Deciphering the Anime Score Puzzle: A Data-Driven Exploration

DS-GA 1001 Capstone

Yirong Bian yb970@nyu.edu
Cindy Lin cyl4981@nyu.edu
Hanyuan Zhang hz1832@nyu.edu
Wanyi Yang wy815@nyu.edu

1 Introduction

1.1 Background

When mentioning anime, the first that comes to people's mind may be *Spirited Away*, which was directed by renowned Japanese director Hayao Miyazaki and won the Academy Award. The growing popularity of anime content around the world is expected to drive the growth of this market over the forecast period. In this report, we aim to explore users' feedback on anime based on the features of the anime. As we explore the dataset with approaches ranging from hypothesis testing to machine learning, we bear the overarching question of what makes a good anime in mind. We try to analyze factors that could affect the quality of the anime and the users' reactions. Moreover, we intend to build a prediction model that could potentially help businesses to project users' feedback for the anime they produce.

1.2 Data Preprocessing

The raw data set is obtained from Kaggle [1], consisting of 35 columns and 17,562 rows. Each row represents information about an anime, meaning there are a total of 17,562 anime being counted, whereas each column represents a feature of the anime. We only use a part of the features, whose details can be seen in Table 1. We notice that NaN values are marked as "Unknown" in this data set and we apply a row-wise removal of the missing value and drop the unnecessary columns for each of the following tasks. We deal with the categorical variables, including genre, producer, rating, type, source, and studio by using one-hot encoding to convert them into dummy variables. It is worth noting that one anime may have more than one genre and producer. We transform the string type numeric data to numeric and extract the year of the feature Premiered as integers. In addition, we divide the anime into five groups and create a new categorical column: extra short anime (0 - 10), short anime (11 - 26), medium anime (27 - 50), long anime (> 50) based on the episodes feature, which will be useful in one of the questions in the inference chapter. In terms of extreme values, we decide not to remove any outliers because they are valid and consistent with reality.

Name	full name of the anime	Genres	comma separated list of genres for this anime
Type	TV, movie or manga	Score	normalized anime score
Premiered	season premiere	Producers	comma separated list of producers
Studios	comma separated list of licensors	Rating	PG-13, R, etc.
Episodes	number of chapters	Popularity	position based on the number of users who have added the anime to their list
Members	number of community members that are in this anime's "group"	Favorites	number of users who have the anime as "favorites"
Watching	number of users who are watching the anime	Completed	number of users who have completed the anime
On-Hold	number of users who have the anime on Hold	Dropped	number of users who have dropped the anime
Plan to Watch	number of users who plan to watch the anime		

Table 1: Feature Table

2 Inference

2.1 Question

- Do newer anime have higher scores than older anime?
- Is adventure anime more popular than action ones?

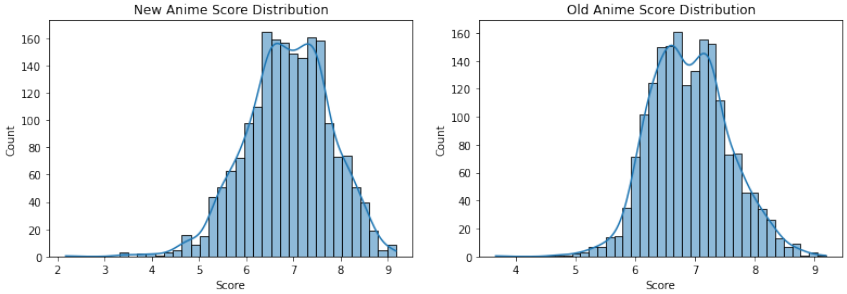
2.2 Approach

- We use the one-sided independent t-test, as it's reasonable to reduce the score to the sample mean as the raw score data is normalized.
- We use the non-parametric one-sided Mann-Whitney U test with the popularity being categorical ordinal data.

In order to cut down false positives, we set the significance level $\alpha = 0.005$.

2.3 Analysis

- To compare the scores of older and newer anime, we split the anime by the median premiered year of 2010. By comparing the old and new anime scores' distribution histograms 1, we find that the distributions are pretty similar. We observed that the score data is already normalized, so we use a one-sided t-test here as our interest lies in detecting a positive shift in scores from older anime to newer anime. The null hypothesis is that the newer anime do not have higher scores than older anime. We got a p-value of 0.79, which is way larger than the alpha level at 0.005, which revealed that we do not have enough evidence to reject the null hypothesis. The sample sizes of the two groups are different, but as we have enough data, we will still believe that the test result is valid. Therefore, we conclude that newer anime do not have higher scores than older anime.



(a) New Animes (b) Old Animes

Figure 1: Score Distribution: Old vs New

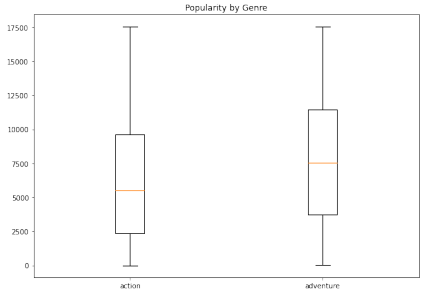


Figure 2: Genre: Action vs Adventure

- For genres, we have 44 genres in total but we only choose two common genres, so that we have enough data in each genre. We chose the Action and Adventure genres to see if they have distributed differently 2. A one-sided Mann-Whitney U test is implemented here because our interest lies in

detecting whether there is a positive shift in the median from action anime to adventure anime. The null hypothesis is that adventure anime are not more popular than action anime. The p-value turns out to be 1.6e-40, which is smaller than the alpha at level 0.005, indicating that we have enough evidence to reject the null hypothesis. The sample sizes of the two groups are different, but as we have enough data, we will still believe that the test result is valid. Based on the analysis, we conclude that adventure anime is more popular than action anime.

3 Prediction

3.1 Question

In this section, we try to use different features to predict the score of different animes. We want to find out the most relevant features that can reveal the quality of animes. Also, we want to investigate how well they can predict the score as we increase the number of features taken as input.

3.2 Approach

In this section, our target output y is the score of each anime.
Step 1: We use all the one-hot encoding features (types, genres, producers, studios, ratings, and episode length), together with the time of premiere and popularity as our input x. With 2462 features in total, we build a multiple linear regression as our benchmark, with RMSE as our evaluation metric.

Step 2: Considering that feature number 2462 is quite huge compared to the number of training data points 9936, there could be an overfitting problem. For instance, there are 1307 producers in our dataset, which are all converted to the one-hot encoding vectors. It is very likely that many producers do not have any contribution to the discriminative power. Hence, we use lasso regression (with RidgeCV to optimize regularization strength) for feature selection. We want to eliminate useless features. In the end, we left 629 selected features and we rank them by their importance in descending order (absolute value of coefficient given by lasso regression).

Step 3: We want to know how the prediction accuracy is going to increase as we keep adding numbers of selected features (in descending order by their importance) while keeping other things fixed. Here we choose ridge regression as our prediction model (with regularization strength fixed to 1).

Step 4: It is reasonable to assume that many of the features are not in a linear relationship with the score. Hence, we build a deep neural network model to investigate if the added non-linearity is further helping us with prediction. Here we take all the 629 selected features as input. We design a relatively lightweight network considering the limited amount of data compared to parameters, with 2 hidden layers of sizes 4 and 8. We also set the dropout rate equal to 0.1 and the weight decay of the Adam optimizer as 1e-5 to prevent over-fitting. After observing the log of training and testing loss, we set the training epoch to 800.

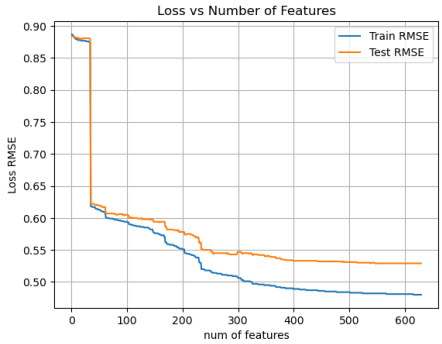


Figure 3: RMSE of Ridge Regression with Different Number of Selected Features

3.3 Analysis

Analysis of step 1: For the simple multiple linear regression model, the prediction result is not satisfactory. We get a training RMSE of 0.448 and a testing RMSE of 4.8e11, which indicates severe overfitting, as a testing RMSE is much larger.

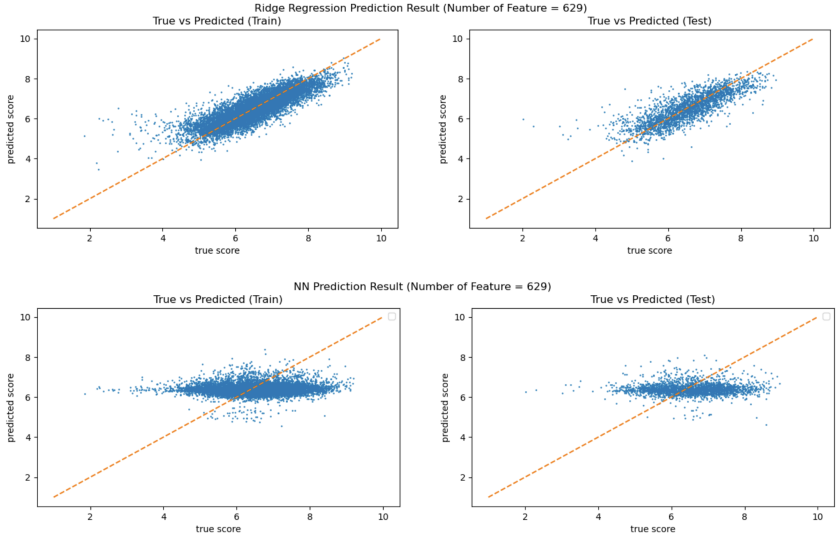


Figure 4: Ridge and NN Prediction Result with All Selected Features

Analysis of step 2: As noted above, we apply lasso regression for feature selection. The optimal lasso regression model (with regularization strength alpha = 1.15e-4) gives us 629 useful features. A feature with high coefficients means it has a huge positive influence on the anime score. Of the top 100 most important features, 97 of them are producers and studios. This indicates who throws the money and who makes the project matter a lot to the quality of animes.

Analysis of step 3: As we can see from Figure 3, the training and testing RMSE gradually decrease as we increase the number of selected features while keeping the other thing fixed. This is quite intuitive as more features give more useful information. It is noteworthy that there is a huge drop in RMSE as we add the 35th most important feature. To further investigate this phenomenon, we observed all the top 50 most important features and find that they are all producers and studios, except the 35th feature is popularity. Since popularity is a different type of feature that provides relatively independent information, we can well explain the occurrence of the huge drop. Also, as we can see from the up two plots from 4, the prediction result of ridge regression with 624 features is quite good, with train RMSE equal to 0.48 and test RMSE equal to 0.529.

Analysis of step 4: As we can see from the below two plots from 4, the prediction result of the NN model with 629 features is surprisingly worse than ridge regression, with train RMSE = 0.927 and test RMSE = 0.938. Most of its prediction scores are centered around the mean score of 6.51. Maybe NN model over-complicates the problem. And a simple penalized linear regression is good enough for this dataset.

4 Classification

4.1 Question

Is it possible to know whether the anime is good or not based on users' actions data such as the number of 'Watching', 'Dropped', etc?

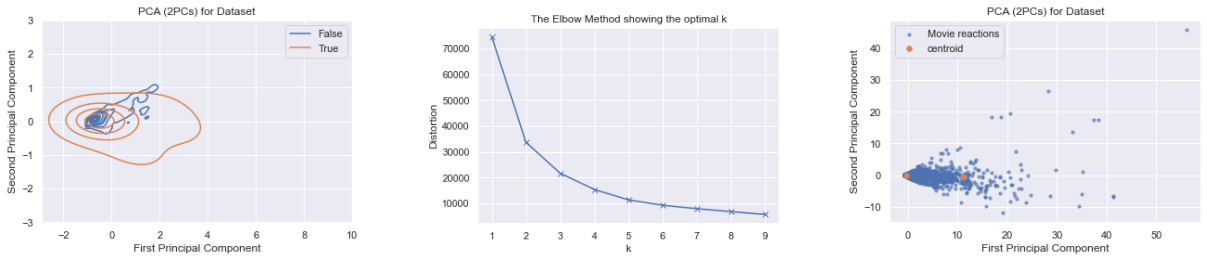
4.2 Approach

We use the number of 'Members', 'Favorites', 'Watching', 'Completed', 'On-Hold', 'Dropped', and 'Plan to Watch' in our cleaned data set as the input X. We use a row-wise deletion to remove the missing values,

which leaves us with 12421 different anime. We apply a median split to the score of each anime in order to transform the contiguous data into categorical $0, 1$ y-labels, where 1 means the anime is high-quality and 0 otherwise. First, we explore our data by doing a PCA dimension reduction after standardization, followed by a K-means unsupervised clustering algorithm. For the classification, we randomly split the data into training and testing data sets with a 0.2 test size and implement simple logistic regression, support vector machine (SVM), decision tree, and random forest model. Additionally, we will use the reduced X input (X after PCA) to avoid overfitting for the logistic regression with cross-validation and SVM model with RBF kernel while using the original X (X before PCA) to train for the decision tree and random forest. To compare the models, we calculate the confusion matrix, AUC, accuracy, and f1-score for each model.

4.3 Analysis

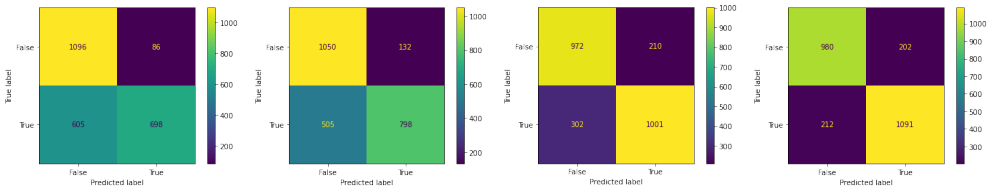
We begin by plotting the density of the data points after doing dimension reduction. The explained variance ratio for each component is $[0.76, 0.10]$. The 2-component reduction is representative of the whole data. From the graph 5a, we can see that the data is not linearly separable and the logistic regression with a linear decision boundary may not perform well. Then, we apply the K-Means clustering algorithm with the elbow method to determine the optimal $k = 2$. Figure 5b is the plot for the elbow method and Figure 5c plots the data and the 2 K-Means centroids. We observe the centroids differ much in the first principal component, which may indicate a tree split is necessary for separating the data set. Then, we train the classification model and Table 2 records the F1-score, AUC, and Accuracy Score of each model. Figure 6 shows the confusion matrix on the testing data set for each of the models. As



(a) PCA (2PC) for Dataset (b) Elbow Method for KMeans (c) KMeans with centroids

Figure 5: Data Exploration by PCA and KMeans

expected, the F1-score does not improve much for the logistic regression model and the SVM model. According to the confusion matrix, both of them do not predict good-quality anime well. For the decision tree model, we can see the F1 score and accuracy score are significantly improved. The random forest decision model has the largest AUC, meaning regardless of the threshold, it is the best model to use. It also yields the best F1-score and accuracy score. By using this model, we can predict 83% of the label accurately. Furthermore, we can see from the visualization of the tree 7, 'Members', 'Favorites', and 'Dropped' are the three most important features in the decision tree, it is again verified by the feature importance generated by the random forest.



(a) Logistic Regression (b) SVM (RBF) (c) Decision Tree (d) Random Forest

Figure 6: Confusion Matrix for Each Model

To answer the question, although it is not possible to fit a linear decision boundary model on the data to differentiate the good or bad quality anime, the random forest is capable of doing the classification with more than 80% accuracy in this situation. We conclude that given users' actions, we are able to predict

Model	F1-Score	AUC	Accuracy Score
Baseline	0.69	0.50	0.52
Logistic Regression	0.67	0.86	0.72
SVM	0.71	0.86	0.74
Decision Tree	0.80	0.87	0.79
Random Forest	0.84	0.92	0.83

Table 2: Classification Model Results

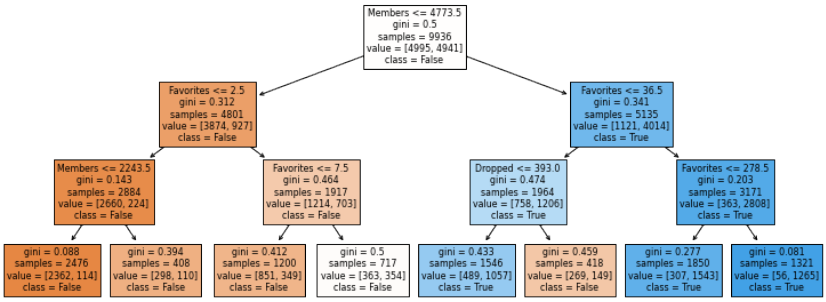


Figure 7: Decision Tree Visualization

the quality of the anime with relatively high accuracy and the most important features are 'Members', 'Favorites', and 'Dropped'.

5 Conclusion and Discussion

5.1 Conclusion

In this capstone project, we explore inference, prediction, and classification approaches to analyze the anime data set. We conclude that there are features that are highly correlated with the score and popularity of an anime. In the inference section, it is interesting to see that the quality of the anime does not improve significantly, as the anime industry and relevant technology develop while genre plays a major role in popularity. We suggest anime producers and investors take the different target audience populations segmented by genre into consideration. In the prediction section, we show that genre and episode length are two major factors that may influence the viewer's enjoyment of anime, while producers and studios are the most important features of the anime quality. By implementing various machine learning models, we find ridge linear regression is the most suitable in this situation. And we are able to predict the quality of anime based on its production background. In the classification section, we find that the random forest has the largest accuracy score in differentiating good and bad quality anime based on the reactions. We also find that the most indicative metric of the anime's quality from the perspective of the users is the number of 'Members' and 'Favorites'.

5.2 Limitations

A number of limitations need to be noted regarding our present analysis:

- Although we have a large data size, we notice that there are some top producers and studios which produce most of the anime in the market. And for some small producers or studios, there is a lot of unknown data. This may cause our model to be biased toward top producers and studios.
- The dataset we are using is from an anime website, so our conclusion may have a bias toward the users. We don't have a comprehensive sample for all audiences.
- We make assumptions that all the data collected are valid and reasonable. However, it is possible that there are errors occurring in the data collection stage.
- Moreover, there are multiple types of anime (e.g. TV, Movie, OVA, ONA, etc), but we don't have time to develop different models for each of them, therefore, our conclusion is a general analysis.

Ideally, if want a dataset without missing values, and to have the anime categorized into only one genre and producer, we may have more significant results regarding how genres and producers can affect the anime scores. Another ideal assumption may be that the outliers are all reasonable instead of occurring due to measuring errors. These two factors could have been changed by expanding the sample size, requiring the users who have completed watching the anime to provide a score, and categorizing each movie into one major genre and one major producer.

5.3 Future Improvements

In our prediction part, we finetune our hyperparameters based on our training set. In our classification part, we don't finetune the hyperparameters. We can improve and further validate our work by finetuning the hyperparameters on the validation set after dividing our dataset into training, validation, and testing parts. Also, in our experiment, we find that our NN model is unstable. A simple modification of our hidden layer size could largely influence inference performance. In the future, we can further investigate the structure of our NN model to make it perform better and more robust.

5.4 Extra Credit

We are curious about whether viewers are more likely to complete the anime with fewer episodes. We use the Kruskal-Wallis test to compare more than 3 groups, as we do not have prior knowledge of the completing rate distribution. To compare the completion of anime with different lengths, we are using the completion rate, which is the number of users who have completed the anime divided by the total number of users in this anime "group" (Members). The null hypothesis is there is no difference in the completion rate of anime with different episode lengths. From Figure 8, we observe that extra short anime have the highest median completion rate while medium anime have the lowest median completion rate. The Kruskal-Wallis test results in a p-value of 2.87e-250, which is smaller than the significance level of 0.005, meaning there's a significant difference between the completion rate of these four groups. Thus, we can reject the null hypothesis and conclude that episode length can result in significantly different completion rates. However, longer episode length does not necessarily mean a lower completion rate because by looking at the boxplot, we can see that anime with medium length have a lower median completion rate than long anime. Then we are interested in whether these two episode classes do have significant differences, so by introducing a one-sided Mann-Whitney U test, we got a p-value of 0.00015, which is also smaller than the significance level of 0.005. Therefore, we can conclude that users are only more likely to complete extra short and short anime but are more likely to complete long anime compared to medium anime.

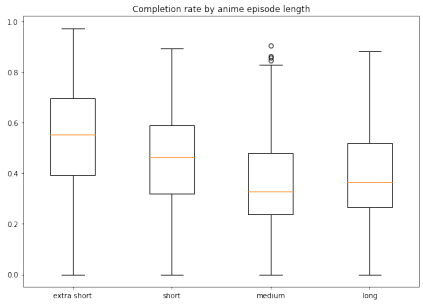


Figure 8: Episode Length

6 Contribution

Wanyi Yang and Cindy Lin are responsible for the inference section. Hanyuan Zhang is responsible for the prediction section and Yirong Bian is responsible for the classification section. Everyone contributes equally to the introduction and conclusion part of the paper.

References

[1] H. Valdivieso, "Anime recommendation database 2020." [Online]. Available: https://www.kaggle.com/datasets/hernan4444/anime-recommendation-database-2020?select=watching_status.csv