Matthew Jaojoco
Dekhtyar
CSC 466-03
10 December 2021

Ethics in Data Mining - Twitch.tv

Introduction

On October 6, 2021, an anonymous hacker posted 125GB worth of data to the forum site 4chan. This data was pulled from Twitch.tv, Amazon's live streaming service platform acquired in 2014. This site has held a virtual monopoly on the live streaming space in recent years and hosts hundreds of thousands of unique livestreams every day with millions of viewers. In the following pages, I will analyze the contents of this leaked data, and reflect on how sensitive information can be mined from this data.

Data Sets Used

The 125 GB included proprietary SDKs, a detailed commit history of Twitch's source code, and an unreleased video game distribution service by Amazon meant to rival Valve's Steam. However, the most powerful data contained in the torrent is provided in the form of streamer payouts from August 2019 to October 2021. This data set provides a list of popular streamers, as well as the income they received from Twitch during the specified time period. This income comes from three different sources: viewer subscriptions, viewer bit donations, and ad revenue. Many streamers have other sources of income such as brand deals and merchandise sales, but these are not included. While the data doesn't provide the exact income a streamer makes, three included sources should prove to be an effective measure of success on the site.

There have been ongoing debates about whether or not login credentials were included as part of the leak, but an official statement from Twitch says they are

“confident that systems that store Twitch login credentials, which are hashed with bcrypt, were not accessed, nor were full credit card numbers or ACH / bank information”, easing the minds of their viewers.

To make the most use out of this data and discover the most meaningful findings, the subscriber counts for all of the streamers in the leak must be included in the data set. While there is no confirmation that the subscriber counts are included in the 125GB torrent, they can be easily obtained from third-party sites such as twitchtracker.com which allows users to look up any streamer and provides the sub counts for every month that they have been a Twitch affiliate, a rank required to receive payouts from streaming.

The Machine Learning Engineer View

A machine learning engineer, if given this data to observe, would notice that knowledge discovery from data techniques and methods could be applied to gain insightful information.

One KDD technique that could prove to be useful here is classification, but to classify a given streamer we must first define the possible classes. The way in which classification can be made useful is by having the classes represent different types of revenue deals that streamers can make with Twitch. Allow me to further explain.

Twitch will often sign exclusivity deals with the streamers on the platform, preventing the streamer from releasing live content on Twitch’s rival platforms such as YouTube Gaming and Facebook Live. These contracts also determine the percentage of sub revenue that goes to the streamer for the duration of the contract. Normally, a streamer gets 50% of the revenue from subscriptions to their channel, but streamers can negotiate for a higher percentage when signing a deal.

In order to predict what type of revenue deal an unsigned streamer would negotiate should they sign a deal with Twitch, categories would have to be created for all the popular revenue splits that Twitch offers their streamers. The exact splits offered to streamers are not publicly released, so the engineer would be left to guess the amount of categories and the values that each category covers. Charles White, known by his screen name moistcr1tikal, has publicly stated that the revenue per sub ranges between $2.50 and $3.50, or 50% to 70% since Charles provided these values assuming all subscribers subscribe at the first of three tiers. With these values in mind, the engineer could experiment with different class splits that cover this range. Since each class is supposed to represent a unique revenue split that Twitch offers its streamers, there should probably only be between two and four different classes. Three seems like the most plausible amount, implying that Twitch offers streamers either 50%, 60%, or 70% of the subscription revenue that they generate.

With the revenue from three different sources included in the database we will have to approximate the percentage of the payout that comes from subscriptions. But since the vast majority of payouts from Twitch come from subscriptions and we are neglecting tier 2 and tier 3 subscriptions which bring in more money, we can approximate by using the value of the payout directly.

Calculating the revenue per subscriber for each viewer can be obtained by simply dividing the total payout by the sum of all the subscribers throughout the 27 month timeframe that the payout covers, since these values are already contained in the data sets. Each signed streamer can then be classified by rounding this value to the closest of the three classes. It would be difficult to verify the accuracy of this classification process since the revenue splits are confidential, but there are a handful of streamers such as

Ludwig, the streamer with the sixth highest sub revenue according to the leak, who have strongly hinted that they signed to get the most profitable split (assumed to be $3.50). These such streamers could be given a ground truth and placed in the top class of $3.50 to test the accuracy of the calculation for revenue per subscription. If there are any streamers in the dataset who are affiliates but not partners, they could also be given a ground truth since they are guaranteed to have the lowest revenue split of $2.50.

Assuming that the revenue splits for the signed streamers are successfully calculated, we can then begin to experiment with different classification methods with the goal of discovering what attributes of a streamer most heavily impact their revenue split, but there are various attributes that could belong to a streamer. One possible way to check similarity between streamers is by comparing the viewers that watch them and interact with their stream, however the number of shared viewers may not prove to be impactful in negotiating revenue split. Another method to include the viewerbase in the similarity calculation is by having the demographics of the viewer base count as attributes. These demographics can include, but are not limited to: gender breakdown of viewerbase, average viewer age, region breakdown or country breakdown of viewerbase, or percentage of viewers that type in chat. These types of values seem more likely to influence the type of split that Twitch will offer. Detailed subscription breakdowns can also be included as attributes to see if having more of a certain subscriber type has a correlation to having a higher revenue split. You would expect all types of subscriptions to be encouraged, with tier 2 and tier 3 subscriptions possibly having a higher impact on contract negotiations. However, there is a type of subscription that may not be desired.

In a 2016 promotional attempt to increase site growth, Amazon began giving a free Twitch subscription to every Amazon Prime subscriber that connected their

Amazon and Twitch accounts. This introduced Twitch Prime subscriptions, renamed Prime Gaming in August 2021, as the fourth type of subscription in addition to the three tier system they've had in place since before Amazon acquired the platform. While Amazon's goal of increasing site growth was accomplished, Twitch loses money from these types of subscriptions so a higher percentage of subscribers that used Prime subscriptions may be undesirable and lead to a lower revenue split.

With the attributes for each streamer decided to be the both percentage of all subscribers and flat amount for each subscriber type, as well as demographics containing info about the viewer base, a data engineer can then create a decision tree by inducing C45 or some other decision tree making algorithm. The attributes closest to the root of the decision tree will show the attributes that have the most impact on revenue split, since at each iteration the C45 algorithm will determine which attribute produces the most information gain. Additionally, the data engineer can check to see if there are any strong association rules in the dataset where the right hand side of the rule is a type of revenue split. These give valuable insight on attributes that Twitch finds valuable in a streamer.

The Data Ethicist View

Knowing the amount of money that streamers make on its own is not harmful to society. Many celebrities sign deals where their earnings become public knowledge including actors and athletes. There was some outrage from the community upon realization of just how wealthy streamers are, with many feeling like the income was not deserved. However, top entertainers have been wealthy since long before live streaming became popular. Many streamers openly share their subcount, and those who do not

still have it listed on third-party sites. Streamers have also talked about the different revenue splits prior to the leak in October 2021, so any curious viewer could have calculated the income a given streamer gets from Twitch.

The classifier discussed in the section above would be an extremely powerful tool for livestreamers if it were to be implemented and made public. The classifier could potentially introduce harm if it encourages certain types of content to be made and alters the virtual ecosystem on the platform. The classifier would give strong hints to streamers about the audience they need to attract and the type of interactions that they need to extract from their viewers in order to get the best contract deal. The association rules mined would also accomplish this. Most top online content creators already keep a close eye on the performance metrics of their videos or streams and take note of ongoing trends in social media, but an algorithm that identifies the attributes Twitch values most in a streamer could very possibly create permanent trends and hurt the platform as all of its top creators become more and more similar in an attempt to draw in the most desired demographics and maximize negotiation leverage with Twitch.

The findings of C45 and association rule mining may also be harmful depending on which attributes are discovered to be valued by Twitch. An example of "harmful" findings would be if the two most desired attributes are having a young viewerbase and receiving a high amount of gifted subs. Then streamers could attempt to cater towards children and get the impressionable audience to spend their parents' money on the stream, since gifted subscriptions have no limit. Users must be at least 13 years old to register for a Twitch account in accordance with COPPA, so while children younger may be watching, the user will be listed at 13 at the youngest. This helps limit the impressionable audience, but users under 13 can still easily get onto the site and 13 is

still younger than the age requirement to work so if they subscribe it most likely isn't from their own wallet. When barely a teenager or still a child, it is difficult to comprehend the value of money and they hardly ever have their own money to spend, so becoming wealthy from catering content towards children is often frowned upon. Catering content towards children isn't intrinsically immoral, but many content creators for children get judged because of aggressive manner in which they force interactions out of their impressionable audience.

Another harmful finding would be if it is discovered that more hours streamed per week has a positive effect on revenue split. This could encourage streamers to live an unhealthy lifestyle. This is not only dangerous for the individual encouraged to sit in front of a screen entertaining and interacting with anonymous viewers for more than 40 hours per week, but dangerous for the impressionable viewers who look up to the streamer they are watching or hardcore fans who feel like they should take in all the content they possibly can. Of course Twitch already values watchtime and other engagement metrics, but excessive consumption of digital media is surely unhealthy for users, especially when the media is of such low production value and the content is shallow, which is generally the case for most streams.

There have already been instances of Twitch being criticized for not caring enough about the people on its site, and having too much of a focus on metrics and revenue. Ludwig, the sixth highest paid streamer on Twitch recently had their contract end and subsequently signed to stream exclusively on YouTube Live. According to him, a main reason that he made the switch was because Twitch failed to fulfill his request to limit the number of gifted subscriptions that a user can give on his channel. He felt uncomfortable receiving such high amounts of money, but Twitch wouldn't consider

doing anything that may limit their revenue even if the alternative is losing one of the top streamers on the site. If the attributes that Twitch values are problematic or immoral, then the extra backlash against Twitch could accelerate the rise of the competing live streaming sites, especially with growing criticism and animosity towards the Amazon owned platform.

The classifier that can be created using the data from the leak along with discovered association rules from the data set may encourage harmful behavior, but these behaviors have been long observed in the online space. Companies have been concerned about driving up user interaction and maximizing revenue since long before Twitch or the recent data leak. Even without the classifier, many streamers already take into consideration many of the attributes previously discussed in order to maximize views and subscriber count. Besides in the case where a certain audience is desirable to Twitch, the changes that a streamer can make to impress Twitch will likely also impress the viewers since Twitch's priorities lie in generating revenue and user watchtime. So if none of the attributes desired by Twitch are deemed problematic, then the harm from mining this data could be negligible.

Conclusion

Overall, the discoveries made from mining this data should have little impact on the viewers of Twitch. Streamers could use these findings to gain a better idea of what metrics they need to reach in order to get the best possible revenue deal when signing with Twitch. If too many streamers hyper focus on the attributes identified by association rules, then Twitch will eventually learn to value these streamers less since they've become more common. Twitch will learn to identify the new trends that the audience likes and the discovered association rules will eventually become untrue.

Whether or not the findings from the data are harmful will largely depend on what channel statistics Twitch uses to determine a streamer's worth if they indeed consider quantifiable metrics when evaluating a streamer. It is possible that they instead collect qualitative feedback on the streamer when signing, and any association rules discovered are just consequential of what is popular on the site at the time.

If the discoveries from mining the leak persuade streamers to chase after a more susceptible and impressionable audience, then it could be said that mining the leak will introduce harm. However discussions have sparked about streams having predatory tendencies before the leak. Any harm that the classifier or association rules cause is not unique. The data mining would simply be revealing immoral tendencies that live streams on Twitch already have. A huge topic on the site over quarantine lockdown for COVID-19 was the morality or lack thereof behind developing these one-sided "parasocial" relationships between streamer and audience members. Some of the more impressionable viewers feel a personal connection to their favorite streamers, and over quarantine it became apparent that these parasocial relationships were a driving force behind subscriptions and donations.

Live streaming naturally encourages competition between streamers and reviewing stream analytics is commonplace for streamers. This can cause some streamers to cross a line and be overaggressive in their attempts to build viewership, but this does not mean that all streamers are predatory. Some creators genuinely want to entertain and try to produce the best content possible. These streamers are the ones who prevent platforms such as Twitch from being completely immoral. They are able to make a living on the platform while being transparent about their relation to the viewer.

So long as it is not encouraged to abuse their viewers, I believe that streamers should be provided with any tools available that can help them grow and effectively spread their content to new audiences. Tools that accomplish this can and have been developed using data mining. One such tool is a clustering of popular Twitch streamers based off of their shared viewerbase. A version of this has been created by a Twitch viewer, but only includes the top 100 streamers at the time. If Twitch were to cluster users themselves, and allow a stream to see the other streamers in their cluster, then they can make improved decision making about collaborations with other streamers that their audience would enjoy. Twitch already provides every streamer with their basic stream statistics including audience demographics, so it would not be too much more difficult to implement KDD techniques to provide much more helpful information to their content creators.