Intent-Satisfaction Modeling: From Music to Video Streaming

GABRIEL BÉNÉDICT, RTL & University of Amsterdam, The Netherlands
DAAN ODIJK, RTL, The Netherlands
MAARTEN DE RIJKE, University of Amsterdam, The Netherlands

Logged behavioral data is a common resource for enhancing the user experience on streaming platforms. In music streaming, Mehrotra et al. have shown how complementing behavioral data with user intent can help predict and explain user satisfaction. Do 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 than in the music streaming setting. Video streaming platforms, in particular, could thus benefit from a better understanding of user intents and satisfaction level. We replicate Mehrotra et al.'s study from music to video streaming and extend their modeling framework on two fronts: (i) improved modeling accuracy (random forests), and (ii) interpretability (Bayesian models). Like the original study, we find that user intent affects behavior and satisfaction itself, even if to a lesser degree, 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. In this study, we find that in video streaming, unsatisfied users provide the main signal: intent influences satisfaction levels together with behavioral data, depending on our decisive vs. explorative grouping.

CCS Concepts: • **Information systems** → *Personalization.*

Additional Key Words and Phrases: Interaction signals; User intents; Session-based Recommendations

ACM Reference Format:
Gabriel Bénédict, Daan Odijk, and Maarten de Rijke. 2023. Intent-Satisfaction Modeling: From Music to Video Streaming. 1, 1 (March 2023), 22 pages. 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 [39]. 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 [53]. 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.).

Authors' addresses: Gabriel Bénédict, RTL & University of Amsterdam, Amsterdam, The Netherlands, g.benedict@uva.nl; Daan Odijk, RTL, Amsterdam, The Netherlands, daan.odijk@rtl.nl; Maarten de Rijke, University of Amsterdam, Amsterdam, The Netherlands, m.derijke@uva.nl.

Manuscript submitted to ACM

1.1 The importance of intent

Implicit and explicit feedback each have their own strengths and weaknesses [18, 30]. Most weaknesses can be avoided through careful survey design for explicit feedback and through granular user tracking for implicit feedback. 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 later that evening.

One way to retrieve context is to explicitly ask users about their current intents, join that survey data to behavioral data for each session, and thus introduce context back into implicit behavioral data. Mehrotra et al. [51] 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. With visualizations and logistic regressions they show that intent together with behavioral data is more predictive of satisfaction than behavioral data alone.

1.2 From music to video streaming

We are interested in generalizing the lessons in [51] from music to video streaming. There are important contextual differences between the two types of platforms that make this generalization far from obvious. See Table 1 (top) for a summary of key differences.

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

	Music [51]	**Video [this paper]**
Content length	3–5 min	45 min–2 hrs
Catalog size[1]	> 70 million	> 5 thousand
Piracy[2]	1 pm	7.5 pm
In-app survey design		
Intent identification	One-on-one interviews w/12 users	User experience specialists
Platform	Mobile	Browser
Timing	Coming back to the homepage	On the homepage for 7 seconds
Intent	One per session	Multiple per session
Survey rate	NA[3]	20%
Response rate	4.5%	3%
Very Satisfied users	33%	44%

First, content length is a difference linked to content type and 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 of platforms each have a few thousand movies and series available at any given time, with little to no content overlap between platforms [33]. Third, the relative scarcity of content and plurality of paid subscription services encourage a strong return to piracy in 2019–2022 [21, 54]. This rise in fragmentation and piracy encourages video streaming actors to (i) quickly and accurately guide *decisive* users to

[1] Similar to the average for the EU competition in the video domain [28] and international competition in the music domain [1, 2, 17, 38].
[2] 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].
[3] From the original study we know that 3 million US Spotify iPhone app users were sampled [51]. We could not find an official number on the US Spotify iPhone app users in 2019.

the content they had in mind within a shallow catalog (compared 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 exist two groups of intents, namely decisive and explorative, and show the essential role they play in video streaming platforms.

We follow Mehrotra et al. [51]'s methodology and adapt it for video streaming, in order to assess whether intent can indeed bring context back to explicit feedback. We adapt the original study to Videoland,[4] a video streaming platform in The Netherlands with over 1 million users. Two key differences in our experimental setup are that we use a browser (instead of a mobile app) and account for multiple intents per session (instead of only one); see Table 1.

This replicability study follows the ACM definition (different team, different experimental setup) [23]. This study is an attempt at replicating and generalizing a large portion of the experimentation pipeline: we cover data collection, survey design, data preprocessing, data enrichment, modeling, and interpretation.

1.3 Insights

In this replicability study of [51], 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:

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

In addition, our *technical* contributions to support replicability 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 is shared at https://github.com/rtlnl/streaming-intent-model.

2 RELATED WORK

Platforms are able to gather implicit feedback with highly granular logged data and explicit feedback via surveys. 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 fusion of explicit and implicit aspects (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 researchers early on [53]. Recently, the use of implicit feedback such as click through rate (CTR) [34, 48, 49] or dwell time [37, 67] has been questioned, in favor of the concurrent use of other behavioral metrics [18, 30, 52]. Wen et al. [66] 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 [51]: temporal (e.g., session length, seconds played), downstream (e.g., number of items played), surface level (e.g., number of slates that

[4] https://www.videoland.com/nl/

were interacted with), and derivative (e.g., total clicks / number of items played). Derivative signals are combinations of the other three signals.

Implicit feedback signals are often used as input for, 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 [65].

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. We argue there are two categories of higher order behavior on streaming platforms: *explorative* versus *decisive* (similar to *fetch, find* and *explore* in the domain of search for video streaming [41]). Decisive behavior refers to a session where the user already knows what she wants to stream and it is typically addressed in search [41]. Exploration can be defined as the experience of finding and consuming content that was previously unknown to the user [25]. In the music streaming domain, surveys have shown that exploration is a complex time-varying personal need [42], nurtures user retention [8], and deeper social connection [45].

A major drawback of surveys is their inherent *response bias*: the response rate of satisfaction surveys is low because users have to deviate from their intent of consuming content in order to provide feedback (our response rate was 3%, compared to 4.5% in [51], 4.6% at Spotify over emails [25], and 2% at Google for individual item surveys [15]).

The willingness to participate in a survey 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 [15, 25, 51]), etc. As a result, datasets collected through surveys have missing-not-at-random (MNAR) data [59]. 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 [15].

Recently, a new type of item-satisfaction survey emerged, e.g., item recommendation satisfaction surveys on YouTube with a Likert scale [68]. Also notable is the trend of the *not interested* button on a recommended item, which is well entrenched in the search & recommendation domain [13], on platforms such as YouTube [69], Twitch [62], and TikTok [61], with all three claiming it will help future recommendations. Such item-surveys suffer even more from response bias and thus motivate a new research field of sparse user-item pairs and debiasing [15].

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

2.3 Connecting implicit and explicit feedback

Typically, evaluation of recommender systems is either done (i) in small-scale lab studies based on explicit feedback, (ii) in offline batch experiments with static test collections again based on explicit feedback, or (iii) through large-scale A/B tests based on implicit feedback. Garcia-Gathright et al. [24] argue for the use of qualitative research in user behavior to provide insight on implicit feedback metrics as a general methodological principle.

An important way of drawing links between implicit and explicit feedback is via the users' current intent [16]. For example, Duan and Zhai [19] study the problem of learning query intent representations for product retrieval. They propose a generative model to discover intent representations from entity search logs and show that the discovered intent representations can be directly used for improving the accuracy of product search and recommendation. Similarly,

Bhattacharya et al. [5] predict user intent from a user's task context and combine it with a frequency-based graphical model to recommend reports to users of a business analytics application.

Recent workshops provide a rich palette of examples of capturing and mining intent from user interactions [7, 50]. Key domains where intent is an important feature for satisfaction prediction include: (i) e-commerce, where, for example, Su et al. [60] uncover different intents, find that different intents lead to different interaction behavior, and try to predict satisfaction from interaction signals, while Hendriksen et al. [32] show that purchase intent prediction for identified (as opposed to anonymous) users can dramatically reduce friction; (ii) movie recommendation, where, for example, Chen et al. [12] capture multiple intents from a (single) user's sequential behavior to guide the recommender to provide results that are diversified based on the intents discovered; (iii) news search and recommendation, where, for example, Lefortier et al. [43] discover that intents may shift dramatically based on real-world events and that user satisfaction may be hurt if the recommender does not shift with the shifting intents; (iv) search in video streaming platforms, where, for example, Lamkhede and Das [41] show that search intents are markedly different from search intents behind web search queries and that new challenges arise from the unavailability of an item that a user is keen to watch; (v) point-of-interest recommendation on maps, where, for example, Omidvar-Tehrani et al. [55] mine implicit intents by iteratively identifying groups of like-minded users and thereby increase user satisfaction; (vi) car GPS trajectories, where, for example, Snoswell et al. [58] use reinforcement learning to discover unobserved behavior intents; and, finally, (vii) advertiser satisfaction prediction, where, for example, Guo et al. [29] jointly model advertiser-side intent and advertiser satisfaction with attention mechanisms and recurrent neural networks. Other key aspects for which intent is an important predictor for user satisfaction include search result page organization [44] and ranking adjustments for different (inferred) needs for result diversity [16].

Identifying intents in search and recommendation can be a mix of supervised and unsupervised tasks that can involve users directly via interviews [51] or research teams internally. In task-oriented dialogue systems, the task of intent is usually addressed as a supervised learning problem [56]. Finally, Lin et al. [47] discover new intents based on a catalog of pre-existing 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 [14]. On music streaming platforms, a study by Mehrotra et al. [51] linked satisfaction with intent via a user survey and behavioral data on a music platform. This study is the most detailed one we found on the topic of intent-satisfaction modeling. This study's individual intents and behavioral data signals (such as *To play music in the background* or *songsPlayed*, respectively) raised questions about possible video domain counterparts.

To the best of our knowledge, there is no open dataset for intent-satisfaction modeling and 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 replicate and generalize [51] from music to video streaming. We generalize to the video domain by proposing video-specific intents and a detailed implementation of the survey design. We replicate models with binarized satisfaction levels as outputs, behavioral data and optionally intent as input, thus testing whether intent can help to 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 REPLICATION 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 replicate the methodology

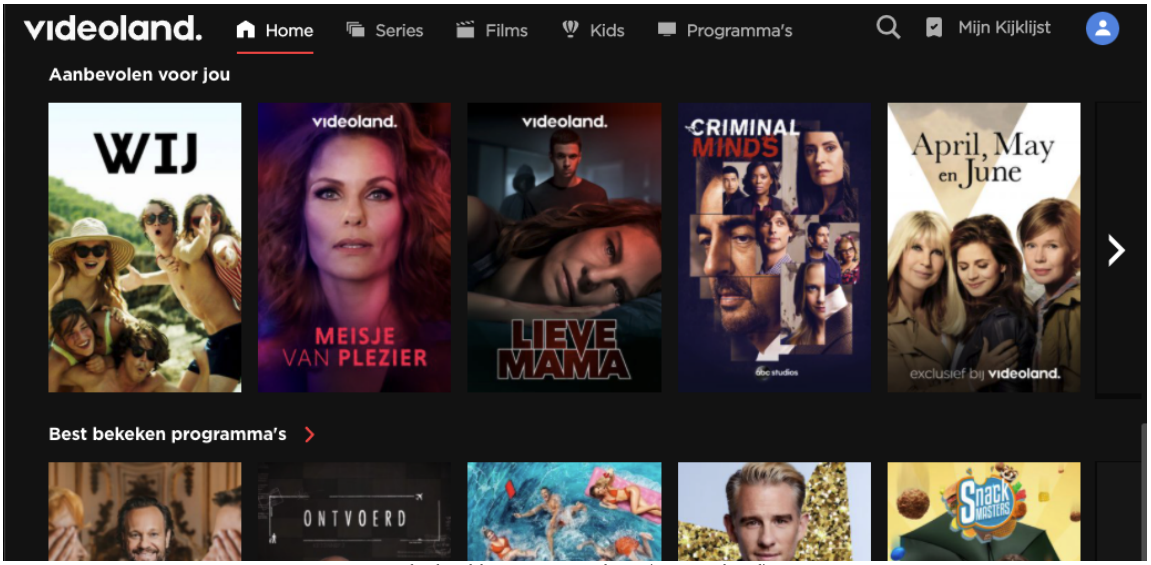


Fig. 1. Videoland homepage with its (personalized) strips.

of [51] and adapt it to video streaming. We compare and contrast two specific music and video streaming settings, before explaining our replication 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 replicability study we contrast a specific music streaming platform, Spotify, which provided the context for [51], 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 streaming platform in The Netherlands 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 replicated paper [51], we focus on a single device type so as to reduce noise. TV is our most used device but is not suited for surveys, due to the laid-back context and difficulty of typing with a remote. We chose our second-most-used device: desktop browser (10% of Videoland sessions), instead of TV or smartphone (as in [51]). 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 [51], the homepage is the focus of our analysis. As detailed in [57], 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 replicated here. For Videoland, the homepage is where most people land (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 recommended [31] and 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, and 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 (corrected for position bias).

3.2 Survey and experimental design

Mehrotra et al. [51] 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. [51] 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 (see Section 5.1, for the results).

The second step, the in-app survey, is the core of [51] and of our replicability study. The major choice here is where and when to show the survey to the user. While replicating 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 must not be deliberate. This forced Mehrotra et al. [51] 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. [51] 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 being 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 [51]).

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. [51]). They refer to, respectively, time related events, streaming related events, and user interface interaction events. Our behavioral variables are similar to the replicated

Table 2. Behavioral variables obtained from traffic data.

	Behavioral metric	Description
Temporal	timeToFirstTrailer	Seconds to the first trailer played
	timeToFirstPlay	Seconds to first content play
	sessionLength	Session length in seconds
Down-stream	numTrailerPlays	Number of trailers played
	numPlays	Number of full content played
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 account icon
	nProfileClicks	Number of clicks on *manage profile*
	nBookmarks	Number of bookmarked items

study, with the exception of *derivative signals* [51], 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 interaction. That last user interaction can be any surface level interaction, but we do not receive a log 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.

3.3.2 In-app survey variables. During the in-app survey (after seven seconds spent on the homepage), we ask two questions.[5] Namely,

(1) "How happy are you with your experience on the homepage today?" with satisfaction levels of 1 to 5 visualized using smiley faces (😠 😒 😐 😊 😍). In [51], 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
(2) "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. Mehrotra et al. [51] allow users to choose only one intent. By letting the user choose one or more intents, we show that a user can have a mixture of intents for the same session (see Section 5.1). Additionally, we add an "others" field, to let users answer with their own words (as in [51]). Mehrotra et al. [51] 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 replicability study. We therefore did not algorithmically extract intents from the "others" field.

[5]See screenshots in Appendix B

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

	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

4 REPLICATION OF SATISFACTION MODELS

In this section we describe our replications of the original satisfaction models with and without intent [51], before describing our own models and the training setup.

4.1 A satisfaction model

Our satisfaction models are exactly aligned with [51]. 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 [51], we construct binarized satisfaction level vectors over all surveyed sessions:

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

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].[6] The most straightforward regression model can estimate satisfaction levels $\mathbf{y}$ via a logit link:

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

with β_0 the intercept, $\{\mathbf{b}_1; \ldots; \mathbf{b}_j; \ldots; \mathbf{b}_J\}$ the behavioral variables and $\{\beta_1; \ldots; \beta_j; \ldots; \beta_J\}$ their respective estimates.

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,

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

with $\{\mathbf{d}_1; \ldots; \mathbf{d}_k; \ldots; \mathbf{d}_K\}$ intents and $\{\delta_1; \ldots; \delta_k; \ldots; \delta_K\}$ their respective estimates.

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

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

[6] In square brackets we include the labels that we use to refer to these models in Table 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 [40], that measures the effect of each intent as a group level effect, with a random intercept δ_k:

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

This time, we clearly model a hierarchical structure in the data and can assess group-level (intent-level) marginal satisfaction effects.[7]

4.2 Further satisfaction models

To achieve higher accuracy, we use XGBoost, a common implementation of gradient boosting decision trees [10], with a logistic regression objective. XGBoost is a strong performer on tabular data, even when compared against recent transformer models adapted to tabular data [6, 27].

For increased model interpretability, we opt for Bayesian satisfaction models with the same specifications as the frequentist versions above:

$$\begin{aligned} \text{logit}(\mathbf{y}^d) &= \beta_0^d + \sum_j \beta_j^d \mathbf{b}_j^d. \\ \beta_j^d &\sim N\left(\mu_j, \sigma_j^2\right). \end{aligned} \tag{6}$$

They allow for the estimation of entire marginal posterior distributions and thus more granular interpretability. We keep to a simple Bayesian logistic regression per intent with with population-level effects only; the focus is on explanation, rather than building a holistic prediction model. We leave more sophisticated models (e.g. varying slope and / or intercept, temporal, neural models) for future work on predicting intent online or offline (see Section 7 on future work).

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 [51] 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 [63] and confidence intervals. The intent-specific models are trained on subsets of the data that contain each specific intent and, thus, each has its 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 [11]). Regarding the Bayesian models, we checked for chain convergence in two ways: (i) visually with chain plots, and (ii) quantitatively with Rhat.[8] We assessed relative goodness-of-fit with leave-one-out cross-validation estimation with Pareto Smoothed Importance Sampling (PSIS) [64]. We evaluate on the same metrics as in [51]: accuracy, precision, recall, and F1 score. To calculate these confusion matrix related metrics, predictions in the $[0; 1]$ range have to be binarized at a certain threshold. Given the imbalance in the data (see Figure 2),

[7] Given that this is a general linear mixed model, we have to approximate log-likelihood. We use the reliable adaptive Gauss-Hermite algorithm that takes the form of a Laplace approximation [36], by setting the integer scalar parameter to 1 [3].
[8] Code and analysis available at https://github.com/rtlnl/streaming-intent-model

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 binarization (*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 REPLICATION

In this section we replicate the data analysis and visualizations from [51] and assess whether the original conclusions generalize from the music to the video domain. We produce three plots in line with [51], 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* (see Table 3). On average, users have 2.18 intents per session. 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.

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 Mehrotra et al. [51]'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).

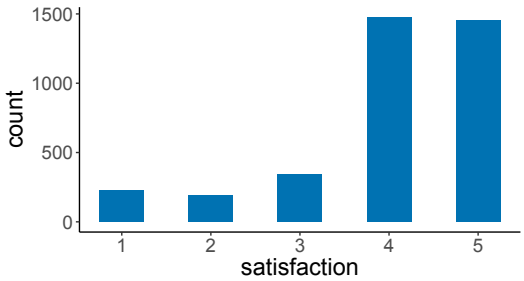


Fig. 2. Our imbalanced dataset: distribution of Likert-scale satisfaction levels for all surveyed users and across intents.

Next, we look at relationships between satisfaction level and intent (Figure 3). We draw a violin plot as in [51]. 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

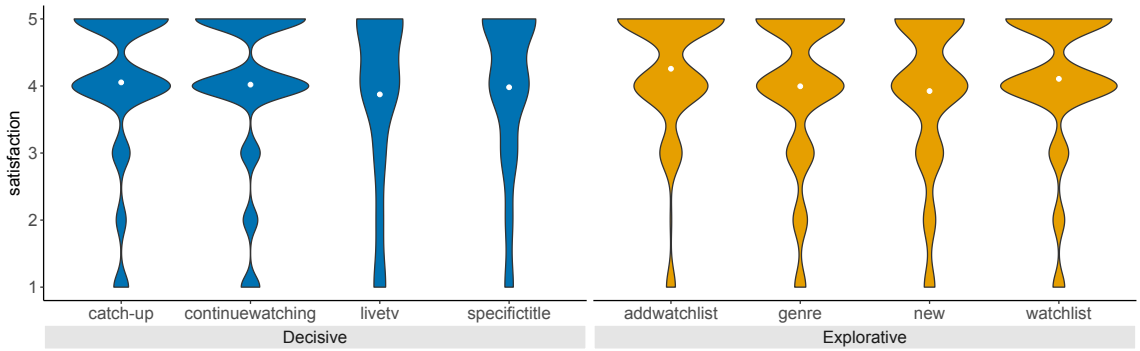


Fig. 3. Satisfaction levels per intent and by intent group (dot indicates the mean).

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

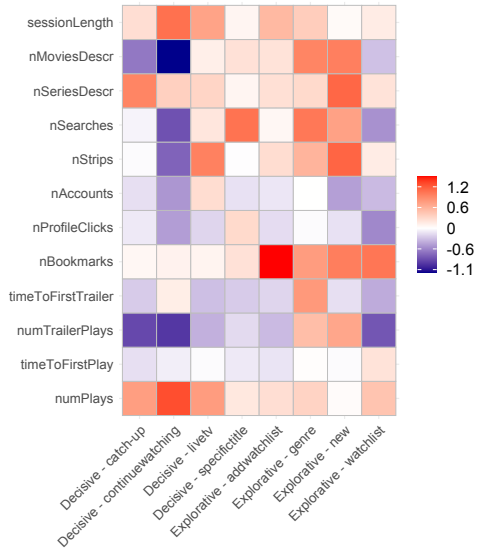
We recontextualize the raw behavioral data with users' revealed intents. The Pearson correlation plot in Figure 4a confirms a few intuitions. Users who intend to continue watching an episode interact the least with the platform, but it does not prevent them from watching a lot of content for long periods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

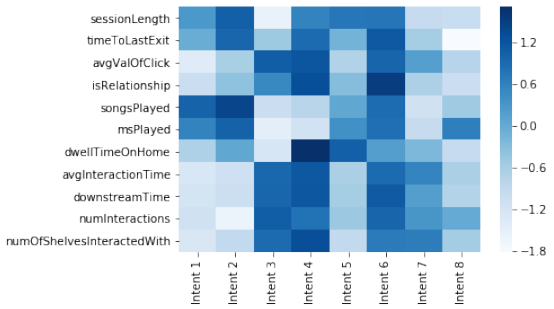
5.3 Upshot: Music versus video streaming

In replicating [51], we collected data in a completely different streaming platform and we adapted the survey design to our context and needs (the main differences are recorded in Table 1). We found Mehrotra et al. [51]'s data analysis to be replicable 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 [51], we find that two intents clearly have a higher amount of dissatisfied users, namely the decisive users looking to watch



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

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 [51]).

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

livetv or a *specifictitle*. Overall, Figure 3 and 4a confirm the learnings from [51], namely that users' satisfaction level and behavior are different depending on their intent.

Like in the original study, our conclusions might be influenced by response bias. For example, we typically observe little use of the bookmarking system on the platform. But our survey-behavioral dataset showed an unusually high number of users adding elements to their watchlist. We assume that users who 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 REPLICATION

We replicate multiple frequentist logistic regression satisfaction models: without intents, with intents, per intent, and with an intent as a hierarchical level, all as in [51]. Going beyond [51], we additionally 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.

6.1 Satisfaction prediction results

Table 4 displays the prediction results with standard deviations using 5-fold test sets. The binarization of intent plays a predominant role in the results (*Overall*, *Satisfied*, *Unsatisfied*). For the *Overall* and *Satisfied* binarizations, 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,

Table 4. Replication of [51] with added mean and standard deviation over 5-fold cross-validation for the three binarizations of the $y \in \{1, 2, 3, 4, 5\}$ satisfaction score (outcome variable) and four metrics (accuracy, precision, recall, F1 score).

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

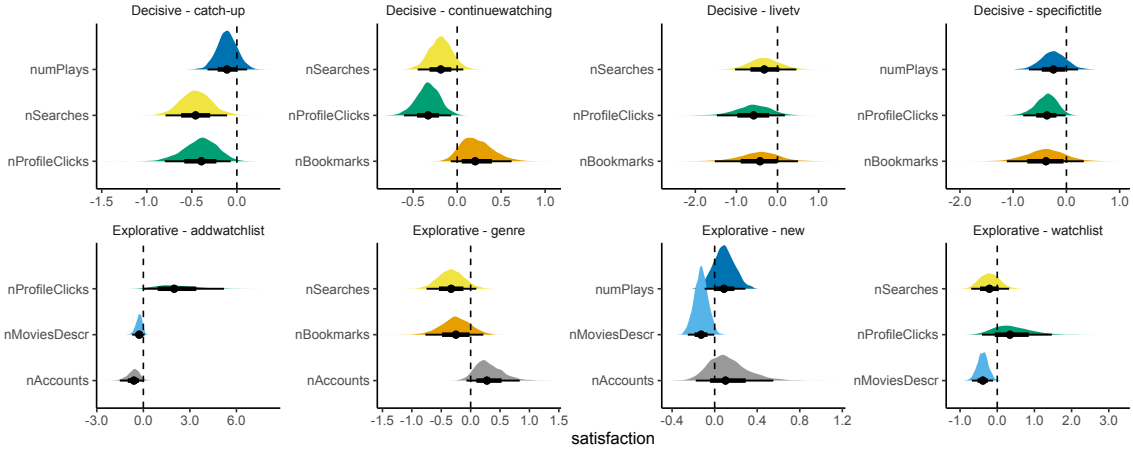


Fig. 5. Marginal effect of behavioral variables on satisfaction; Table 2 provides descriptions of our behavioral variables. One Bayesian fit per intent (median and IQR in thicker marks, 99.8% of the probability density function in thinner line).

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. 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 (*XGB 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[9] (similarly to the frequentist variable importance analysis in [51]).

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

[9]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/rtlnl/streaming-intent-model).

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 [51]. 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 [9], 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 [63]. (iii) We ran XGBoost and Bayesian models, for prediction accuracy and interpretability.

The conservative measures (i) and (ii), 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 motivates future research with a focus on dissatisfied users.

Overall, we could replicate the main finding of [51], namely that at least for unsatisfied users intent seems to impact satisfaction levels.

7 CONCLUSIONS

We have replicated and generalized Mehrotra et al. [51]'s work on intent-based satisfaction modeling, from music to video streaming. We have replicated the full experimental setup, from data collection – behavioral data and enrichment with an in-app survey – to computations. We provide our code for data preprocessing, visualization of the interactions between intents, satisfaction and behavioral data in line with the visualizations in [51]. Finally, we extended the modeling section with XGBoost models as standard tabular data benchmarks and per intent Bayesian models for interpretability.

7.1 Findings

Table 5 summarizes our findings in comparison to the replicated study [51]. We have found that in video streaming, as in music streaming, intent influences satisfaction levels together with behavioral data, although to a lesser degree than the original replicated study [51]. 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) Decisive users are not so keen on using the platform's personalized features and thus deserve special attention in the user experience design.

7.2 Broader impact

More broadly, this study reveals that it is possible to replicate a survey across different domains, device types and with smaller sample sizes. We hope this real-world small-sample replicable 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 replicated counterpart [51]) now show that it is not enough to look at behavioral

[10] This is probably due to the sampling methodology. In [51], the unsatisfied minority class is oversampled; while in the current study, the data is modelled as is.

Table 5. Overview of conclusions from Mehrotra et al. [51] compared to the current work. A checkmark indicates that the conclusion holds in the replicability study

Mehrotra et al. [51]	**Our work**
8 key user intents for music	8 different intents for video
No particular grouping	Grouped in decisive and explorative
Imbalance in satisfaction levels	✓
For unsatisfied users intent impacts satisfaction	✓
2 intents with more dissatisfied users	✓
Intent influences satisfaction levels	✓ (albeit to a lesser extent)
Level of satisfaction is not linked to amount of signal in behavioral data	Unsatisfied users are more prone to reveal their intent via behavioral data[10]
Intents important when predicting user satisfaction	✓
Different interaction signals important across intents	✓
Shared learning across intents improves satisfaction model	✓ (albeit to a lesser extent)
Users explore by playing	Users explore by interacting with the platform.
Blind exploration phase	Active exploration phase
Call for using user-level idiosyncrasies	Calls for exploratory user hand-holding
Listen "blindly," without much interaction	Tendency to use the user interface for inspiration

data alone to measure user satisfaction. Surveying and later predicting intents on each streaming platform help to better guide users to their goal or give users new perspectives.

7.3 Limitations

Our small-sample 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.

7.4 Further models

We focused on predictability (XGBoost) and interpretability (Bayesian intent model). For predictability, there is little evidence that improvement is possible with more sophisticated models, given the performance of XGBoost in the tabular data domain even in recent years [6, 27]. If we were to add a time aspect, such as sessions of the same user across time (i.e. longitudinal tabular data); we would consider a transformer neural network architecture [46]. For interpretability, we could consider fitting a single Bayesian model with all intents and variables, given a bigger sample. If intents are modeled as latent hierarchical effects, the model can be useful for daily user data, where intent is not available (because no survey was shown). We could thus extend the model to all user data and predict satisfaction given behavioral data and unobserved intent.

7.5 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, share insights on user intent and eventually try to predict them, given user behavior. 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.

ACKNOWLEDGMENTS

This work was supported by many people. We are very grateful to Rishabh Mehrothra who helped in re-contextualizing the study at Spotify and in articulating some of our overarching conclusions. At RTL, UX researcher Daniella Mittemeyer provided precious insights into the behavior and needs of our users, and helped design and run the survey. Web analysts Brenda Noordeloos and Roland Goudriaan helped write the code for data retrieval. Data analyst Isabella van der Vlies helped sharpen some of our conclusions. Finally, at the University of Amsterdam, we thank Ming Li and Romain Deffayet for reviewing the manuscript before submission.

This research was (partially) funded by Bertelsmann SE & Co. KGaA and by the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl.

All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

APPENDIX

A IMPLEMENTATION RESOURCES

To support replicability of our work, in the video or music streaming domain and beyond, we share[11] the following resources: (i) code for behavioral data retrieval (BigQuery); (ii) code for satisfaction modeling; and (iii) a detailed implementation of the in-app survey design. We cannot share individual user data, for GDPR compliance. However, to enable others to run our code, we include simulated behavioral and survey data in the repository, replicating the distributions in our dataset.

Our repository contains the libraries we use, the data preparation steps, visualization code for the plots in this paper and some additional distribution plots. Finally, the repository contains the modeling code to reproduce our cross-testing across different test sets and chain plots of marginal posterior distributions, to check for collinearity between sampling of different chains and variables.

B SURVEY FORM

Figure 6 shows the survey pop-ups in the original language. See Section 3.3.2 for translations.

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.

[11] https://github.com/rtlnl/streaming-intent-model

(a) Survey pop-up 1 on the bottom-right of the Videoland homepage, after 7 seconds.

(b) Survey pop-up 2 on the bottom-right of the Videoland homepage, after 7 seconds.

Fig. 6. Pop-up 2 shows after "next" is clicked on pop-up 1. For a translation, see Section 3.3.2.

[3] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48. https://doi.org/10.18637/jss.v067.i01

[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. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE. https://doi.org/10.1109/icdmw.2017.8

[6] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. 2021. Deep Neural Networks and Tabular Data: A Survey. *arXiv preprint arXiv:2110.01889* (2021).

[7] 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. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. Association for Computing Machinery, New York, NY, USA, 899–900. https://doi.org/10.1145/3336191.3371883

[8] 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.

[9] 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

[10] 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

[11] 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.

[12] Wanyu Chen, Pengjie Ren, Fei Cai, Fei Sun, and Maarten de Rijke. 2020. Improving End-to-End Sequential Recommendations with Intent-aware Diversification. In *CIKM 2020: 29th ACM International Conference on Information and Knowledge Management*. ACM, 175–184.

[13] 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

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

[15] 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

[16] Aleksandr Chuklin, Pavel Serdyukov, and Maarten de Rijke. 2013. Using intent information to model user behavior in diversified search. In *34th European Conference on Information Retrieval (ECIR'13)*. Springer.

[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. In *The World Wide Web Conference on - WWW '19*. ACM Press. https://doi.org/10.1145/3308558.3313592

[19] Huizhong Duan and ChengXiang Zhai. 2015. Mining Coordinated Intent Representation for Entity Search and Recommendation. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM. 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. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM. 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. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM. 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] Liyi Guo, Rui Lu, Haoqi Zhang, Junqi Jin, Zhenzhe Zheng, Fan Wu, Jin Li, Haiyang Xu, Han Li, Wenkai Lu, and et al. 2020. A Deep Prediction Network for Understanding Advertiser Intent and Satisfaction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. ACM. https://doi.org/10.1145/3340531.3412681

[30] Qi Guo and Eugene Agichtein. 2012. Beyond Dwell Time. In *Proceedings of the 21st international conference on World Wide Web - WWW '12*. ACM Press. https://doi.org/10.1145/2187836.2187914

[31] 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

[32] Mariya Hendriksen, Ernst Kuiper, Pim Nauts, Sebastian Schelter, and Maarten de Rijke. 2020. Analyzing and Predicting Purchase Intent in E-commerce: Anonymous vs. Identified Customers. In *eCOM 2020: The 2020 SIGIIR Workshop on eCommerce*. ACM.

[33] 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

[34] Jeff Huang, Ryen W. White, Georg Buscher, and Kuansan Wang. 2012. Improving Searcher Models Using Mouse Cursor Activity. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '12*. ACM Press. https://doi.org/10.1145/2348283.2348313

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

[36] Ke Ju, Lifeng Lin, Haitao Chu, Liangliang Cheng, and Chang Xu. 2020. Laplace Approximation, Penalized Quasi-likelihood, and Adaptive Gauss-Hermite Quadrature for Generalized Linear Mixed Models: Towards meta-analysis of binary outcome with sparse data. *BMC Medical Research Methodology* 20 (06 2020). https://doi.org/10.1186/s12874-020-01035-6

[37] Youngho Kim, Ahmed Hassan, Ryen W. White, and Imed Zitouni. 2014. Comparing Client and Server Dwell Time Estimates for Click-level Satisfaction Prediction. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM. https://doi.org/10.1145/2600428.2609468

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

[39] 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

[40] 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

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

[42] 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

[43] Damien Lefortier, Pavel Serdyukov, and Maarten de Rijke. 2014. Online Exploration for Detecting Shifts in Fresh Intent. In *CIKM 2014: 23rd ACM Conference on Information and Knowledge Management*. ACM, 589–598.

[44] Damien Lefortier, Pavel Serdyukov, Fedor Romanenko, and Maarten de Rijke. 2014. Blending vertical and web results: A case study using video intent. In *36th European Conference on Information Retrieval (ECIR 2014)*. Springer.

[45] Tuck W. Leong and Peter C. Wright. 2013. Revisiting Social Practices Surrounding Music. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/2470654.2466122

[46] Jeffrey Lin and Sheng Luo. 2022. Deep learning for the dynamic prediction of multivariate longitudinal and survival data. *Statistics in Medicine* 41, 15 (2022), 2894–2907. https://doi.org/10.1002/sim.9392 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.9392

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

[48] Yiqun Liu, Ye Chen, Jinhui Tang, Jiashen Sun, Min Zhang, Shaoping Ma, and Xuan Zhu. 2015. Different Users, Different Opinions. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM. https://doi.org/10.1145/2766462.2767721

[49] 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. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM. https://doi.org/10.1145/3132847.3133001

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

[51] 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. In *The World Wide Web Conference on - WWW '19*. ACM Press. https://doi.org/10.1145/3308558.3313613

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

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

[54] MUSO. 2022. Muso Discover Q1 2022 Digital Piracy Data Insights. *MUSO* (2022). https://www.muso.com/magazine/muso-discover-q1-2022-digital-piracy-data-insights

[55] Behrooz Omidvar-Tehrani, Sruthi Viswanathan, Frederic Roulland, and Jean-Michel Renders. 2020. Sage: Interactive State-aware Point-of-interest Recommendation. In *Workshop on State-Based User Modelling (SUM '20)*. 13th ACM International Conference on Web Search and Data Mining. https://www.k4all.org/wp-content/uploads/2020/01/WSDMSUM20_paper_SAGE_Interactive_State_aware-Point_of_Interest_Recommendation.pdf

[56] Jiahuan Pei, Pengjie Ren, and Maarten de Rijke. 2021. A Cooperative Memory Network for Personalized Task-oriented Dialogue Systems with Incomplete User Profiles. In *The Web Conference 2021*. ACM.

[57] 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

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

[59] Harald Steck. 2010. Training and Testing of Recommender Systems on Data Missing Not At Random. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10*. ACM Press. https://doi.org/10.1145/1835804.1835895

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

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

[62] 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.

[63] 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

[64] Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. 2015. *Pareto Smoothed Importance Sampling*. arXiv:stat.CO/1507.02646v8

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

[66] Hongyi Wen, Longqi Yang, and Deborah Estrin. 2019. Leveraging Post-click Feedback for Content Recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems*. ACM. https://doi.org/10.1145/3298689.3347037

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

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

[69] Youtube. 2022. Manage Your Recommendations and Search Results. https://support.google.com/youtube/answer/6342839?hl=en&co=GENIE.Platform%3DAndroid. Accessed on 11.02.2022.