A Pirate's Discretion: Modeling Differences Between Legal and Illicit Anime-Streaming Communities

By

Bryan Wang

Advised by: Iqbal Zaidi

Submitted to Princeton University

Department of Economics

In Partial Fulfillment of the Requirements for the A.B. Degree

April 12, 2024

Acknowledgements

I would certainly first like to thank my advisor Professor Iqbal Zaidi for guiding me along this rocky, uncertain, but altogether exciting adventure that is the senior thesis. Thank you for your wisdom and advice ever since my junior paper. It has been an honor and pleasure to work with you, and I hope your future students enjoy their time with you as much as I did.

Thank you to my family: Mom, Dad, and Andrew. Thank you all for your support throughout these four difficult but fruitful years. You've all celebrated with me, mourned with me, and challenged me. You've seen me succeed and fail and nonetheless welcome me when I fly back into SFO. I love you guys.

Thank you to my very good friends: To Brian and James for serving with me, praying for me, and encouraging me to pursue the Lord. To Andrew for all the funny memes we've shared, the good times we've had living together for (approximately) two years, and the hard times we've endured. To Katie for the many many things you do for me and our friends and for still being my friend despite how often we seem to butt heads. To Ellie for helping me figure out which models I should use and making me laugh all the time with jokes I find ridiculous but also tell. To Sunny, Daniel, and Nathan for showing me what it means to know and love God. Thank you also to my Manna thesis fairies, Daniel and David: I'd be a stick without you guys feeding me these past few weeks.

Most of all, praise be to God for all that He has done: For every good thing that I have, for I received them all from You; for the heartache, ailments, and discouragements that You ordain, through which You teach me, grow me, discipline me, and love me; for the cross upon which my Savior hung, silent like a lamb that is led to the slaughter. Thank you for staying on that cross, Jesus, and conquering the grave. What a friend you've been these four years.

Table of Contents

Abstract

Market segmentation allows businesses to divide their target market into distinct sections and fine-tune their commercial approach to each one. In the anime-streaming market, however, these distinctions are often blurred, particularly over the distinction of the legality of streaming means—either of legal or illicit streaming sites—attenuating market insights that stakeholders might expect to successfully guide internal and customer-facing decisions. This paper proposes and tests the hypothesis that differences between the two kinds of streaming sites and their user bases are statistically and practically significant, refactoring the way entertainment leaders select and support shows for broadcast and streaming in the anime market.

To do this, the paper takes three data sets representing the legal anime-streaming community (Crunchyroll), the illegal anime-streaming community (AniWave), and the whole, non-sectioned anime community (MyAnimeList), and studies the intersection of all three sources across key feature variables. Using LASSO regression and random forest trees, two powerful machine learning models, it then predicts the commercial success of shows on both the legal and illegal streaming sites and highlights the different variables significant to each platform. The best-performing model on the Crunchyroll data, LASSO, achieved an r-squared of 0.494. Random forest trees, the best-performing model on the AniWave data, achieved an r-squared of 0.596. The most significant variables for the Crunchyroll data, moreover, regardless of model, are controversy (the extent to which AniWave and Crunchyroll ratings disagree about any show) and the number of episodes in a show; for AniWave, the show's success according to MyAnimeList, and again, the number of episodes.

Introduction

Japanese animation, colloquially referred to as “anime,” has in recent decades spread far beyond the borders of the single island nation to the shores of countries both near and to the opposite side of the globe. Out of all television sub-genres, anime ranked third highest in global demand share, claiming 4.74% of international watch time (Parrot Analytics, 2021)—an impressive feat of outsized popularity, given Japan’s 1.53% share of the world population (Worldometer, 2024).

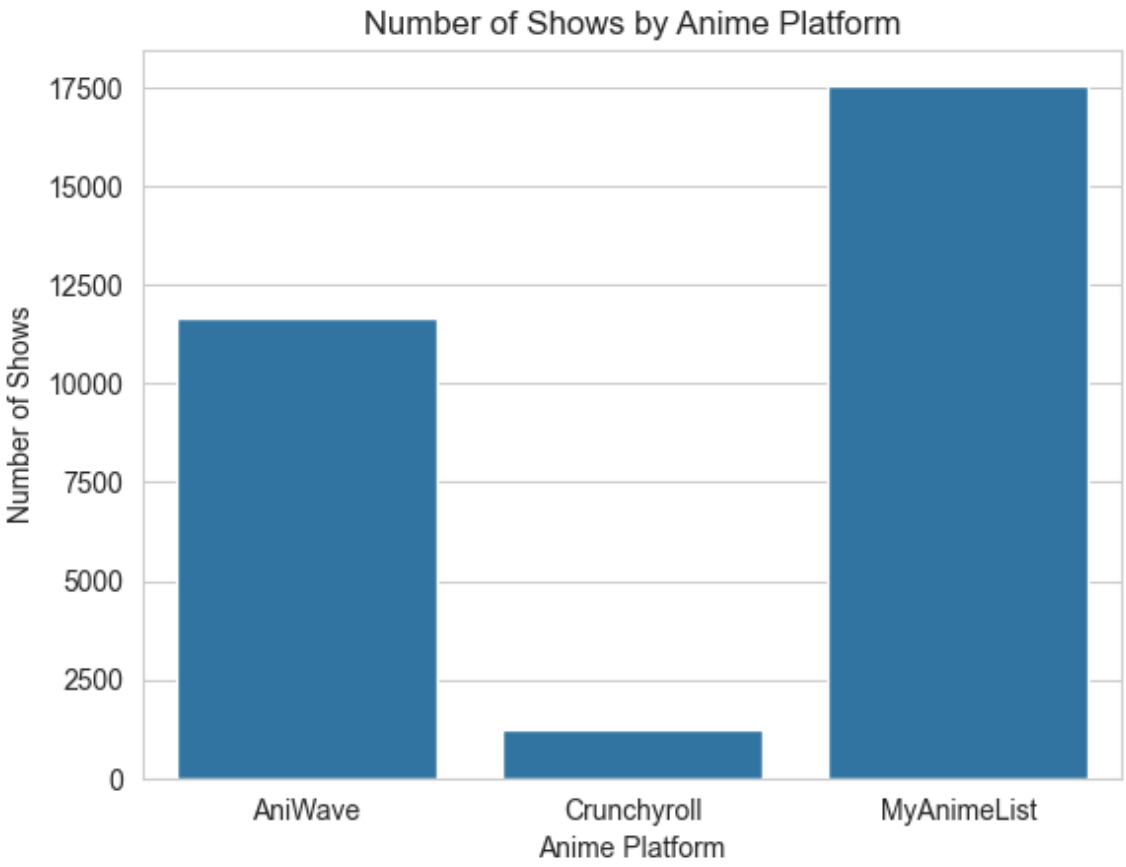
Indeed, nowhere else is anime’s international influence more palpable than in the United States, with franchises like Pokémon, Dragon Ball, and Naruto hemmed and stitched to both casual streetwear and luxury brands (Sinha, 2023), adored by Hollywood stars like Michael B. Jordan and Keanu Reeves (Barnes 2023)—even making its way to Capitol Hill when Hillary Clinton, in her 2016 presidential election campaign, famously declared before a Virginia rally how she hoped to galvanize young voters to “Pokémon Go to the polls” (White, 2016).

The numbers quantify anime’s penetrating influence in everyday American life: In 2022, the anime market reached an evaluation of 25.8 billion USD and is projected to reach 62.7 billion USD by 2032, growing with a CAGR of 9.4% (DataHorizzon Research, 2023). Among all US Netflix users, moreover, 74% have watched at least one anime show and 27% watch anime content every day (Lindner, 2023).

However, while streaming platforms like Netflix offer one point of entry into the world of anime, they were not the first nor the only medium. Due to licensing difficulties and lagging investments by American television networks, particularly in the late 1900s when anime was first introduced to the United States, the main way that American anime fans watched their favorite shows was through “fansubs”—English-subtitled shows independently produced by fans fluent

in Japanese (Stone, 2018). These fansubs were then distributed pro bono across the internet, ignoring copyright laws and bypassing the capital-heavy, arduous process of licensing negotiations (Sevakis, 2012).

Even as legal means of anime access have exponentially grown through the 2000s, much of the original grassroots culture remains on these pirating sites, preserving a sense of nicheness and community among older fans. Moreover, while streaming platforms have made tremendous progress securing licensing rights to old and simulcast shows alike, many titles remain inaccessible by sites like Netflix and Hulu—even by anime-specialty sites such as Crunchyroll, America's preeminent anime provider. The following visualization highlights disparities between the number of shows available on Crunchyroll, AniWave, an anime-pirating website, and MyAnimeList, an anime database documenting every type of anime media commercially available.



Although functioning as an anime database rather than a streaming service makes this feat much easier, MyAnimeList boasts the record of 17562 shows, followed by AniWave's 11675 shows—66.48% of the MyAnimeList collection. The drop from AniWave's second to Crunchyroll's third is even steeper, however, with only 1255 shows on the legal streaming platform. During anime's nascent experience in the United States, perhaps pirating has become a necessary yet liminal evil—first, as the only way fans can watch certain anime shows, and second, to demonstrate enough demand for anime to attract the attention and capital of American entertainment leaders. Similarly, this was how foreign media like Charles Dickens' novels rose to stardom in the United States—through publishing companies that, unfettered by toothless copyright laws, ripped off from the English author and disseminated his work at a discounted price (The Dickens Society, 2023).

Looking forward, however, as the US streaming infrastructure matures and anime is increasingly merged into the mainstream, the salutary effects of pirating shrink, all the while its costs persist and gradually eclipse the dimming benefits: In 2021, The Content Overseas Distribution Association estimated anime-piracy losses to be 15 billion USD, over five times the estimated losses from only two years earlier in 2019. While the pandemic-era lockdowns certainly contributed to this uptick and rates might have fallen since the national reopening, nonetheless, the expanded pirating infrastructure (in websites, supply chains, etc.) continues through the present, exacerbating the already enormous losses far beyond the size they would have been bar lockdowns (The Content Overseas Distribution Association, 2023, as cited in Peters, 2023).

Thus, it appears facile to study the anime market as though all participants were an average of what are likely two distinct types of anime fans: First, the type that consumes their anime content through illegal streaming services, and second, the type that tolerates advertisements or pays to legally watch their shows. This study then sets out to accomplish two things: To accentuate key differences between legal and illegal streaming users, and with that knowledge, to highlight advantages illegal sites might have over their legal competitors and how legal sites might adapt to win over more of the anime market. Moreover, broadening our understanding of both communities can help legal sites predict show success—both for their current user base and unreached segments of the anime market claimed by pirating services.

Literature Review

With Anime's rapid expansion into new international markets, entertainment industry leaders across all time zones have become increasingly attentive to emerging trends in these nascent but promising media. However, due to anime's broad international viewership and Japanese origins, not much research has been done on specifically the US anime market nor piracy behavior unique to English-speaking fans.

AlSulaim and Qamar (2021) also use machine learning methods to predict anime series' levels of success, building a convolutional neural network to classify show reviews as either positive or negative. Like this study, they also use data from MyAnimeList but approach the data with an emphasis on sentiment analysis for binary classification rather than taking on the more precise regression task set out in this study. In addition, AlSulaim and Qamar stop short of studying how having positive reviews affects a show's probability of commercial success, limiting the implications of their results to only an intermediate research step. Thus, while their model performs well by typical metrics (accuracy of 0.95, precision of 0.96, recall of 0.98, and an F1-score of 0.97), their results do not provide much practical guidance other than a simple "good" or "bad"—for only show comments rather than actual show success.

Moreover, while MyAnimeList is the largest, most exhaustive community of English-speaking anime fans, it does not segment users by their means of viewership, whether through legal or illicit services, preventing AlSulaim and Qamar from making any conclusions about the behaviors and preferences of either kind of user (only for an erroneously averaged user) which we hypothesize to be meaningfully different.

Armenta-Segura and Sidorov (2023) conduct a related experiment also using MyAnimeList data. Rather than looking at viewer comments for the shows, however, they look

at shows' synopses and several other qualities such as show genre to predict, like AlSulaim and Qamar, whether a show will be successful (defining success as having a rating of 7 out of 10 or higher). Their workflow is split into two stages: First, vectorization of the text, converting the synopses into a format intelligible to machine learning methods; and second, the machine learning methods that output a binary classification result. Three techniques are evaluated at each of the two stages, resulting in nine unique modeling combinations.

The three vectorization techniques are derived from two variants of N-grams, a textual representation method dividing a corpus into basic units of words/characters called tokens—the first variant with n=1, meaning that each token represents a single word, and the second variant with n=1,2, and 3 and tokens representing words, as well as n=3 and tokens representing characters—and dependency trees, directed graphs representing words as nodes and their relationships as edges. The three machine learning methods are Support Vector Machines (SVMs), Logistic Regression, and Naïve Bayes. These methods, although functioning by different mathematical processes, all measure how the appearance and ordering of certain words and characters in the show synopses data predict the success or failure of a show.

While the approach is sophisticated and utilizes powerful natural language processing techniques, the results unfortunately engender little practical value, with all models displaying an accuracy score between 50 and 60 percent: In the context of binary classification, this means that the models perform only as well or just marginally better than randomly guessing as to whether a show will or will not be successful. Improvements in the methodology as well as technological advancements in natural language processing might improve these results, but the greater lesson appears to be that there is little connection between the synopsis and success of a show.

Haraguchi (2022) poses a similar question to this study and scrutinizes the different factors that might affect the preference for illicit over legal anime-streaming services. Using a studied framework for leisure research, he creates a 31-question survey, segmenting the questions to inform five determinants of "involvement," one's propensity to use or have an attachment to a particular good or service—in this case, for pirating websites. These determinants are: "Attraction," a composite variable of how important and pleasurable one finds a particular thing or activity; "sign," how using the good or participating in the activity affects a person's reputation; "risk probability," the probability of making a mistake due to choosing the good or activity; "risk consequence," the magnitude of that aforementioned mistake; and "centrality," how deeply embedded the good or activity is to one's network of social connections. The author then hypothesizes that sign and centrality would dissuade one from pirating sites—sign, because of the negative associations with pirating, and centrality, since illegal streaming, if only slightly, violates one's moral obligations which are directly related and influenced by centrality. Attraction, risk probability, and risk consequence, he posits, encourage piracy because of, for attraction, the tantalizing selection of shows on the pirating sites available (sometimes exclusively), and for risk probability and consequence, because there is no fee, no financial risk of using the platform, unlike Crunchyroll and other legal streaming means.

Using a partial least squares model to classify survey respondents as either using or not using pirating sites, the author is directionally correct about the relationship between feature and target variables insofar as the polarity of the coefficients for each feature variable, whether they are positive or negative (except for the relationship between centrality and moral obligation, where there seems to be a slight negative correlation). Out of the five predictors, the model highlights only two as significant: Risk consequence (most influenced by the size of the

subscription fee of legal platforms) and moral obligation. Surprisingly, attraction does not pass the 5% significance threshold, signaling that the pirating platforms might not be much better quality/experience-wise than legal alternatives. However, this study was conducted on Japanese college students who have nearly exhaustive access to anime shows via legal means, narrowing differences across many quality metrics between legal and illicit platforms. In the United States, on the other hand, due to licensing complications and the additional effort needed to subtitle/dub shows in English, pirating sites currently offer so much more than legal sites, making them far more attractive for English-speaking audiences purely on the grounds of the breadth of their selection.

Upon review, there appear to be several significant gaps in the literature. The first gap is the subject of research. Most studies, particularly AlSulaim and Qamar and Armenta-Segura and Sidorov, exclusively review the omnibus MyAnimeList data which dissolves meaningful differences between platform users. These studies make no mention of pirating sites or their users which we hypothesize to be significant in number as well as significantly different from those who patron legal sites.

The second gap is the prediction task. Rather than regression, many models in the literature (again, as with the same two previous studies) are focused on binary classification: whether a show is successful or unsuccessful, if it receives positive or negative reviews, etc. While these models might be useful in other contexts, for the particular use case that both papers mention—consulting television networks and streaming platforms about what shows to invest in—they do not seem to likely offer specific, useful recommendations.

The third gap is localization. The attractiveness of pirating is most likely a function of the availability of shows on legal websites, the main competitor of the pirating sites, which is itself

likely a function of the physical location/language of the legal websites. In Japan, for instance, there are better partnerships between legal streaming services and anime producers and no need to subtitle/dub shows, allowing these legal sites to offer a wider selection of anime content. In distant English-speaking countries like the United States, however, these sourcing advantages do not exist, engendering smaller, less competitive show libraries on the legal sites.

This study attempts to fill these gaps in the literature, initiating novel research into an overlooked segment of one of the planet's fastest-growing entertainment sectors.

Data

Data Scraping

The first leg of the race started with data scraping an anime piracy site. I began experimenting with several candidates such as AniWave, Aniwatch, and Gogoanime, some of the more popular websites of their kind, and ultimately decided to use AniWave because of its large user base and simple website architecture. Very quickly, however, I ran into two problems:

Problem 1: Pop-up ads

Traditionally, one of the drawbacks of these sites is the many virus-ridden pop-up ads that appear whenever a user clicks on something on the page. This became a problem for my first approach which was to use a WebDriver, an artifact of the Selenium python package that allows me to automate/simulate normal human behavior on a website: Typing in the search bar, clicking buttons, etc. My initial approach was to search for shows one by one, access their profiles, and scrape the data of interest. However, whenever I tried to simulate a click, rather than get the desired outcome, I would be taken to a pop-up ad that interrupted the workflow and could potentially inflict my computer with malware.

Solution 1: Exploit website architecture

I instead went into AniWave's gallery-view list of all their shows ordered alphabetically. One way of navigating this list is by clicking through a series of numbered pages, each page containing 30 show titles (with 403 pages there were over 12,000 shows). Most critical for my data scraping workflow was that each page had an associated URL unique on the final numerical

character. For instance, the first page had the URL “https://aniwave.to/az-list?page=1”; the second page, “https://aniwave.to/az-list?page=2”. I was then able to copy the entire URL link up until the concluding number and then use a simple for loop to iterate through all 403 pages. From there, I scraped the URL of each show on each page and accessed them directly—without having to simulate any clicking activity, bypassing the pop-up ad problem.

Problem 2: Cloudflare

Cloudflare is a security system that protects websites from bot activities irrespective of motive, whether it be for malicious or academic or any other purpose. The Cloudflare security system on AniWave then detected the WebDriver and prevented it from further accessing the website. Using the base version of the package was insufficient to get the information I wanted.

Solution 2: undetected_chromedriver

I looked through online forums and found a developer who had created a stealthier version of the Selenium WebDriver called undetected_chromedriver. While the documentation for this package was sparse (having only the backing of a single developer) I was able to figure out how to use the necessary functions through trial and error and the help of ChatGPT, synthesizing and explaining the example code to quickly create intelligible documentation on how to use the package.

Problem 3: Insufficient application memory

My original web scrapping approach consumed far too much application memory because of the large amount of data that was being actively handled. I first tried to use a different computer, hoping that would resolve the issue, only for the program to crash again.

Solution 3: Batching

After consulting with a friend, I realized that while I was overextending my RAM (where application memory is stored), I could access more storage by saving active data to nonvolatile storage like my computer's hard disk. I then decided to web scrape in batches: Scraping 1000 shows, saving that data, clearing the huge temporary DataFrame where I stored the show data batch, and repeating until I had exhausted every show on the website.

After a preliminary cleaning, my final product was a data set with 11675 shows, including the title, type (TV, movie, etc.), score (on a 0-10 scale), number of score votes, start date, and studio for each show.

Data Sets

Other than data personally obtained through data scraping, we also examine two data sets representing different segments of the anime community—these being 1) Crunchyroll, the preeminent legal anime-streaming website in the English-speaking world (although they also offer subtitled and dubbed versions in other languages) with both a free but content-restricted version and advertisements punctuating each episode, as well as a fee-based premium version; and 2) MyAnimeList, the preeminent online anime forum and ranking database, where fans record/rate the shows they've watched and discuss all things anime-related across countless discussion threads. Given the site use cases and the size/demographics of their user bases, we use

Crunchyroll to represent the entire legal anime-streaming community and MyAnimeList to represent the entire anime community, agnostic to how its users might have viewed the shows they praise or criticize.

The Crunchyroll data set comes from Kaggle and contains 41 different features about 1255 different anime shows. 29 out of these 41 features, however, are dummy variables on genre, each variable representing a different discrete value of the whole range of possible genres (for instance, genre_action indicates whether a show belongs to the action genre). Other notable variables include, most importantly, the name and overall rating of the show (a continuous variable on the interval 1-5), the number of episodes, the number of ratings, and the quantity of 1,2,3,4, and 5-star ratings, represented by five separate rating variables (Filardi, 2019). Since the data was collected in 2019, however, the data set does not fully account for all the shows and preferences on Crunchyroll in 2024. However, given that the company was founded relatively long ago in 2006 and, being the first streaming platform to focus exclusively on Anime content, has from inception been the leader of its industry, the 2019 Crunchyroll community resembles its later cross section in 2024.

The MyAnimeList data also comes from Kaggle and contains 35 features about 17558 instances of anime content (before preliminary cleaning). The data collection date is only 4 years ago and given the platform's even earlier launch date in 2004, we have confidence that the data still represents the current anime-fandom landscape in 2024. Moreover, while there are fewer features in this data set than in the Crunchyroll data set, the MyAnimeList data includes a wider range of interesting and likely significant features for predicting show success: Along with critical features in the Crunchyroll dataset (show name, score, genres, etc.) the MyAnimeList

data set also documents the English-translated and original Japanese name, the media type (TV show, movie, etc.), the producers, and the animation studio of each show (Valdivieso, 2020).

Data Cleaning

Renaming columns:

The first stage of data cleaning was to implement a variable naming convention. For instance, originally, all three data sets referred to the title of a show by various column identifiers: “anime,” “Name,” and “title.” I then standardized the names of these platform-invariant variables, simplifying future merge steps in the data processing.

There were also variables measuring the same quantity but with unique values per data set, such as the overall rating and the number of ratings for each show. I renamed these platform-variant variables to have the same prefix and appended the platform name as the suffix (for instance, “rating_aw” to denote a show rating on AniWave).

Handling missing values:

After ensuring naming consistency across my three data sets, I then moved on to handling missing values: Many entries in the AniWave and MyAnimeList data sets had missing values and represented those missing values differently: AniWave, for instance, simply showed the string “?” when something was missing; MyAnimeList, on the other hand, inputted the string “unknown.” I first applied unique filters to each data set to identify the missing entries, and from there, simply dropped the entire row if it had a missing start date or rating. Had I placed particular significance on studying every show in existence, this dropping method would have eliminated important parts of my data; however, since I planned to merge the three data sets at

the end, so long as Crunchyroll, the data set with the least number of missing values, did not contain the dropped titles in the AniWave/MyAnimeList data, I would not affect the final data product. This was indeed the case since most of the shows with missing values were old and obscure and thus lacking proper documentation—shows that Crunchyroll did not typically offer.

Other cleaning:

Along with the previous two data cleaning tasks, I also had a variety of smaller challenges to overcome. First, rating standardization: I had to ensure that ratings from the three platforms were all on the same scale to compare them more easily; second, media-type filtering: To maximize the chances of success in the following step, merging data by record linkage, I kept only the TV shows—the most significant media type—and dropped all anime movies, OVAs, ONAs, and other media types. This reduced the number of entries with similar names (if they were a movie and television show from the same franchise, for instance) increasing merge success through record linkage; third, keeping only relevant columns: I identified and dropped all irrelevant non-data columns such as the URL of each show on Crunchyroll, the website-specific unique identifiers, and other such variables.

Record Linkage

Record Linkage is a data-matching technique for linking records from different sources that describe the same real-world entity. In a perfect world where all data is consistently formatted and standardized, I could have easily merged the data on the show names and retained all entries from the smaller dataset. However, due to inconsistencies in translation preference (whether a show is identified by its English or Japanese name), other naming differences, and

simple human error, record linkage is needed to distinguish what might be true matches from genuinely different entries. Record linkage performs this matching process by taking a pre-selected column from each data set, calculating a match score for every possible observation pairing on those columns, and assigning each pairing either a 0 or 1 to indicate the absence or presence of a match. We can also compare across multiple column pairs and then declare the observation pairing with the highest number of matches as the true match.

We can customize the recordlinkage class to render optimal comparison models for idiosyncrasies among the different data pairings. Of these customizations, the two most significant factors are the comparison method and the similarity threshold a potential pairing must cross to qualify as a match—or in other words, how similar two entries must be to be considered a match.

To evaluate which method-threshold combination yielded the best merge, for each sensible combination, I calculated the merge success rate (the length of the merged data set divided by the length of the Crunchyroll data set—the smallest data set) and isolated the duplicated rows in the merged data set (the mistaken merges). I then optimize both metrics, maximizing the merge success rate while minimizing the number of incorrect, duplicated matches.

I chose to use "jarowinkler" for the comparison method. The similarity threshold, however, posed a far more challenging decision to make: If set too low, the program would match one entry with many entries in the other data set rather than the single one it was supposed to match with. If set too high, the program would miss too many matches, demonstrated in a low merge success rate. Ultimately, I found that setting the threshold to 0.95 (on a scale from 0-1) provided the best results. For start date, however, I set the threshold to 1, requiring a perfect

match, because I was previously able to standardize the start dates by casting them all into Python's datetime format: A true match would then share the exact same start date, character for character.

After first merging the AniWave data with Crunchyroll and then with MyAnimeList, I was able to retain 51.234% of the Crunchyroll shows. Ultimately, after the many cleaning and merging steps, my final product was a data set with 623 anime shows, each with the combined information of the three data sources, and most importantly, each source's unique rating and popularity score for all 623 shows.

Feature Engineering

Finally, after obtaining my cleaned composite data set, I created additional feature variables to supplement my modeling task.

Show age:

To calculate the age of each show, I took the difference in number of days between the show start date and January 1, 2024. Ideally, I would have collected data from all three sources on the same day and then used that day to take the difference; however, given that the Crunchyroll and MyAnimeList data were scraped at different times, I chose January 1, 2024, as the benchmark date, approximately the date I collected the AniWave data. However, even if the dates are displaced a bit too late or early, so long as they are displaced by a constant number of days (which, for Crunchyroll and MyAnimeList, are the number of days between the true collection date and January 1, 2024) the date offset should not affect the strength of my model because the *relative* dates are retained, even after adding a constant.

Controversy:

Controversy is defined as the amount of disagreement over the rating of a show between AniWave and Crunchyroll (the illegal and legal streaming communities). This is measured by taking the squared difference between the AniWave and Crunchyroll ratings:

$$(rating_{aw} - rating_{cr})^2$$

Show source:

Show source is a collection of one-hot encoded dummy variables derived from the "source" variable from the MyAnimeList data set. Originally, "source" was a single variable that took on one of several possible sources that could inspire an anime television show adaptation: Manga, web manga, digital manga, 4-koma manga, book, novel, light novel, visual novel, card game, game, music, original, other, or unknown. Because source is a nominal variable, we must first convert it into this collection of binary variables or implement some other conversion technique to include it in our regression models.

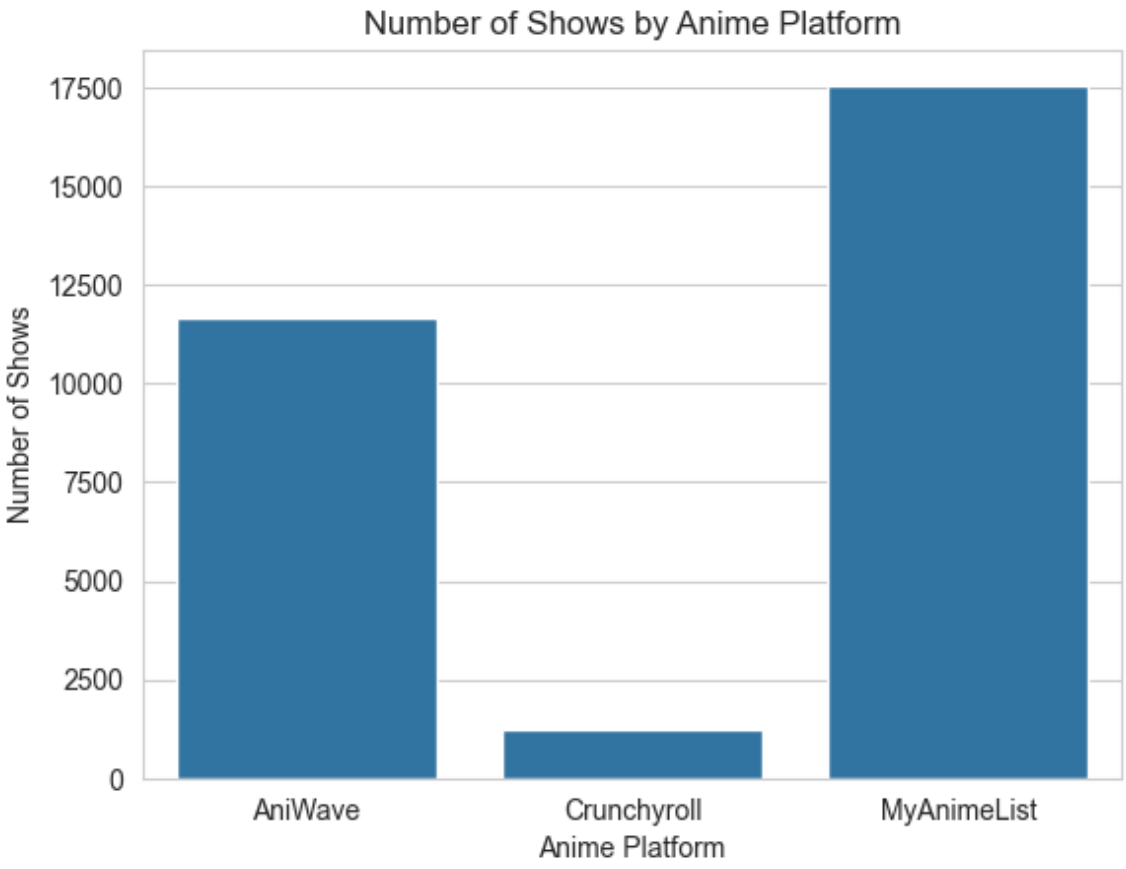
Methodology

Exploratory Data Analysis

We begin our analysis by comparing basic biographical details among the three anime platforms: The number of shows offered, the average show ratings, and the top 20 shows from each site.

Number of Shows by Site

As first mentioned in the introduction, the number of shows offered across the three sites differs drastically, as the bar chart from before demonstrates:

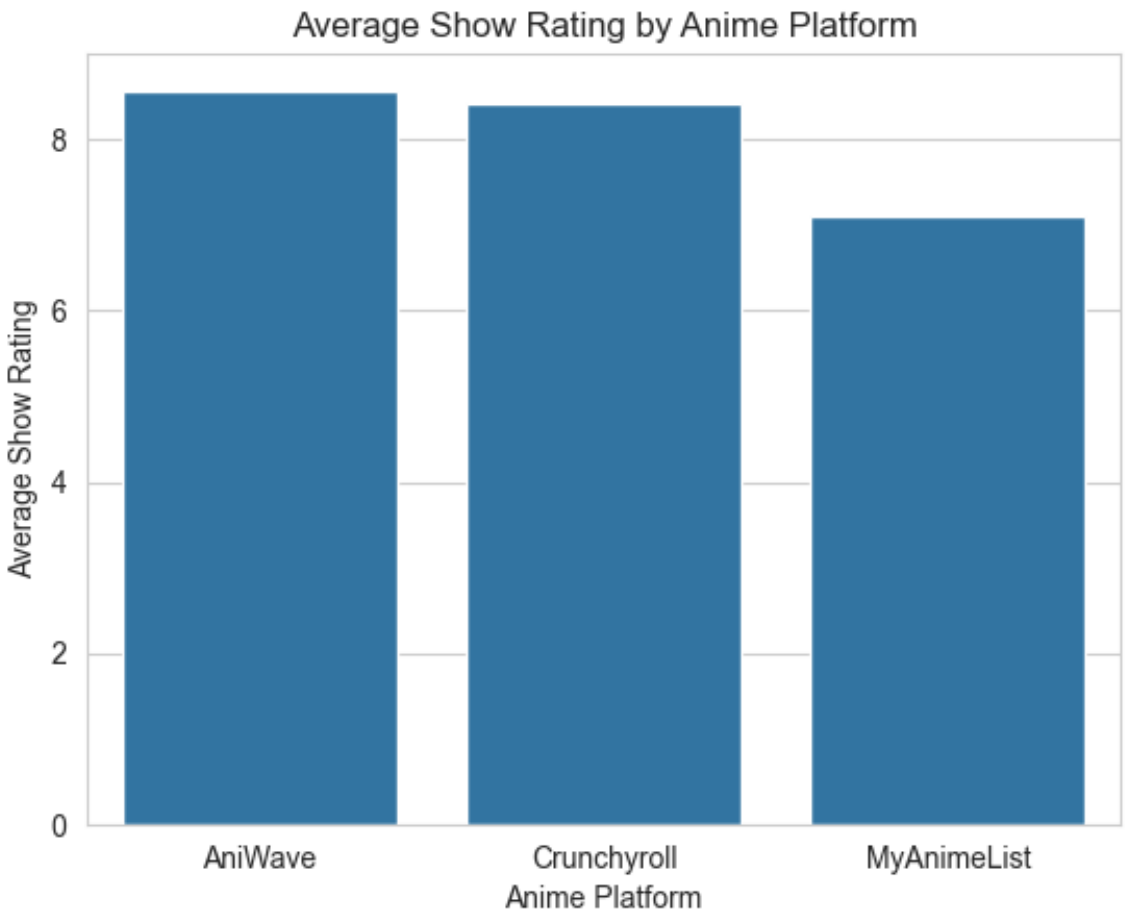


AniWave offers a broad selection of 11675 titles—66.47% of all 17562 titles recorded on MyAnimeList—and 9.3 times that of Crunchyroll's 1255 titles. Granted, Crunchyroll lists multi-season shows under a single title, whereas AniWave and MyAnimeList separate them by different titles. Still, even if we grossly overestimate that every show on Crunchyroll is a three-season franchise, tripling its true size, AniWave would nonetheless boast over 3 times as many titles as its primary legal competitor.

Looking at this first table, pirating sites like AniWave have a clear advantage over legal sites in the sheer quantity of shows available. This is likely due to the licensing costs and negotiations requisite for legal sites before they can offer a particular show to their patrons—a step that pirating sites disregard, albeit illegally.

Average Ratings by Site

All three sites share a similar rating system: AniWave has a five-star system and allows users to leave half-star ratings, displaying overall scores on a continuous 0–10 interval, truncated at the second decimal place. Crunchyroll also has a five-star system but doesn't allow for half-star ratings, choosing to present scores on a continuous 0–5 interval, also truncated at the second decimal place (although on the website it appears to display the score to the first decimal place). MyAnimeList also has a 1–10 interval for its rating system, but rather than representing scores by stars, it guides users toward a drop-down selection of discrete 1–10 rating options. Thus, to maintain parity across our three data sets, we double the Crunchyroll ratings to expand its rating interval from 0–5 to 0–10. By the end of our data reformatting, all platforms are on a continuous scale of 0-10 and are truncated at the second decimal place. We calculate the average show rating across the three platforms and visualize their differences in the following bar chart:



AniWave and Crunchyroll seem to be equally generous in their show ratings, averaging scores of 8.574 and 8.431, respectively, yet the average MyAnimeList score falls at a much lower 7.099. What explains this score discrepancy? It could be that AniWave and Crunchyroll, both being streaming platforms, elicit a unique sampling bias in their ratings that MyAnimeList, as a show-rating platform rather than a streaming service, is largely resilient against: Since AniWave and Crunchyroll are primarily streaming services, only those who either really loved or disliked the show would likely go on to use the secondary rating function. MyAnimeList, on the other hand, is built around this rating/review function and is therefore more likely to attract a wider spread of users, even those with only middling opinions. In other words, the platforms'

different focuses—how central ratings/reviews are to the platforms' core function—can engender sampling bias.

Another explanation could be that the same numerical rating can mean different things among the different platforms. For instance, while AniWave and Crunchyroll have a standard, noncommittal rating system, MyAnimeList appends a small description to each score option: A 10 is described as a "masterpiece," 9 as "great," 8 as "very good," 7 as "good," 6 as "fine," and 5 as "average"—the very last of the non-negative ratings. These descriptions, if synchronized with how users organically rate shows (If a 5 truly represented what people consider to be an average show), would pose minimal effect on show ratings. However, there appear to be incongruities that potentially engender response bias in the MyAnimeList ratings: While people have different intuitions on what qualifies as an 8 or 6, as "very good" or "fine," at least for 5 – "average," there is a mathematical definition. According to our most recent graph of *average* show ratings, however, while MyAnimeList insists that a 5 should signify an average show, the platform's true average rating is slightly over 7: Thus, it could be that MyAnimeList, by its score labeling system, discounts nominal scores across its platform so that, for example, one user might rate a show on Crunchyroll ~8.4, the platform average, but then rate the same show ~7 on MyAnimeList because of the biasing score descriptions.

It is possible, perhaps even likely, that both the sampling and response biases affect the data and add noise to our regression models. Sampling bias is perhaps the most deleterious: If the three data sources are unrepresentative of their respective populations, then the data might misinform us about population characteristics and sabotage our show success predictions. The response bias, on the other hand, would largely disappear after normalization, maintaining the reliability of our models.

Altogether, even in basic details about the number of shows offered and the average score recorded by these streaming platforms, we begin to see significant differences between AniWave and Crunchyroll as both services and communities. These analyses also demonstrate a recurring problem throughout this study, particularly in the data processing stages: Due to, first, the international nature of anime and inconsistencies in translation, and second, the lack of general standardization across anime-specialized platforms because of their grassroots origins, it is very difficult to massage the data into a standardized format necessary for cross-platform comparisons.

Most Popular Shows by Site

Along with the number and associated scores of shows on each platform, it would also be helpful to know *what* specific shows are most successful among the three anime communities. We begin by looking at the top twenty most popular shows from each site:

Top 20 Shows (AniWave, Crunchyroll, MyAnimeList)

Rank	AniWave	Crunchyroll	MyAnimeList
1	ONE PIECE	Naruto Shippuden	Death Note
2	**Redo of Healer (2021)**	Shugo Chara	Shingeki no Kyojin
3	**The Eminence in Shadow (2022-2023)**	BLEACH	Fullmetal Alchemist: Brotherhood
4	Black Clover	Naruto	Sword Art Online
5	JUJUTSU KAISEN	Skip Beat!	One Punch Man
6	Naruto: Shippuden	REBORN!	Boku no Hero Academia

7	Boruto: Naruto Next Generations	Gintama	Tokyo Ghoul
8	JUJUTSU KAISEN Season 2	La Corda d'Oro ~primo passo~ and ~secondo passo~	Naruto
9	Attack on Titan Final Season	Eyeshield 21	Steins;Gate
10	Chainsaw Man	Hayate the Combat Butler! (S1 e S2)	No Game No Life
11	Demon Slayer: Kimetsu no Yaiba Swordsmith Village Arc	Natsume Yuujinchou	Kimi no Na wa.
12	Demon Slayer: Kimetsu no Yaiba Entertainment District Arc	Myself; Yourself	Hunter x Hunter (2011)
13	**Oshi No Ko (2022-2023)**	Blue Exorcist	Boku no Hero Academia 2nd Season
14	Bleach	Durarara!!	Code Geass: Hangyaku no Lelouch
15	Horimiya	Time of Eve	Angel Beats!
16	**I Got a Cheat Skill in Another World and Became Unrivaled in The Real World, Too (2023)**	Tegami Bachi Letter Bee	Shingeki no Kyojin Season 2
17	SPY x FAMILY	School Days	Toradora!
18	**Hell's Paradise (2023)**	Sword Art Online	Naruto: Shippuden

19	Attack on Titan Final Season Part 2	The World God Only Knows	Noragami
20	Mushoku Tensei: Jobless Reincarnation Season 2	Demon King Daimao	Mirai Nikki

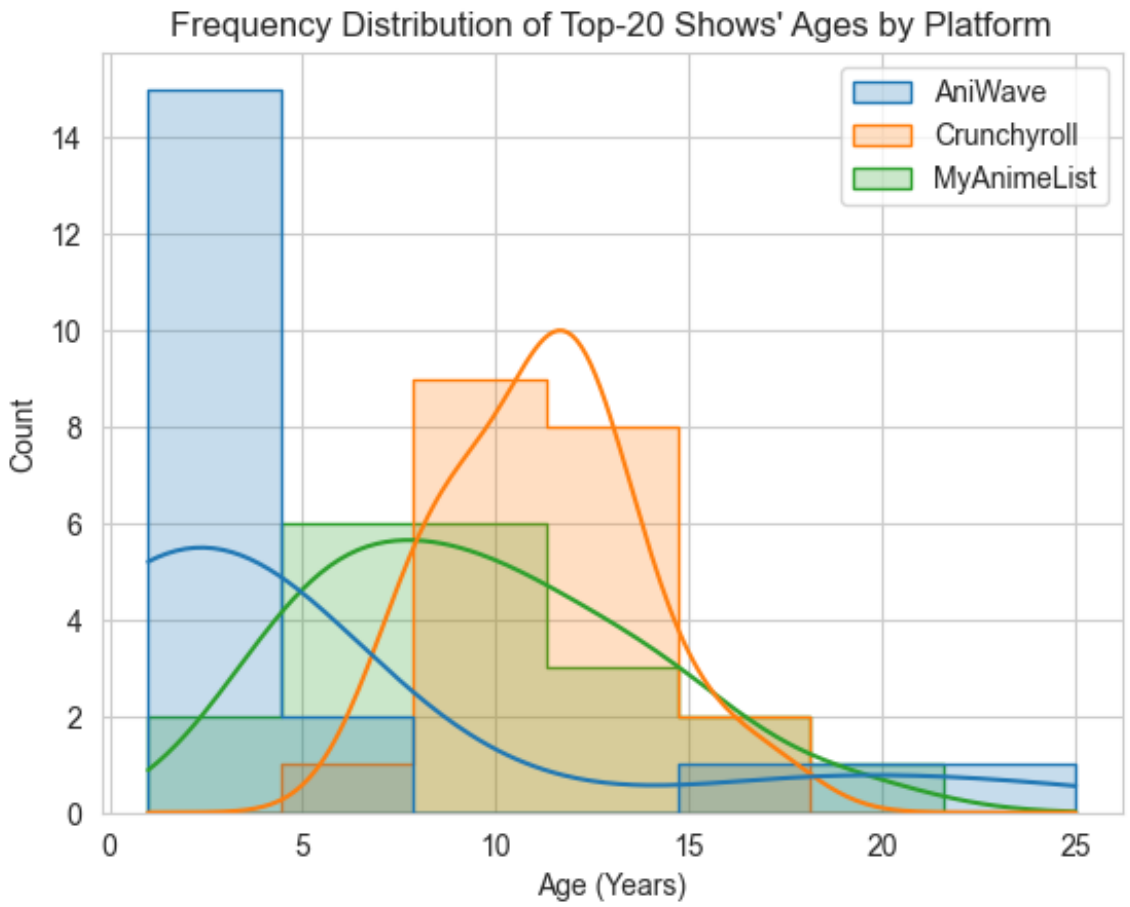
As the table illustrates, each site seems to prefer a very different set of shows, Naruto: Shippuden being the only shared entry among all three top-20 lists. Beyond that, there are no remarkable similarities; rather, there are far more differences: Between AniWave and Crunchyroll, for instance, as of Spring 2024, the five bolded titles in the AniWave column are unavailable on Crunchyroll, meaning that if AniWave users migrated to Crunchyroll, they would likely lose access to ~25% of their favorite shows. This reinforces our hypothesis that show selection is one of the most attractive features of anime piracy as well as a critical weakness of legal streaming platforms.

Moreover, apart from a few high-profile classic franchises (*One Piece, Bleach,* etc.), the most popular shows on AniWave are far more recent than those on Crunchyroll and MyAnimeList, even after factoring in the older collection dates for the Crunchyroll and MyAnimeList data (2019 and 2020, respectively). We then calculate the age of each show in the above table in years, taking the difference from when the show first aired and when the data was collected, and obtain the following summary statistics:

Top-20 Shows' Ages (in Years): Summary Statistics by Platform

Platform	Mean	Median
AniWave	5.35	2.5
Crunchyroll	11.25	11.5
MyAnimeList	9.55	9.5

The most striking result from this table is perhaps in the third column, the median: Crunchyroll has the oldest median show age of 11.5 years, closely followed by MyAnimeList's 9.5 years—then, finally, by AniWave's 2.5 years median, a steep drop from the previous two site statistics. Moreover, we see that while Crunchyroll and MyAnimeList appear to have negligible differences between their mean and median, a 0.25-year difference for Crunchyroll and 0.05 for MyAnimeList, that difference is far greater for AniWave, its mean being 2.85 years greater than—over twice the amount of—its median. Because the AniWave mean is so much higher than its median, we expect a dramatic right skew in the AniWave data whereas, in the Crunchyroll and MyAnimeList data, we would expect the frequency distribution to look more symmetrical. We visualize these distributions in the following histogram:



Crunchyroll and MyAnimeList both seem to have symmetrical bell-shaped curves—although the most popular Crunchyroll shows are more concentrated around the median, as demonstrated in the tall spike shape in the middle, whereas the MyAnimeList show ages are more evenly spread apart, as expressed in their flatter shape. AniWave users, however, as we inferred from the summary statistics, behave very differently: 15 out of the 20 shows are in the youngest bin (shows that are 0–5 years old), followed by another two in the following 5-year bin, then three outliers after the 15-year mark. Indeed, the five shows in AniWave's top 20 list unavailable on Crunchyroll—*Redo of Healer* (2021), *The Eminence in Shadow* (2022-2023), *Oshi No Ko* (2022-2023), *I Got a Cheat Skill in Another World and Became Unrivaled in The Real World, Too* (2023), and *Hell's Paradise* (2023)—all fall into the youngest bin. Thus, we

infer two things about AniWave users and extrapolate to the broader anime-pirating community: First, they are fans who sit at the edge of the anime industry, eagerly anticipating new shows, hoping to be the first to see and discuss a new episode with other fans. For these more enthusiastic fans, pirating sites are like specialty stores with shelves upon shelves of exclusive, rare items; legal sites like Netflix, bloated department stores full of things they might have little interest in, or if like Crunchyroll, far smaller and more cramped than the competition. Therefore, even if legal streaming sites were completely free, it seems unlikely that many viewers would entirely migrate from the illegal to the legal platforms.

Second, because they are such fans, these AniWave users have likely already seen the big mainstream shows, and whether a site carries these shows might not matter much to the AniWave users. Examining the history of AniWave and the broader anime-pirating landscape illuminates this claim: AniWave, although currently the largest anime library on the internet, was launched only in 2016, well after the most popular series like *Dragon Ball* and *Naruto* had made their way into the American pop culture canon (For instance, the popular cartoon television channel, Nickelodeon, aired English-dubbed episodes of *Dragon Ball Z Kai*, a *Dragon Ball* sequel, from May 2010 to February 2013) (Fandom, n.d.). Moreover, following the 2020 shutdown of KissAnime, the once most popular anime-pirating website, ex-KissAnime users migrated to other platforms like AniWave: Many of these users had already watched, rated, and left comments on older, more mainstream shows on KissAnime, yet these records did not transfer over with them to the new platforms. This explains why older, enormously successful shows don't make it to the AniWave top shows list: AniWave users don't use the website to watch these shows since they've previously watched them on other sites or even on mainstream American television. Rather, as the top-20 list indicates, they visit AniWave to watch new shows, particularly shows that are unavailable on Crunchyroll or other legal platforms.

Modeling

Overview

The EDA yields several important details informing what model best suits the data and the prediction task. First, the data, originally just 623 shows long, narrows to an even smaller training set of 436 shows after dividing the data upon a 70-30 train-test split: It is a relatively small data set. Second, several key variables seem likely to share a strong linear relationship—for instance, the rating variables from the three anime platforms among themselves. For this modeling step, we specifically focus on the AniWave and Crunchyroll data to underscore differences between their represented streaming communities. We will use the MyAnimeList data as feature variables for both models and later return to it when further investigating our model results.

Given both the relatively small sample size and strong linear relationships between key variables, a simple linear model might be the best candidate. Thus, I will first implement the linear model, expecting already decent performance, and then proceed to use more complex machine learning models. At the end of this modeling stage, I will compare the models' performances and highlight the best modeling approach for each data set. To conclude, I will perform a feature importance analysis and interpret the synthesized modeling results, focusing on their implications for both streaming platforms and their user bases.

Model Variables

Target Variable

1. *Weighted score*:

a. To calculate the shows' overall success, we take the rating and popularity of each show, normalize these metrics to eliminate bias, and calculate the weighted average of the two scores, weighing the rating score at 20% and the popularity score at 80%: *Weighted score = normalized rating * 0.20 + normalized popularity * 0.80.* Popularity is weighed more heavily than rating because of how we define "overall success"—primarily by its commercial success rather than its artistic quality. In other words, as long as the show draws a large audience and attracts more overall demand for its host platform, the show will have, from a pure profit-seeking perspective, fulfilled its purpose nearly irrespective of the overall show rating.

Feature Variables

1. Weighted score (other):
    1. When constructing our models to explain a particular data source, we use the weighted show scores of the two other anime platforms as feature variables. For instance, when predicting the weighted score of AniWave shows, we use the weighted scores reported by Crunchyroll and MyAnimeList in our regression model.
2. Show Age:
    1. Show age is calculated by the number of days between the show start date and January 1, 2024.
3. Controversy:

    1. Show controversy represents the squared difference in show ratings between AniWave and Crunchyroll: $(rating_{aw} - rating_{cr})^2$. The greater the difference, the more "controversial" the show.
4. Episodes:
    1. Episodes denotes the number of episodes in a show.
5. Genres:
    1. Genres is a collection of one-hot encoded dummy variables, each variable representing a unique genre (For instance, if a show is classified as a romance/action show, it has the value 1 for the romance and action variables and the value 0 for all other genre variables). This is useful since many shows are categorized by more than one genre.
6. Sources:
    1. Sources is another collection of one-hot encoded dummy variables, each variable representing the original media that inspired the show. Many anime shows are adaptations from already-successful manga franchises (Japanese comics), light novels (short novels typically geared toward younger audiences), video games, etc.

Fine-tuning Hyperparameters

To select the optimal hyperparameters, I use the GridSearchCV() function from scikit-learn (sklearn), a Python library for machine learning, to perform a grid search on an initial batch of potential hyperparameters and then use 10-fold cross-validation to calculate performance scores without further shrinking our training data to create a validation set: For each model, using all possible combinations of hyperparameters from the lists given, the function performs 10

rounds of cross-validation and calculates the average model performance. Then, once we know the performance of each combination of hyperparameters, we select the highest-performing combination and use it in the final fine-tuned model. We reuse this fine-tuning process for all models, changing only the hyperparameters that differ from model to model.

Model 1: LASSO Regression

About the Model

Least Absolute Shrinkage and Selection Operator (LASSO) regression is a modification of the standard linear regression: Adding a penalty term to the loss function, LASSO discourages large variable coefficients, reducing some coefficients down to zero. It is useful for regularization and feature selection, effectively removing features irrelevant to the prediction task. The base OLS model is as follows:

$$rating_{aw/cr} = w_1\left(weighted\ score_{cr/aw}\right) + w_2(weighted\ score_{mal}) + w_3(episodes) + w_4(show\ age) + w_5(controversy) + c_{6-28}(*\ genres) + c_{29-32}\ (*\ sources)$$

In an ordinary OLS model, we would optimize the mean squared error (cost function). However, with LASSO, we add an additional penalty term that penalizes large coefficients. Therefore, the model is pressured to set the coefficients for irrelevant variables to zero and perform other regularization, as previously mentioned. The cost function would then become this equation:

$$Cost\ Function\ = \frac{1}{n}\sum_{i=1}^{n}\left(h_\theta(x)^i - y^i\right)^2 + \lambda \sum_{i=1}^{n}|slope|$$

$$\lambda = Hyperparameter$$

(Dev, 2023)

With 32 feature variables, most of which are one-hot-encoded genre and source variables, LASSO helps reduce the dimensionality of the data set, guarding our model against overfitting so it can accurately predict scores of shows outside of the training set.

Hyperparameters

1. Alpha: The primary hyperparameter for LASSO regression is the penalty term, alpha.
    a. AniWave model: 0.01
    b. Crunchyroll model: 0.001

From our fine-tuning experiments, it appears that the optimal penalty for AniWave is a whole order of magnitude greater than that for Crunchyroll. This might indicate that the relationships between the predictor variables and the target variable in the Crunchyroll data have more noise than those in the AniWave model, allowing AniWave to have a simpler model with more penalty than Crunchyroll.

Lasso Regression - Performance

AniWave

1. MSE: 0.0012858024584301323
2. R-squared: 0.4449417039952387

Crunchyroll

1. MSE: 0.002429309282147904
2. R-squared: 0.4944873587814178

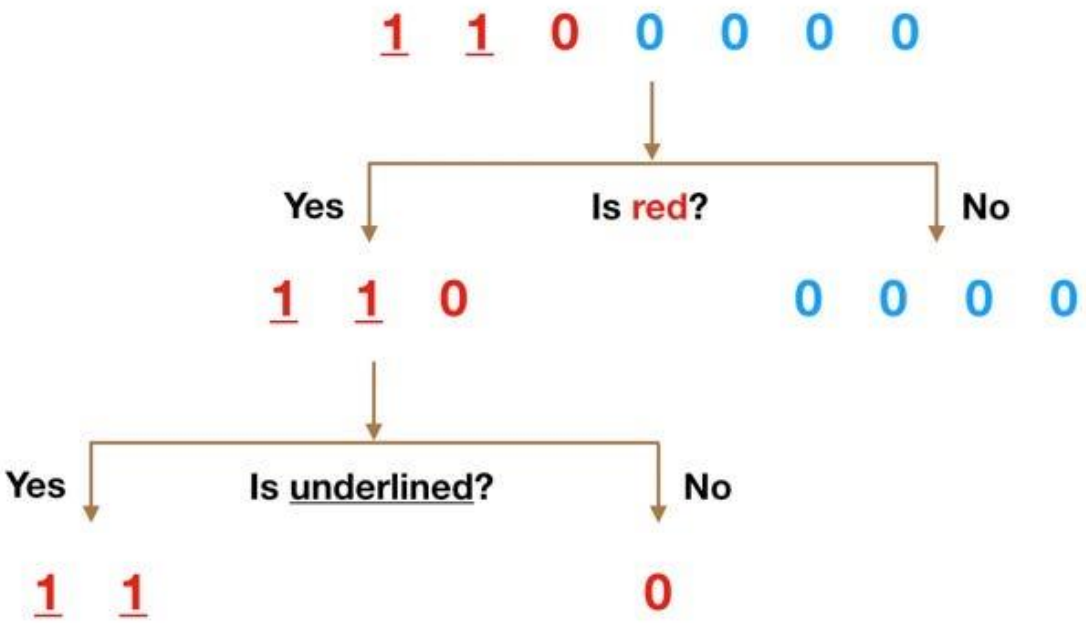
While the models perform slightly differently in each run-through, both the AniWave and Crunchyroll models perform similarly well by the r-squared metric—between 0.45 and 0.5, meaning that our models explain 45-50% of the variation in their respective data sets. While this is not a shockingly high r-squared value, it is still substantial given the expected difficulty of the estimation task: It is difficult, perhaps facile, to reduce a creative work to a few or even many features. Indeed, two shows can have the same setting, characters, and themes but lead to completely different levels of commercial success due to unquantifiable, elusive characteristics—both endogenous to the show such as writing style and animation quality, as well as exogenous such as consumer preferences during the launch of the show. For instance, in 2017, the Claymation children's cartoon, *Pingu in the City*, with only six staff members and a single voice actor, summited the rating charts on MyAnimeList—its popularity fueled by viral memes about the show's protagonist, Pingu the penguin. Seemingly arbitrary events such as these are not so rare in the anime fandom, rendering predictions of show success far more difficult than perhaps that of shows in other fandoms.

Still, using this LASSO model, legal streaming sites such as Crunchyroll would be able to, with some confidence, assess the commercial success of a potential new show with AniWave users and the broader anime-pirating community.

Model 2: Random Forest Trees

About the Model

To understand random forest trees, we must first understand its predecessor decision trees. A decision tree is a network of splits in the data made by "decision nodes"—true/false statements about features. For instance, with a data set of integers between 1 and 100, a decision node could be "> 70": All instances that evaluate true on this inequality would be segregated to the right side, and all that evaluate false, to the left. We then continue with these decision nodes until we've reached a satisfactory level of predictive power—when we observe that if an instance has these set of qualities mapped out by our decision tree, we can predict that so and so will happen or is true. After all the splits are made by the decision nodes, the remaining collections of samples end up in what are called leaf nodes. The following graphic illustrates this process:



(Yiu, 2019)

Random forest trees, on the other hand, looks at a collection of decision trees and takes a "vote": In a simple binary classification case, it would take the majority opinion of all the trees and output the verdict as the final prediction. Along with classification, random forest trees are also able to perform regression and output a continuous prediction, which is how we will use it in this study.

Hyperparameters:

1. n_estimators: Determines the number of decision trees that we use to make our final prediction.
    a. AniWave model: 100
    b. Crunchyroll model: 50
2. max_depth: Determines the maximum depth of the individual decision trees. Depth is measured by the largest number of decision nodes preceding the lowest leaf node.
    a. AniWave model: None
    b. Crunchyroll model: 30
3. min_samples_split: Determines the minimum number of samples required for a decision node to split the sub-sample. For instance, if the minimum number is 4 but there are 3 nodes in the sub-sample, there will not be a split.
    a. AniWave model: 2
    b. Crunchyroll model: 10
4. min_samples_leaf: Determines the minimum number of samples needed to be a leaf node. For instance, if the minimum number is 5 but there are 10 samples left, then that would not qualify as a leaf node and would require further splitting.

a. AniWave model: 1

b. Crunchyroll model: 2

The two fine-tuned models appear to have different values for all four hyperparameters. While it is difficult to make conclusions about the data based on the hyperparameters for random forest trees, it appears as though the Crunchyroll model takes on more unusual hyperparameter values, deviating further from the default hyperparameters than the AniWave model deviates.

Random Forest Trees - Performance

AniWave:

1. MSE: 0.0009350730277568754
2. R-squared: 0.59634542769468

Crunchyroll:

1. MSE: 0.0030638380874803656
2. R-squared: 0.3624488675649852

From these performance results, we see that the random forest trees model works well for AniWave, scoring an r-squared value of 0.596. On the other hand, the model performs far worse for Crunchyroll and has an r-squared value of 0.362. This is also far lower than the r-squared value of the Crunchyroll LASSO regression model, 0.494: The steep drop in r-squared from the less complex LASSO regression to the more complex random forest model suggests that the random forest model is too complex for and is overfitting on the Crunchyroll data, degrading the quality of predictions made on Crunchyroll show data outside of the training set.

On the other hand, it appears that a more complex model is well suited for the AniWave data, as demonstrated in the jump from the AniWave LASSO model's performance, r-squared = 0.445, to the random forest model's performance of r-squared = 0.596. While further study should precede any dogmatic conclusions made about the data sets, from our modeling results, it seems likely that the relationships between variables in the AniWave data are both stronger and more complex than those in the Crunchyroll data.

Indeed, these implications corroborate a critical component of my overall hypothesis, that the typical user of anime-pirating sites has unique preferences and complex reasons for using these sites—more so than do users of legal anime-streaming services.

Feature importance

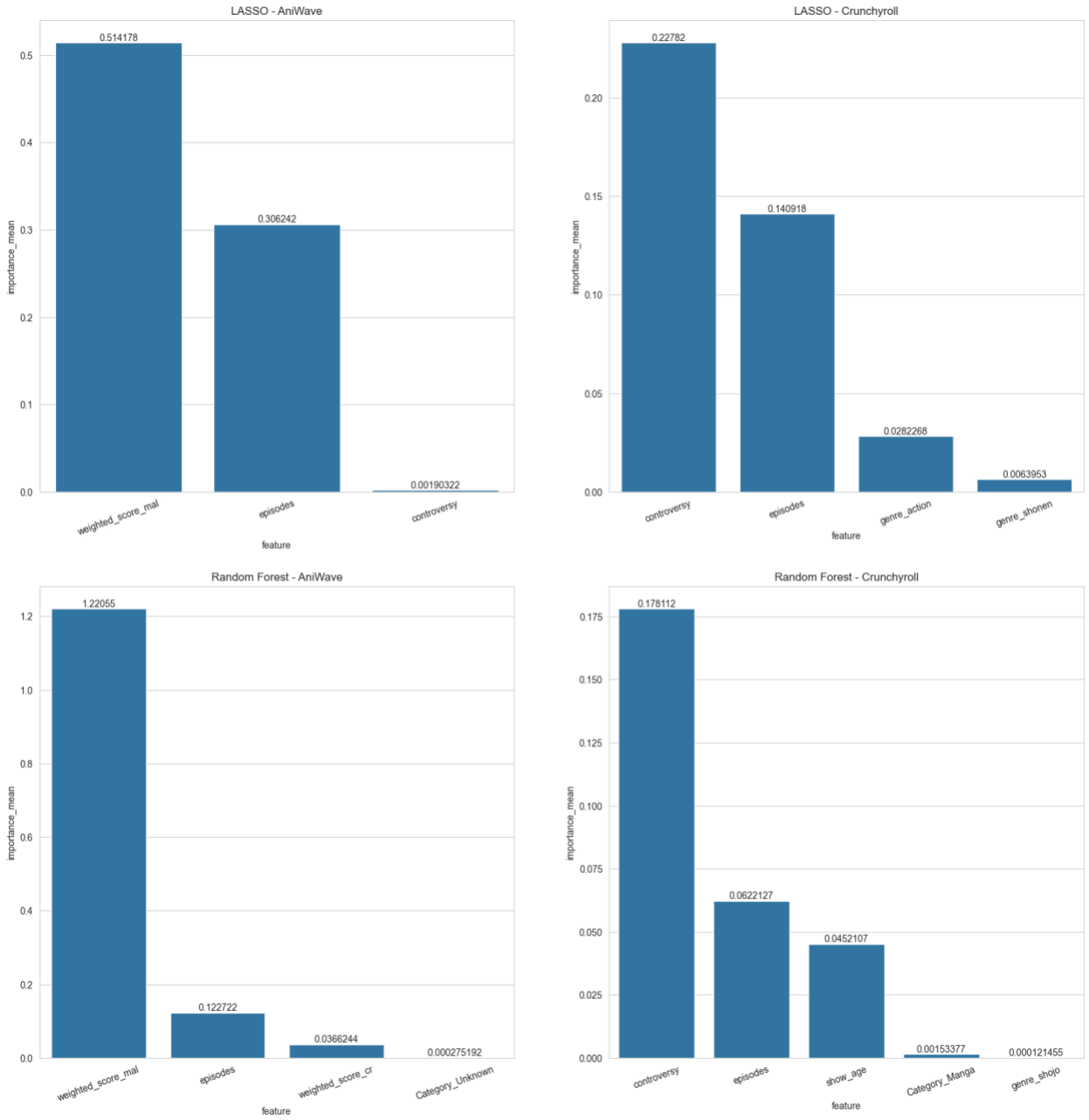
Having fit our models to the two data sets and then evaluated the goodness of their predictions, we now investigate feature importance and consider what variables are significant to our models and the resulting implications about the data. We do this through sklearn's permutation feature importance function, a model-agnostic technique for measuring feature importance. At a high-level view, this approach measures importance by 1) iterating through and randomly shuffling each feature variable n times, breaking the relationship between the feature and target variable which mimics dropping the feature from the model, then 2) measuring the decrease in r-squared with the permuted feature. A large decrease indicates that the feature variable is very important to the model and individually explains that much of the variation in the data.

Next, we apply two layers of filtering: First, we filter for variables that are significant at the 5% level (we calculate this confidence interval using the n-sized score distributions, one

distribution for each feature variable that we permute n times); second, we select only the top ten most important features to compare in our analysis.

We examine feature importance first on the LASSO models and then on the random forest models.

Permutation Feature Importance – LASSO and Random Forest Models

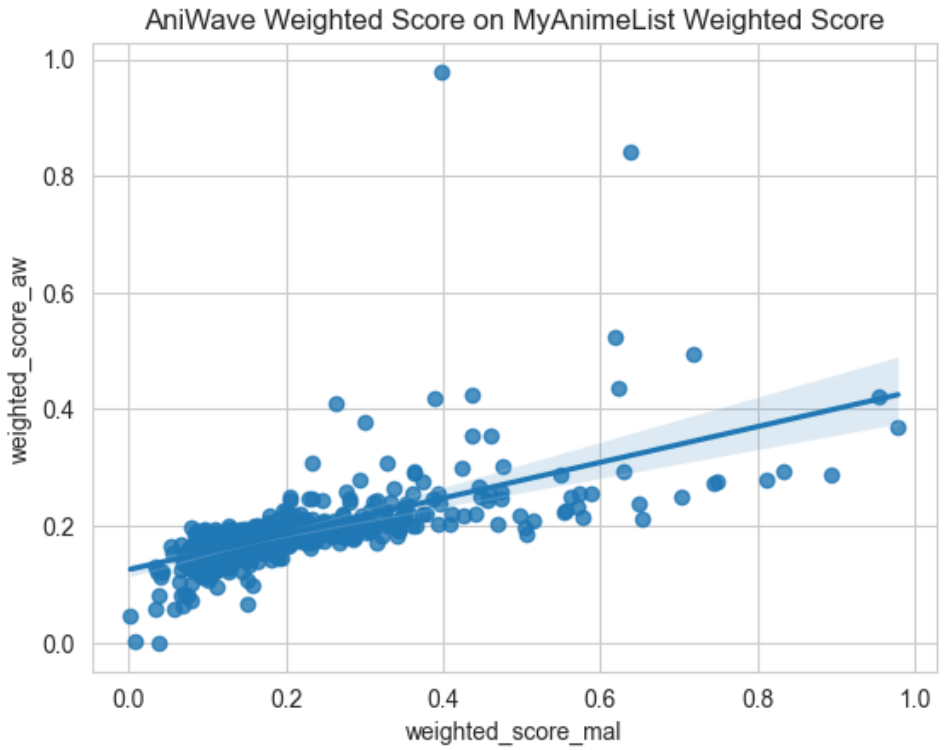


LASSO Models – Feature Importance

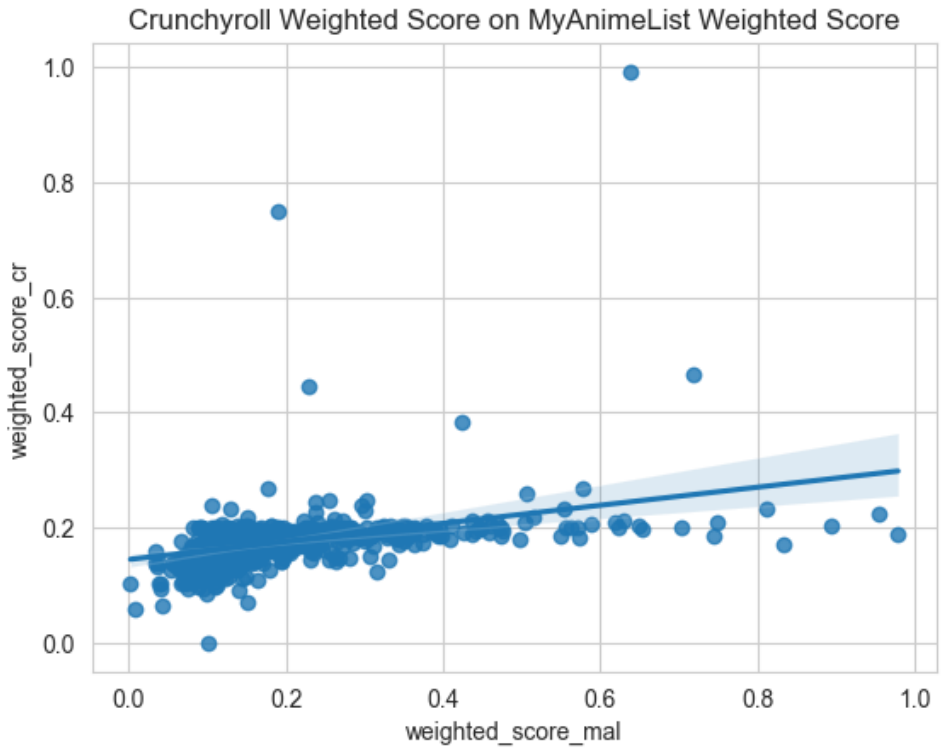
First looking at the LASSO - AniWave model, we see that the MyAnimeList weighted score (weighted_score_mal) is by far the most important feature variable in the model, so much so that, without it, the r-squared of the model falls by over 0.5. One caveat to balance this finding: Although permutation importance simulates removing a feature variable, it does so imperfectly, introducing some degree of inaccuracy. Moreover, the mean importance score for weighted_score_mal has an error of +/- 0.052 (5.2%)—which might further cause the reported importance score to diverge from the true importance score. Still, regardless of some amount of inaccuracy, permutation importance allows us to confidently compare the *relative* importance of the feature variables within and across models, making it perfect for this study. That aside, we look to the other significant feature variables: The episodes feature has a score of 0.306 (r-squared drops by 0.306), and controversy, a score of 0.0019 (r-squared drops by 0.0019).

Next, looking at the LASSO – Crunchyroll model, we see that controversy is the most important feature (0.22782). This is then followed by episodes (0.062), the first feature to reappear, and then by two genre variables: The action genre dummy variable, genre_action (0.028), and then the shonen genre dummy variable, genre_shonen (0.0063).

Surprisingly, the MyAnimeList weighted score feature is not even significant in the Crunchyroll model despite its high feature importance score in the AniWave model. This seems to indicate that the AniWave and MyAnimeList communities have far more overlap with each other than they have with Crunchyroll. We see this through a simple univariate linear regression, regressing the AniWave and Crunchyroll weighted scores on the MyAnimeList weighted score.



R-squared: 0.404



R-squared: 0.153

Indeed, given the close relationship between AniWave and the wider anime community on MyAnimeList, it is likely that Crunchyroll users, those who pay to legally stream anime, are in fact the niche, and anime pirates, the norm.

Random Forest Trees Models – Feature Importance

Our feature importance results for the random forest trees require further qualification. Beginning with the Random Forest - AniWave model, the most startling result is the 1.220 feature importance score for the MyAnimeList weighted score feature. We would typically expect our R-squared value to be between 0 and 1; however, R-squared can fall into the negatives when the model performs worse than the mean of the target variable as a predictor. The margin of error for this feature importance score is also much greater than that of any other feature of any other model, indicating that there is much more variance in the distribution of performance scores when weighted_score_mal is permuted. While a variety of other factors could have engendered this improbable statistic, it appears directionally correct in affirming the importance of the MyAnimeList weighted score in models predicting AniWave's weighted scores.

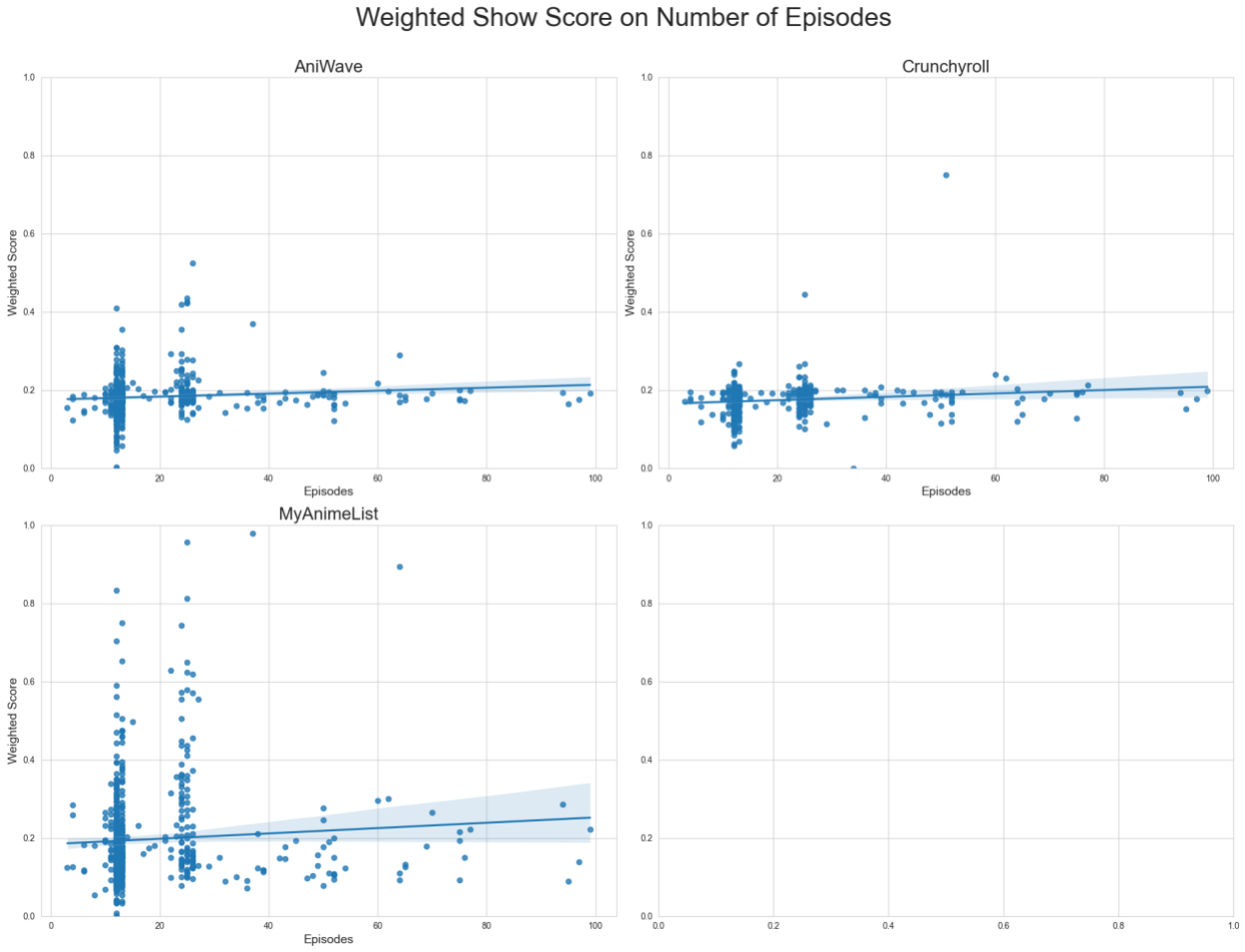
Besides weighted_score_mal, we see again that episodes is the second-most important feature for our AniWave model variant, with a feature importance score of 0.123. The Crunchyroll weighted score, weighted_score_crunchyroll, debuts with a feature importance score of 0.037. Category_unknown, the source dummy variable indicating that a show has unknown/undocumented origins, also registers as statistically significant but has such a small feature importance score that it is practically insignificant for our prediction task.

Then, moving on to the Crunchyroll variant for our random forest model, we see that controversy and episodes are again the most important feature variables, with feature importance scores of 0.178 and 0.062, respectively. Show age is also significant and has a score of 0.0452. The final two feature variables—the source dummy variable, Category_Manga, and the genre dummy variable, genre_shojo—while also statistically significant, have such low feature importance scores that they, like the Category_Unknown feature in the AniWave model, have little practical significance.

Examining Significant Variables

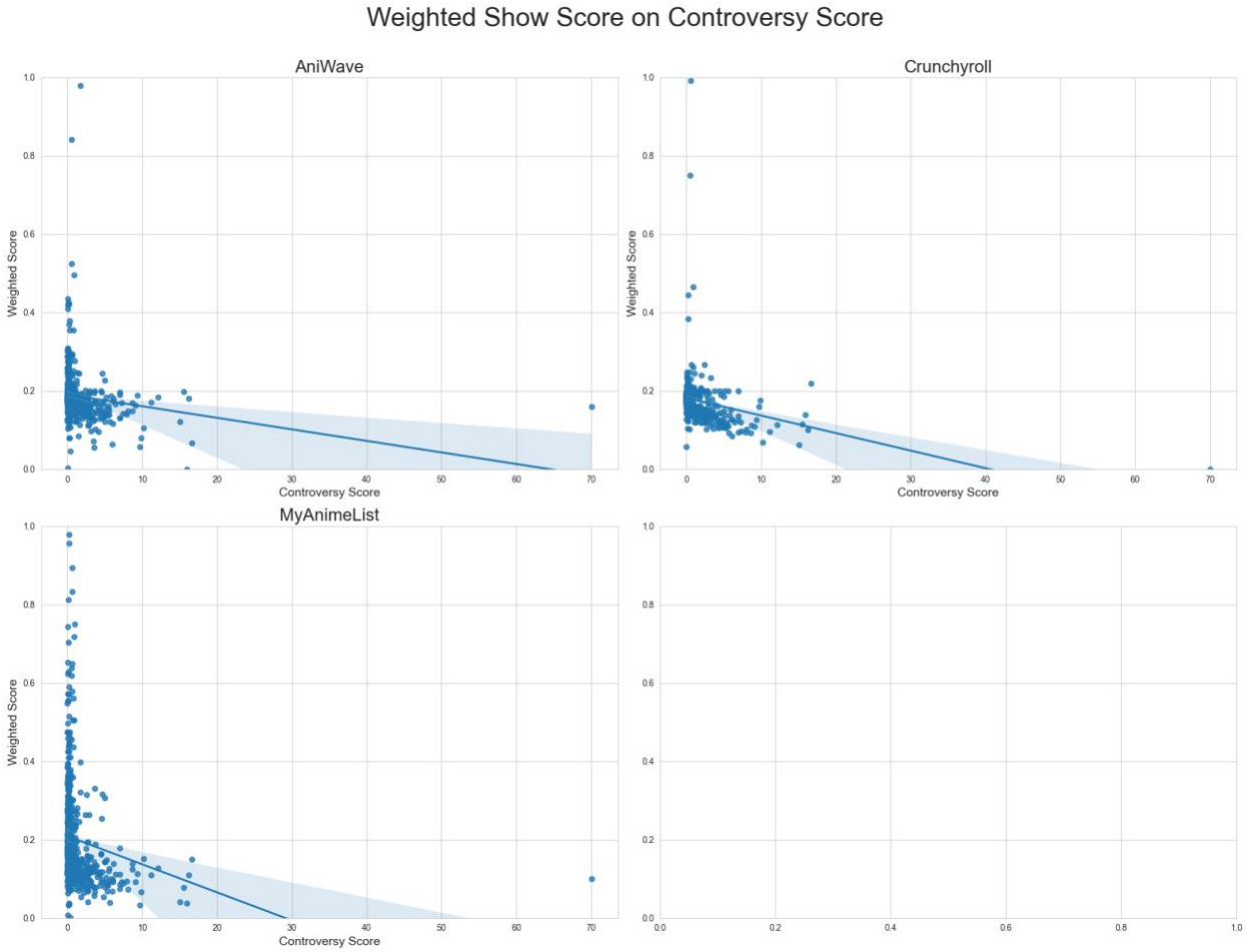
Finally, we further explore some of the most significant/important variables from our feature importance section: Episodes, controversy, and the genre dummy variables.

Episodes



For the episodes feature variable, for each streaming platform, we plot the platform weighted score against the number of episodes for each show and fit a linear regression model to the data, obtaining the three plots above. In all three plots, there seems to be a direct relationship between the feature and target variable: The more episodes there are in a show, the higher the weighted score. However, the slopes for all three lines are small, revealing the limited predictive benefit of the episodes feature variable.

Controversy



We see that the controversy feature variable—the squared difference between the AniWave and Crunchyroll ratings for a show—has an indirect relationship with each of the target variables: The more that AniWave and Crunchyroll disagree about the rating of a show, the generally worse it performs. The slope appears steepest for the MyAnimeList and Crunchyroll data and less so (but still moderately) steep for the AniWave data.

Looking at the top 20 most controversial shows in the table below helps explain why this might be the case: The most controversial shows are also extremely niche, attracting a small group of fans whose few unique opinions dominate the entire show rating.

Top 20 Most Controversial Shows (AniWave - AW, Crunchyroll - CR)

Anime	Controversy	Rating (AW)	Rating (CR)	Votes (AW)	Votes (CR)
Space Warrior Bladios	70.0569	8.37	0	39	0
11eyes	16.6464	5.08	9.16	353	**2137**
Gasaraki	16.2409	9.03	5	47	2
Love Rice 2	15.9201	2.93	6.92	7	13
HEYBOT!	15.5236	9.66	5.72	43	14
Nazotokine	15.0544	7.02	3.14	30	7
Street Fighter II:The Animated Series	12.1104	9.14	5.66	80	6
The Master of Ragnarok & Blesser of Einherjar	11.1556	7.94	4.6	**2271**	217
Rozen Maiden:Zurückspulen	10.1761	6.51	3.32	51	155
Hagane Orchestra	9.8596	5.66	8.8	14	5
The Glass Mask Year 3 Class D	9.6721	4.89	8	11	2
BBK/BRNK	9.3636	9.3	6.24	74	57
Active Raid	9.1204	8.44	5.42	90	52
Norn9	8.7025	8.55	5.6	333	5
Taboo Tattoo	8.6436	7.42	4.48	1571	148
TWOCAR	8.0656	7.94	5.1	66	11

Cerberus	7.9524	7.48	4.66	598	67
Bloodivores	7.3984	7.54	4.82	589	94
God Mars	7.2361	8.69	6	30	1
Hundred	6.9696	8.74	6.1	2428	177

The first thing that we observe is the final two columns: the Votes (AW) column (representing the number of ratings for a show on AniWave) and the Votes (CR) column (the same statistic but on Crunchyroll) have values far lower than the mean number of votes across all shows on AniWave and Crunchyroll—1920.524 and 436.346 votes, respectively. The average number of votes in this top-20 table, in contrast, is 436.25 votes on AniWave and 158.50 votes on Crunchyroll. Thus, we see that show controversy is closely related to show popularity which is also 80% of our final weighted show score: The more controversial shows tend to be less popular and therefore have lower weighted scores.

Comparing the average number of votes between platforms also demonstrates how much more popular AniWave is than Crunchyroll: The average show on AniWave, as calculated earlier, receives 1920.524 ratings, but on Crunchyroll, only 436.25 ratings, more than four times less than AniWave. Using the number of ratings as a proxy variable for overall site traffic, this comparison affirms what we earlier suspected, that AniWave and other pirating sites attract and serve most anime fans—especially the most passionate fans—whereas Crunchyroll has access to only a small fragment of the market.

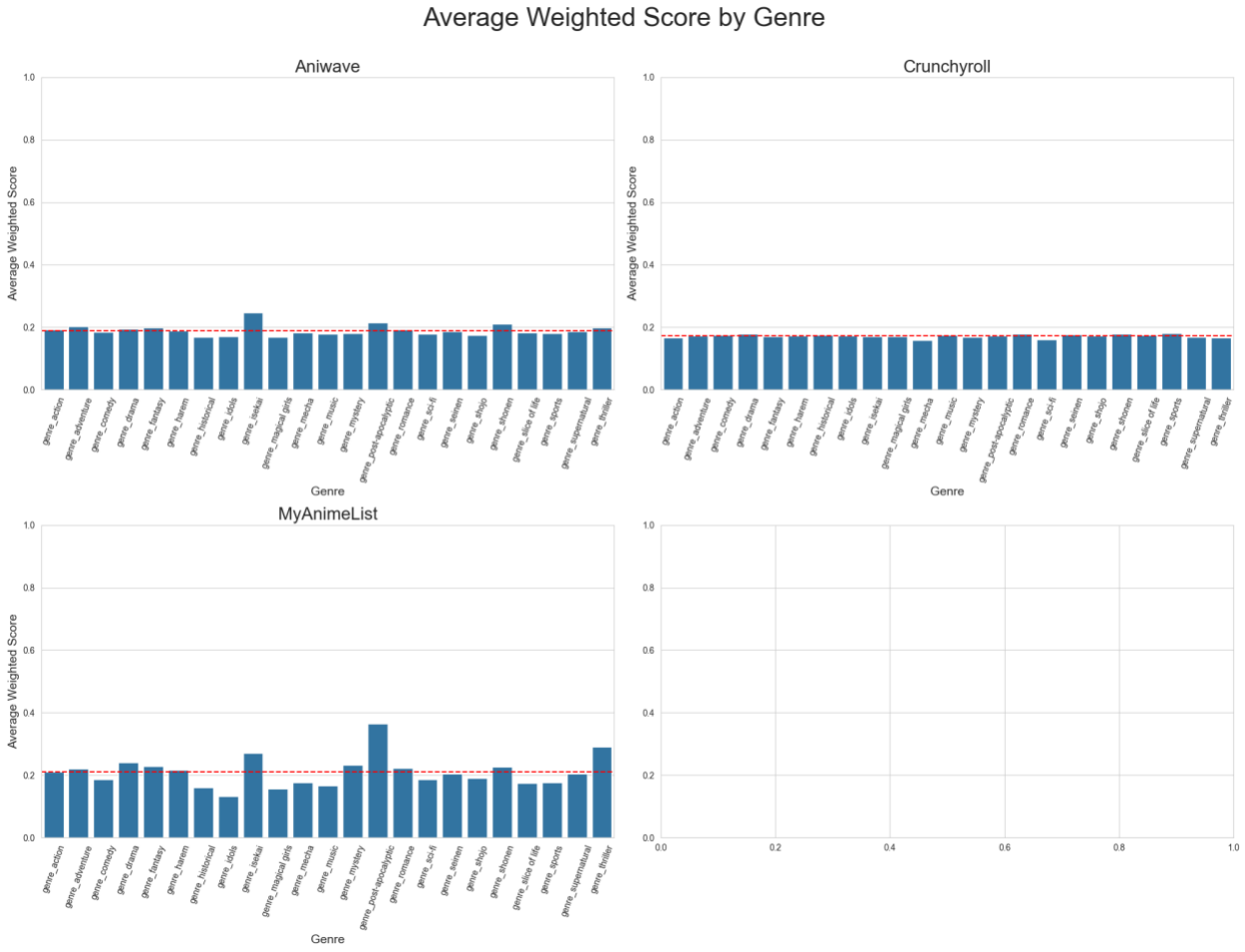
Finally, we also observe that out of the 20 shows in this table, 16 of them were rated more favorably on AniWave than on Crunchyroll. To quantify this difference in ratings, we calculate the average rating for the top 20 most controversial shows and find that AniWave, with

an average rating of 7.517, is over two points more generous toward these obscure shows than Crunchyroll is with its average rating of 5.437. Expanding this list from the top 20 most controversial shows to the top 100, we see that the average ratings are 7.850 for AniWave and 6.664 for Crunchyroll. Although the gap has narrowed, the average AniWave rating is still higher than Crunchyroll's.

Comparing these average ratings for obscure shows to the average ratings for *all* shows on AniWave and Crunchyroll—8.574 and 8.431, respectively—we see that the former pair of averages is far lower than the latter, which our linear models from earlier predict: The higher the controversy score, the lower the rating and popularity scores, and therefore the lower the weighted show score.

These statistics seem to indicate that AniWave hosts a greater proportion of users with niche, unique preferences and who are more likely to favorably rate obscure shows, once again affirming our hypothesis about the higher density of "mega fans" on AniWave than are on Crunchyroll.

Genre



Finally, we investigate the genre dummy variables by grouping each platform data set by genre and calculating the mean weighted show score for each genre. The above collection of bar charts visualizes these statistics and how the genre-specific averages compare to the average of the average genre scores, the red-dotted line, allowing us to see which genres are more or less successful on each platform.

We first look at the Crunchyroll results and see that all bars stand at approximately the same height, indicating that there is no obvious preference for any genre among Crunchyroll users: Genre does not inform the success of a show. The MyAnimeList chart, on the other hand, has a very different shape, showing that genre might be an important variable for predicting

show success, or at the very least, that members on the platform have some preference for and against shows of certain genres. AniWave seems to be in the middle, with most of its bars at approximately the same height but with a few jumping above and a few others falling below the mean of means line. However, these over-performing and under-performing genres in AniWave appear to mimic genre performance in MyAnimeList: In both data sets, the isekai, post-apocalyptic, and shonen genres surpass the mean of means, whereas the historical, idols, magical girls, mecha, and several other genres fall below. In other words, AniWave's preference for genres is a similar but more moderate version of MyAnimeList's, with shorter peaks and shallower troughs, once more strengthening our finding that AniWave and MyAnimeList share more overlap with each other in community, preferences, or both, than they do with Crunchyroll.

Discussion

Overall, these differences seem to indicate that AniWave's greatest advantage over Crunchyroll is the breadth of its show library—particularly in its recent show offerings, many of which are unavailable on Crunchyroll, allowing the pirating site to attract a larger, more passionate, and more cosmopolitan community with variegated preferences. Crunchyroll, on the other hand, attracts a much smaller part of the market, and perhaps because of both its limited show offerings and concomitantly mildly opinionated user base, demonstrates a more egalitarian distribution of success across its shows when grouped by show genre.

We also observe that, for the AniWave and Crunchyroll data separately, the important features for predicting show success generally remain the same regardless of which model we use to make the prediction: For AniWave, the MyAnimeList weighted score and number of episodes per show appear to be the most important features for both models, whereas for Crunchyroll, the controversy score and number of episodes appear to be most important. While there is likely some overlap between the AniWave and Crunchyroll communities, studying their differences broadens our understanding of both user and platform. We summarize some of these differences below.

Difference 1: Preference for older/newer shows

Very early in the exploratory data analysis, we saw that the most popular shows on AniWave were far younger than those on both Crunchyroll and MyAnimeList, even after adjusting for the different dates when the data were collected: The mean show age for Crunchyroll's top 20 most popular shows, 11.25 years, is more than twice as old as that of AniWave's 5.35 years; the median, 11.5 years versus 2.5 years, over 4 times as old.

Difference 2: Diversity in preferences

As we also saw, not only are the most popular shows on AniWave far more recent, but a significant proportion of them are also unavailable on Crunchyroll—25% of the top 20. Indeed, comparing the whole show libraries of both platforms, we recall that AniWave offers 9.3 times as many titles as Crunchyroll, making the pirating site (and others like it) a necessary service for those who want to watch any one of these many shows absent from legal streaming libraries. Moreover, looking back at the show controversy feature variable, we saw that AniWave users are more likely to favorably rate controversial shows, shows that also tend to be relatively obscure. From these two observations, we see the connection between AniWave's large supply of shows, many of which are obscure and difficult to watch legally, and the demand for these shows. even the ones that typically go unnoticed by most fans.

The genre dummy variables also reveal differences in the kinds of shows that the typical AniWave/Crunchyroll user prefers: While Crunchyroll users don't seem to prefer any genre over another (the average show in each genre all perform equally well) there seems to be small preferences in the AniWave data, as revealed in the average ratings for some genres rising above or falling below the average of the average genre ratings. Moreover, we observed that these preferences among AniWave users mimicked those of MyAnimeList users, showing how AniWave might be more representative of the broader, diverse anime community found on MyAnimeList.

Difference 3: Significance in the broader anime community

We find further evidence that AniWave is more synchronized with the broader anime community of MyAnimeList through the univariate linear regression of AniWave's weighted

show scores on MyAnimeList's weighted show scores and comparing the regression's performance to that of the regression of Crunchyroll's weighted show scores on MyAnimeList's. Looking at their r-squared values, we see that, in the former model, the MyAnimeList score explains 40.4% of the variation on the AniWave data, but in the latter, only 15.3%.

We also see that AniWave simply attracts and serves more users than Crunchyroll, given that the average show on AniWave receives 1920.524 ratings, whereas the average show on Crunchyroll receives just 436.25 ratings. It thus seems likely that AniWave is not only more representative of MyAnimeList, but because of its large size, has considerable influence over the MyAnimeList ratings as well.

Limitations

Had I more time, the very first thing I would enhance is my Crunchyroll and MyAnimeList data. While it is likely that show-consumption preferences of Crunchyroll and MyAnimeList users have not changed in the years since the data was collected in 2019 and 2020, obtaining the 2024 cross sections of both sites' data might divulge further insights about new, high-performing shows on the two platforms—because, we concluded, AniWave users were most attracted to these new shows. It would then be interesting to see how these same shows successful on AniWave perform on Crunchyroll and MyAnimeList. However, when our Crunchyroll and MyAnimeList data sets were collected, these shows had yet to be aired.

Similarly, because our modeling section only looked at shows in the intersection of the three datasets, many shows on AniWave were omitted from the final regression and feature importance analyses. Indeed, these missing shows provide AniWave with its distinct advantage because there are so many of them—shows found on AniWave but absent from Crunchyroll and

other legal streaming services. Their exclusion likely impacted results to some extent, particularly in the feature selection analyses for the AniWave models. From the exploratory data analysis, we saw and expected to see in the modeling results that show age is an important determinant of the success of a show on AniWave, as demonstrated in the extremely young median show age of the top 20 AniWave shows (2.5 years) in contrast to that of the top-20 Crunchyroll and MyAnimeList shows (11.5 and 9.5 years). However, many of these young shows did not even exist in 2019 when the Crunchyroll data was scraped, necessarily omitting them even if they are presently available on Crunchyroll in 2024.

Moreover, recalling Haraguchi's research on the factors of pirating involvement—he found that the "attraction" attribute of anime-pirating sites was not a significant predictor of one's participation on such platforms; yet, in English-speaking markets, the greater abundance of these pirating-exclusive shows likely elevates the attractiveness of such illicit services. This zoomed-in view of pirating-exclusive shows might have yielded fascinating research, yet it would require a level of specificity beyond the scope of this study, which is simply to prove and highlight high-level differences between illegal and legal anime-streaming communities.

In the same vein about data: Although AniWave is one of—if not, the most popular illegal anime-streaming service on the internet—it is certainly not the only service of its kind, and obtaining data from other pirating sites might have improved the performance of the models and strengthened their external validity. Understanding a website's architecture, writing the web-scraping code, and simply running the code on thousands of show titles is, however, a very time-intensive process that would have had to be restarted and adapted to each website I wanted to scrape. Had I more time, however, I might have considered other platforms such as Gogoanime or Aniwatch that also boast large user bases. AniWave alone should still be representative of all

anime-pirating site users, although it would only be beneficial to have more data from other analogous sites.

For modeling, I also wanted to investigate other potential hyperparameter values for my random forest trees model. This was much simpler with my single-hyperparameter LASSO regression model; for random forest trees, however, since I was fine-tuning four hyperparameters, it would have been much more costly to add even a single additional value since the grid search iterates through every possible combination of the four parameters, and for each combination, performs n rounds of cross-validation, amplifying the already-amplified number of model evaluations. With only 3-4 potential values for 4 hyperparameters and using 10-fold cross-validation, the grid search took over four minutes to run all 1080 of the tests (3*3*3*4 combinations of hyperparameters run 10 times each) and find the optimal combination of hyperparameters. I then had to repeat this time-intensive process each time I updated my model or obtained a new training/test split, which was often, and adding additional hyperparameter values would have significantly prolonged my modeling process. Had I more time, I would have increased the number of potential values for each hyperparameter and done more rounds of grid searches to find the most optimal combination of values for my models.

Finally, I also wanted to use other metrics and techniques to evaluate model performance—particularly for my random forest model because of the unexpectedly large r-squared value that I consistently obtained for the MyAnimeList weighted score predictor. As previously mentioned, the technique I used, permutation feature importance, only simulates the dropping of the permuted feature variable, producing a less accurate result. I could have instead used drop-column feature importance which veritably removes the entire feature column from the model to measure the change in r-squared/importance of the dropped feature, although this

would also have been more costly since it would require the model to refit the data each time a feature was dropped. Both the permutation and drop methods, however, calculate feature importance using changes in r-squared; alternatively, I could have considered SHAP values as another metric, one that is both easy to interpret and more accurate since it considers every possible combination of feature variables with and without the particular feature whose importance is being evaluated, accounting for the interactions between the variables that simply dropping the variable overlooks. Unfortunately, I learned about SHAP values late into my thesis research and didn't have the time to learn the whole theory behind and the programming implementation for SHAP values. While these alternatives would likely tell the same story as the permuted feature importance results, they would provide me with more precise and accurate results with which to quantify differences between the AniWave and Crunchyroll data.

Conclusion

In the literature review, we identified three critical gaps in the current research corpus about the anime fandom, those being: The subject of research, the prediction task, and problems with localization. Through this study, although not exhaustively, we have begun to fill in the gaps.

For the subject of research: We investigate characteristics of the sparsely studied illegal anime-streaming community, hypothesizing that its members have distinct preferences and viewing behaviors from their legal-site counterparts. Through our exploratory data analysis, modeling, and feature importance steps, we were able to explicate these differences, using our results to make preliminary recommendations for legal platforms about the types of shows that would be most successful with both their current userbase and in the market segment claimed by illegal streaming.

For the prediction task: We were able to give more robust recommendations through two regression models: LASSO regression and random forest trees. Unlike previous studies that exclusively studied the MyAnimeList data and derived all its target and predictor variables from the one source, this study enhances the data quality by collecting new data from the pirating site AniWave and studying the two in tandem with a third source, the Crunchyroll data set. This allowed us to interpolate significant features from all three datasets into our model, enabling the fine-grained regression task of this paper over the lower-resolution binary classification tasks of previous research.

Finally with the problem of localization: Unlike previous literature studying anime pirating in Japan—the results of which, due to differences in the pirating landscape across languages, have problems with external validity—we rather focus on anime pirating in the

United States/English-speaking countries, exploring data from AniWave, Crunchyroll, and MyAnimeList, three US-based anime websites, thereby producing results and recommendations pertinent to American broadcasting and streaming enterprises.

Still, much work remains in studying the nascent but rapidly growing anime fandom and the melting pot of people, opinions, and behaviors it brings together. As mentioned previously in the discussion section, it would be exciting to see research isolated on shows that, for one reason or another, do not make their way onto legal streaming sites like Crunchyroll. These reasons do not appear obvious: As we saw earlier, for instance, the second most popular show on MyAnimeList, *Redo of Healer*, is missing from Crunchyroll's show library, discrediting the seemingly likely and simple hypothesis that only nonpopular shows are passed over by legal streaming sites. In addition, it would also be interesting to conduct a sentiment analysis study on AniWave show comments, similar to what AlSulaim and Qamar did with the MyAnimeList comments, and compare those results with Crunchyroll show comments, exploring a new vein of potential differences between the legal and illegal anime-streaming communities. Still, by collecting important data and performing analyses about costly yet sparsely studied anime pirating, this paper takes a critical step toward understanding and modeling the dynamics of this young but burgeoning entertainment medium watched by millions in the United States and around the world.

References

Alsulaim, S., & Ali Mustafa Qamar. (2021). Prediction of Anime Series' Success using Sentiment Analysis and Deep Learning. *2021 International Conference of Women in Data Science at Taif University (WiDSTaif)*. https://doi.org/10.1109/widstaif52235.2021.9430244

AniWave. (n.d.). Aniwave.to. https://aniwave.to/home

Armenta-Segura, J., & Sidorov, G. (2023). *Anime Success Prediction Based on Synopsis Using Traditional Classifiers*. https://rcs.cic.ipn.mx/2023_152_9/Anime%20Success%20Prediction%20Based%20on%20Synopsis%20Using%20Traditional%20Classifiers.pdf

Barnes, E. (2023, November 3). *19 Celebrities Who Are Surprisingly Into Anime*. Ranker. https://www.ranker.com/list/celebrities-who-like-anime/erik-barnes

DataHorizzon Research. (2023, September 18). *Anime Market to Reach USD 62.7 Billion by 2032 | CAGR: 9.4% | DataHorizzon Research*. Yahoo! Finance. https://finance.yahoo.com/news/anime-market-reach-usd-62-120000412.html?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAALKzcX3v7tfArBJU-nnvea7eq9PUwsMtZtEWQHPkewA8KcC7h-JL2zU4zcbqBpeLgnMk-Yoedm1V_L_ZmhTXnrroq6Q2WPBJL6zMMiDgchQPedNFaNjvWWHc5GiC6xwErHdS0QhDz9eHPyNMy9B3yJUzx0RPLjX4qhoWeOlkx7ei

Dev, S. (2023, January 20). *Ridge Regression and Lasso Regression: A Beginner's Guide*. Medium. https://medium.com/@devsachin0879/ridge-regression-and-lasso-regression-a-beginners-guide-b3e33c77678

Fandom. (n.d.). *Dragon Ball Z Kai*. Nickelodeon. Retrieved April 12, 2024, from https://nickelodeon.fandom.com/wiki/Dragon_Ball_Z_Kai#:~:text=Dragon%20Ball%20Z%20Kai%20(known

Filardi, F. (2019). *Crunchyroll animes database*. Kaggle. https://www.kaggle.com/datasets/filipefilardi/crunchyroll-anime-ratings

Haraguchi, K. (2022). Effects of Enduring Involvement on Intention toward Digital Piracy: The Case of Japanese Anime. *Journal of Student Research*, *10*(4). https://doi.org/10.47611/jsr.v10i4.1403

Lindner, J. (2023, December 16). *Anime Popularity In America Statistics [Fresh Research]*. Gitnux. https://gitnux.org/anime-popularity-in-america-statistics/#:~:text=Highlights%3A%20Anime%20Popularity%20In%20America%20Statistics&text=Approximately%2074%25%20of%20US%20Netflix

Parrot Analytics. (2021). *From niche to mainstream: Anime's journey around the world*. Parrot Analytics. https://www.parrotanalytics.com/academy/from-niche-to-mainstream-animes-journey-around-the-world

Peters, M. (2023, April 23). *Anime and Manga Piracy Tied to $15 Billion Loss, New Report Reveals*. ComicBook. https://comicbook.com/anime/news/anime-manga-piracy-2022/#:~:text=It%20was%20there%20fans%20learned

Sevakis, J. (2012, June 11). *All About Licensing: Part I*. Anime News Network. https://www.animenewsnetwork.com/feature/2012-06-11

Sinha, E. (2023, February 15). *7 Times Anime Had Major Influence On The Fashion World For Good - Elle India*. Elle. https://elle.in/7-times-anime-had-major-influence-on-the-fashion-world/

Stone, M. (2018, September 19). *A Nerdy History Lesson: The Early Days of Anime in America*. Geeks. https://vocal.media/geeks/a-nerdy-history-lesson-the-early-days-of-anime-in-america

The Dickens Society. (2023, January 9). *Charles Dickens's Unwitting Victory over American Literary Pirates*. The Dickens Society. https://dickenssociety.org/archives/3658

Valdivieso, H. (2020). *Anime Recommendation Database 2020*. Kaggle. https://www.kaggle.com/datasets/hernan4444/anime-recommendation-database-2020

White, D. (2016, July 14). *Donald Trump and Hillary Clinton Want to Catch Voters With Pokémon Go*. Time. https://time.com/4407067/donald-trump-hillary-clinton-pokemon-go/

Worldometer. (2024). *Japan Population (2024) - Worldometer*. Worldometer. https://www.worldometers.info/world-population/japan-population/#:~:text=Japan%20population%20is%20equivalent%20to,(and%20dependencies)%20by%20population.

Yiu, T. (2019, June 12). *Understanding Random Forest*. Medium; Towards Data Science. https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Pledge

This paper represents my own work in accordance with University regulations.

Bryan Wang