Intent-based Satisfaction Modeling: From Music to Video Streaming

Gabriel Bénédict
RTL & University of Amsterdam
Amsterdam, The Netherlands
g.benedict@uva.nl

Rishabh Mehrotra
Sharechat
London, United Kingdom
rish@sharechat.co

Daan Odijk
RTL
Amsterdam, The Netherlands
daan.odijk@rtl.nl

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
m.derijke@uva.nl

ABSTRACT

Logged behavioral data is a common resource to enhance user experience on streaming platforms. In music streaming, Mehrotra et al. 2019 have shown how complementing behavioral data with user intent can help predict and explain user satisfaction. We look at whether their findings extend to video streaming. Compared to music streaming, video streaming platforms provide relatively shallow catalogs. Finding the right content demands more active and conscious commitment from users. Video streaming platforms in particular could thus benefit from a better understanding of user intents and satisfaction level. We reproduce Mehrotra et al.'s study from music to video streaming and extend their modeling framework on three fronts: adding user characteristics (demographics), improved modeling accuracy (random forests) and interpretability (Bayesian models). Like the original study, we find that user intent affects behavioral data and satisfaction itself, based on data analysis and modeling. By proposing a grouping of intents into decisive and explorative categories we highlight a tension: decisive video streamers are not as keen to interact with the user interface as exploration-seeking ones. Meanwhile, music streamers explore by listening. To enable reproducibility, we provide code for the retrieval of behavioral data on a popular analytics platform and the implementation of an intent-based satisfaction model.

CCS CONCEPTS

• **Information systems** → *Personalization.*

KEYWORDS

Interaction signals; User intents; User satisfaction

ACM Reference Format:
Gabriel Bénédict, Rishabh Mehrotra, Daan Odijk, and Maarten de Rijke. 2022. Intent-based Satisfaction Modeling: From Music to Video Streaming. In *ACM SIGIR 2022, July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Personalized content and experiences on music, video and other types of content platforms, rely on user data as feedback [38]. Such input often has the form of interaction data on a website or from a dedicated app and is then used as implicit feedback from the user [49]. For paid-subscription platforms whose longer term goal is retention, this type of implicit feedback might not be enough [19]. In the short term, retention propensity translates to some form of satisfaction that is highly subjective, time-varying and might form a signal hidden in the implicit feedback data. The literature lists two possible ways to approximate a measure of short-term satisfaction [4]: (i) seek explicit feedback via surveys (e.g., in-person, in-app, in-email), or (ii) obtain implicit feedback from user behavior on the website or app (e.g., content consumption, time on site, time on homepage, etc.).

The importance of intent. Implicit and explicit feedback each come with their own strengths and weaknesses [18, 29]. For explicit feedback most weaknesses can be avoided through careful survey design and for implicit feedback through granular user tracking. However, we identify one irreducible weakness: *missing context* from behavioral data. For example, someone might watch a few trailers during a session and never play a full movie/episode. This could be interpreted as an unsuccessful session. It could also be that the user did not have time to watch the full content and instead was selecting content for a family watching session in the evening.

One way to retrieve the context is to explicitly ask users about their current intents. One can then bring context back to implicit behavioral data by joining behavioral and survey data for each session. Mehrotra et al. [47] use a survey to retrieve users' current intent and satisfaction level, before collecting said user's interaction signals on a music streaming platform. They then show that satisfaction models are more accurate when intent is included as a variable. Their methodology revolves around an in-app user survey and behavioral data on the streaming platform. With visualizations and logistic regression modeling they show that intent together with behavioral data is more predictive of satisfaction than behavioral data alone.

From music to video streaming. We are interested in generalizing the lessons in [47] from music to video streaming. There are important contextual differences between the two types of streaming that make this generalization far from obvious. See Table 1 (top) for

a summary of key differences. First, content length is an obvious difference that has important behavioral consequences [35]. Second, the music streaming domain has settled around half a dozen actors that each provide about the same deep catalog of music. But the opposite is happening in video streaming, where a plethora or platforms each have a few thousand movies and series available at any given time, with little to no content overlap between platforms [32]. Third, the relative scarcity of content and plurality of paid subscription services encourage a strong return to piracy in 2020–2021 [21]. This rise in fragmentation and piracy encourages video streaming actors to (i) quickly and accurately guide *decisive* users to the content they had in mind within a shallow catalog (relative to music), and (ii) provide a customized and seamless user experience for its *explorative* users looking for inspiration (via recommendations, personalized newsletters, etc.), in contrast with its illegal video streaming counterpart. To mirror this situation, we formulate the assumption that there exists two groups of intents, namely decisive and explorative, and show the essential role they play in video streaming platforms.

We follow Mehrotra et al. [47]'s methodology and adapt it for video streaming, in order to assess whether intent can indeed bring context back to explicit feedback. More precisely, we adapt the original study to Videoland, a video streaming platform in the Netherlands with a little over 1 million users. Two key differences in our experimental setup are that we use a browser (instead of mobile) and have multiple intents per session (instead of only one); see Table 1 (bottom).

This reproducibility study follows the ACM definition (different team, different experimental setup) [23] and is inspired by the early initiatives at ECIR [31].[1] Previous IR endeavors have focused on benchmarking on pre-existing data [3, 14–16, 34, 60]; this paper is an attempt at reproducing and generalizing a bigger portion of the experimentation pipeline: we cover data collection, survey design, data preprocessing, data enrichment, modeling, and interpretation.

Insights. In this reproducibility study of [47], we find that for the most part, the conclusions drawn for the music streaming domain also hold in the video streaming domain, both on the data analysis and modeling front. In particular, our contributions in terms of *generalization* are:

(i) a proposal of typical intents for a video streaming platform;
(ii) the in-app survey design for a medium size streaming platform (~1 million users), which involves some small sample adjustments; and
(iii) in addition to Mehrotra et al.'s frequentist logistic regression model, we test Bayesian multilevel models for visualisation and explanations, along with random forests for improved accuracy.

In addition, our *technical* contributions to support reproducibility of work on intent-based satisfaction modeling are: (i) a detailed implementation of the in-app survey design; (ii) code for behavioral data retrieval from Google Analytics using BigQuery; and (iii) code for satisfaction modeling, all of which are made available at https://github.com/gabriben/streaming-intent-model.

[1] We adopt a lenient interpretation of the ACM definition: Rishabh Mehrotra was involved with the paper that we are reproducing and is one the authors for this paper. He helped in re-contextualizing the original study at Spotify and reaching some of our overarching conclusions, over conference calls, and as a proof reader.

Table 1: Contrasting music streaming and video streaming (top), and key differences in experimental setup (bottom).

	Music [47]	Video [this paper]
Content length	3–5 min	45 min–2 hrs
Catalog size[2]	> 70 million	> 5 thousand
Piracy[3]	1 pm	7.5 pm
Platform	Mobile	Browser
Intent	One per session	Multiple per session

2 RELATED WORK

With highly granular logged data and through explicit user feedback, platforms are able to gather implicit and explicit feedback, respectively. In-app surveys (Section 2.2) are only as granular as the number of questions asked to the user but are valuable to retrieve hidden signals that are unavailable in logged data (Section 2.1). Even more powerful is the combination of explicit and implicit aspects to complement each other (Section 2.3), in our case to assign intent and satisfaction levels to raw behavioral data.

2.1 Implicit feedback

In the context of interactive platforms, logged data (time on page, number of pages seen, etc.) has caught the attention of IR researchers early on [49]. Recently, the use of implicit feedback such as click through rate (CTR) [33, 44, 45] or dwell time [36, 61] has been questioned, in favor of the concurrent use of other behavioral metrics [18, 29, 48]. Wen et al. [59] highlight that, in the music domain, many users click a song but consume only a fraction of it, before skipping to the next. In the same domain, implicit feedback signals have been classified into four categories [47]: temporal (e.g., session length, seconds played), downstream (e.g., number of content played), surface level (e.g., number of slates that were interacted with), and derivative (e.g., total clicks / number of content played). Derivative signals are formed with combinations of the other three different signals.

Implicit feedback signals are often used as input or for the evaluation of a search or recommendation model. For example, comparing recommendation predictions with what users actually watched on different metrics and directly relating these metrics to satisfaction levels [58].

2.2 Explicit feedback

In the case of explicit feedback, the services of a representative sample of a user population are enlisted to obtain information on a task, such as recommendation accuracy [4]. A survey can help reveal behavioral traits that are not apparent in the logged data. Arguably, there are two higher order categories of behaviors on streaming platforms: *explorative* versus *decisive* (similar to *fetch*, *find* and *explore* in the domain of search for video streaming [40]). Decisive behavior refers to a session where the user already knows what she wants to stream and is typically addressed in search [40].

[2] Similar to the average for the EU competition in the video domain [28] and international competition in the music domain [1, 2, 17, 37].
[3] Average number of accesses to pirate sites per month and per internet user in the EU+UK in 2017–2020 for the respective video and music domains [26].

Exploration can be defined as the experience of finding and consuming content that is previously unknown to the user [25]. In the music streaming domain, surveys have shown that exploration is a complex time-varying personal need [41], nurtures user retention [7] and deeper social connection [42].

A major drawback of surveys is their inherent *response bias*: satisfaction surveys response rate is low because users have to deviate from their intent of consuming content in order to provide feedback (our response rate was 3%, 4.5% in [47], 4.6% at Spotify over emails [25], 2% at Google for individual item surveys [13]).

The willingness to participate is dependent on hidden factors such as time-on-hand, satisfaction with the platform in the first place (see the satisfaction distribution in Figure 2 and in [13, 25, 47]), etc. As a result, the dataset has missing-not-at-random (MNAR) data [53]. If data is available on who was shown the survey but did not respond, MNAR can be corrected for with inverse propensity scoring or multi-task neural networks [13].

Recently, a new type of item-satisfaction survey emerged, e.g., on YouTube with a Likert scale item recommendation satisfaction survey [63]. Also notable is the trend of the *not interested* button on a recommendation item, which is well entrenched in the IR domain [11], on platforms such as YouTube [62], Twitch [56], and TikTok [55], with all three claiming it will help future recommendations. Such item-surveys suffer even more from response bias and thus motivated a new research field of sparse user-item pairs and debiasing [13].

A fruitful way to address the two major drawbacks of explicit feedback, response bias and sparsity, is to complement a user survey with that user's logged interaction data, as we discuss next.

2.3 Merging implicit and explicit feedback

Merging implicit and explicit feedback is a field that draws learnings from explicit feedback over the long term on short-term implicit feedback signals [24]. One way of drawing the link between implicit and explicit feedback is via the users' current intent.

User intent as a feature for satisfaction prediction has received attention in product search [19] and for recommendation [5], as well as in recent workshops [6, 46] and more particularly in the domain of conversational IR [46], search in video streaming platforms [40], search on online shopping platforms [54], point-of-interest recommendation on maps [50], reinforcement learning to discover unobserved intents for car GPS trajectories [52], and advertiser satisfaction prediction [54]. Identifying intents in the first place can be a mix of supervised and unsupervised tasks that can involve users directly via interviews [47] or research teams internally. Lin et al. [43] propose to discover new intents based on preexisting human-identified intents.

In the domain of entertainment, a seminal study at Pinterest found that not only intent was related to satisfaction, but that, using a simple logistic regression classifier, intent can be predicted quickly during a session [12]. On music streaming platforms, a study linked satisfaction with intent via a user survey and behavioral data on a music platform [47].

To the best of our knowledge, no study of the effect of intent on satisfaction has been published yet for the video streaming domain. In this work we consider both implicit and explicit feedback to reproduce and generalize [47] from music streaming to video streaming. We generalize to the video domain by proposing video-specific intents and detailed implementation of the survey design. We reproduce models with binarized satisfaction levels as outputs, behavioral data and optionally intent as input, thus testing whether intent can help better predict satisfaction levels. We use (hierarchical) logistic regression as in the original study and further look at random forest models to optimize for accuracy and Bayesian models for interpretability.

3 REPRODUCTION SETUP FOR VIDEO STREAMING

Our aim is to verify if on a video streaming platform, like in the music streaming domain, behavioral data coupled with intent predicts satisfaction more accurately than behavioral data alone. To this end, we reproduce the methodology of [47] and adapt it to video streaming. We compare and contrast two specific music and video streaming settings, before explaining our reproducibility design choices. We then describe our available data, acquired via in-app survey and behavioral data on the platform. Finally, we describe our satisfaction prediction model, with or without intent as input.

3.1 From music streaming to video streaming

For our reproducibility study we contrast a specific music streaming platform, Spotify, which provided the context for [47], and a specific video streaming platform, Videoland. Spotify is one of the largest music streaming platform with 180 million paid subscribers and over 70 million tracks. The most salient differences with Videoland, a Dutch streaming platform with a little over 1 million users, are listed in Table 1. Videoland has a few thousand titles (movies, series, TV programs) with a mix of in-house productions, rotating external content, and live TV (RTL TV channels).

After a two weeks free trial, Videoland requires users to subscribe to one of three tiers. Both Spotify and Videoland require users to log in to use their platform on smart TVs, smartphones or computer browsers (and other devices for Spotify such as smart speakers). This guarantees access to identifiable behavioral data.

At Videoland, behavioral data varies greatly between device types (smart TVs, smartphones or computer browsers). Like in the reproduced paper [47], we focus on a single device type so as to reduce noise. Behavioral and survey data engineering is most mature on desktop web, which naturally focuses the scope of this study on desktop browser-based streaming platforms (10% of Videoland sessions), instead of TV or smartphone (as in [47]). We conduct in-app surveys with Usabilla and retrieve behavioral data via Google Analytics and BigQuery.

To manage both survey and behavioral data privacy, Videoland displays consent banners, uses a consent management system, and user preferences to allow individual user tracking limits, in accordance with GDPR regulations [20].

Like in [47], the homepage is the focus of our analysis. As detailed in [51], at Spotify, each strip is either personalized or editorial and the order of strips is purely personalized for each session, at the time of the study reproduced here. For Videoland, the homepage is where most people land on (71% of users, during the survey period) and it is where the platform puts most effort on guiding the user to their desired content. It is populated with personalized [30] and

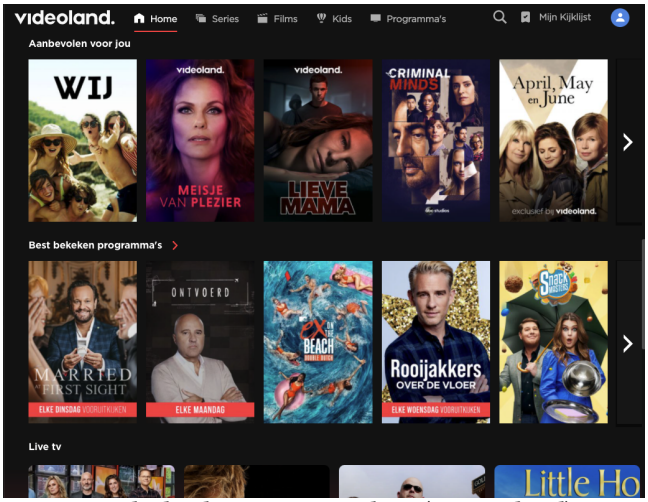


Figure 1: Videoland Homepage with its (personalized) strips.

editorial content. The homepage provides direct access to a search bar and a genre catalog at the top, a "continue watching" slate, a few live TV slates, a mix of editorial and personalized slates (see Figure 1). The homepage layout (i.e., the strip order) is changed daily by human editors, aided with slate popularity models (debiased for position bias).

3.2 Survey and experimental design

Mehrotra et al. [47] perform intent surveys in two stages: (i) intent identification, and (ii) a large-scale in-app survey. The first stage is intended as a way to discover intents of users. Mehrotra et al. [47] held in-depth one-on-one interviews with twelve users on-site. To discover intents on Videoland, we collaborate with our user experience specialists, who have conducted numerous in-app, email, on-phone, and on-site interviews and surveys on topics surrounding intent. With them, we identified eight intents in two groups, described in the next section. In our in-app survey, we allow users to specify other intents that we might have missed in an "others" field.

The second step, the in-app survey, is the core of [47] and of our reproducibility study. The major choice here is where and when to show the survey to the user. While reproducing the work on a different platform, we need to reconsider this choice below.

When opening the Spotify mobile app, the user does not always land on the homepage. Thus, the reason for presence on the homepage can be deliberate or not, this forced [47] to add an intent "Homepage is the first screen shown (i.e., default screen)". On the Videoland web app, most users land on the homepage (72% of users, during the survey period). Another fraction lands on the page of a content item. At Spotify, users switch back and forth between pages and tend to see the homepage in the middle of the session. On Videoland, most users start with the homepage, select and watch content, before closing the web app. This difference is strongly linked to the content type: listening to music can result in a lengthy session with dozens of music plays, whereas video streaming sessions tend to be dedicated to one movie or one series (thus little interest in returning to the homepage in the middle of a session).

Mehrotra et al. [47] show the in-app survey whenever a user comes back to the homepage from another page. While it is desirable to survey users in the middle of a session in order to measure their satisfaction, this particular setup is not possible at Videoland. One possibility would have been to show the survey in between series episodes, but this was quickly discarded as highly intrusive by our user experience researchers. We opt for the next best approach: showing the survey after having been on the homepage for seven seconds (the mean survival time of a user on the homepage, whether the user left the platform or clicked on an item). We look at the impact of that choice in Section 7.

Our survey, and thus the study as a whole, was conducted between November 18, 2021 and January 20, 2022. For every user logging in, there was a 20% chance of being surveyed. Each user is shown the survey at most once to avoid pushing the survey several times to the same user (in line with [47]).

3.3 Data collection

Next, we show the variables gathered at the session-level from two sources, namely interactions on the platform and an in-app survey.

3.3.1 Behavioral variables. Behavioral variables are obtained on the website at the session level (see Table 2) and can be grouped into temporal, downstream, and surface level signals (cf. [47]). They refer to, respectively, time related events, streaming related events, and user interface interaction events.

Our behavioral variables are similar to the reproduced study, with the exception of *derivative signals* [47], which are absent from our study. They are ratio combinations of other signals and therefore would exhibit high collinearity with some other variables in a regression model.

Note that we measure sessionLength as the difference between last and first user feedback. That last user feedback can be any surface level feedback, but we do not receive a feedback when a user closes her Videoland browser tab. Additionally, by default, Google Analytics creates a new session after 30 minutes of inactivity.

The remainder of the implicit feedback signals are exact measures. We complement the behavioral variables with survey data to reveal user satisfaction and intent. Again, see Table 2.

3.3.2 In-app survey variables. During the in-app survey (after seven seconds spent on the homepage), we ask two questions. Namely,

(i) "How happy are you with your experience on the homepage today?" with satisfaction levels of 1 to 5 illustrated by smiley faces (😞 😕 😐 😊 😍). In [47], this question was answered on a numeric Likert scale from 1 to 5. We opted for emojis because our user experience specialists reported better results due to the more intuitive cues. We then ask
(ii) "Why are you using the homepage today?" with eight multiple choice answers (see Table 3).

We divide intents into two main groups: *decisive* and *explorative*. Decisive users tend to arrive on the platform knowing what they want to watch. The exploration-seeking group indicates the opposite: the user is expecting the platform to help them decide what to watch.

Table 2: Behavioral variables obtained from traffic data.

Group	Behavioral metric	Description
Temporal	timeToFirstTrailer	Seconds to the first trailer played
	timeToFirstPlay	Seconds to first content play
	sessionLength	Session length in seconds
Downstream	numTrailerPlays	Number of trailers played
	numPlays	Number of played trailers
Surface level	nStrips	Number of strips seen
	nSearches	Number of content searches
	nSeriesDescr	Number of series description pages
	nMoviesDescr	Number of movies description pages
	nAccounts	Number of clicks on the account icon
	nProfileClicks	Number of clicks on the profile management icon
	nBookmarks	Number of bookmarked items

Mehrotra et al. [47] allow users to choose only one intent. By letting the user choose one or more intents, we can test the hypothesis that any user can have a mixture of intents for the same session. Additionally, we add an "others" field, to let users answer with their own words (as in [47]). Mehrotra et al. [47] analyzed the others field with a Bayesian non-parametric model (dd-CRP), in order to extract salient intents from free text. In the results section we report on the lack of signal in that data in our reproducibility study. We therefore did not algorithmically extract intents from the "others" field.

3.3.3 User-level variables. As is mentioned in [47]'s future work section, different user segments might interact differently with the platform. We extract user-level variables in order to test that hypothesis. These variables include age and gender.

Together with the behavioral and survey data, user metadata forms the input to our satisfaction model.

4 REPRODUCTION OF SATISFACTION MODELS

In this section we describe our reproductions of the original satisfaction models with and without intent [47], before describing our own models and the training setup.

4.1 A satisfaction model

Our satisfaction models are exactly aligned with [47]. We start with the simplest possible satisfaction model and iteratively add complexity.

Each session on Videoland is linked to its corresponding survey data and a satisfaction level $y \in \{1, 2, 3, 4, 5\}$, in increasing order of satisfaction. Following [47], we construct dichotomized satisfaction level vectors over all surveyed session

$$\mathbf{y}_{overall} = \mathbb{1}_{\hat{y} \geq 4}, \quad \mathbf{y}_{satisfied} = \mathbb{1}_{\hat{y}=5}, \quad \mathbf{y}_{dissatisfied} = \mathbb{1}_{\hat{y}=1}, \quad (1)$$

with $\mathbb{1}_{(\cdot)}$ an indicator function, allowing for the use of binary satisfaction prediction models and to focus on different user groups.

A logistic regression model [w/o intent].[4] The most trivial regression model can estimate satisfaction levels $\mathbf{y}$ via a logit link

$$logit(\mathbf{y}) = ln\left(\frac{\mathbf{y}}{1-\mathbf{y}}\right) = \beta_0 + \sum_j \beta_j \mathbf{b}_j, \quad (2)$$

[4] In brackets we include the labels that we use to refer to these in Table 4.

Table 3: Possible intents to be selected by survey respondents.

Group	Intent	Description
Explorative	new	I am looking for something new to watch
	genre	I am looking for a genre (e.g., action, comedy)
	watchlist	I want to look at my watchlist
	addwatchlist	I want to add something to my watchlist
Decisive	continuewatching	I want to continue watching a series/film where I left off
	livetv	I want to watch live TV
	catch-up	I want to catch-up on an episode I missed
	specifictitle	I am looking for a specific title

with β_0 the intercept and $\{\mathbf{b}_1, \ldots, \mathbf{b}_i, \ldots, \mathbf{b}_J\}$ the behavioral variables.

Adding intent [w intent]. The model that we have just described does not include context: a user might be interested in adding elements to their watchlist for a later viewing session, but does not have time to watch content. In that case, a low number of minutes seen and a low number of video plays need not be bad indicators. As a next iteration, context and thus intents can be added as parameters,

$$logit(\mathbf{y}) = \beta_0 + \sum_j \beta_j \mathbf{b}_j + \sum_k \delta_k \mathbf{d}_k, \quad (3)$$

with $\{\mathbf{d}_1, \ldots, \mathbf{d}_k, \ldots, \mathbf{d}_K\}$ intents.

One regression per intent [catch-up, ...]. Alternatively, one could consider fitting one model per intent d, reverting back to Eq. 2:

$$logit(\mathbf{y}^d) = \beta_0^d + \sum_j \beta_j^d \mathbf{b}_j^d. \quad (4)$$

This formulation is insightful to assess satisfaction levels of different session groups but ignores possible interaction effects between intents. It is also problematic in our small sample setting: some intents are only represented by a few hundred datapoints. This formulation does not measure the relative effect of a certain intent over another.

A global intent model [multiLevel]. We revert back to a single frequentist multilevel model [39], that measures the effect of each intent as a group level effect, with a random intercept δ_k:

$$\begin{aligned} logit(\mathbf{y}) &= \delta_k + \sum_j \beta_j \mathbf{b}_j \\ \delta_k &\sim N(\mu_\delta, \sigma_\delta^2). \end{aligned} \quad (5)$$

This time, we clearly model a hierarchical structure in the data and can assess group-level (intent-level) marginal satisfaction effects.

4.2 Further satisfaction models

To achieve higher accuracy, we use XGBoost, a common implementation of gradient boosting decision trees [9], with a logistic regression objective. XGBoost is a strong performer on tabular data, even against recent transformer models adapted to tabular data [27].

For model interpretability, we opt for Bayesian satisfaction models with the same specifications as the frequentist versions above. They allow for the estimation of entire marginal posterior distributions and thus more granular interpretability.

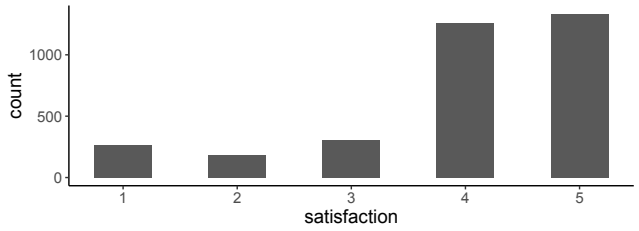


Figure 2: Our imbalanced dataset: distribution of Likert-scale satisfaction levels for all surveyed users and across intents.

4.3 Training, evaluation and hyperparameter tuning

We recall the available data: behavioral data, user metadata, and survey data (intent and satisfaction level). The original study [47] does not compute uncertainty intervals and we did not have access to their training regime, we thus opted for our own. The data is split into training and test sets in $k = 5$ folds, in order to provide out-of-sample estimates [57] and confidence intervals. The intent-specific models are trained on subsets of the data that contain each specific intent and have thus each their specific 5-fold split.

For XGBoost we split each training set into a training and a validation set (with an 80/20% ratio) to tune the hyperparameters: *mphmax_depth* $[3; 10]$, *min_child_weight* $[1; 10]$, *subsample* $[0.5; 1]$, and *colsample_bytree* $[0.5; 1]$ (see documentation [10]). Regarding Bayesian models, we checked for chain convergence in two ways: (i) visually with chain plots, and (ii) quantitatively with Rhat.[5]

We evaluate on the same metrics as in [47]: accuracy, precision, recall, and F1 score. To calculate these confusion matrix related metrics, predictions in the $[0; 1]$ range have to be dichotomized at a certain threshold. Given the imbalance in the data (see Figure 2), we refrain from using a heuristic 0.5 threshold, and instead use a threshold-moving technique at inference time, based on the F1 score, to balance precision and recall for each model and at each Likert-Scale dichotomization (*Overall*, *Satisfied* and *Unsatisfied*) [22, p. 53–55]. This is an inference-time task and we distinguish it from hyperparameter tuning to be done on validation sets.

5 DATA ANALYSIS REPRODUCTION

In this section we reproduce the data analysis and visualizations from [47] and assess whether the original conclusions generalize from the music to the video domain. We produce three plots in line with [47], two of which are focused on survey results. The last plot mixes behavioral and survey data. For comparison purposes, the visualizations are kept similar to the original study.

5.1 Survey results

The response rate was 3%, with a survey rate of 20% from logged-in users after 7 seconds on the home page, we ended up with about 3, 350 sessions. 21% of these users responded to the first (satisfaction) but not to the second (intent) question and are thus not modelled in Section 6, leaving a total of 2, 632 survey responses in our datasets. The most selected intents were continuewatching, and only 3.6% users added a remark in the "other" section. We thus decided to read them all. They were for a minor part bug reports, enunciating an existing intent in the list, some grateful or ungrateful comments, or asking for content to appear on Videoland. Given the lack of signal on intent in the "others" section we decided to leave it out of this study.

[5] https://github.com/gabriben/streaming-intent-model

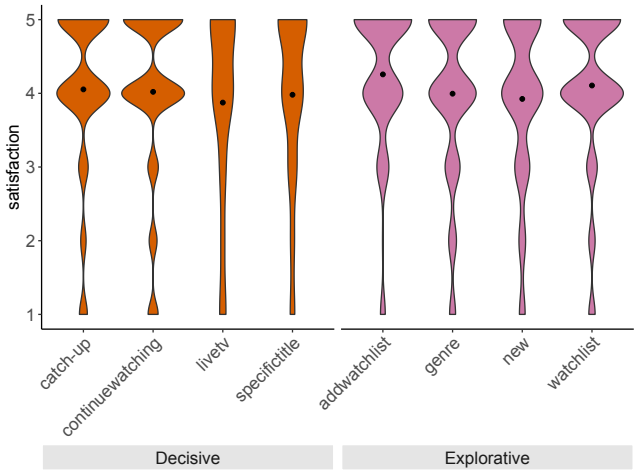


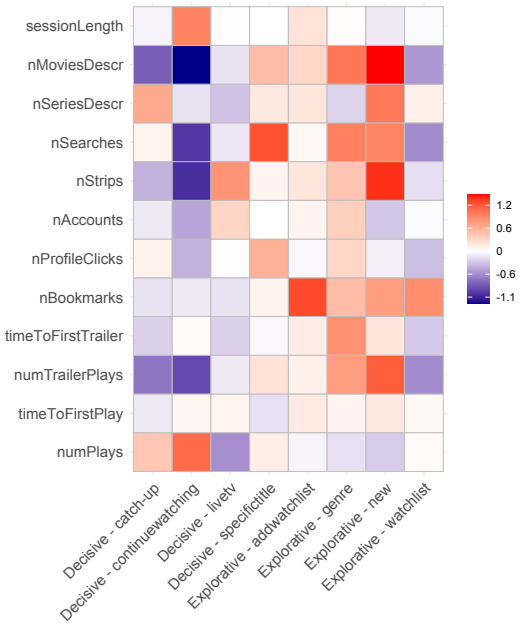
Figure 3: Satisfaction levels per intent and by intent group (dot indicates the mean).

Figure 2 displays the satisfaction levels across all sessions and reveals that most users who answered the survey are satisfied with the platform. This is in line with [47]'s setup, which let users rate their satisfaction with numbers from 1 to 5 instead of emojis in our case. Also note that quite satisfied users ($y \geq 4$) are overrepresented compared to their less satisfied neighbors ($y < 4$). This might be a sign of MNAR in our dataset (see Section 7 for a discussion on the topic).

Next, we look at relationships between satisfaction level and intent (Figure 3). We draw a violin plot as in [47]. From left to right, we notice that decisive users looking for live TV or a specific title have the most spread out satisfaction distribution; users who add content to their watchlists have the lowest representation of satisfaction levels 1 and 2; users who are looking for inspiration via new genres or new titles are the least satisfied (i.e., they have the highest concentration of levels 1, 2 and 3). Following our earlier discussions of rising fragmentation and piracy in the video streaming domain, it might be necessary to look closely at these unsatisfied decisive users and in particular those looking for a specific title, for which piracy or an alternative platform is the most natural substitute. In the following section we further investigate these intents in relation with the interaction data.

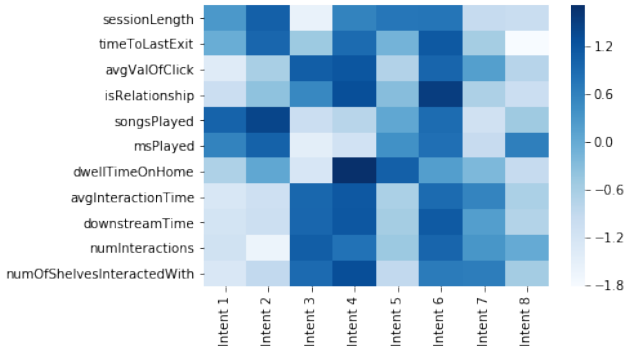
5.2 Correlation between survey and behavioral data

We recontextualize the raw behavioral data with users' revealed intents. The Pearson correlation plot on Figure 4a confirms a few intuitions. Users who intend to continue watching an episode interact the least with the platform, but it does not prevent them from watching a lot of content for long periods of time. Users who are looking for something new to watch interact with a number of



(a) On our video streaming platform. Red and blue respectively indicate positive and negative correlation.

Intent	Definition
Intent 1	Homepage is the first screen shown (i.e. default screen)
Intent 2	To quickly access my playlists or saved music
Intent 3	To discover new music to listen to now
Intent 4	To play music that matches my mood or activity
Intent 5	To Find X
Intent 6	To find music to play in the background
Intent 7	To save new music or follow new playlists to listen to later
Intent 8	To explore artists or albums more deeply.



(b) On a music streaming platform (table and visualization taken from [47]).

Figure 4: Pearson Correlation (×10) plots between intents (x-axis) and behavioral data (y-axis).

features on the platform and watch a lot of trailers. They do not tend to find more content to watch than other users (as indicated by the lack of correlation with numPlays and sessionLength). For comparison, in music streaming, at Spotify, users even tend to play less songs for less time (negative correlations) when they desire "to discover new music to listen to now" (see Figure 4b).

We note one salient difference with the original interaction plot at Spotify: users whose intent is "to explore artists or albums more deeply" comparatively play songs for a longer time and do not have a particularly high number of interactions with the user interface. In other words, *in the music domain, users explore by playing. In the video domain, users explore by interacting with the platform.* The main reason is probably that a song listener can afford to listen and try out full 10–15 songs while a user watches a single movie or series episode.

Taking a step back, these disparities highlight the differences between the blind exploration phase in the music domain (low interaction) and the more tedious, active exploration phase in the video domain. Thus, it seems that the video medium itself calls for exploratory user hand-holding. We emphasize the need to provide a thoroughly thought out and personalized user experience to a video streamer looking for inspiration, otherwise the video platform risks loosing the customer to piracy or another video streaming platform.

5.3 Upshot: Music versus video streaming

In reproducing [47], we collected data in a completely different streaming platform and we adapted the survey design to our needs (main differences in Table 1). We found [47]'s data analysis to be reproducible in several aspects. We observe the same imbalance in satisfaction levels, with levels 4 and 5 overly represented. Satisfaction by intent is less comparable, since we formulated video streaming intents. Unlike in [47], we find that two intents clearly have a higher amount of dissatisfied users, namely the decisive users looking to watch *livetv* or a *specifictitle*. Overall, Figure 3 and 4a confirm the learnings from [47], namely that users' satisfaction level and behavior is different depending on their intent.

Like in the original study, our conclusions might be influenced by response bias. For example, we were used to observe little use of the bookmarking system on the platform. But our survey-behavioral dataset showed an unusually high amount of users adding elements to their watchlist. We assume users that use the watchlist are more likely to respond to the survey or maybe even that some users discovered the existence of the watchlist button after seeing it as an intent option in the survey: the average of 0.03 bookmarks per session for all sessions during the survey period jumps to 0.09 for our surveyed cohort who made it to the second question and saw the bookmarking intents.

6 MODEL REPRODUCTION

Next, we reproduce multiple frequentist logistic regression satisfaction models: without intents, with intents, per intent and with an intent as a hierarchical level, all as in [47]. Additionally, going beyond [47], we report on XGBoost predictions with and without intents; we then fit one Bayesian logistic regression per intent and report on marginal posterior distributions for each behavioral metric.

Table 4: Reproduction of [47] with added mean and standard deviation over 5-fold cross-validation for the three dichotomizations of the $y \in \{1, 2, 3, 4, 5\}$ satisfaction score and four metrics (accuracy, precision, recall, F1 score).

	Overall ($\mathbb{1}_{\hat{y}\geq 4}$)				Satisfied ($\mathbb{1}_{\hat{y}=5}$)				Unsatisfied ($\mathbb{1}_{\hat{y}=1}$)			
Method	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
w/o intent	0.81 ± 0.03	0.81 ± 0.03	1.00 ± 0.00	0.89 ± 0.02	0.46 ± 0.04	0.43 ± 0.04	0.95 ± 0.03	0.59 ± 0.04	0.87 ± 0.02	0.16 ± 0.08	0.26 ± 0.13	0.19 ± 0.10
w intent	0.81 ± 0.03	0.81 ± 0.03	1.00 ± 0.00	0.89 ± 0.02	0.47 ± 0.04	0.43 ± 0.04	0.94 ± 0.03	0.59 ± 0.04	0.92 ± 0.03	0.28 ± 0.17	0.20 ± 0.15	0.21 ± 0.12
catch-up	0.82 ± 0.04	0.82 ± 0.04	1.00 ± 0.01	0.90 ± 0.03	0.41 ± 0.07	0.40 ± 0.07	0.97 ± 0.04	0.56 ± 0.08	0.83 ± 0.05	0.10 ± 0.11	0.24 ± 0.25	0.13 ± 0.12
continuewatching	0.81 ± 0.03	0.81 ± 0.03	1.00 ± 0.01	0.89 ± 0.02	0.41 ± 0.04	0.40 ± 0.04	0.97 ± 0.02	0.56 ± 0.04	0.92 ± 0.03	0.42 ± 0.31	0.19 ± 0.15	0.26 ± 0.20
livetv	0.79 ± 0.16	0.80 ± 0.14	0.97 ± 0.08	0.87 ± 0.10	0.45 ± 0.17	0.44 ± 0.18	0.94 ± 0.12	0.58 ± 0.18	0.23 ± 0.11	0.11 ± 0.12	0.50 ± 0.53	0.18 ± 0.20
specifictitle	0.87 ± 0.22	0.87 ± 0.22	1.00 ± 0.00	0.91 ± 0.14	0.60 ± 0.22	0.60 ± 0.22	1.00 ± 0.00	0.73 ± 0.16	0.22 ± 0.33	0.03 ± 0.11	0.10 ± 0.32	0.05 ± 0.16
addwatchlist	0.95 ± 0.16	1.00 ± 0.00	0.95 ± 0.16	0.97 ± 0.11	0.55 ± 0.50	0.55 ± 0.50	0.60 ± 0.52	0.57 ± 0.50	0.70 ± 0.35	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
genre	0.83 ± 0.32	0.83 ± 0.32	0.90 ± 0.32	0.86 ± 0.31	0.42 ± 0.29	0.38 ± 0.31	0.70 ± 0.48	0.48 ± 0.36	0.33 ± 0.38	0.15 ± 0.34	0.20 ± 0.42	0.17 ± 0.36
new	0.74 ± 0.07	0.74 ± 0.07	1.00 ± 0.00	0.85 ± 0.05	0.41 ± 0.04	0.37 ± 0.05	0.96 ± 0.04	0.53 ± 0.05	0.86 ± 0.07	0.14 ± 0.16	0.28 ± 0.32	0.17 ± 0.17
watchlist	0.84 ± 0.08	0.84 ± 0.08	1.00 ± 0.01	0.91 ± 0.05	0.42 ± 0.09	0.40 ± 0.10	0.94 ± 0.06	0.56 ± 0.11	0.75 ± 0.07	0.06 ± 0.05	0.32 ± 0.33	0.10 ± 0.09
multiLevel	0.81 ± 0.03	0.81 ± 0.03	1.00 ± 0.00	0.89 ± 0.02	0.45 ± 0.04	0.42 ± 0.04	0.96 ± 0.02	0.59 ± 0.04	0.87 ± 0.02	0.16 ± 0.08	0.26 ± 0.13	0.19 ± 0.10
XGBoost w/o intent	0.83 ± 0.03	0.83 ± 0.03	0.99 ± 0.01	0.90 ± 0.02	0.63 ± 0.04	0.53 ± 0.04	0.78 ± 0.07	0.63 ± 0.04	0.86 ± 0.03	0.23 ± 0.13	0.38 ± 0.19	0.28 ± 0.15
XGBoost w intent	0.82 ± 0.02	0.83 ± 0.02	0.98 ± 0.01	0.90 ± 0.01	0.57 ± 0.06	0.49 ± 0.05	0.83 ± 0.10	0.61 ± 0.06	0.91 ± 0.03	0.31 ± 0.15	0.40 ± 0.23	0.33 ± 0.16

6.1 Satisfaction prediction results

Table 4 displays the prediction results with standard deviations using 5-fold test sets. The dichotomization of intent plays a predominant role in the results (*Overall*, *Satisfied*, *Unsatisfied*). For the *Overall* and *Satisfied* dichotomizations, the effect of adding intent to the model is not clear: *w/o intent* versus either *w intent* or its random-effects counterpart *multiLevel*. The per-intent models do not deliver satisfying results over the global model. We also find that, contrary to expectations, XGBoost does not always perform best; we believe that this is due to the linearity in the data, which is accurately modeled by logistic regression.

Turning to classifiying *Unsatisfied* users, differences between results are more stark, especially for Accuracy (non-overlapping standard deviations). This implies that dissatisfied users are the ones who deliver the most signals to researchers. Hence, we focus on dissatisfied users. In general, XGBoost does not deliver consistently higher results here either. Notably, continueWatching (when a user decisively continues watching a show she started) is the best performing per-intent model. That is, continueWatching users that are dissatisfied have very recognizable behavior. Finally, for predicting dissatisfied users, adding intents to either the plain logistic model (*w/o intent*) or the XGBoost model (*XGBoost w/o intent*) leads to performance increases. This confirms the important role of intent in user satisfaction across the music and video domains at least for dissatisfied users.

In the following, we analyze intent specific models in more detail, via their Bayesian counterparts.

6.2 Bayesian marginal posteriors

Figure 5 examines the role of implicit feedback in satisfaction prediction, with intent factored out (given one model per intent). This figure displays marginal posterior distributions of each behavioral metric, given each of eight intent models. Note, for example, that one unit increase in the nStrips coefficient corresponds to a one unit increase in log odds ratios for satisfaction. We kept the three variables with the highest absolute median posterior draws[6] (similarly to the frequentist variable importance analysis in [47]).

[6]We withdrew divergent draws (Rhat > 1.05) and confirmed they did not prevent other estimates to converge with chain plots. Distributional outliers shown in the descriptive statistics plots. https://github.com/gabriben/streaming-intent-model.

Given the small-data context (around 3,000 observations), we refrain from interpreting exact odds ratios. Instead, we focus on marginal posterior distributions whose IQR does not overlap with the zero effect line. Overall for decisive intents, the more a user dwells on different pages and interacts with them instead of playing full videos or trailers, the more their satisfaction is hurt: notably nSearches, nProfileClicks and nBookmarks have negative coefficients in three out of four decisive intents (see the top row of Figure 5). The conclusions are more mixed for explorative users. We see that users who were looking for inspiration via genre pages are rather dissatisfied if they have to do searches instead, but are happy to spend time looking at series descriptions.

6.3 Upshot: Music versus video streaming

We fully re-implemented the predictive models used in [47]. We complemented the original study in three ways: (i) We dealt with imbalanced data by tuning inference-time thresholds [22] instead of oversampling the dataset once with SMOTE [8], thus refraining from duplicating datapoints. (ii) We computed uncertainty intervals by computing out-of-sample estimates on a rotation of five-fold different test sets [57]. (iii) We ran XGBoost and Bayesian models, for prediction accuracy and interpretability.

These conservative measures, together with a smaller dataset could be what lead to less noticeable differences across models than in the original study. It is also possible that our study expresses a reality, namely that in the video setting only dissatisfied users see their satisfaction vary with their intent. This speaks to the intuition that users responding with a 1/5 on the satisfaction scale are the ones sending the strongest signal. This is a motivation to future research with a focus on dissatisfied users. Overall, we could reproduce the main finding of [47], namely that at least for unsatisfied users intent seems to impact satisfaction levels.

7 CONCLUSIONS

In this paper we have reproduced and generalized Mehrotra et al. [47]'s work on intent-based satisfaction modeling, from music to video streaming. We have reproduced the full experimental setup, from data collection of behavioral data and enrichment with an in-app survey to the computations. We provide our code for data

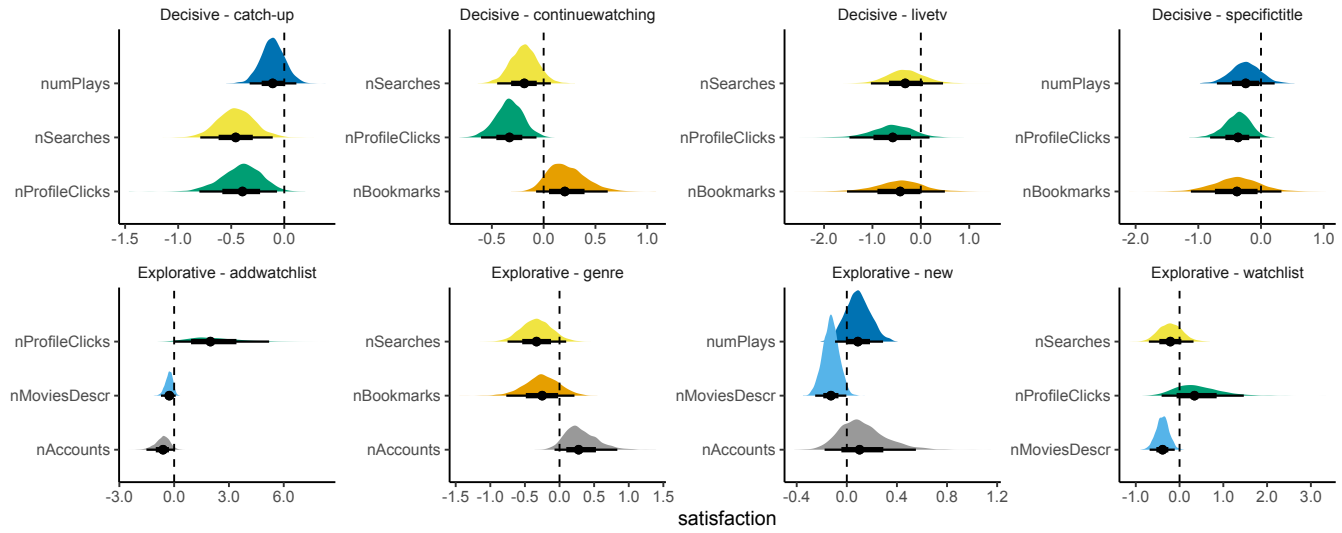


Figure 5: Marginal effect of behavioral variables on satisfaction. One Bayesian fit per intent (median and IQR in thicker marks, 99.8% of the probability density function in thinner line).

preprocessing, visualization of the interactions between intents, satisfaction and behavioral data in line with the illustrations in Mehrotra et al. [47]. Finally, we extended the modeling section with XGBoost models as standard tabular data benchmarks and per intent Bayesian models for interpretability.

Findings. We have found that in video streaming, as in music streaming, intent influences satisfaction levels together with behavioral data. The video context also allowed us to draw new conclusions: (i) Unsatisfied users are more prone to reveal their intent via their behavior on the website (see Table 4). (ii) By introducing a differentiation between explorative and decisive intents, we highlight the tendency of video streamers to use the user interface for inspiration (Figure 4a and 5), whereas music streamers listen "blindly", without much interaction on the interface (Figure 4b), thus highlighting the higher relevance of behavioral data in the video context. (iii) We have also found that decisive users are not so keen on using the platform's personalized features and thus deserve special attention in the user experience design.

Broader impact. More broadly, this study reveals that it is possible to reproduce a survey across different domains, device types and with smaller sample sizes. We hope this real-world small-sample reproducible scenario further encourages human-scale studies in general and in the academic domain, where respondent recruitment is also prone to response bias. With regards to intents, two studies (this paper and its reproduced counterpart [47]) now show that it is not enough to look at behavioral data alone to measure user satisfaction. Surveying and later predicting intents on each streaming platform help better guide users to their goal or give users new perspectives.

Limitations. Our small-scale study also comes with its limitations. We surveyed respondents after seven seconds on the homepage. This means that there is a chance that the survey has influenced certain behaviors. Regarding response bias and MNAR, ideally we would have used the data on users who were shown the survey but did not answer. For future research we propose to track that data.

Looking ahead. As to future work, we hope that this study and the materials that we share encourage researchers working in other domains to investigate and share insights on user intent. We compared the setting of short songs versus long videos and revealed disparities related to the medium itself. This leaves open the effect of intent on platforms focused on longer audio content such as podcasts, short video content like TikTok, or emerging live streaming platforms like Twitch. Understanding intents and their groupings (decisive, explorative, and maybe others) on different platforms could allow for experiences tailored to unobservable time-varying user needs as opposed to relying more on direct user feedback (clicks, scrolls, etc.). Finally, as we pointed out in our discussion of related work, a lot of previous work has highlighted explorative users; decisive users are somewhat neglected in the literature. Our study highlights the need for further research into algorithmically balancing the interests of decisive and explorative users.

IMPLEMENTATION RESOURCES

To support reproducibility of our work, in the video or music streaming domain and beyond, we share[7] the following resources: (i) code for behavioral data retrieval (BigQuery); (ii) for satisfaction modeling; and (iii) a detailed implementation of the in-app survey design.

For GDPR compliance, we cannot share individual user data without receiving explicit consent from these said users with an extensive description of the future usage of that data. However, for others to be able to run our code, we include simulated behavioral and survey data in the repository, reproducing the distributions in our dataset.

[7] https://github.com/gabriben/streaming-intent-model

REFERENCES

[1] Amazon. 2020. Amazon Music has more than 55 million customers worldwide. https://www.aboutamazon.com/news/entertainment/amazon-music-has-more-than-55-million-customers-worldwide.

[2] Apple. 2022. Apple Music. https://www.apple.com/apple-music/. Accessed on 03.02.2022.

[3] Timothy G. Armstrong, Alistair Moffat, William Webber, and Justin Zobel. 2009. Improvements That Don't Add up: Ad-Hoc Retrieval Results since 1998. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. Association for Computing Machinery, 601–610.

[4] Amin Beheshti, Shahpar Yakhchi, Salman Mousaeirad, Seyed Mohssen Ghafari, Srinivasa Reddy Goluguri, and Mohammad Amin Edrisi. 2020. Towards Cognitive Recommender Systems. *Algorithms* 13, 8 (2020). https://doi.org/10.3390/a13080176

[5] Biswarup Bhattacharya, Iftikhar Burhanuddin, Abhilasha Sancheti, and Kushal Satya. 2017. Intent-Aware Contextual Recommendation System. *2017 IEEE International Conference on Data Mining Workshops (ICDMW)* (Nov 2017). https://doi.org/10.1109/icdmw.2017.8

[6] Sahan Bulathwela, María Pérez-Ortiz, Rishabh Mehrotra, Davor Orlic, Colin de la Higuera, John Shawe-Taylor, and Emine Yilmaz. 2020. *SUM'20: State-Based User Modelling*. Association for Computing Machinery, New York, NY, USA, 899–900. https://doi.org/10.1145/3336191.3371883

[7] Steven Caldwell Brown and Amanda Krause. 2016. A psychological approach to understanding the varied functions that different music formats serve. In *Proceedings of the 14th International Conference on Music Perception and Cognition*. 849–851. http://icmpc.org/icmpc14/ 14th Biennial International Conference on Music Perception and Cognition, ICMPC14.

[8] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16 (Jun 2002), 321–357. https://doi.org/10.1613/jair.953

[9] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[10] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, Mu Li, Junyuan Xie, Min Lin, Yifeng Geng, and Yutian Li. 2021. *xgboost: Extreme Gradient Boosting*. https://CRAN.R-project.org/package=xgboost R package version 1.5.0.2.

[11] Zhixiang Chen, Xiannong Meng, Binhai Zhu, and R.H. Fowler. 2000. WebSail: from on-line learning to Web search. In *Proceedings of the First International Conference on Web Information Systems Engineering*, Vol. 1. 206–213 vol.1. https://doi.org/10.1109/WISE.2000.882394

[12] Justin Cheng, Caroline Lo, and Jure Leskovec. 2017. Predicting Intent Using Activity Logs. *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion* (2017). https://doi.org/10.1145/3041021.3054198

[13] Konstantina Christakopoulou, Madeleine Traverse, Trevor Potter, Emma Marriott, Daniel Li, Chris Haulk, Ed H. Chi, and Minmin Chen. 2020. Deconfounding User Satisfaction Estimation from Response Rate Bias. *Fourteenth ACM Conference on Recommender Systems* (Sep 2020). https://doi.org/10.1145/3383313.3412208

[14] Ryan Clancy, Nicola Ferro, Claudia Hauff, Jimmy Lin, Tetsuya Sakai, and Ze Zhong Wu. 2019. The SIGIR 2019 Open-Source IR Replicability Challenge (OSIRRC 2019). In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

[15] Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. 2021. A Troubling Analysis of Reproducibility and Progress in Recommender Systems Research. *ACM Transactions on Information Systems (TOIS)* 39, 2 (2021), 1–49.

[16] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 101–109.

[17] Deezer. 2022. Deezer About Page. https://www.deezer.com/en/company. Accessed on 03.02.2022.

[18] Paolo Dragone, Rishabh Mehrotra, and Mounia Lalmas. 2019. Deriving User- and Content-specific Rewards for Contextual Bandits. *The World Wide Web Conference on - WWW '19* (2019). https://doi.org/10.1145/3308558.3313592

[19] Huizhong Duan and ChengXiang Zhai. 2015. Mining Coordinated Intent Representation for Entity Search and Recommendation. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (Oct 2015). https://doi.org/10.1145/2806416.2806557

[20] European Commission. 2022. General Data Protection Regulation. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

[21] Brian Feldman. 2019. Piracy Is Back. *New York Magazine* (2019). https://nymag.com/intelligencer/2019/06/piracy-is-back.html

[22] Alberto Fernández. 2018. *Learning from Imbalanced Data Sets* (1st ed. 2018. ed.). Springer International Publishing, Cham.

[23] Nicola Ferro and Diane Kelly. 2018. SIGIR Initiative to Implement ACM Artifact Review and Badging. *SIGIR Forum* 52, 1 (jun 2018).

[24] Jean Garcia-Gathright, Christine Hosey, Brian St. Thomas, Ben Carterette, and Fernando Diaz. 2018. Mixed methods for evaluating user satisfaction. *Proceedings of the 12th ACM Conference on Recommender Systems* (Sep 2018). https://doi.org/10.1145/3240323.3241622

[25] Jean Garcia-Gathright, Brian St. Thomas, Christine Hosey, Zahra Nazari, and Fernando Diaz. 2018. Understanding and Evaluating User Satisfaction with Music Discovery. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Jun 2018). https://doi.org/10.1145/3209978.3210049

[26] Francisco Garcia-Valero, Michal Kazimierczak, Carolina Arias-Burgos, and Nathan Wajsman. 2021. Online Copyright Infringement in the European Union. *European Union Intellectual Property Office* (December 2021). https://doi.org/10.2814/505158

[27] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. 2021. Revisiting Deep Learning Models for Tabular Data. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=i_Q1yrOegLY

[28] Christian Grece and Marta Jiménez Pumares. 2020. Film and TV content in VOD catalogues – 2020 edition. *European Audiovisual Observatory* (December 2020). http://diversidadaudiovisual.org/wp-content/uploads/2021/02/Report-Film-and-TV-content-in-VOD-catalogues-2020-Edition.pdf

[29] Qi Guo and Eugene Agichtein. 2012. Beyond dwell time. *Proceedings of the 21st international conference on World Wide Web - WWW '12* (2012). https://doi.org/10.1145/2187836.2187914

[30] Mateo Gutierrez Granada and Daan Odijk. 2021. Recommendations at Videoland. In *Fifteenth ACM Conference on Recommender Systems*. Association for Computing Machinery, New York, NY, USA, 580–582. https://doi-org.proxy.uba.uva.nl/10.1145/3460231.3474617

[31] Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr. 2015. ECIR 2015: 37th European Conference on Information Retrieval. *SIGIR Forum* 49, 2 (dec 2015).

[32] Alex Hern. 2021. Streaming was supposed to stop piracy. Now it is easier than ever. *The Guardian* (2021). https://www.theguardian.com/world/2017/mar/12/netherlands-will-pay-the-price-for-blocking-turkish-visit-erdogan

[33] Jeff Huang, Ryen W. White, Georg Buscher, and Kuansan Wang. 2012. Improving searcher models using mouse cursor activity. *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '12* (2012). https://doi.org/10.1145/2348283.2348313

[34] Dietmar Jannach, Gabriel de Souza P. Moreira, and Even Oldridge. 2020. Why Are Deep Learning Models Not Consistently Winning Recommender Systems Competitions Yet? A Position Paper. In *Proceedings of the Recommender Systems Challenge 2020 (RecSysChallenge '20)*. ACM, 44–49.

[35] Jebara, Tony. 2019. Personalization of Spotify Home and TensorFlow. https://www.oreilly.com/radar/personalization-of-spotify-home-and-tensorflow/. Accessed on 13.02.2022.

[36] Youngho Kim, Ahmed Hassan, Ryen W. White, and Imed Zitouni. 2014. Comparing client and server dwell time estimates for click-level satisfaction prediction. *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval* (Jul 2014). https://doi.org/10.1145/2600428.2609468

[37] KKBox. 2022. KKBox About Page. https://www.kkbox.com/about/en/about. Accessed on 03.02.2022.

[38] Hyeyoung Ko, Suyeon Lee, Yoonseo Park, and Anna Choi. 2022. A Survey of Recommendation Systems: Recommendation Models, Techniques, and Application Fields. *Electronics* 11, 1 (Jan 2022), 141. https://doi.org/10.3390/electronics11010141

[39] Jennifer L. Krull and David P. MacKinnon. 2001. Multilevel Modeling of Individual and Group Level Mediated Effects. *Multivariate Behavioral Research* 36, 2 (Apr 2001), 249–277. https://doi.org/10.1207/s15327906mbr3602_06

[40] Sudarshan Lamkhede and Sudeep Das. 2019. Challenges in Search on Streaming Services. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Jul 2019). https://doi.org/10.1145/3331184.3331440

[41] Jin Ha Lee and Rachel Price. 2015. Understanding Users of Commercial Music Services through Personas: Design Implications. In *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015, Málaga, Spain, October 26-30, 2015*, Meinard Müller and Frans Wiering (Eds.). 476–482. http://ismir2015.uma.es/articles/12_Paper.pdf

[42] Tuck W. Leong and Peter C. Wright. 2013. Revisiting social practices surrounding music. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Apr 2013). https://doi.org/10.1145/2470654.2466122

[43] Ting-En Lin, Hua Xu, and Hanlei Zhang. 2020. Discovering New Intents via Constrained Deep Adaptive Clustering with Cluster Refinement. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 05 (Apr 2020), 8360–8367. https://doi.org/10.1609/aaai.v34i05.6353

[44] Yiqun Liu, Ye Chen, Jinhui Tang, Jiashen Sun, Min Zhang, Shaoping Ma, and Xuan Zhu. 2015. Different Users, Different Opinions. *Proceedings of the 38th*

International ACM SIGIR Conference on Research and Development in Information Retrieval (Aug 2015). https://doi.org/10.1145/2766462.2767721

[45] Rishabh Mehrotra, Ahmed Hassan Awadallah, Milad Shokouhi, Emine Yilmaz, Imed Zitouni, Ahmed El Kholy, and Madian Khabsa. 2017. Deep Sequential Models for Task Satisfaction Prediction. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (Nov 2017). https://doi.org/10.1145/3132847.3133001

[46] Rishabh Mehrotra, Ahmed Hassan Awadallah, and Emine Yilmaz. 2018. LearnIR: WSDM 2018 Workshop on Learning from User Interactions. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (Feb 2018). https://doi.org/10.1145/3159652.3160598

[47] Rishabh Mehrotra, Mounia Lalmas, Doug Kenney, Thomas Lim-Meng, and Golli Hashemian. 2019. Jointly Leveraging Intent and Interaction Signals to Predict User Satisfaction with Slate Recommendations. *The World Wide Web Conference on - WWW '19* (2019). https://doi.org/10.1145/3308558.3313613

[48] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a Fair Marketplace. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (Oct 2018). https://doi.org/10.1145/3269206.3272027

[49] Masahiro Morita and Yoichi Shinoda. 1994. Information Filtering Based on User Behavior Analysis and Best Match Text Retrieval. *SIGIR '94* (1994), 272–281. https://doi.org/10.1007/978-1-4471-2099-5_28

[50] Behrooz Omidvar-Tehrani, Sruthi Viswanathan, Frederic Roulland, and Jean-Michel Renders. 2020. Sage: Interactive state-aware point-of-interest recommendation. *Workshop on StateBased User Modelling (SUM '20)* (2020). https://www.k4all.org/wp-content/uploads/2020/01/WSDMSUM20_paper_SAGE_Interactive_State_aware-Point_of_Interest_Recommendation.pdf

[51] Oguz Semerci, Aloïs Gruson, Clay Gibson, Ben Lacker, Catherine Edwards, and Vladan Radosavlijevic. 2019. Homepage Personalization at Spotify. *RecSys* (September 2019). https://fr.slideshare.net/OguzSemerci/homepage-personalization-at-spotify

[52] Aaron J. Snoswell, Surya P. N. Singh, and Nan Ye. 2021. LiMIIRL: Lightweight Multiple-Intent Inverse Reinforcement Learning. arXiv:cs.LG/2106.01777

[53] Harald Steck. 2010. Training and testing of recommender systems on data missing not at random. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10* (2010). https://doi.org/10.1145/1835804.1835895

[54] Ning Su, Jiyin He, Yiqun Liu, Min Zhang, and Shaoping Ma. 2018. User Intent, Behaviour, and Perceived Satisfaction in Product Search. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (Feb 2018). https://doi.org/10.1145/3159652.3159714

[55] TikTok. 2020. How TikTok recommends videos #ForYou. https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you. Accessed on 11.02.2022.

[56] Twitch. 2022. Removing recommendations you are not interested in. https://help.twitch.tv/s/article/Removing-recommendations-you-are-not-interested-in?language=en_US. Accessed on 11.02.2022.

[57] Aki Vehtari, Andrew Gelman, and Jonah Gabry. 2016. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* 27, 5 (Aug 2016), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

[58] Wenjie Wang, Fuli Feng, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2021. Clicks can be Cheating: Counterfactual Recommendation for Mitigating Clickbait Issue. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Jul 2021). https://doi.org/10.1145/3404835.3462962

[59] Hongyi Wen, Longqi Yang, and Deborah Estrin. 2019. Leveraging post-click feedback for content recommendations. *Proceedings of the 13th ACM Conference on Recommender Systems* (Sep 2019). https://doi.org/10.1145/3298689.3347037

[60] Wei Yang, Kuang Lu, Peilin Yang, and Jimmy Lin. 2019. Critically Examining the "Neural Hype": Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking Models. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, 1129–1132.

[61] Xing Yi, Liangjie Hong, Erheng Zhong, Nanthan Nan Liu, and Suju Rajan. 2014. Beyond clicks. *Proceedings of the 8th ACM Conference on Recommender systems - RecSys '14* (2014). https://doi.org/10.1145/2645710.2645724

[62] Youtube. [n.d.]. Manage your recommendations and search results. https://support.google.com/youtube/answer/6342839?hl=en&co=GENIE.Platform%3DAndroid. Accessed on 11.02.2022.

[63] Youtube. 2019. YouTube survey FAQs. https://support.google.com/youtube/thread/1920627/youtube-survey-faqs?hl=en. Accessed on 11.02.2022.