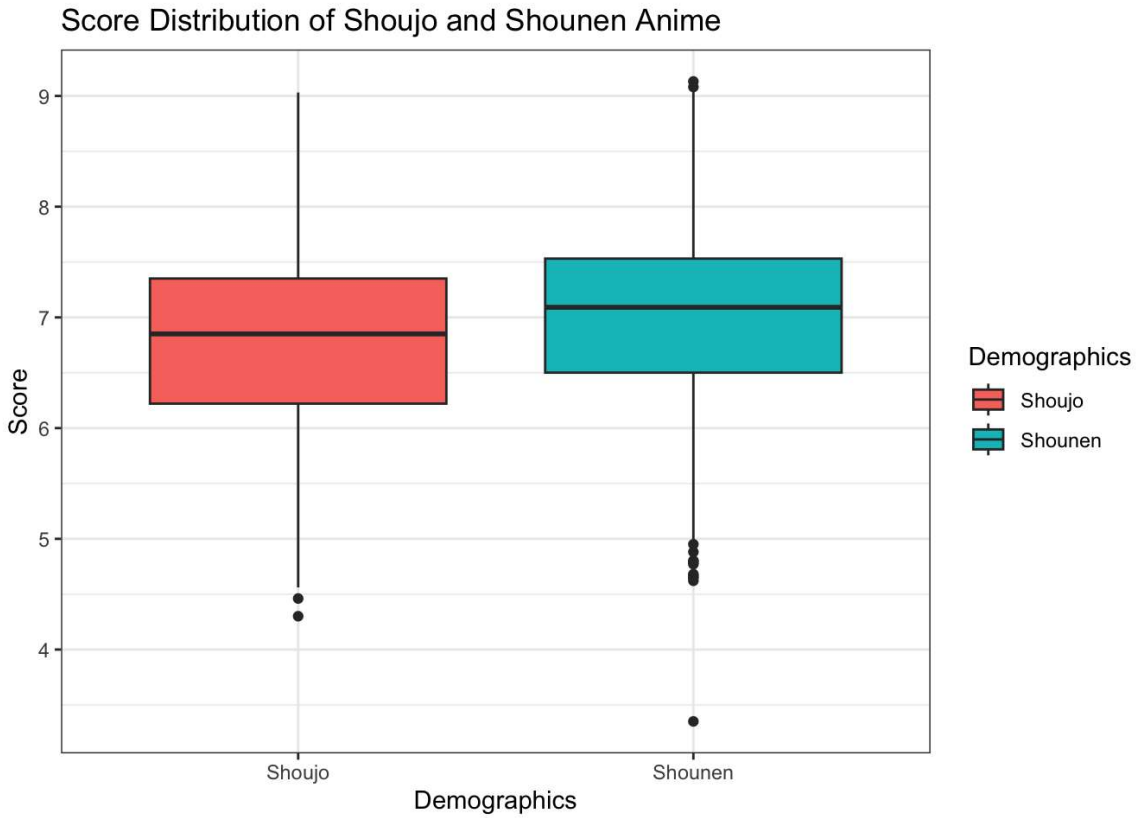
Show 3 **entries** **Search:**

	ID	Title	Synopsis	Type	Episodes	Start_Aired	Source	Genres	Them
1	16498	Shingeki no Kyojin	Centuries ago, mankind...	TV	25	7-Apr-13	Manga	Action, Drama	Gore, Milita Survi
2	1535	Death Note	Brutal murders, petty...	TV	37	4-Oct-06	Manga	Supernatural, Suspense	Psych
3	5114	Fullmetal Alchemist: Brotherhood	After a horrific alchemy...	TV	64	5-Apr-09	Manga	Action, Adventure, Drama, Fantasy	Milita

Showing 1 to 3 of 2,701 entries

Previous 1 2 3 4 5 … 901 Next

These variables are key elements that make up the anime database, collectively portraying a detailed profile of each anime. **ID** is a unique identification code assigned to each anime for easy tracking and reference. **Title** refers to the anime's original name. **Synopsis** provides a brief overview of the anime's content, while **Type** describes the format of the anime, such as TV series or film. **Episodes** indicates the number of episodes in the anime, and **Start_aired** records the date when the anime was first broadcasted. **Source** refers to the original material of the anime, for example, whether it is adapted from a manga or an original creation. **Genres** and **Themes** relate to the category and central ideas of the anime, such as science fiction, romance, or adventure. **Demographics** clarifies the primary audience the anime is aimed at, like boys or girls. **Minutes** typically refers to the duration of a single episode in minutes. **Rating** is the suitability level of content for different age audiences. **Score** reflects the overall evaluation of the anime by the audience, and **Ranked** shows the anime's relative position within the platform based on the score. **Members** refers to the number of users following the anime, and **Favorites** records the number of users who have marked the anime as their favorite. Together, these data provide a comprehensive picture of each anime, helping us understand and analyze anime works from multiple dimensions.



We aim to gain insights from the box plot above depicting the distribution of scores for shoujo and shounen anime, as our ultimate focus involves exploring predictive models. We aspire to better understand and anticipate the scores assigned to anime specifically tailored for teenage audiences. Both have their median around 7. Additionally, the points above and below the boxes represent outliers, and we can see more outliers for Shounen anime, especially some with particularly low scores. While the average score for Shounen anime might be higher, there is also greater variability in its scores.

Results

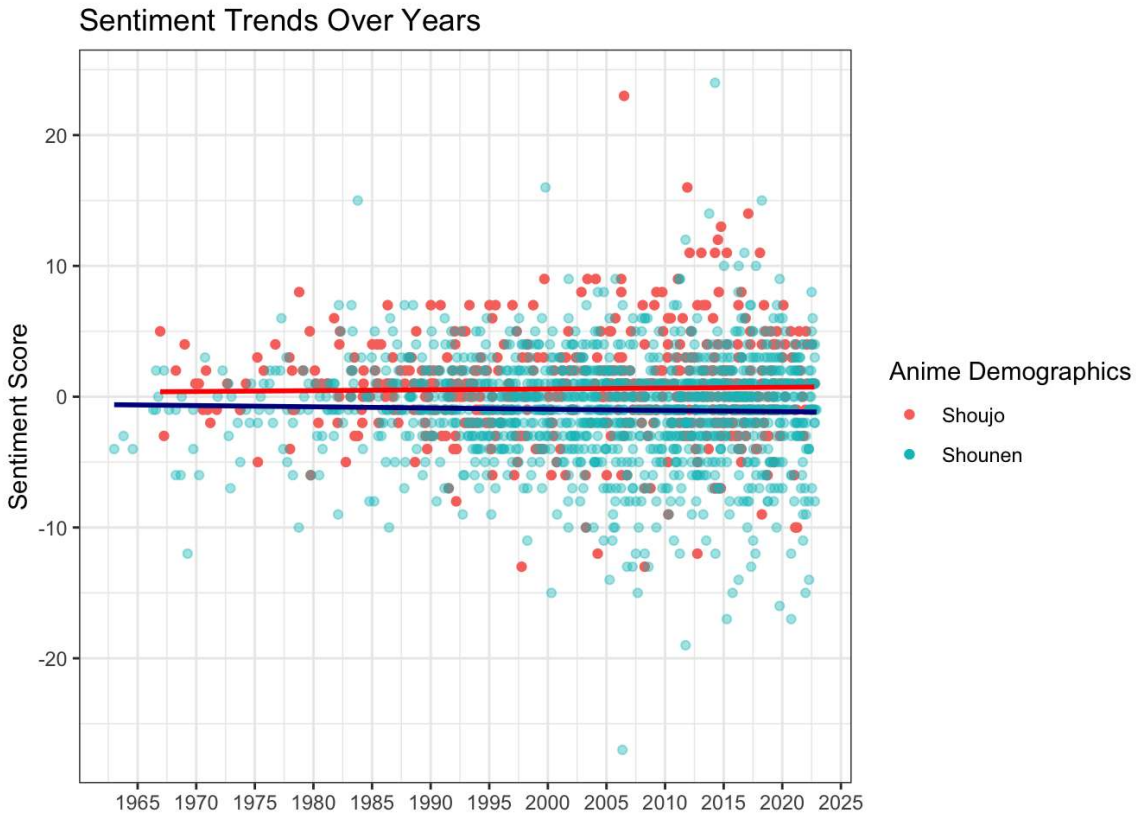
Part I Sentiment Analysis

In Japanese, “shoujo” literally means girls and “shounen” means boys. Shoujo anime are typically geared towards young girls and women, while “shounen” anime are geared towards young boys and men, typically between the ages of 12 and 18. We focused on the **Demographics** variable from the complete anime dataset that contains more than 20,000 entries. This filtered dataset serves as a repository for us to gain meaningful insights into the world of both shoujo and shounen anime, as we intend to explore young women’s and men’s preferences and interests in anime.

After applying the tidytext library to perform an innerjoin() with the Bing lexicon, which categorizes words into two groups: positive +1 and negative -1, we aggregated the scores to produce an overall sentiment score year-on-year for the text from the **Synopsis** variable. Our aim is to analyze the trend in sentiment over time.

It seems that from 1966 to 2022, the sentiment scores for both shoujo and shounen anime appear to remain relatively stable, which suggests that both types of anime tend to maintain a relatively consistent emotional tone over time. However, the graph also shows that shoujo anime has a slightly higher positive sentiment score than shounen anime. This indicates that shoujo anime may have more elements of happiness, joy, and love, while shounen anime may have more elements of sadness, anger, and violence. These differences reflect the different preferences and interests of their target demographics.

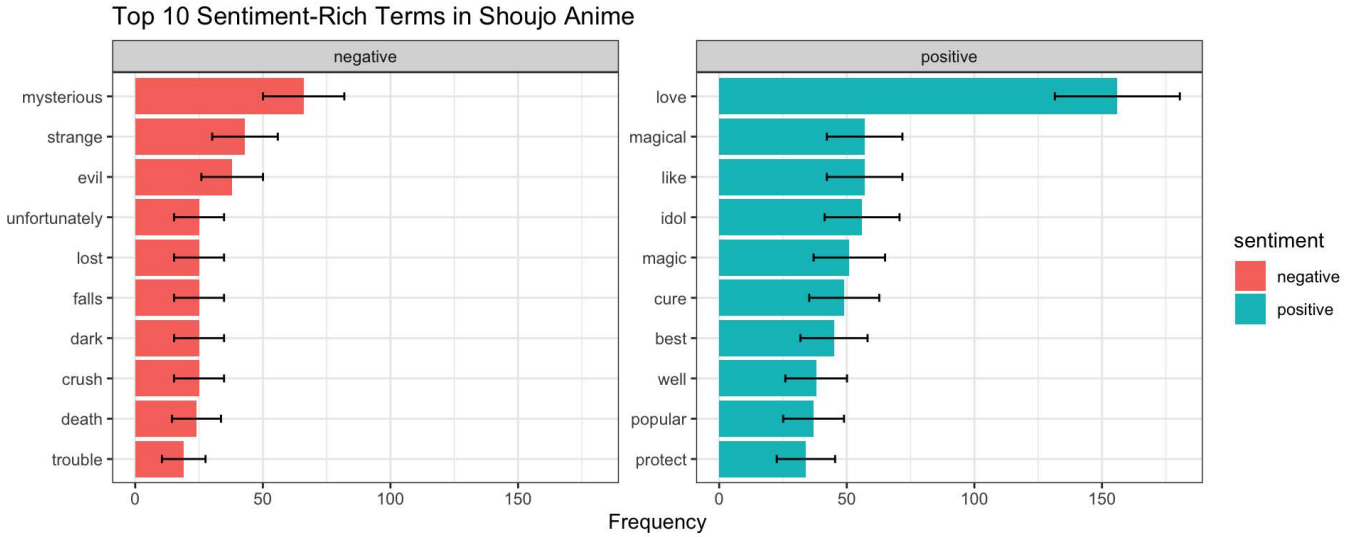
It’s also interesting to explore the extreme spikes in negative sentiment scores. This suggests that something specific may have influenced the emotional tone of shoujo anime in those particular years. Case in point is the anime “Pretty Guardians Sailor Moon Eternal The Movie Part 2”, released in 2021. The portrayal of a post-solar eclipse Earth being shrouded in a dark force suggests a potentially dire and challenging situation for the characters, hence a strong negative sentiment.

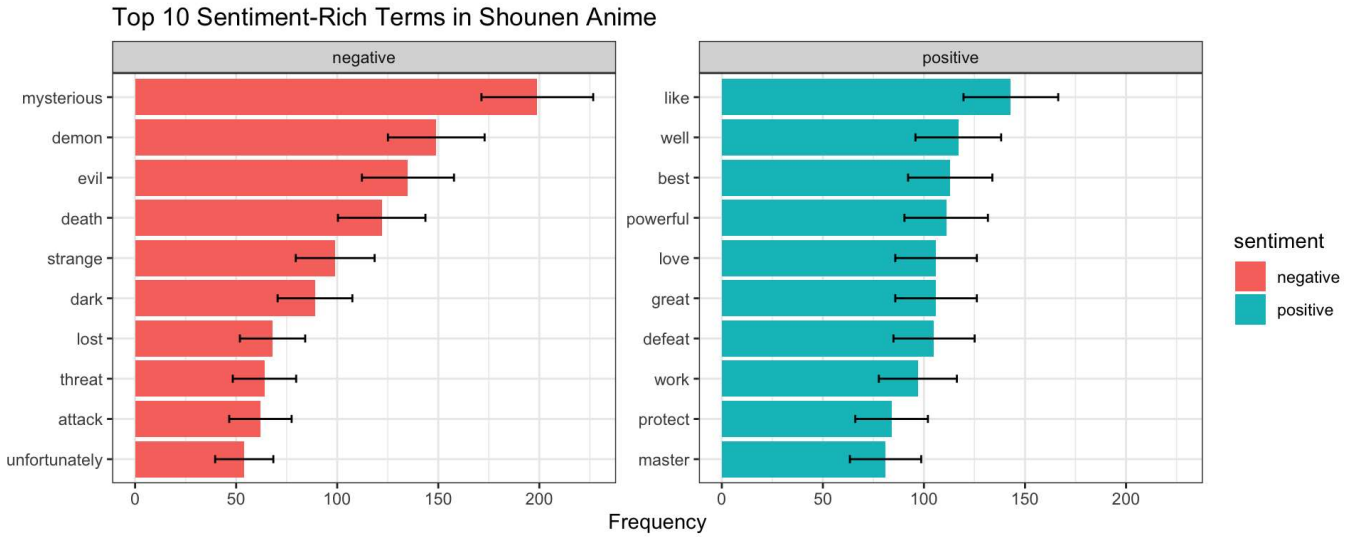


From the bar plots of the top 10 positive and negative terms for both anime, with error bars indicating the 95% confidence intervals, the highest frequency of the term “love” suggests a strong emphasis on romantic themes within the shoujo anime. “Love” significantly outranks the other term, more than twice as frequent as the second term “magical”. This corresponds with our preconceived notion that young women often have a preference for romance in their anime choices. It indicates that romance remains a potent and popular theme for shoujo anime audiences. The rest of the top terms suggest themes of magic, admiration, healing, excellence, and popularity.

In shounen anime, the top positive terms suggest themes of personal growth, power, victory, protection, and mastery. These themes are common in shounen anime, which often feature young male protagonists who strive to become stronger, protect others, and overcome challenges. The relatively even distribution of these terms indicates a balance of these themes in shounen anime.

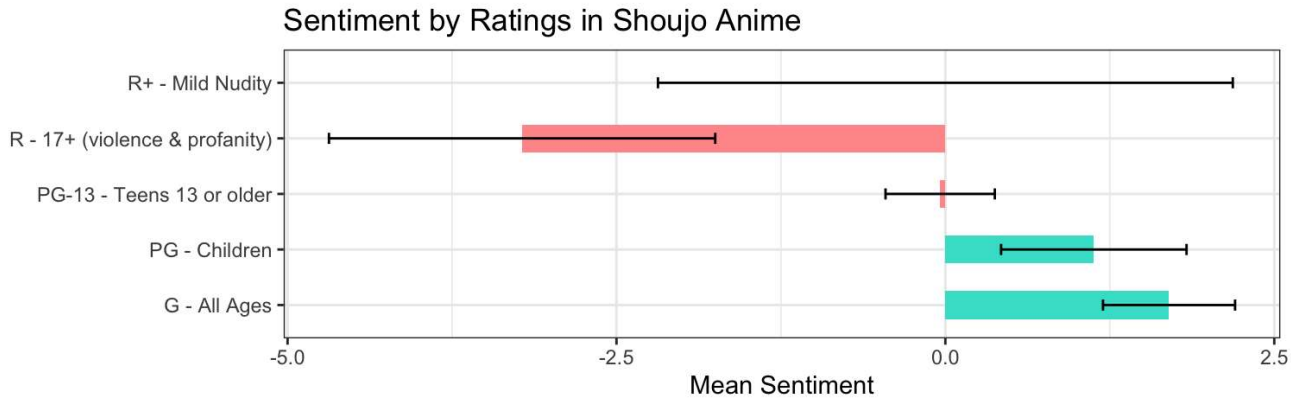
The term “mysterious” appearing as the top negative term in both shoujo and shounen anime might be due to the nature of storytelling in these anime. The term “mysterious” might be categorized as negative because it often implies uncertainty, confusion, or potential danger. But in the context of anime, “mysterious” elements can make the story more engaging and exciting for the viewers. Its frequent appearance could be read as a sign of the richness in plot and character development.

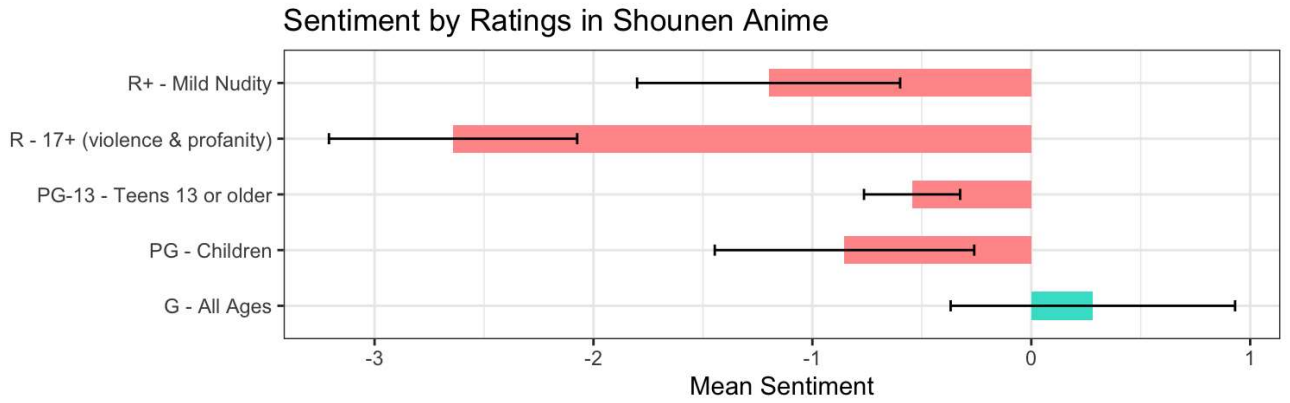




We then calculated the sentiment score based on their ratings. It's no surprise that "R - 17+ (violence & profanity)" **Ratings** has the lowest sentiment score for both shoujo and shounen anime, as this rating normally has more violent, profane, or disturbing content that can affect the viewer's emotions negatively.

We are suprised to find that the "PG-13 - Teens 13" and "PG-13 - Teens 13" **Ratings**, which are supposed to have more enjoyable or relatable content that can elicit more positive emotions, reflect negative sentiment scores within shounen anime.

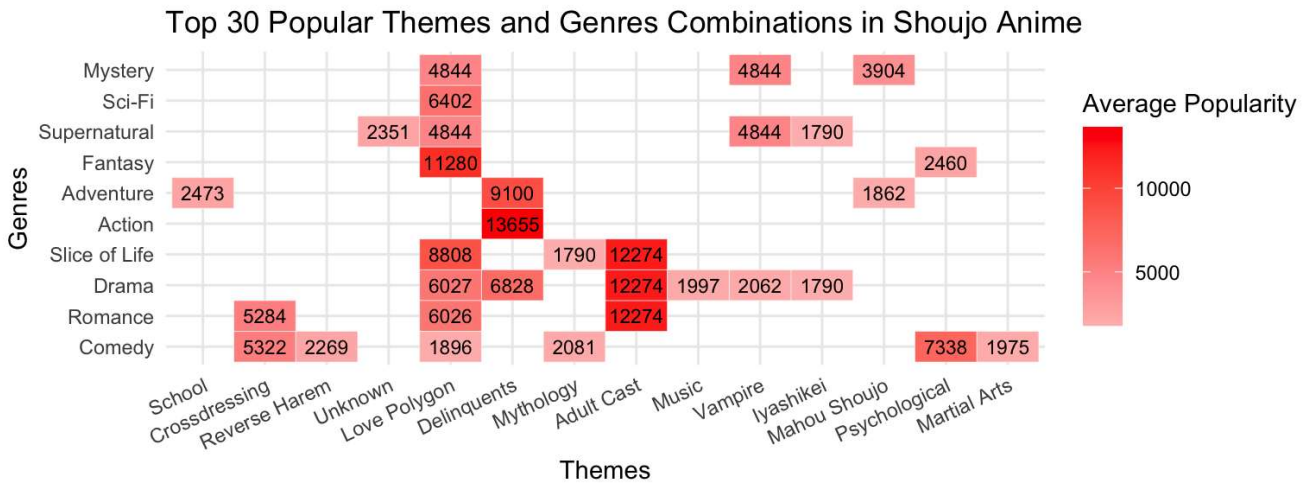


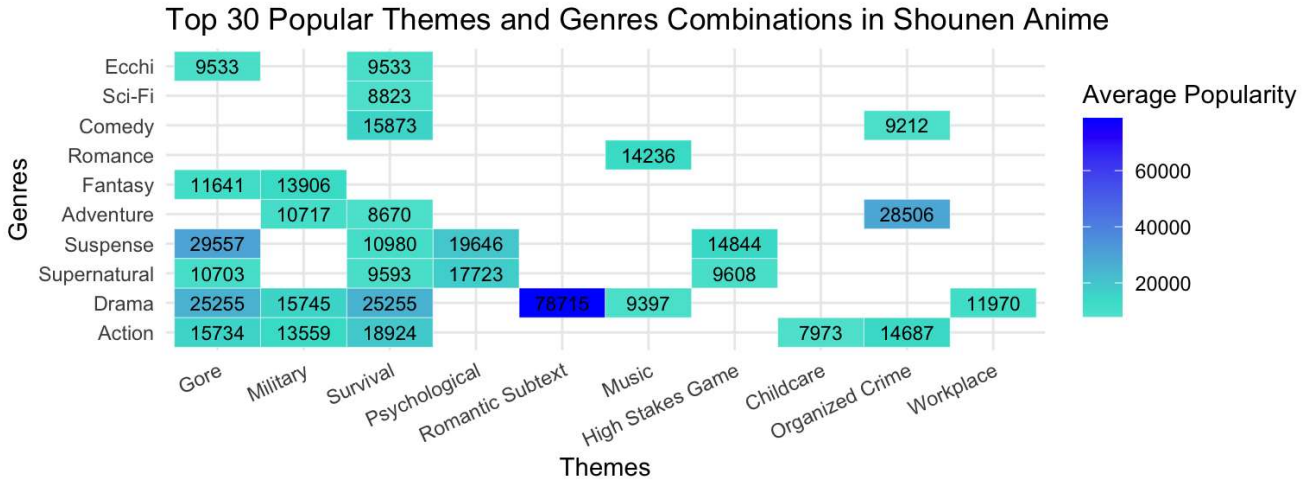


We proceeded to draw two heatmaps based on the average popularity score from the **Favorites** variable for different theme and genre combinations. This involves grouping the data by the **Themes** and **Genres** variables. The **Favorites** variable represents the number of users who have marked the anime as their favorite anime.

The most popular combinations of theme and genre for shoujo anime are “Action” and “Delinquent”. This combination suggests that viewers of shoujo anime appreciate narratives that involve conflict, action, and characters who are rebellious or non-conforming. This could indicate a desire for stories that break away from traditional norms and expectations, reflecting the complexities and challenges of adolescence. Shoujo anime viewers also love any combination of “Adult Cast” with either “Drama”, “Romance”, or “Slice of Life”. It might due to their vested interest in mature and realistic narratives. They might appreciate stories that deal with adult relationships, emotional conflicts, and everyday life experiences.

As expected from young male audiences, popular combinations for shounen anime viewers include “Suspense” and “Gore”, and “Adventure” and “Organized Crime”, all of which suggest a preference for narratives that balance emotional depth with thrilling action, survival challenges, and criminal intrigue. However, the most favored combination is “Drama” and “Romantic Subtext”, with over 70,000 viewers adding it to their favorites. This is much higher than any other combination by a wide gap. While it may come as a surprise, it appears that young male viewers do have a strong preference for narratives that delve into emotional depth and subtle romantic elements. This is despite the fact that shounen anime is typically associated with action and adventure themes. Our guess is that overtly romantic themes can sometimes fall into cliche story lines. Romantic subtext allows for a more subtle and nuanced exploration of romantic elements. This can appeal to a broader audience, including those who may not typically seek out explicit romance stories.





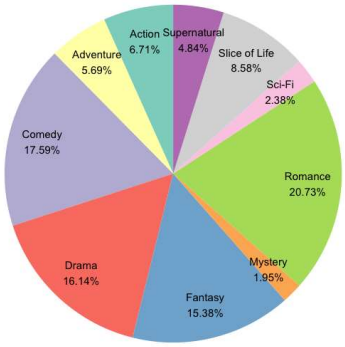
From the bar chart below of the top 10 genres and its corresponding sentiment score, Shounen anime have mostly negative scores. It is expected as shounen anime often involves themes of conflict, violence, tragedy, and hardship. These themes can introduce more negative words and emotions into the synopsis text, such as “death”, “fight”, “kill”, “betrayal”, “suffering”. These words can lower the overall sentiment score of the anime, and consequently, the genre.

As we look at the graphs, we discover that shoujo anime are more fascinating to investigate. The prominence of the romance genre further consolidated my preconception that romantic relationships and themes hold a central position in shoujo anime. The popularity of fantasy genre indicates a desire for escapism and immersion in imaginative worlds.

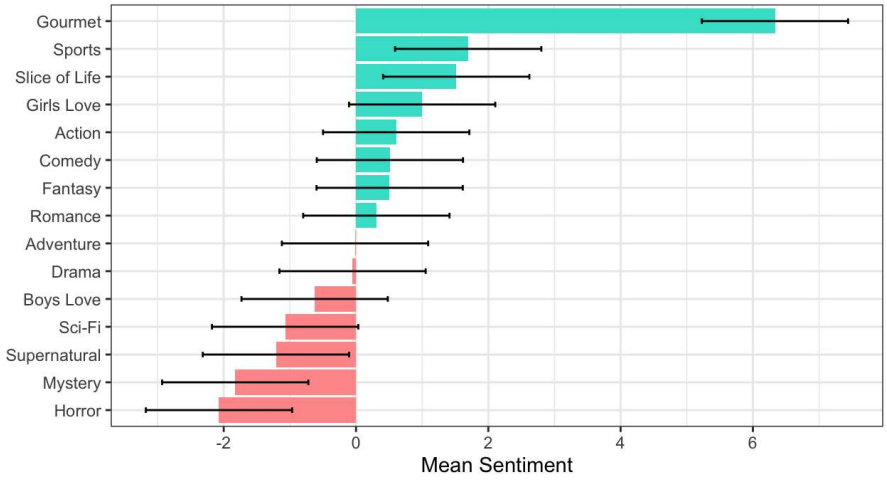
However, when we have the scope expanded to the top 15 genres, we find something interesting in terms of sentiment within each genre. Although “Boys Love” and Girls Love” anime are both typically created with a female audience in mind, focusing on the romantic relationships between same-sex characters, BL have a negative sentiment, whereas GL tends to be positive overall.

Upon research, this might be the answer: BL anime often explores themes of unrequited love or complex relationships. This can introduce more elements of heartbreaks and emotional turmoil. BL anime also often includes elements of sexual tension and can delve into more explicit themes. These complexity could potentially result in a lower sentiment score. This finding is also corroborated by the fact that most of the BL genre shoujo anime belong to the “R - 17+ (violence & profanity)” rating. In contrast, GL anime place a stronger emphasis on emotional and spiritual connections between female characters, which often evoke a sense of warmth, understanding and fulfillment.

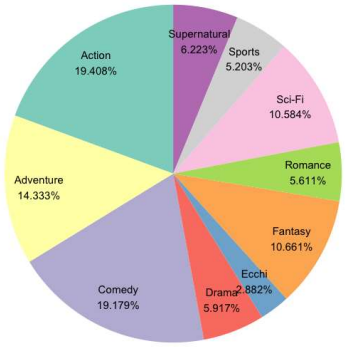
Top 10 Genres in Shoujo Anime



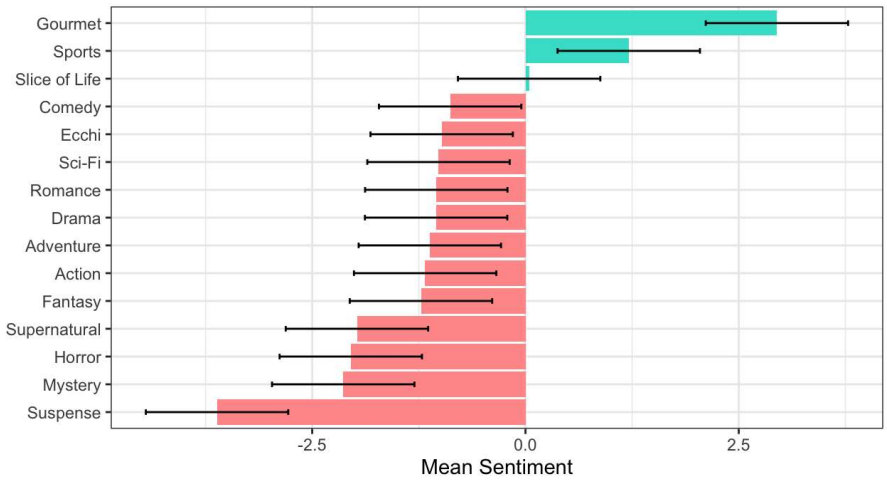
Mean Sentiment by Genre in Shoujo Anime



Top 10 Genres in Shounen Anime



Mean Sentiment by Genre in Shounen Anime



Part II Regression Model

To develop a regression model that predicts the score of an anime targeting adolescent viewers, we initiated the data preparation stage by focusing exclusively on Shounen and Shoujo demographics. To enhance the dataset for modeling efficiency, we eliminated extraneous columns, retaining only 11 essential features. Subsequently, to uphold data integrity, we removed rows containing null values, resulting in a refined dataset comprising 2449 entries out of the original 2771.

Following this, we standardized the numerical data to ensure consistency in scale. With the processed dataset, we are poised for the application of machine learning algorithms. The preprocessed features are optimized for model training, and the target variables have undergone log transformation to improve predictive accuracy.

We selected three models — Linear Regression, Random Forest Regressor, and XGBoost Regressor. We employed a 5-fold cross-validation to ensure the robustness of our models. After assessing and comparing their out-of-sample Root Mean Squared Errors (RMSE), we arrived at the following conclusions:

Show 4 **entries**

Model	Parameters	Best.RMSE
Null Model	Null	0.11276
Linear Regression	Null	0.10069
Random Forest	max_depth: 30, min_samples_split: 5, n_estimators: 500	0.07184
XGBoost	learning_rate: 0.1, max_depth: 3, n_estimators: 200	0.07156

The null model serves as a baseline or “naive” prediction that doesn’t involve any learning from the features. It acts as a point of comparison for the more complex models. The RMSE for the null model is calculated using the mean of the target variable (`y_train`) as predictions for the test set.

The initial linear regression model, with its default settings, produced a RMSE of 0.10069. This performance is expected given the simplicity of linear regression and lack of hyperparameter tuning.

Next, we explored more sophisticated models, such as the Random Forest Regressor, an ensemble machine learning technique that leverages multiple decision trees to enhance predictive accuracy and mitigate overfitting. After an extensive search involving 27 different combinations and 135 model fits, we identified the optimal configuration with a `max_depth` (depth of each decision tree in the forest) of 30, `min_samples_split` (number of samples required to split an internal node) of 5 and `n_estimators` (number of trees in the forest) set to 500. This improved the RMSE to 0.07184, showcasing the benefits of using multiple decision trees to reduce overfitting and enhance prediction accuracy.

Similarly, we tuned the XGBoost model with 27 candidate parameter sets and 135 fits. Unlike Random Forest, which constructs multiple decision trees independently and merges their outputs, XGBoost handles intricate patterns and refines predictive accuracy through iterative improvement of weak learners. The best-performing configuration had a `learning_rate` (step size) of 0.1, `max_depth` of 3, and `n_estimators` of 200, resulting in an RMSE of 0.07156. While this improvement is slight, it underscores XGBoost’s efficiency in handling complex patterns through its gradient boosting framework.

Both Random Forest and XGBoost models outperformed the baseline linear regression model, with XGBoost leading by a narrow margin. This indicates the superior predictive capabilities and robustness of these models against variations in the dataset. The results underscore the importance of hyperparameter tuning in unlocking optimal performance from complex models like Random Forest and XGBoost, even for those not familiar with these specific algorithms.

Nevertheless, even though the linear model may not be optimal, we can still examine the coefficients in the linear regression formula to understand the impact of each feature on the score.

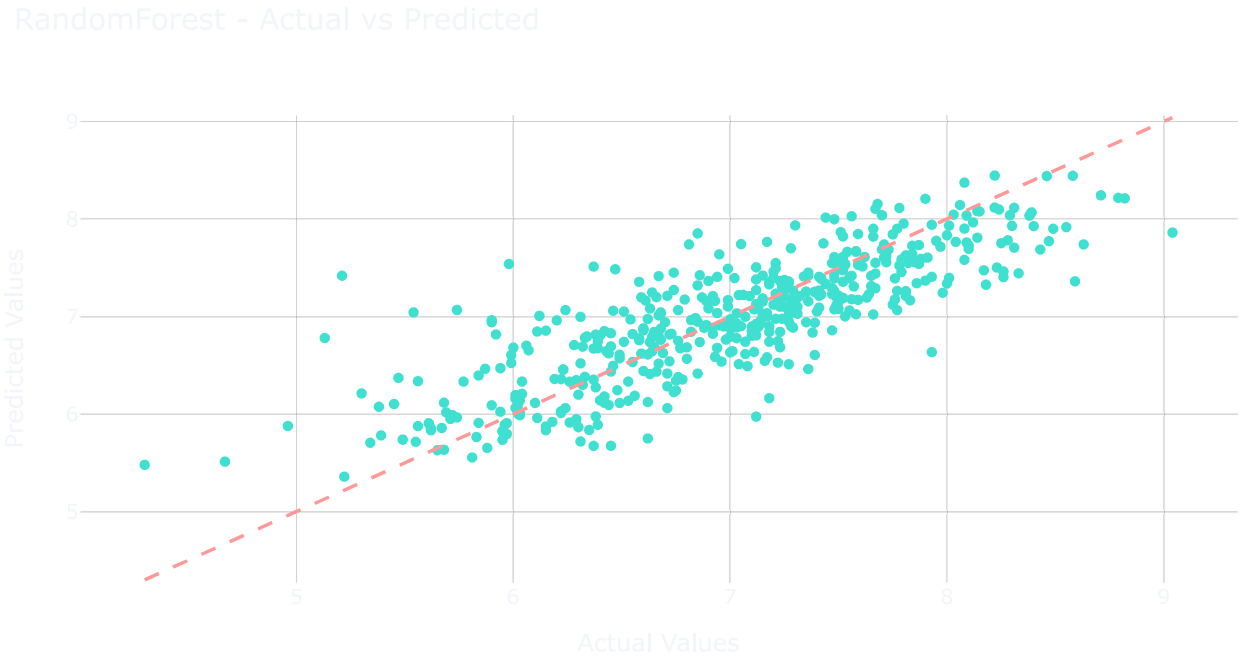
$$\begin{aligned}\hat{\mathbf{Score}} = \mathbf{1.8584} \\ &+ (\mathbf{0.0229} \times \mathbf{Type}) \\ &+ (\mathbf{0.0159} \times \mathbf{Episodes}) \\ &+ (\mathbf{0.0204} \times \mathbf{Source}) \\ &+ (\mathbf{-0.0034} \times \mathbf{Demographics}) \\ &+ (\mathbf{-0.0005} \times \mathbf{Duration\_Minutes}) \\ &+ (\mathbf{-0.1134} \times \mathbf{Rating}) \\ &+ (\mathbf{-0.0505} \times \mathbf{Scored\_Users}) \\ &+ (\mathbf{-0.0950} \times \mathbf{Members}) \\ &+ (\mathbf{-0.2069} \times \mathbf{Favorites})\end{aligned}$$

Analyzing the linear regression formula reveals that **Favorites**, **Rating**, and **Members** exert a notable impact on the **Score**. These variables exhibit substantial coefficients, indicating a significant influence on the predicted **Score** values as their respective values increase or decrease. **Duration_Minutes** has a coefficient of 0.0005, implying baisally no observable direct effect on the **Score**.

We can also observe the performance of the model from the visualization of the model:

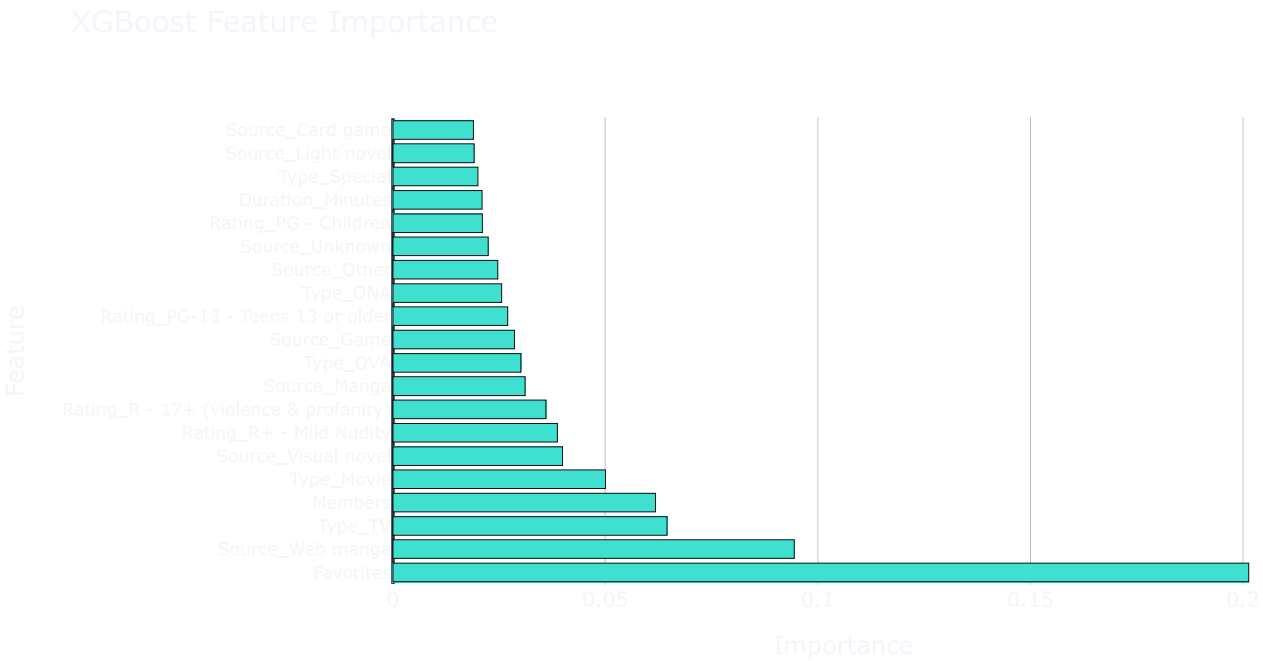
The scatter plot for the Linear Regression model displays the predicted versus actual values, revealing the model’s accuracy in estimating the **Score**. The red dashed line represents perfect predictions, where the predicted values match the actual values exactly. Points scattered around this line indicate the variance from the perfect prediction. While many predictions are reasonably close to the line, indicating a decent fit, there is a visible spread, especially for actual values on both ends, suggesting that the model has limitations in capturing the trend for anime with these scores.

For the Random Forest regressor, the scatter plot shows a tighter clustering of points around the red dashed line compared to the Linear Regression model, which suggests a better fit to the data. The concentration of points near the line indicates that the Random Forest model has a higher predictive accuracy. Nonetheless, some spread persists, particularly at the lower and higher ends of the scale, which may imply that certain extreme values are not as precisely predicted by this model.



The visualization for the XGBoost model depicts a similar pattern to the Random Forest, with predictions closely hugging the line of perfect prediction. The cluster of points around the dashed line suggests that XGBoost provides a robust prediction across a range of scores. The model appears to handle both the lower and higher ends of the score spectrum a bit better.

The XGBoost model outperforms the linear model in terms of RMSE, it might be more beneficial to focus on understanding feature importance rather than trying to interpret the coefficients.



The bar chart highlights which variables are most influential when predicting the score of anime.

The **Favorites** feature has the highest importance, suggesting it is the most significant predictor of an anime's score. This is followed by **Members**, **Source_Manga** and **Scored_Users**, which also have high importance scores.

The next set of features includes **Source_Web_Manga**, **Type_Movie**, **Rating_R+ (mild nudity)**, **Type_TV**, and **Type_OVA**, each carrying moderate importance in the model. These features point to the source material of the anime, its type, and content rating as influential in how it is scored. Other features like **Duration_Minutes**, **Episodes**, and various demographic and source-related variables have lower relative importance.

The chart demonstrates that community engagement metrics (**Favorites**, **Members**, **Scored_Users**) are the most impactful on an anime's score, followed by content-related features (**Source**, **Type**, **Rating**), with some nuances based on the specific demographic and source type.

In our final step, we employed the XGBoost model to predict scores for anime entries with missing scores in our dataset. To assess the accuracy of these predictions, we compared them with scores from IMDb. Moreover, we delved into viewer comments to gauge the plausibility of our predictions. Leveraging the random library in Python, we employed our predictive models to evaluate scores for five randomly chosen anime, repeatedly sampling from a pool of over 200 unscored anime titles. This iterative approach indicate a satisfactory level of score accuracy in our predictions matching that of IMDb. We selected three well-known anime for inclusion in our report, enhancing the reliability of our comparative assessment for our audience.

This result indicates that our model captures the key factors influencing scores rather effectively. Notably, there are numerous positive reviews on these anime titles, which reinforces the credibility of our model. These favorable assessments align with our predicted scores, highlighting the XGBoost model's ability to reflect viewer preferences accurately.

Show 3 **entries**

Title	Demographics	Predicted_Score	IMDb_Rating
Akuyaku Reijou nanode Last Boss wo Kattemimashita	Shoujo	7.18189	6.9
Bishoujo Senshi Sailor Moon Cosmos Movie	Shoujo	6.886652	7.7
Detective Conan: Love Story at Police Headquarters - Wedding Eve	Shounen	6.260386	6.3

Showing 1 to 3 of 3 entries Previous 1 Next

It's important to note that, despite both sites using the same score scale ranging from 0 to 10, this comparison should be viewed as relative rather than absolute. The reason lies in the inherent differences between the rating systems and user communities of our dataset from MyAnimeList.net and IMDb. While the scores may align to some extent, various factors such as cultural nuances, viewer expectations, and rating criteria may contribute to a nuanced interpretation. Therefore, one should approach this comparison with an understanding that the scores are relative and not necessarily directly interchangeable between the two platforms.

A more effective way to evaluate our model's accuracy is to select a specific anime from our dataset with missing scores, calculate its percentile within our dataset, and then perform a similar percentile calculation for the same anime on IMDb. Comparing these two percentiles allows us to assess our model's accuracy more precisely. However, it's worth noting that IMDb lacks specific demographic filters for anime fans, making it challenging to filter out anime targeting only adolescents. Despite this limitation, we find this method more convincing, as it offers a nuanced comparison by normalizing scores within their respective platforms. It takes into account the diverse user base on IMDb and the distinct preferences of anime enthusiasts, especially those in the adolescent demographic. This approach, using percentiles, could provide a more comprehensive and meaningful evaluation of our model's performance.

Conclusion

In our journey to unravel the world of anime, we embarked on an exploratory data analysis of a comprehensive anime dataset. Our journey marches on deconstructing the intricacy of shoujo and shounen anime that target primarily adolescents. Our initial focus was conducting a sentiment analysis of the synopses, which served as a window into the emotional landscape of shoujo and shounen anime. The result revealed a consistent emotional tone over the years, with shoujo anime displaying a slightly higher positive sentiment. This consistency in emotional expression speaks volumes about the enduring appeal of these genres and their resonance with target demographics. However, our curiosity was piqued by occasional spikes in negative sentiment scores in certain years. These anomalies hinted at specific influences or events that might have temporarily shifted the emotional tone during those periods.

Furthermore, the prominence of romance in shoujo anime and themes of personal growth in shounen anime illustrate the varied emotional landscapes these type of anime navigate. Building on these initial insights, we decided to delve deeper into the interplay between themes and genres. Both shoujo and shounen anime viewers appreciate narratives that break away from traditional norms and expectations.

Last but not least, our findings on the sentiment associated with different themes, such as the contrast between “Boys Love” and “Girls Love”, offer nuanced insights into the complex emotional fabric of anime storytelling. These results underscore the rich diversity and intricate emotional dynamics within the anime genre, providing valuable perspectives for creators, viewers, and researchers alike.

In our sentiment analysis project, we used the Bing lexicon and the tidytext package in R to calculate sentiment scores for each shoujo anime based on the synopsis text. These scores, representing the overall emotional tone, were used to compare and group anime. However, this method has limitations like ignoring context, sarcasm, or word intensity, so bear in mind that the scores are not absolute measures of emotion.

For the second part, we focus on predicting anime scores for teenage audiences. Our approach involved a sophisticated interplay of data preparation and the application of advanced machine learning techniques. The key was to develop a regression model adept at forecasting the scores of anime specifically targeting adolescent viewers. This task necessitated a thorough evaluation and refinement of our dataset, leading us to identify and focus on those variables most influential in determining anime scores.

The rigorous analysis led us to deploy several predictive models, with the XGBoost model emerging as a standout for its remarkable accuracy in mirroring actual IMDb scores and viewer feedbacks. This high degree of predictive precision not only validates the efficacy of our model but also underscores the critical importance of community engagement metrics and content characteristics in shaping an anime’s success. Our study reveals that these factors play a central role in captivating the teenage audience, a demographic known for its distinct preferences and viewing habits.

In conclusion, the insights from our predictive modeling provide a valuable tool for understanding and forecasting the preferences of teenage anime viewers. The alignment of our model’s predictions with real-world ratings and feedback highlights its potential as a strategic asset in entertainment analytics. This research not only enriches our comprehension of the factors influencing anime ratings but also sets a promising path for future advancements in the field, opening doors to more targeted and refined analysis tailored to the evolving tastes of the anime audience.