Predicting User Scores on My Anime List

Erik Ridd

Data Mining and Machine Learning

New College of Florida

Sarasota, FL 34243

erik.ridd16@ncf.edu

Abstract

Anime has become a large industry both in Japan and worldwide. MyAnimeList is a website that compiles a database of anime information and allows users to rate and share the shows they like to watch. As the industry grows and more shows come out every year, it would be useful to predict what ratings a show may receive on MAL for the fans to choose shows to watch and for producers to gauge reception. To answer this question, a REPTree, a ridge linear regression model, and an MultilayerPerceptron were generated and trained with data on anime shows and their MAL rating to try and predict the ratings. The MultilayerPerceptron proved ill-suited to the data, and the linear regression produced the most accurate predictive model. We concluded that while it was the most effective of the three at a 69% correlation coefficient it still was not accurate enough to be truly useful for professional use.

I. Introduction

Anime (a term used to describe Japanese animation) is an industry that has steadily grown in both use and popularity over the past century, becoming the global phenomenon that exists today. Ever since Osamu Tezuka's "*The Mighty Atom*" (known in America as "*Astro Boy")* appeared on the small screens in 1963, dozens of anime have debuted each year both in Japan and in the worldwide market.[1] These range from short 13-52 episode shows like "*Full Metal Alchemist Brotherhood*" to long running anime like the ever-popular "*Dragon Ball*" and the 1,787 episode juggernaut "*Doreamon*".

With the thousands of anime shows that have been released, there arose a number fan communities and a desire to share opinions.. In 2006, MyAnimeList (MAL) was founded, a website where people can discuss, rate, discover, and learn information about anime and, keep track of the shows they have watched, are watching, or plan to watch. This website has grown to be quite popular, with an estimated 120 million page views a month in 2015.[2] It keeps a running list of the top rated anime by its users (including shows, movies, original video animations, etc.) and stores information on them.

Using this information, I hope to answer this question: Is it possible to predict what rating an anime tv show will receive from the users of MAL based on the features of said show (such as genre, studio that made it, premier season, etc.)? With how large of an industry anime has become, this question is important to both the creators and the fans. For the creators, if there is a way to predict what opinions anime fans may have of a series (positive, negative, middling, etc.) then the creators may be able to better tailor their series to be successful. Essentially, if a creator can predict whether the audience will react positively, they then can make more shows that will find more success. For the fans, if there is a way to predict what anime may receive high ranks or scores, they then may be more accurately able to identify which shows may be of quality and worth the time investment to watch. This will allow them to be able to make more accurate decisions on shows released in the future to choose to watch next, as they can now with the ratings of already released shows on MAL.

II.1 Background

This project will make use of three algorithms with a focus on tackling this problem from a regression perspective. Regression is a technique for prediction analysis, where the values of something- such as a car's price- can be determined based on its features of it- such as a car's horsepower or miles per gallon.

The first algorithm is a regression decision tree. These are structures similar to flowcharts, and are used to map how different 'choices' lead to different outcomes. The algorithm will create an initial node, the data value or question, and split into different branches. These branches lead to more nodes, which split further into more branches and nodes until they reach an answer. This creates a map of paths that will show which choices lead to each result. There are different variations of decision trees, based on how they choose which questions to use for nodes. The regression tree variant determines node selection by calculating which will give the largest gain in information about the variables it is analyzing.

[1] "The History of Anime." Cooper, Lisa Marie, 2017

[2] "Exclusive Interview with the Founder of MyAnimeList, a Colossal Site with 120 Million Monthly PVs." Tokyo Otaku Mode News, 2015

The second algorithm is a linear regression algorithm. Linear regression makes predictions based on linear functions of the data's features, meaning that it treats the data as a sequence of points on a line. The features of the data act as the axis of the graph, with the line from the linear function representing the actual values that result from the features. The predicted value is found based on where the features of the data we are predicting falls on that line. The ridge method of linear regression will be used here. Ridge attempts to find the right features to predict values with the least mean square error (the average of the difference between predicted and actual values squared). It also regularizes the data so that the slopes of the lines used for the parameter is as close to 0 as possible to prevent overfitting the data.

The final test algorithm is a neural network. As the names imply, they are pattered after the way the neurons in a brain work and can help make predictions on both linear and non-linear data (a combination of features that do not fit a linear function). Neural networks work by generalizing linear models over multiple stages of processing. They take the sum of input features and assign weights that are estimated estimate based on the features to find 'hidden data' composed of a combination of weighted sums. This is done by 'lifting', where a significant part of the data is compiled into a non-linear combination of the original features. A group of the weighted, lifted inputs are then used as a new feature space, the 'hidden data'. The neural network then repeats this process on the 'hidden data' to find the outputs. Backpropagation, the process of tuning the weights of the hidden data, is also done to improve accuracy.

The three algorithms will be evaluated using 10-fold cross validation. This is a means to evaluate how well the other algorithms worked. The cross validation takes a section of the training data to be a "fold", and then runs each fold's features through the model and evaluates how accurate the model was at predicting the values of the fold. The average of the results of all 10 fold run-throughs are calculated and used to evaluate the accuracy of the test algorithm

.II.2 Related Works

MAL has been a popular data source in the past. Some data-mining projects have used it, such as tumblr user "The Bunny Advocate"'s "Mapping the Anime Fandom"[3]. This article compared each anime and genre to one another based on shared fans (determined by users including the two shows in their lists). The fans of each show were also analyzed, looking at values like gender, recency of the show, and age. The goal was to determine the factors that caused different groups of anime fans to watch the same shows, by the genre, studio, or something else. The map was represented as "beachballs in a pool", essentially representing each anime as a ball bouncing in a pool with strings connecting them to other anime, with the strength of the string depending on the number of shared fans. This was done by running a simulation until the anime natural gravitated towards and connected to the shows they had the most fans in common with. "The Bunny Advocate" concluded that age of fans were directly related to the age of the show, that action anime and romance anime had the biggest divide in fans while psychological, crime, and thriller genres had the most overlap, and that male viewers tended to watch more recent and more risqué shows while female viewers tended to watch more psychological shows.

Another study was reported in "Anime Reviews and Scores"[4] by Yisong Tao of the NYC Data Science Academy also used MAL. Here, they analyzed how often anime were reviewed by users over time and how the age rating affected the rating of the anime. This tackles a similar question to mine, though more focused in specific features and scope. Tao analyzed each feature of an anime in relation to score individually, looking at the type of production (tv show, OVA, movie) and the age rating. Tao then modelled a "network" of anime based on type and rating, finding that TV shows, OVA's, and movies tended to be closely related. A random forest was then used to make a prediction model, examining the type of the show, directors and cast, number of favorites by users, number of viewers, and the age rating. Tao found that the type of anime and the number of favorites seemed to have the most correlation with score. This suggests that scores are for all anime are closely tied to their viewers and their reactions, but also that the type of production is a large factor.

III. Data

I will be using the publicly available data "My Anime List" for this project, compiled by Matěj Račinský[5] and released on Kaggle. Račinský used a web crawler script to gather the data on all the anime listed on MAL and all the public information of the current users as of May 2018. I will focus on the anime data, of which Račinský gathered the following features for each: titles in Japanese, English localized titles, literal translated titles, synonyms for the title, type of anime (show, movie, OVA), source material for the story, number of episodes, status [of airing], airing (whether or not it was currently airing), dates in which the anime aired on tv, duration of episodes, age rating, user rating on MAL (score), number of people it was scored by, its rank on MAL, popularity (how many people had it on their "currently watching" list, number of members who had it on a list, number of members who had favorited it, background information, premier season, broadcast timeslot, related information, producer, licensor, studio (the production studio that animates and creates the series), the genres of the anime, opening theme song, and ending theme song.

[3] "Mapping the Anime Fandom." The BunnyAdvocate, Faye, The Bunny Advocate

[4] "Anime Reviews and Scores." Tao, Yisong, Data Science Central, 31 Mar. 2017

[5] "MyAnimeList Dataset." Račinský, Matěj, June 2018

It is also important to understand how the rating of an anime on MAL is calculated. The user rating is calculated using the following formula[6], where S = the average score of the Anime, v = the number of votes, and C = the mean score of all anime:

$$User\ Rating = \left(\frac{v}{v+50}\right) * S + \left(\frac{50}{v+50}\right) * C$$

Users votes are only counted if they have listed themselves as having watched 1/5th of a series (to prevent ratings of the entire series on one episode alone). We know that this score is calculated based on the individual ratings of those who watch the series, but begs the question of why people rate the anime as they do. Since we know how the number of members who rate, favorite, and watch the anime influence the rating already through this formula, those features were determined not to be important to this study.

This data was cleaned and limited down to the features we desired to test for predicting the score. I removed the anime that did not fall into the TV category, as this study is interested only in anime TV shows. Additionally, the licensors, producer, studio, and genre categories were split into separate columns for each possible licensor, producer, studio, and genre with a binary 0 for false and 1 for true based on if that series had one of the features. This was done because the original categories listed multiple features under one banner; a show may have one licensor or 5. It was determined using the binary true/false method for whether the entry was the anime's genre, studio, etc. would be the best way to represent this. The duration category and premiered category were also altered. Duration only gives the integer value of the minutes an episode lasts instead of "# min. per ep", and the year was dropped from premiered to be only the premier season. This was because seasons occur repeatedly in the future whereas each year only comes once.

The final, cleaned version of the anime list has the following features: The English title of the anime series, the user rating, the premier season, the number of episodes the show has (with MAL flooring this number to 0 for shows with abnormally large number of episodes, upwards of 500, or are currently airing so the number cannot be predicted), the duration of episodes in minutes, the source material of the story, and then the Boolean values for whether the show has each of the ratings, source materials, licensors, producers, studios, and genres.

The features that included numbers of MAL members were removed for the reasons previously given, as they are used to calculate the score. The background information was removed as there was far more missing information than there was background information present, and the air dates information was removed as it was redundant with the more general Season. The one category that was removed that may be important is the broadcast slot. While this data is likely important to the rating, there was unfortunately too much missing data for it to be useful. Since the worldwide broadcast slots are likely different than in Japan, with much of anime being released through streaming, it's absence will not be too problematic.

IV. Methodology

I chose to approach this problem using the algorithms detailed in the background section. For creating a model of the data that could be used to make a prediction of the scores, I chose to run the data through three algorithms: a decision tree, a linear regression model, and a neural network. I evaluated the models they produced using the 10-folds cross validation algorithm to measure their accuracy. To ensure consistency, I split the data into two sections- the training data consisting of 3621 anime that was used to train the algorithms and create the model, and a set of test data made up of 500 anime that was to be used to test the accuracy of the algorithms. The test and training data were the same on all three algorithms, to ensure that strange issues caused by the use of slightly different collections of test and training data between the three would not arise in comparing them.

The Weka explorer GUI was used to perform these tasks. Weka is a collection of java-based tools and algorithms for regression, clustering, association, pre-processing, classification, and visualization.[7] These can be called from Java, or used in a graphical user interface like the explorer used for this project. I chose to use Weka due to it having built in implementations of all three algorithms that it would automatically perform 10-folds cross validation on after generating.

For the decision tree, I used Weka's REPTree implementation. This version of the decision tree is regression based, and uses information gain divided by variance to decide where to split on, and utilized reduced-error pruning with back-fitting to prune the tree.[8] This pruning breaks the tree down to only the important splits by removing the nodes that don't improve on the previous node in finding values or ones that lower the tree's accuracy. The decision to use a regression tree was made because regression trees can handle continuous data and real values, such as the ratings of the anime that I am analyzing.

The Weka ridge linear regression implementation was also used. This implementation is a standard linear regression model, fitting the features of the data to a line (or in the case of multiple features like mine, a hyperplane).[9] This is performed by estimated coefficients that fit the data well based on their mean square error, and it regularizes them to choose coefficients that are not large. The regularization of the data helps prevent overfitting, which should in theory help make the model better for

[6] "How Are Top Anime/Manga Scores Calculated?" My Anime List.

[7] "An Introduction to Weka." Open Source For You, Shah, Palak.

[8] "Class REPTree API." Generated Documentation

[9] "How To Use Regression Machine Learning Algorithms in Weka." Brownlee, Jason, Machine Learning Mastery

general predictions of anime show ratings rather than specifically this dataset.

For the neural network, the Weka implementation of a multilayer perceptron was used. This is a standard implementation, generating hidden layers from the input features and weighing them to find the output. It can be given instructions on what to look for, and you can tune the amount of time it will spend adjusting the weights it uses for lifting. Larger time spent increases the accuracy, but also increases performance time. The neural network can help discover non-linear relationships between the features that the linear regression and the decision tree may miss.

Evaluating the algorithms was done in two steps, first by a 10-folds cross validation and then by running the test data through the model. The cross validation was completed automatically by Weka, creating 10 folds of the training data and returning the accuracy of the models in predicting their scores. The test data was also run through the model, giving an ideal of the model's performance with data it wasn't trained on. Both reported back the correlation coefficient, mean absolute error, root mean squared error, relative absolute, and the root relative squared. I will be evaluating the effectiveness of the algorithms based on the correlation coefficient (the percentage of variance explained by the model), the mean absolute error (average distance between predicted value and actual), and the root mean square error (square difference between predicted and actual value). Higher correlation coefficients are ideal, while lower absolute and square errors are better to indicate less distance between values.

V: Results

The baselines I would like to see are around an 80% or higher correlation coefficient and a distance of 0.3000 to 0.4000 or lower. This would mean that 80% of all variance between predicted and actual values is explained, with a short distance between the two points

The final REPTree built a tree of 100 nodes, with 96 final ones. In the cross-validation, the correlation coefficient was 0.6379 (63.79%), the mean absolute error was 0.5612 (56.14%), and the root mean squared error was 0.7485 (74.85%). In the evaluation using the test data, the correlation coefficient was 0.6007 (60.07%), the mean absolute error was 0.5862 (58.62%). and the root mean squared error was 0.8107 (81.07%). It seems to have performed decently, with the cross validation explaining about 63% of the variance in the model (60% for the test data). The root mean squared error was quite large though, with 81% square difference between the predicted and actual values in the test data. The mean absolute error was better at 58%. Overall, the cross validation evaluation did performed better than the testing data, indicating that the decision tree worked well with the training data but was less adept at predicting values for new data that it was not trained on.

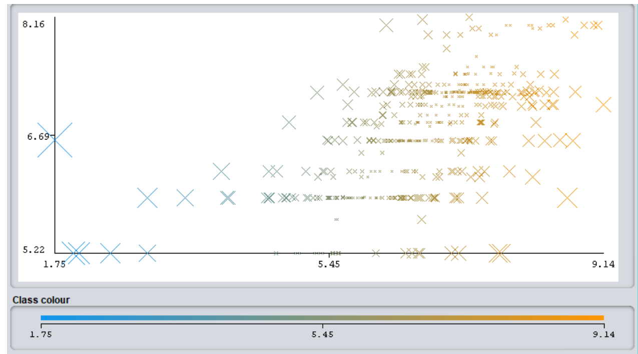

Figure 1: Classifier errors of the REPTree. X is the score and Y is the predicted score

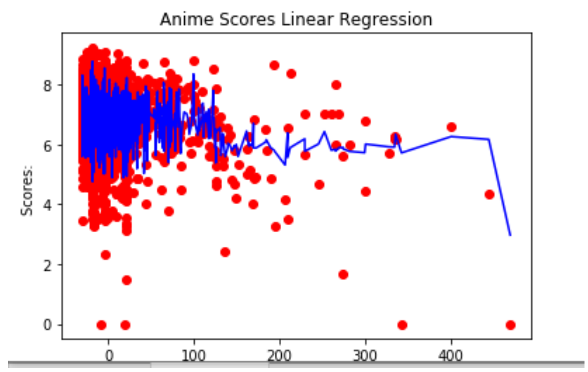

Figure 2: Model of the linear regression using PCA for the x axis. The red points are the actual values, the blue the predicted values

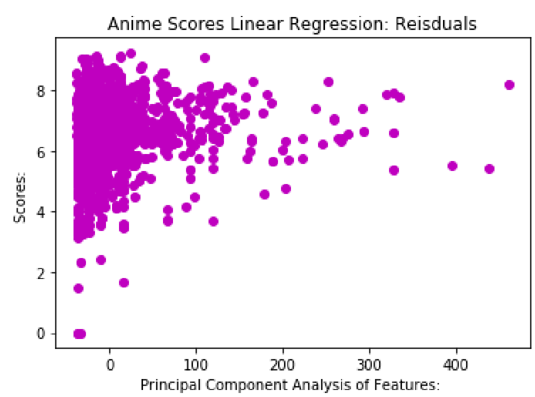

Figure 3: Model of the residuals (difference between actual and predicted values) in the linear regression, represented as purple points.

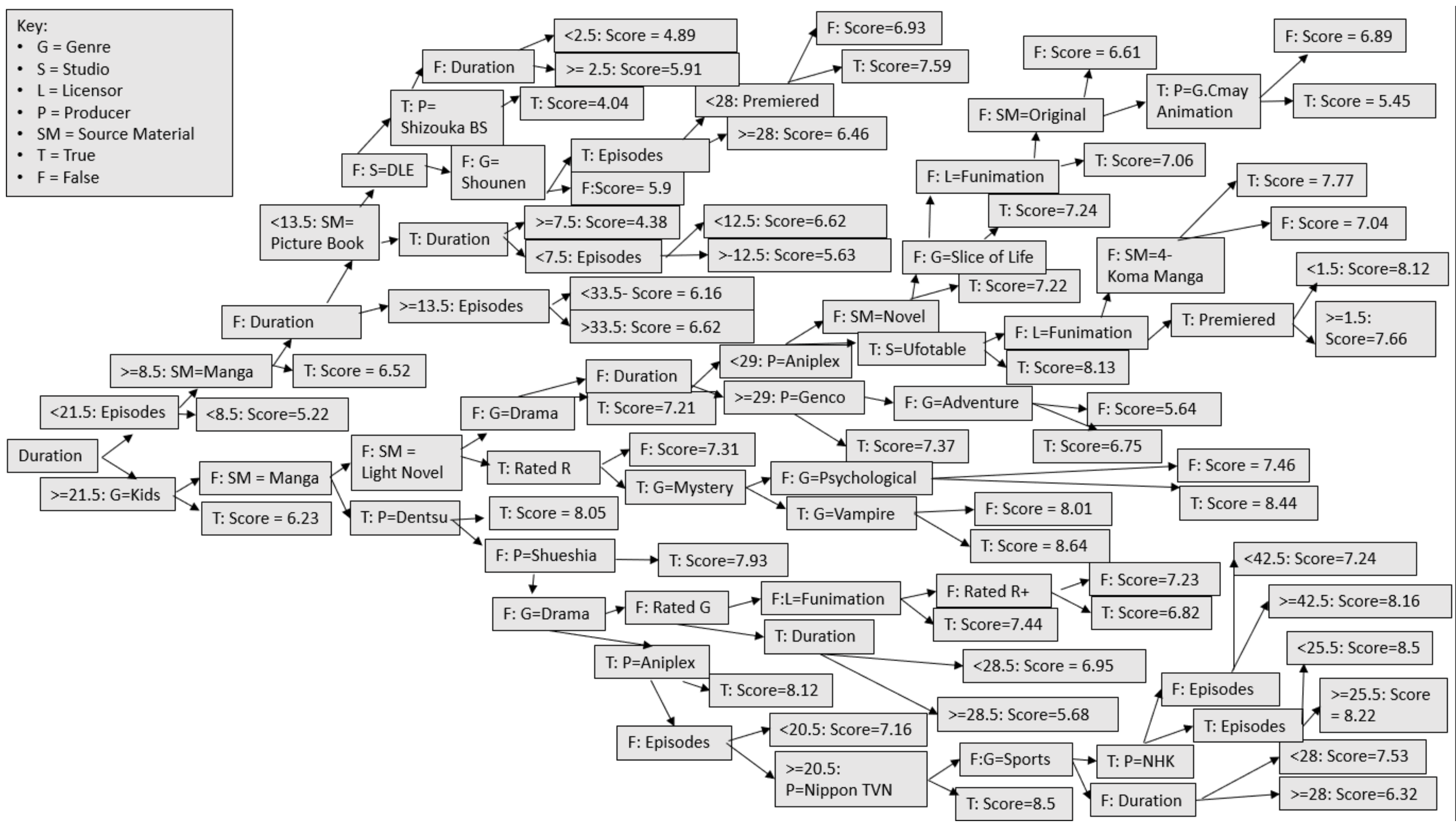

Figure 4: REPTree Model of the MyAnimeList shows

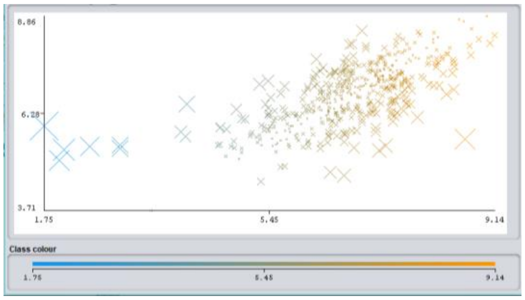

Figure 5: Classifier errors of the linear regression model

In the linear regression, the algorithm singled out the episodes, duration, and all the possible source materials as important features. Of the ratings, only PG-13, R, and Unrated were represented as particularly important in the output. From the bulk of the features the algorithm selected 149 producers out of 1424 (10.4%), 18 licensors out of 97 (18.5%), 96 studios out of 613 (15.6%), and 23 genres out of 47 (48.9%).

The results of the cross-validation revealed a correlation coefficient of 0.6709 (67.09%), a mean absolute error of 0.5353 (53.53%), and a root mean squared error of 0.7341 (73.41%). In the test data evaluation the correlation coefficient was 0.693 (69.3%), the mean absolute error was 0.5428 (54.28%), and the root mean squared error was 0.7357 (73.57%). Overall, the linear regression performed well in both evaluations, with a correlation coefficient explaining slightly less than 70%. It is also of note that the algorithm performed slightly better with the test data it was not trained on with the cross validation, though the errors were slightly worse.

The neural network, on the other hand, proved ineffectual and ill-suited to the data. Due to the width of the data, caused by the boolean features, generation of the multilayer perceptron in Weka proved to be slow, running backpropagation for over 24 hours and eventually crashing the program due to filling all available memory, even on lower test and tuning settings. Neural networks can take large amounts of time and memory to handle wide data, and it is likely that even with more available memory the neural network would still prove to be impractical for this experiment. This and algorithm were not compatible as a result.

VI: Conclusion

Of the three algorithms, the neural network is the only one that did not yield quantitative results. This was due to the width of the data, which the neural network was not designed to handle. Therefore it can easily be ruled out as a potential model for predicting the scores of an anime on MAL. Even assuming one was able to overcome the memory issues, since the scores of on MAL are updated twice daily the slow speed of the neural network would make it's output outdated as soon as it was completed.

This leaves linear regression and REPTree as the models to choose from. Of the two, the linear regression is easily the better model. It had a much higher correlation coefficient than the REPTree in both the cross-validation and the test data analysis, explaining about 5%-9% more of the variance at 67%/ 69.9% versus 63.8%/60% respectively. The linear regression also had slightly lower mean absolute errors (at 53%/54%) and root square error (at than the REPTree's (at 56%/58% for mean absolute and 74%/81% for root square). Because the linear regression had better scores in all three categories, as well as the better correlation coefficient for the test data compared to the decrease seen with the REPTree, the linear regression is the ideal model of the three for predicting the anime scores.

However, I believe the experiment only partially answered the question of if it is possible to predict the average score an anime may receive on MyAnim List based on its features. While the linear regression did perform the best of the three, the results still were not accurate. At best it explains 69.3% of the variance between the actual and predicted points, while still having a fairly large distance between them with over 0.5000. It can be effectively for one application, where users use it to predict the scores of anime and gauge whether or not to watch them. There will be times the model gets it wrong, but it will give a good baseline. For the other application, where studios and producers can predict the rating of a show they are making, I think a more accurate model would be necessary. The linear regression model gives a good starting point for predicting scores but does not have a high enough accuracy for making sound business decisions. A model that is more accurate would be needed for any industry use.

VII: Future Work

In the future, it may be a worthwhile approach to redo this experiment with slightly different features, perhaps focusing entirely on genre or episode count, while testing different algorithms to find one that performs more accurately than the linear regression. I also would be interested in testing the effectiveness of the linear regression model and the regression decision trees in predicting the scores of anime movies, since those were left out of this dataset. Additionally, there is merit in trying this experiment again on western animated shows, and to see if there are any similar trends in western-produced and Japan-produced animated shows.

VIII: Acknowledgments

I would like to thank professor John Doucette for his advice on designing the project and providing help when needed, and to thank Hunt Sparra for his aid in proof-reading and critiquing drafts of the report. I would also like to thank the work of Matěj Račinský in collecting and releasing the MyAnimeList data and the staff of MyAnimeList for creating their database and creating a community for anime fans.

IX: References

"Class REPTree API." Generated Documentation, 4 Sept. 2018, weka.sourceforge.net/doc.dev/weka/classifiers/trees/REPTree.html.

Brownlee, Jason. “How To Use Regression Machine Learning Algorithms in Weka.” Machine Learning Mastery, 22 June 2016, machinelearningmastery.com/use-regression-machine-learning-algorithms-weka/.

Cooper, Lisa Marie. “The History of Anime.” Right Stuf Anime, 2017, www.rightstufanime.com/anime-resources-global-history-of-anime.

“Exclusive Interview with the Founder of MyAnimeList, a Colossal Site with 120 Million Monthly PVs.” Tokyo Otaku Mode News, 6 July 2015, otakumode.com/news/5590fbd763cd06585662ce9d/Exclusive-Interview-with-the-Founder-of-MyAnimeList-a-Colossal-Site-with-120-Million-Monthly-PVs.

Müller, Andreas C., and Sarah Guido. Introduction to Machine Learning with Python: a Guide for Data Scientists. O'Reilly, 2017.

My Anime List. “How Are Top Anime/Manga Scores Calculated?” MyAnimeList.net, myanimelist.net/info.php?go=topanime.

Račinský, Matěj. “MyAnimeList Dataset.” Kaggle: Your Home for Data Science, June 2018, www.kaggle.com/azathoth42/myanimelist/home.

Shah, Palak. “An Introduction to Weka.” Open Source For You, 10 Jan. 2017, opensourceforu.com/2017/01/an-introduction-to-weka/.

Tao, Yisong. “Anime Reviews and Scores.” Data Science Central, 31 Mar. 2017, www.datasciencecentral.com/profiles/blogs/anime-reviews-and-scores.

The BunnyAdvocate, Faye. “Mapping the Anime Fandom.” The Bunny Advocate, 22 Feb. 2018, bunnyadvocate.tumblr.com/post/171165531592/mapping-the-anime-fandom?is_related_post=1.