Jointly Leveraging Intent and Interaction Signals to Predict User Satisfaction with Slate Recommendations

Rishabh Mehrotra
Spotify
rishabhm@spotify.com

Mounia Lalmas
Spotify
mounial@spotify.com

Doug Kenney
Spotify
dkenney@spotify.com

Thomas Lim-Meng
Spotify
tommeng@spotify.com

Golli Hashemian
Spotify
golli@spotify.com

ABSTRACT

Detecting and understanding implicit measures of user satisfaction are essential for enhancing recommendation quality. When users interact with a recommendation system, they leave behind fine grained traces of interaction signals, which contain valuable information that could help gauging user satisfaction. User interaction with such systems is often motivated by a specific need or intent, often not explicitly specified by the user, but can nevertheless inform on how the user interacts with, and the extent to which the user is satisfied by the recommendations served. In this work, we consider a complex recommendation scenario, called *Slate Recommendation*, wherein a user is presented with an ordered set of collections, called *slates*, in a specific page layout. We focus on the context of music streaming and leverage fine-grained user interaction signals to tackle the problem of predicting user satisfaction.

We hypothesize that user interactions are conditional on the specific intent users have when interacting with a recommendation system, and highlight the need for explicitly considering user intent when interpreting interaction signals. We present diverse approaches to identify user intents (interviews, surveys and a quantitative approach) and identify a set of common intents users have in a music streaming recommendation setting. Additionally, we identify the importance of shared learning across intents and propose a multi-level hierarchical model for user satisfaction prediction that leverages user intent information alongside interaction signals. Our findings from extensive experiments on a large scale real world data demonstrate (i) the utility of considering different interaction signals, (ii) the role of intents in interpreting user interactions and (iii) the interplay between interaction signals and intents in predicting user satisfaction.

ACM Reference Format:
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 *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3308558.3313592

1 INTRODUCTION

An increasingly larger proportion of users rely on recommendation systems to pro-actively serve them recommendations based on diverse user needs and expectations. Developing a better understanding of how users interact with such recommender systems is important not only for improving user experience but also for developing satisfaction metrics for effective and efficient optimization of the recommendation algorithm. This is especially true in the case of online streaming services like Pandora, Spotify and Apple Music, wherein the system could gauge user satisfaction and adapt its recommendations to better serve user needs.

Since obtaining explicit feedback from users is prohibitively expensive and challenging to implement in real world systems, commercial systems rely on exploiting *implicit* feedback signals derived from user activity. When users interact with the recommendations served, they leave behind fine-grained traces of interaction patterns, which could be leveraged for predicting how satisfied was their experience, and for developing metrics of user satisfaction.

Prior work have studied implicit feedback signals (e.g., clicks, dwell time, mouse scrolling) in the case of web search, and verified their effectiveness in predicting user satisfaction, both on traditional desktop [9, 11, 26, 27] and mobile setting [17, 39]. Furthermore, past work in web search systems has also highlighted the importance of considering tasks and intents when interpreting implicit feedback signals [14, 38]. On the other hand, the identification and effectiveness of corresponding implicit feedback signals as well as the role of user intents has remained understudied in the context of recommendation systems, especially in the mobile context.

While search systems have access to explicit queries from users, based on which one could extract tasks, and interpret interaction signals, thereby differentiating between success and failure; recommender systems, on the other hand, lack such explicit indicators of user intent and clear indicators of success. Indeed, the interpretation of signals varies with goals; for example, scrolling can indicate negative experience when the goal is to quickly listen to music now, but can also indicate a positive experience when the goal is to browse the diverse collection of music the system has to offer. Furthermore, interpreting interaction signals becomes especially hard in the context of complex recommendation settings, like *Slate recommendations*, a scenario typical to many music streaming services, where users are recommended a set of collections (called *slates*), with different purposes (to explore new music, or quickly jump to recently played music, etc), and heterogeneous content (playlists, artist profiles, other audio content, etc). Thus, there is a need for a

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313592

detailed, holistic view of user interactions with such recommender systems, to establish their utility in predicting satisfaction.

In this work, we consider the use case of music streaming via slates of recommendation, and aim at understanding the relationship between interaction signals, user intents and user satisfaction. We consider the case of users interacting with a mobile based music streaming app, Spotify, and investigate the different interaction signals that can be extracted in slate recommendation setting, and how these interaction signals vary across different intents. Since the list of possible intents user might have is hitherto unknown, we adopt a mixed methods approach to understand user intents, and leverage insights from (i) face-to-face interviews, (ii) large scale in-app survey and (iii) non-parametric clustering techniques to identify the list of possible intents. We identify eight user intents and verify their validity using large scale log data.

To predict user satisfaction, we jointly leverage insights from the extracted interaction signals and intents. While interaction signals have been directly used to predict satisfaction in search [9, 24, 27], we instead hypothesize that interpreting interaction signals without factoring in user intent would lead to noisy, unreliable estimates of satisfaction in our context. We demonstrate that traditional approaches of using a single, global model for satisfaction trained on aggregate data without considering intent groups does not perform well in predicting satisfaction. We show that significant performance gains are obtained in switching from a single global prediction model to separate satisfaction models for each intent.

Further, we identify issues with the global and per-intent models. While the former fails to capture intent-specific intricacies, separate per-intent models ignore information and insights from other intents. To address these, we show the importance of shared learning across intents and propose a *multi-level hierarchical model* that allows the estimation of intent-specific effects while at the same time learning from data across all intents, thereby compromising between the overly noisy per-intent model and the oversimplified global model that ignores intent groups. The proposed multi-level model facilitates incorporating both individual session level and intent group level effect on user satisfaction, thereby allowing for variability in user interaction behavior across intents.

Extensive experiments on a real world large scale music streaming data from Spotify highlight the benefit of explicitly considering user intent and demonstrate the effectiveness of different interaction signals in predicting user satisfaction in a slate recommendation setting. We contend that our findings provide a valuable framework for fine-grained analysis of user interaction behavior for developing metrics and gauging user satisfaction. To the best of our knowledge, the present study is among the first to identify the benefit of incorporating user intent information in understanding user interactions and satisfaction in recommendation settings, and to demonstrate the significant gains obtained over the popularly used single global intent-agnostic prediction models.

2 RELATED WORK

The current research builds upon four areas: (i) slate recommendations, (ii) user interactions for satisfaction prediction, (iii) user intent modeling and (iv) mixed methods approach to analysis.

Slate Recommendation. Recommendation of slates is a common problem arising in various contexts including search [33], ads and recommendations [8]. Several approaches have been proposed to generate slates, including List Conditional Variational Auto-Encoders [12] and Slate-MDPs [32]. A different approach [33] investigates the evaluation of policies that recommend slates, and then introduces a pseudo-inverse estimator for off-policy evaluation in combinatorial contextual bandits. Different from proposing a new approach to generate slates, our work is concerned with investigating user interaction and gauging user satisfaction with recommended slates.

Satisfaction & User Interaction Signals. With respect to interaction signals, implicit feedback signals (e.g., clicks, dwell time, mouse scrolling) and their sequences have been extensively studied in web search and their effectiveness in predicting satisfaction [9, 11, 24, 26, 27, 29] has been verified, using techniques such as Markov models or deep neural models. A detailed overview regarding the evaluation of interactive information retrieval systems with users can be found in Kelly *et al.* [13].

Scrolls and cursor movements do not exist in mobile search; however, the swipe interaction performs a similar function. Past work leveraging user interaction on mobile phones have used signals such as swipes, dwell times on landing pages and zooms for detecting search result relevance [10] as well as for gauging user attention and satisfaction [25]. User interactions with touch based devices have also been used to detect good abandonment [39] and user satisfaction with intelligent assistants [17]. Building upon such work, we leverage swipe interactions and some signals that result from swipe activity to predict satisfaction.

In the context of recommendation systems, temporal features, including variants of dwell time have been considered for understanding satisfaction [16] and relevance [41]. Specifically focusing on evaluating recommender systems, past work has investigated preference elicitation [19], explanations [34] and user centric combinations of explicit and implicit preferences [18]. Past work has also investigated the influence of personal and situational characteristics towards explaining user experience with recommender systems [20, 22].

Our work also investigates implicit feedback signals, similarly to those works, but in a less explored context, that of music recommendation. Additionally, we aim to investigate the hypothesis on how the interpretation of such interaction signals and their role in predicting satisfaction changes across different user intents.

User Intent Modeling. Related to our consideration of user intents, the role of query intents and search tasks have been extensively studied in the search community, with past work aimed at leveraging intents in predicting query and task satisfaction [27, 37, 42]. Recent work reported in [1] considers the role of intents in recommendations and propose an intent-aware recommendation model. In music recommendation, recent efforts have been made around initial exploration of music listening intents associated with common activities [35], as well as studies of why and when people of different ages and in different countries listen to music [36]. These investigations suggest that intent modeling can potentially improve recommendation quality.

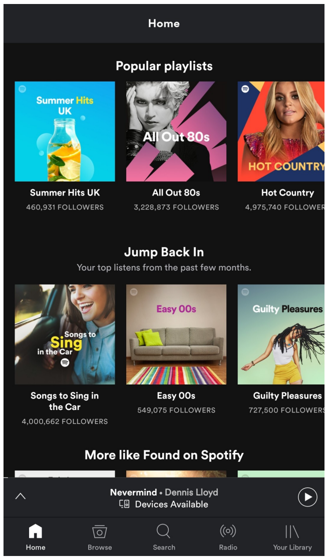


Figure 1: An example homepage of Spotify, a music streaming app.

In web search, tasks have been considered as an analogous abstraction to intents. Prior work has emphasized understanding the role of task level differences while leveraging implicit feedback. Specifically, White *et al.* [38] investigated how individual and task differences impact the effectiveness of algorithms based on implicit relevance feedback. Additionally, Kelly *et al.* [14] considered the effects of task on display time, and quantified the potential impact of this relationship on the effectiveness of display time as implicit feedback.

Mixed methods. Finally, our mixed methods approach of leveraging interviews along with qualitative and quantitative insights follow recent works investigating user expectations, behaviors, and satisfaction in the context of music discovery [5], image [40] and product search [30].

3 PROBLEM FORMULATION

Our goal is to extract and leverage user interaction data to understand and predict user satisfaction in a slate recommendation setting. We consider the Homepage of a major commercial music streaming service, and base our experiments around user interaction with the resulting recommendation page.

3.1 Key concepts

We first define the key concepts used throughout the paper.

- **User Session:** A session is defined as a sequence of user actions with the recommendations rendered, including active browsing of content and passive listening to music. A user session ends when the user has not interacted with or streamed from the Homepage for more than 30 minutes.
- **Slate Recommendation:** A slate is a collection of playlists grouped together under a common theme. Figure 1 shows two slates, *Popular Playlists* and *Jump Back In*, each with 3 playlists. The Slate Recommendation task is to recommend a collection of slates.

Table 1: Description of user interaction signals used.

Signal Type	Signal	Description
Temporal	session length	duration of entire session in seconds
	ms played	total milli-seconds streamed
	dwell time on Homepage	session duration minus downstream time spent streaming content
	avg interaction time	time spent on the Homepage interacting with slates
	time to success	time until first stream
Downstream	songs played	number of songs played
	relationship built	binary signal indicating whether the user saved or downloaded any track or album
	downstream time	total streaming time in any playlist reached via Homepage
Surface Level	no of interactions	total number of clicks on Homepage
	nSlates	no of slates interacted with
	didScroll	binary variable indicating whether or not the user scrolled to view additional recommendations
	max depth	maximum depth reached on the Homepage (number of slates vertically scrolled)
	no of exits	number of exits from Homepage to any playlist
Derivative	avg value of click	total no of clicks / number of stream events
	abandoned	binary feature denoting if the session was abandoned without any interaction
	intentional	binary feature indicating if the user intentionally came to the Homepage from other app feature

- **Homepage:** Also referred to as the home tab, the homepage is the first landing page of the music streaming app, which surfaces a number of playlists, organized as different slates. Figure 1 gives an example of the homepage of Spotify, the music streaming app used in this study.
- **User Satisfaction:** User satisfaction can be viewed as a subjective measure of the utility of recommendations, and we posit that user satisfaction is conditioned not only on the recommendations served, but also on user intent. We rely on implicit signals to derive satisfaction estimates and then use user reported satisfaction scores as gold standard estimate of user satisfaction.

The slate recommendation algorithm aims to find the optimal, ordered subset of items (playlists), a.k.a. slate, given the page layout to serve users recommendations so as to maximize user satisfaction. A Homepage is a collection of such recommended slates, which the user can interact with. Since users come to the app with different goals at different points in time, it becomes important for system designers to understand user interaction to gauge user satisfaction with the slates of recommendations rendered on the app Homepage.

In the rest of this section, we shed light on the different interaction signals we extract from user interaction with the Homepage of Spotify and briefly motivate the need for considering user intent for understanding and predicting user satisfaction in our context.

3.2 Extracting User Interactions Data

The Homepage rendered for a user is rich enough to allow him or her to interact with it in a myriad of ways, including clicking on playlists, scrolling vertically to view more slates, scrolling horizontally to view more playlists in a specific slate, pausing to read and visually absorb content, clicking and consuming content via streaming, among others. While past work on understanding user interaction in mobile search setting have proposed few signals [17, 39], we additionally propose a number of new signals resulting from the specific *Slate Recommendation* scenario considered in this work.

We use back-end logs of user interactions and extract four different types of interaction signals for each user session:

(1) **Temporal signals**: these focus on the temporal aspects of user interaction, including time spent interacting with the slates, session length, dwell time, etc.
(2) **Surface level signals**: these capture aggregate user interaction on the surface of the Homepage, including total number of clicks, total number of slates the user interacted with, maximum depth the user reached by vertical scrolling, and total number of exits the user had from the Homepage.
(3) **Downstream signals**: these capture downstream user engagement resulting from the Homepage, i.e., user interaction with the playlists, including streaming songs, saving or downloading tracks, viewing Artist profile pages.
(4) **Derivative signals**: these are derived using interaction features like whether the session was abandoned, or whether the session was intentional, with the user going to the Home tab from elsewhere in the app. The average value of click is derived by the considering all clicks in a session and dividing by the total number of streams observed in that session.

Table 1 provides a detailed description of the different interaction signals extracted for each user session.

3.3 Role of User Intent

In the use case of a slate recommendations on a surface like some of the big music streaming apps, users use the Homepage for different needs at different times. With differing needs and intents, users would interact with the app differently, and hence leave different traces of behavioural signals.

Indeed, a user might just want to quickly play some background music, and would not spend much time in finding music to play; on the other hand, a different user might want to carefully find music to save for future listening. In both cases, the way the user interacts with the music app differs, thereby leading to different interaction signals. Since implicit measures of user satisfaction rely heavily on interpreting interaction signals, we hypothesize that understanding and identifying user intent is important to predict user satisfaction in our context, as has been done in other scenarios, particularly in web search [3, 15, 21, 27, 31].

To this end, we discuss different ways of identifying user intents in a recommendation setting in Section 4, and leverage the intents identified along with the extracted interactions signals to predict user satisfaction in Section 5.

4 INTENT IDENTIFICATION

Users use recommendation platforms with different intents. The space of possible intents a user might have is hitherto unknown. In this section, we focus on identifying the different intents users might have when interacting with a music recommendation system.

Table 2: Demographic details of the participants interviewed.

Home Usage	Participant	Gender	Age	Job
Rarely visit Home	Participant 1	Female	31	HR Analyst
	Participant 2	Female	23	Accountant
	Participant 3	Male	20	Student
	Participant 4	Female	27	Construction Manager
	Participant 5	Female	27	Student
	Participant 6	Female	36	Nightlife Manager
Frequently visit Home	Participant 7	Female	28	Student
	Participant 8	Male	31	Finance Manager
	Participant 9	Male	20	Student
	Participant 10	Male	35	Makeup Artist
	Participant 11	Female	32	Attorney
	Participant 12	Male	25	Marketer

4.1 Interviews

To discover and better understand user intents when interacting with the Homepage composed of a set of slate recommendations (playlists), we conducted in-depth 1:1 in-person interviews with users of the streaming app. We describe the interview process and discuss our findings.

Approach. We interviewed 12 users living in the city of New York who were found to have varying degrees of Homepage usage with the Spotify app in the 4 weeks prior to the study. We selected participants from both iOS and Android users with a range of levels of engagement with the app (low, medium, high) and the Homepage (rarely visit, frequently visit), with engagement defined as the number of days they actively used the app in the past month. We emailed a screener survey to a list of Spotify users with these specifications to recruit and schedule the interviews. Table 2 provides a detailed demograpic overview of the participants. With each user, we conducted 1:1 face-to-face interviews spanning 45 minutes each and provided them a $100 gift card upon completion of the interview.

Our goal was to understand and discuss their intents and satisfaction when using the Homepage. We asked them questions around (i) What intents they have when coming to the Homepage; (ii) How satisfied they are with their experience, and (iii) How they behave in relation to intent and satisfaction. The participants were encouraged to provide detailed responses to these questions which were recorded (audio and video).

Findings. We performed a detailed post-hoc analysis of the data from the interviews by going through the video recording of the interviews and the interview transcripts. We grouped together findings that appeared in more than one interview in order to create themes. As per our initial hypothesis, different participants reported to using the Homepage of Spotify for different goals at different times. We were able to categorize user intents into two main themes, Passively Listening, and Actively Listening, and to identify the following overarching intents:

(1) **Theme 1: Passively Listening**
  (a) To find music to play in the background
  (b) To play music that matches their mood or activity
  (c) To quickly access their playlists or saved music

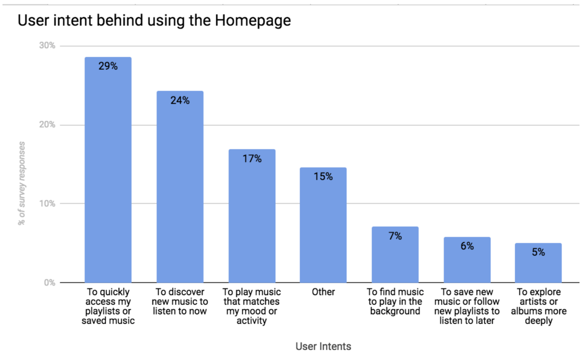


Figure 2: Distribution of user intents.

(2) **Theme 2: Actively Engaging**
  (a) To discover new music to listen to now
  (b) To save new music or follow new playlists to listen to later
  (c) To explore artists or albums more deeply

Additionally, some users pointed out that they use the Homepage because it is the default page that opens upon loading the music app, without having a clear intent when they go on it. We therefore added *Because it is the first screen shown* as an additional intent to the above, thereby making it a total of seven identified intents.

These intents differ from those reported in [36], since they correspond to user intents in using a homepage consisting of slates of recommendations, whereas the intents identified in [36] are user intents for listening to music in general.

Importantly, users often alluded to the fact that their behavior on the Homepage varied with their intent. For example, *scrolling* can be an indicator of exploration, but *scrolling* back and forth can signal a struggle. Not *scrolling* could mean satisfaction with the recommendations presented, but could also indicate dissatisfaction with what was presented. This motivates the need for interpreting interaction signals differently for different intents.

4.2 In-App Survey

Following up on the intents identified from the face-to-face interviews, we performed a large scale survey to validate and understand the prevalence of the intents identified from the interviews. The survey allowed us to collect data from users about their intent and satisfaction during a given session with the Homepage.

Methodology. We selected a random sample of 3 million iPhone users of the Spotify app residing in the US, and presented them with a brief survey at the end of their Homepage session. To not interfere with their natural interaction, and to not bias the user interaction data collected from a session, we trigger the survey only after the user ended their session and moved to another feature of the app (e.g. Search, My Library). To avoid inundating users with the survey repeatedly, the in-app survey is trigged such that each user participates in the survey at most once. Furthermore, we assume that a user has one overarching intent when they begin their Homepage session, an assumption we intend to relax as part of future work.

To gauge user satisfaction we asked the following question: *How satisfied or dissatisfied were you with your experience on the Home screen today?* and presented them with 5 options ranging from *Very satisfied* to *Very Dissatisfied*, along with an additional option of "I wasn't on the Home screen today". Further, to understand their intent behind using the Homepage, we asked them the following question: *Why were you on the Home screen today?* and showed them the six intents identified during the interviews.

As the listed intents may not capture all possible intents users might have, we added *Other* as an option. The *Other* intent option was accompanied by a free text block where users were asked to specify their intent; this allows us to identify any intent not identified during the interviews. Parsing the free flow natural language text and identifying common intents from it is a non-trivial task, which we address in Section 4.3.

Survey Results & Findings. We briefly discuss findings about user intents from the large scale survey, and leave the discussion of findings about user satisfaction for later (Section 6.2). The response rate was 4.5%, with over 116000 users from the 3 million users targeted responding to the survey, which is similar to the response rates we observed in our past surveys.

Figure 3 presents the distribution of intents across user sessions, which indicates how prevalent the intents are. We observe a fairly even distribution of intents across passively listening and actively engaging, which the top two intents: (i) *To quickly access their playlists or saved music* and (ii) *To discover new music to listen to now*, covering 29% and 24% of sessions respectively. A sizeable portion of the sessions involve the user saving music for later consumption or exploring artists in detail, with over 20000 user sessions dedicated to these intents.

Given that the majority of the sessions could be linked to one of the intents identified during interviews, it validates the identified intents. However, since over 15% of users opted for *Other*, indicating that their intent is not covered in the presented list of intents, we further investigate the presence of other intents from the text entered by users. The next section describes a computational approach to analyze the responses to identify new intents.

4.3 Bayesian Non-parametric Model

To ensure that the intents identified sufficiently cover the space of all possible intents users might have when interacting with the Homepage of the Spotify app, we investigate the text entered by the users when they selected the *Other* intent category in the survey responses. Simple clustering on the text is not much useful since the number of possible intents in not known apriori. As a result, we resort to Bayesian nonparametric clustering [7].

The distance dependent Chinese restaurant process (dd-CRP) [2] was recently introduced to model random partitions of data, and is a commonly used method for non-parametric clustering [2, 28]. To extract intents, we consider the dd-CRP model in an embedding-space setting and place a dd-CRP prior over the intents. While a detailed description of non-parametric clustering is beyond the scope of this work, we briefly describe the major steps below. Interested readers are referred to Blei *et al.* [2] who introduce the dd-CRP model, and to Mehrotra *et al.* [28] who apply dd-CRP model to identify query clusters.

Let z_i denote the ith intent assignment, the index of the text with whom the ith text is linked. Let d_{ij} denote the distance measurement between texts i and j, let D denote the set of all distance measurements between texts, and let f be a decay function. The distance dependent CRP independently draws the text assignments to intent cluster conditioned on the distance measurements,

$$p(z_i = j|D, \alpha) \propto \begin{cases} f(d_{ij}) & \text{if } j \neq i \\ \alpha & \text{if } j = i \end{cases}$$

Here, d_{ij} is an externally specified distance between texts i and j, and α determines the probability that a customer links to themselves rather than another customer. The monotonically decreasing decay function $f(d)$ mediates how the distance between two texts affects their probability of connecting to each other. Cosine similarity between the word2vec embedding representation of two input text describe the distance function (d_{ij}).

Given a decay function f, distances between texts D, scaling parameter α, and an exchangeable Dirichlet distribution with parameter λ, N M-word queries are drawn as follows,

(1) For $i \in [1, N]$, draw $z_i \sim dist - CRP(\alpha, f, D)$.
(2) For $i \in [1, N]$,
  (a) If $z_i \notin R^*_{q_{1:N}}$, set the parameter for the ith text to $\theta_i = \theta_{q_i}$. Otherwise draw the parameter from the base distribution, $\theta_i \sim Dirichlet(\lambda)$.
  (b) Draw the ith terms, $w_i \sim Mult(M, \theta_i)$.

We employ a Gibbs sampler, wherein we iteratively draw from the conditional distribution of each latent variable, given the other latent variables and observations. The Gibbs sampler iteratively draws from

$$\begin{aligned} p(z_i^{new}|z_{-i}, x) \propto p(z_i^{new}|D, \alpha) \\ p(x|t(z_{-i} \cup z_i^{new}), G_0) \end{aligned} \tag{1}$$

The first term is the dd-CRP prior and the second is the likelihood of observations (x) under the partition, and $t(z)$ is the intent-cluster formed from the assignments z. We employ a Dirichlet-Multinomial conjugate distribution to model the likelihood of text terms.

Input texts from users are assigned to intent clusters by considering sets of texts that are reachable from each other through the intent cluster assignments. Notice that many configurations of text assignments might lead to the same intent-cluster assignment. Finally, text assignments can produce a cycle, e.g., text 1 linking with 2 and text 2 linking with 1. This still determines a valid intent-cluster assignment; all texts linked in a cycle are assigned to the same intent cluster.

Upon running the dd-CRP model on the 15% response with user entered text, we obtained 5 different clusters of user responses. We manually investigated these clusters and identified the following intents: (i) To Find X, where X could be any specific playlist a user was trying to find; (ii) To explore or casually browse; (iii) All of the above, wherein the users wanted to select all options presented to them, (iv) I wasn't on Home tab, and (v) Miscellaneous.

Among these five newly identified intents, *I wasn't on Home tab* indicates that the users did not participate in a Homepage session, so we discarded these sessions. Further, Miscellaneous cluster mainly consisted of incoherent text entered by users, thereby rendering it futile. Similarly, we discard sessions tagged with *All of the above* cluster, since it does not present us with new intent and only had few sessions. Finally, the *casually browsing* intent was a very small cluster. In summary, we identified one additional intent: *To Find X* to be added to the pool of the seven intents identified earlier.

Table 3: Final list of user intents identified when interacting with the slated recommended by the Homepage of a music streaming app. Although Intent 1 is not an actual intent, we still refer to it as Intent for ease of writing.

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.

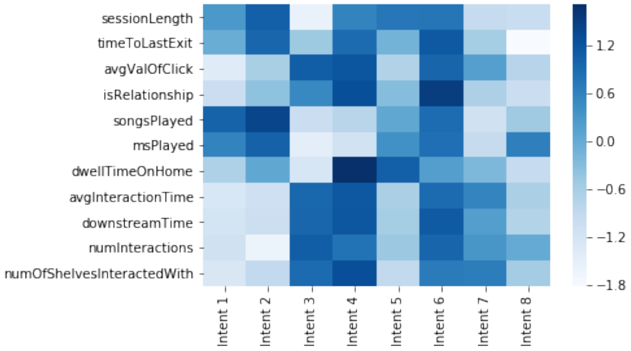


Figure 3: Heatmap of interaction signals across the different user intents. Signals are prevalent to varying extents in different intents, thus highlighting varying user interactions across different intents.

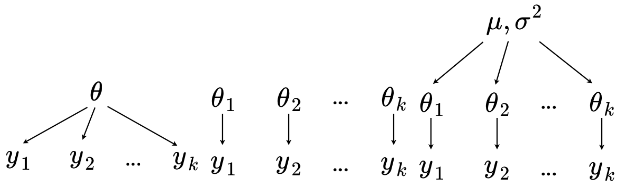
Users often use the Homepage to find something they're looking for, which could be any specific playlist or any specific shelf. The extracted additional intent caters to such search-like use-cases.

Although the rigorous process of non-parametric clustering applied to user entered responses only gives us one new intent, it reaffirms the exhaustiveness and coverage of the six intents identified before.

4.4 Analysis of User Interactions & Intents

Based on the interviews, in-app survey and non-parametric clustering of textual user responses, we end up with eight user intents, as described in Table 3. The diversity of these intents highlights the variety in what users intend to achieve when using the Homepage of the music app.

We hypothesize that the way users interact with the recommended slates of playlists would differ across these intents. Figure 3 presents a heatmap of a subset of interaction signals (see Table 1) with the different intents. We observe that the prominence of interaction signals significantly differs across intents, with certain signals like interaction time on the Homepage significantly lower for intent 2 (to quickly access music). These differences in interaction signals highlight the fact that users indeed behave differently when having different intents. Since implicit measures of satisfaction highly depend on interaction signals, this places



(a) Global Model (b) Per-Intent Model (c) Multi-level Model

Figure 4: Different modeling assumptions for intent aware SAT prediction model: (a) Global Model for all data across all intents; (b) Per-Intent Model; and (c) Hierarchical Multi-level Model with shared parameters.

special emphasis of explicitly considering intents when predicting satisfaction.

5 PREDICTING SATISFACTION

Our main goal is to understand and predict user satisfaction (SAT) using interaction data. To this end, we extracted detailed user interaction signals, and identified different intents users might have. In this section, we leverage the extracted signals and intents and present techniques to predict user satisfaction using the signals. We present three approaches for satisfaction prediction, covering the spectrum of intent-level granularity, i.e. global model for all intents to a separate model for each intent and a shared model across intents.

5.1 Global Model

We begin by describing the most conventional approach to satisfaction prediction, referred to as *Global Model.* The simplest technique would be to treat all user sessions as a homogeneous collection of data, with the user intent featuring simply as a categorical variable along with user interaction signals. Almost all existing work on predicting user satisfaction from interaction signals [9, 11, 26, 27] have employed such an approach to satisfaction (SAT) prediction.

For the Global prediction model, we perform traditional Logistic Regression and Gradient Boosted Decision Trees classifiers. Given a dataset with n user sessions ($x_i \in X$), each of which is represented by m features, $D = \{(x_i, y_i)\}$ ($|D| = n, x_i \in R^m, y_i \in R$), the logistic regression classifier assigns a probability to each outcome as

$$P(y_i = c|x_i) = logit^{-1}(\theta^T x_i + b) \tag{2}$$

where $c \in \{-1, 1\}$ and is trained with conditional maximum likelihood estimation, wherein we choose θ that maximize the (log) probability of the labels given the observations x_i. Thus, the objective function we aim at maximizing is:

$$L(w) = -\Sigma_{i=1}^{n} \sigma(\theta^T x_i + b)^{y_n} (1 - \sigma(\theta^T x_i + b)^{(1-y_n)} \tag{3}$$

where σ is the sigmoid function. Alternatively, tree boosting type of models (e.g. Gradient Boosted Regression Trees) optimize a tree ensemble model using additive functions to predict the output. We train both Logistic Regression and GBDT models using all the interaction features extracted for each session, along with a categorical variable describing user intent for the current session. As shown in Figure 4 (a), the same parameter is learned and shared across all intents and sessions.

5.2 Per-Intent Model

A key issue with a Global Model is that it ignores the variations in the interaction signals across different intents. Indeed, a user with an intent to quickly access her playlist to play music (Intent 2) would know what she is looking for, and not spend time in detailed exploration, unlike a user interested in exploring an artist more deeply (Intent 8). Whereas the categorical variable for the session intent would slightly help the model account for intent specific traits, a better approach would be to train a separate model for each intent, referred as *Per-Intent Model* illustrated in Figure 4 (b).

Let η_i denote the intent for session i, and $K = 7$ be the total number of intents, i.e. $\eta_i \in [1, K]$. We divide the session data ($D = \{(x_i, y_i)\}$) into K different intent specific groups and train K different SAT prediction models, one for each intent:

$$P(y_i = c|x_i) = logit^{-1}(\theta_k^T x_i + b_k) \tag{4}$$

with (θ_k, b_k) representing the parameters of the k-th intent SAT model.

While the Per-Intent model is able to capture intent specific intricacies of the different interaction signals, it suffers from a few drawbacks. It prohibitively collect data for each intent separately, and there may not be enough data to learn a meaningful model for each intent separately. This problem gets compounded when platform changes encourages users to use the platform for new intents. Furthermore, an intent specific model fails to benefit from shared learning of parameters and hence fails to leverage insights from the interplay between interaction signals and satisfaction data of other intents. We next present a hierarchical multi-level approach to address these shortcomings.

5.3 Multi-level Model

Both the Global and Per-Intent SAT prediction models introduced suffer from various disadvantages. The Per-Intent SAT prediction model works by using just the local intent specific information, and assumes that the data is sampled from separate models for each intent, thereby ignoring information and insights from other intents. Furthermore, it can be rendered futile for intents with small labeled data. At the other extreme, the Global SAT prediction model ignores intent-level variations in the user interaction data and inadvertently suppresses variations that can be important.

To address these issues, we resort to *Multi-level Modeling* [6, 23], which represents a compromise between the two extremes of a single global prediction model and a separate prediction model for each intent. Multi-level models, illustrated in Figure 4 (c), are regression and classification models in which the constituent model parameters are given probability models. This implies that the model parameters are allowed to vary by group. This approach can be used to handle clustered or grouped data (e.g. different user intents), wherein each level is (potentially) a source of unexplained variability. Multi-level modeling provides a coherent model that simultaneously incorporates both individual and group level models.

Benefits of Multi-level Modeling

Modeling the satisfaction prediction model as a multi-level model offers numerous benefits. First, leveraging multilevel models allows us to account for the intent level grouping of the user session.

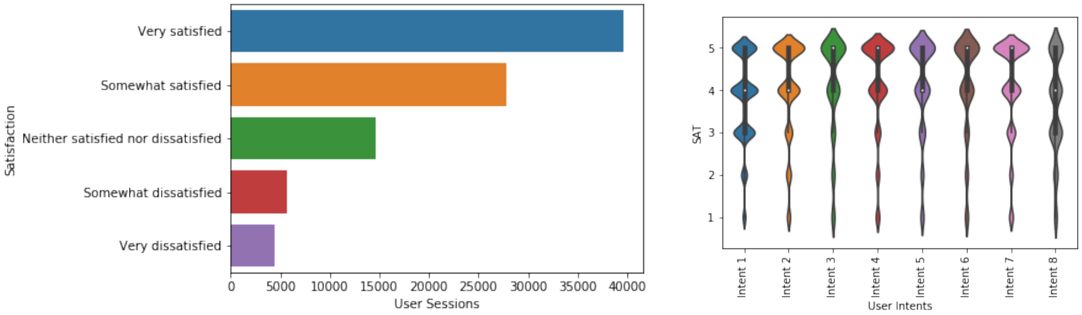


Figure 5: Analysis of user survey responses. Left: Distribution of user satisfaction labels obtained. Right: Distribution of SAT labels across the different user intents. SAT levels 1 to 5 refer to different levels from Very Dissatisfied (1) to Very satisfied (5).

Second, they facilitate incorporating both individual session level as well as intent group level effects on user satisfaction, thereby allowing for variability in user interaction behavior across different intents. Third, by assuming that the intent group level effects come from a common distribution shared across all intents, they facilitate information sharing across different intents. This can help in improving the accuracy and predictive performance for intents with relatively little data.

Intent-aware Multi-level Model

We model the satisfaction prediction problem with a multi-level logistic regression model with the different user intents serving as groups for the hierarchical model. The Multi-level Model can be written as:

$$P(y_i = c|x_i) = logit^{-1}(\theta^T_{\eta_i} x_i + b_{\eta_i}) \tag{5}$$

with η_i denoting the intent group for the i-th user session, and x_i denoting the user interaction signals extracted for the session. In our hierarchical model, parameters are viewed as a sample from a population distribution of parameters.

$$\begin{pmatrix} \theta_\eta \\ b_\eta \end{pmatrix} = \begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix} + \begin{pmatrix} \gamma^{int}_{0,\eta} \\ \gamma^{int}_{1,\eta} \end{pmatrix} \tag{6}$$

$$\begin{pmatrix} \gamma^{int}_{0,k} \\ \gamma^{int}_{1,k} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma^{int} \right) for\ k = 1, \dots, K \tag{7}$$

Thus, we view them as being neither entirely different (as was the case in the Per-Intent model) or exactly the same (as was the case in the Global Model). Figure 4 pictorially depicts the main difference between the parameter sharing across the three different satisfaction prediction models investigated in this work.

6 EXPERIMENTAL EVALUATION

We perform an evaluation to compare the different satisfaction prediction models, consider the role of different interaction features and present results on a large scale real world dataset.

6.1 Dataset

We work with real world user data from Spotify, a large music streaming service, and conduct a large scale in-app survey to collect judgments about intents and user satisfaction. We trigger the in-app survey to a random sample of over 3 million users, and observe a response rate of 4.5%. In total, we received responses from over 116000 users, resulting in over 200K judgments about intents and satisfaction combined. For each session, we collected back-end logs of user interactions with the slates of recommendation and extracted data for all the different interaction signals mentioned in Table 1. For each session, we use the intent selected by the user as the session intent and use the satisfaction prediction label given by the user to train and evaluate our satisfaction prediction models.

Owing to the *response bias*, i.e., the respondents of the survey are likely to be users with a positive bias toward the system, we would end up with a biased dataset for training, which would lead to unreliable prediction estimates. We mitogate this by oversampling the minority class by synthetic minority over-sampling technique [4]. This results in a balanced dataset with a healthy distribution of labels across both satisfaction and dissatisfaction cases.

6.2 Analysis of Survey Results

We begin with an analysis of the survey responses by understanding the degree to which users are satisfied overall, and across different intents. Figure 5 (left) presents the overall distribution of user satisfaction. We observe that most users are either *Very satisfied* or *Somewhat satisfied* with the slates recommended to them in their current session with over 33% user sessions being judged as very satisfying. On the other extreme, over 12000 user sessions were reported to be very or somewhat dissatisfying. A relatively larger number of user sessions were judged as neutral.

To gauge satisfaction for each intent, we plot the satisfaction label distribution separately for each intent in Figure 5 (right). Similar to overall satisfaction, we observe that more user sessions are satisfying than they are dissatisfying. Users with intents 3, 6 and 7 were more satisfied with the slate recommendations shown to them, than users pursuing other goals, which highlights the efficacy of the system in facilitating users to discover new music for now (Intent 3), playing music in background (Intent 6) and saving music for future (Intent 7). However, the distribution of the different levels of satisfaction differs across intents. As can be seen in the violin plot, a substantial number of user sessions were judged as neutral on satisfaction scale for Intent 8, whereas the distribution is skewed more towards the top-most level of satisfaction for Intents 3 and 6.

Considering such differences in distribution of satisfaction across intents helps system designers in identifying those intents where recommendation system under-performs, an insight which instead would have been hidden when considering overall user satisfaction.

Table 4: Comparing satisfaction prediction performances across different prediction models: Global, Per-Intent and Multi-level. * indicates statistical significant ($p \leq 0.05$) using paired t-tests compared to the Global + intent Model.

		Overall			Dissatisfaction			Satisfaction		
Method	**Accuracy**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Global w/o intent	0.55	0.56	0.56	0.56	0.57	0.48	0.53	0.54	0.63	0.59
Global w/ intent	0.57	0.58	0.58	0.57	0.59	0.52	0.55	0.56	0.63	0.60
Intent 1	0.675	0.68	0.68	0.67	0.66	0.73	0.69	0.70	0.62	0.65
Intent 2	0.663	0.67	0.67	0.67	0.66	0.71	0.68	0.68	0.62	0.65
Intent 3	0.672	0.67	0.67	0.67	0.69	0.67	0.68	0.65	0.68	0.67
Intent 4	0.678	0.68	0.68	0.68	0.67	0.71	0.69	0.69	0.64	0.66
Intent 5	**0.814***	**0.81***	0.81	0.81	0.80	0.84	**0.82***	0.83	0.79	0.81
Intent 6	0.760	0.77	0.76	0.76	0.73	0.83	0.77	0.81	0.69	0.75
Intent 7	0.7771	0.77	0.77	0.77	0.75	0.81	0.78	0.79	0.74	0.76
Intent 8	0.769	0.78	0.77	0.77	0.71	**0.87***	0.78	0.85	0.67	0.75
Multi-Level	0.804	0.80	0.79	0.79	0.81	0.79	0.79	0.79	**0.81***	**0.81***

This further motivates the need for incorporating intent information when developing and interpreting user satisfaction metrics.

6.3 Predicting User Satisfaction

We next investigate the extent to which the proposed interaction signals and the predictive models are able to predict user satisfaction. We consider the satisfaction label provided by the user during an in-app survey as ground truth label, and consider a user to be satisfied if their response was either *Very Satisfied* or *Somewhat Satisfied*. In all other cases, the user session was labeled as dissatisfied. Such binarization of user satisfaction isn't ideal, but we contend that it is a good starting point and has been extensively used in prior industrial research on understanding and predicting user satisfaction [17, 27, 39].

In line with prior work on predicting binary user satisfaction, we use four different evaluation metrics: accuracy, precision, recall and F-score. Table 4 presents detailed results comparing the Global, Per-Intent and Multi-level Model on these four metrics. We additionally divide the overall prediction performances and separately consider the performance in predicting satisfaction and dissatisfaction, since understanding them (SAT and DSAT) independently is useful in different use-case scenarios and at different stages of system development.

We observe that the Global Model is not able to predict user satisfaction much better than random, giving an overall accuracy of 57%. The Global Model without intent as a feature (i.e. Global w/o intent) performs worse, which highlights that incorporating intent information is useful in predicting satisfaction. The Per-Intent Model, on the other hand, gives much better prediction results than the Global Model. We observe an increased performance across all eight intents, with performance gains ranging from 10% to 24% for different intents. This confirms the main hypothesis of our work – considering intent information is crucial in accurately understanding and predicting user satisfaction. Since the interaction signals differ across different intents, the Global Model is unable to find a good universal relation between such signals and satisfaction. The Per-Intent Model creates a separate local intent-specific model and is able to capture the interplay between interaction signals and satisfaction to a much better extent.

We observe performance improvements in not just the accuracy, but across all metrics considered for both predicting satisfaction and predicting user dissatisfaction. Furthermore, we observe that the Multi-level Model performs significantly better than the Global Model (with intent) with over 20% improvement in prediction accuracy over the Global Model. Additionally, the Multi-level Model outperforms all but one Per-Intent Model, with performance improvement ranging from 4-14% in terms of prediction accuracy across different intents, while giving comparable performance to the last intent. This shows that the Multi-level Model is able to leverage insights from other intents, which in turn help boost the performance of other intents.

Additionally, we also observe that some intents are easier to predict than others, with intent 1 being the hardest to predict satisfaction for. Indeed, intent 1 sessions are of users who do not particularly have any intent in mind, and had a homepage session because it was the first page shown to them, so they could have any intent while interacting. This makes it harder for the system to predict satisfaction. It is important to note that an intent-specific model works best for certain intents (e.g. Intents 5 & 8) which suggests that the intent specific interaction data is powerful enough in itself, and doesn't gain much from shared learnign across intents. However, the shared learning across intents via the multi-level model is more generally beneficial to most intents and gives improved performance for 7 out of 8 intents. Moreover, maintaining separate intent level models is a demanding engineering task, while the shared multi-level model provides a sweet middle-ground whcih is relatively easier to manage and support from a large scale engineering perspective.

6.4 Importance of Interaction Features

We next investigate the different types of interaction signals considered and analyse their importance in prediction satisfaction. Table 5 presents accuracy results for different groups of interaction signals considered, for all three models. We omit displaying results for the Global - intent Model due to space constraints. We bold the best

Table 5: Comparing the importance of different interaction signals using different satisfaction prediction models using prediction accuracy. * and $^{\&}$ indicates statistical significant ($p \leq 0.05$) using paired t-tests compared to Downstream and Temporal baselines respectively.

Interaction Features	**Global**	**Intent 1**	**Intent 2**	**Intent 3**	**Intent 4**	**Intent 5**	**Intent 6**	**Intent 7**	**Intent 8**	**Multilevel**
Downstreams	0.528	0.556	0.568	0.555	0.525	0.599	0.593	0.641	0.551	0.605
Surface	0.522	0.548	0.546	0.512	0.572	0.550	0.556	0.544	0.595	0.576
Temporal	0.554	0.655	0.639	0.641	0.515	0.756	0.732	0.735	0.754	0.746
Metrics	0.541	0.602	0.593	0.578	0.663	0.711	0.662	0.639	0.712	0.723
Downstreams + Surface	0.523	0.564	0.565	0.560	0.602	0.582	0.586	0.597	0.561	0.598
Downstreams + Temporal	0.560	0.672	0.655	0.649	0.579	0.772	0.746	0.747	0.752	0.781
Temporal + Metrics	0.569	0.660	0.638	0.650	0.668	0.792	0.718	0.725	0.743	0.795
Downstream + Surface + Temporal + Metrics	**0.571**$^{*\&}$	**0.675**$^{*\&}$	**0.666**$^{*\&}$	**0.672**$^{*\&}$	**0.675**$^{*\&}$	**0.811**$^{*\&}$	**0.760**$^{*\&}$	**0.771**$^{*\&}$	**0.769**$^{*\&}$	**0.804**$^{*\&}$

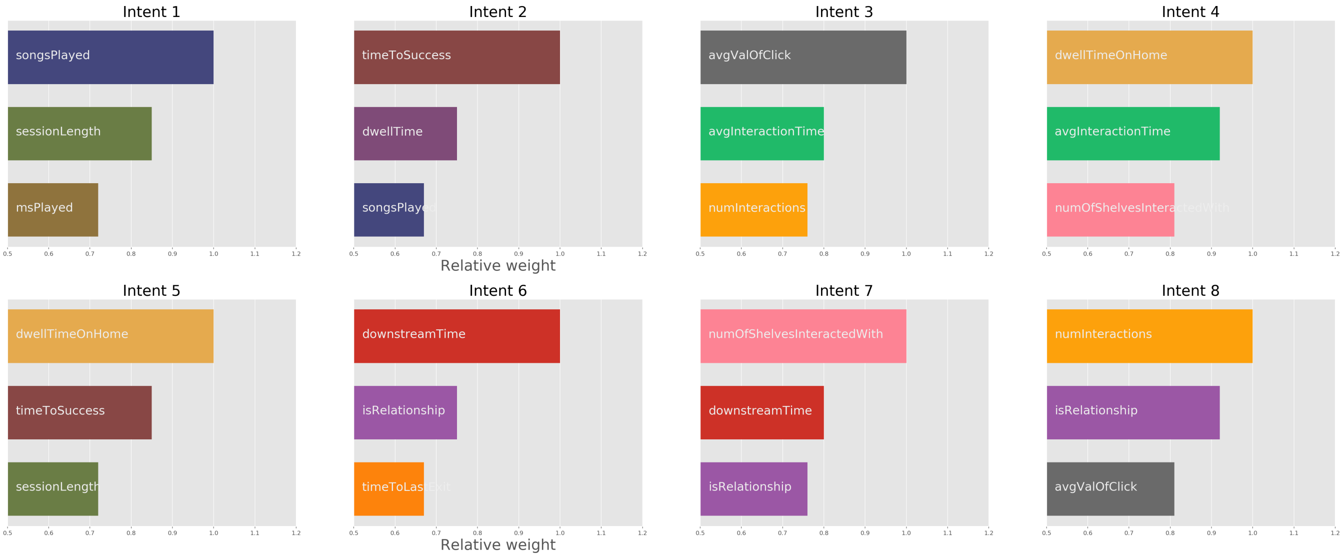


Figure 6: Relative weights of the top 3 predictive features across all eight intents.

performing features for each prediction model. We observe that whereas temporal signals are more important than other types of signals for most intents, surface signals such as number of slates interacted with and depth are important for Intent 4 (play music matching mood). This highlights that different types of signals are important for different intents, which supports the analysis presented in Figure 3 that showed interaction signals to differ based on user intent. Additionally, we observe that the Downstream group of signals performs better for Intent 7 (save new music or follow playlist), than it does for other intents, with 6-10% better predictive performance. Indeed, downstream signals help capture user activity related to saving music and exploring artists in detail, which is what Intent 7 is focusing on.

Finally, we observe a sharp increase in predictive performance when combining downstream and temporal signals, with over 10% performance jumps when compared to any individual signal group. We also observe that even in the case of different signal groups, Per-Intent and Multi-level models performs significantly better than the Global Model. Finally, combining different signals helps in improving satisfaction prediction accuracy by over 10%-15% for most intents.

Importance of signals across intents. To gain insights into which interaction signal is most useful across different intents, we present the top three signals and report their relative feature weight across all eight intents in Figure 6. As expected, we observe that the signals are important to varying degree across different intents. For the case where the user wants to quickly access saved music, signals like time to success and dwell time are most informative, while they are not informative for cases where the user wants to play music in the background. For intents where users wish to explore artists in more detail, signals involving the users building relationships (i.e. saved or downloaded tracks) are shown to be more important.

The variation of the most informative signals across the different intents highlight the fact that different signals are indeed important to different extent for different use cases. These results clearly emphasize the need for modeling intent while interpreting user interactions for predicting user satisfaction.

7 CONCLUSIONS

Given the *query-less paradigm* of slate recommendation, it becomes non-trivial to understand user intents. Based on a mixed-methods approach composed of interviews, in-app survey and non-parametric clustering, we identified eight key user intents, and experimentally demonstrated the importance of explicitly considering these intents when predicting user satisfaction. Our results also indicate that different interaction signals are important to varying extent across intents. Furthermore, the significant improvement in prediction results advocate not only the need for grouping user sessions into intent groups for predicting satisfaction, but also for shared learning across all intents via the Multi-level Model. We contend that the methodology adopted in this work to identify user intents: interviews, in-app survey and quantitative modeling would provide firm starting grounds for researchers to think about user intents in other recommendation scenarios where it is prohibitively challenging to extract unspecified user intents.

Limitations & Future Work. While our findings are more generally applicable beyond just music streaming domain, a few interesting challenges remain, which serve as useful directions for future work. While we leveraged the intent labels provided by survey responders, these labels are not readily available at run time for user sessions. This motivates the need for developing an intent prediction module to predict the intent for each user session. The current study also operates under the assumption of each session having one intent, which may not always be correct, since user intents might change within a session. Development of intent boundary detection or intent switching prediction modules would help in eliminating this assumption. Additionally, the model would need to cater to intents changing over time, and across contexts. Finally, we did not incorporate user specific intricacies in our formulation, but we believe that user interaction is conditional on user groups, and differ user segments interact differently. Incorporating such user level idiosyncrasies would be a fruitful direction.

REFERENCES

[1] Biswarup Bhattacharya, Iftikhar Burhanuddin, Abhilasha Sancheti, and Kushal Satya. 2017. Intent-Aware Contextual Recommendation System. In *Data Mining Workshops (ICDMW), 2017 IEEE International Conference on*. IEEE, 1–8.

[2] David M Blei and Peter I Frazier. 2011. Distance dependent Chinese restaurant processes. *The Journal of Machine Learning Research* 12 (2011), 2461–2488.

[3] Olivier Chapelle, Shihao Ji, Ciya Liao, Emre Velipasaoglu, Larry Lai, and Su-Lin Wu. 2011. Intent-based diversification of web search results: metrics and algorithms. *Information Retrieval* 14, 6 (2011), 572–592.

[4] 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 (2002), 321–357.

[5] 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, 55–64.

[6] Andrew Gelman and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

[7] Samuel J Gershman and David M Blei. 2012. A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology* 56, 1 (2012), 1–12.

[8] Carlos A Gomez-Uribe and Neil Hunt. 2016. The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)* 6, 4 (2016), 13.

[9] Qi Guo and Eugene Agichtein. 2012. Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior. In *Proceedings of the 21st international conference on World Wide Web*. ACM, 569–578.

[10] Qi Guo, Haojian Jin, Dmitry Lagun, Shuai Yuan, and Eugene Agichtein. 2013. Mining touch interaction data on mobile devices to predict web search result relevance. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 153–162.

[11] 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*. ACM, 195–204.

[12] Ray Jiang, Sven Gowal, Timothy A Mann, and Danilo J Rezende. 2018. Optimizing Slate Recommendations via Slate-CVAE. *arXiv preprint arXiv:1803.01682* (2018).

[13] Diane Kelly et al. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval* 3, 1–2 (2009), 1–224.

[14] Diane Kelly and Nicholas J Belkin. 2004. Display time as implicit feedback: understanding task effects. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 377–384.

[15] Eugene Kharitonov, Craig Macdonald, Pavel Serdyukov, and Iadh Ounis. 2013. Intent models for contextualising and diversifying query suggestions. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, 2303–2308.

[16] 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, 895–898.

[17] Julia Kiseleva, Kyle Williams, Ahmed Hassan Awadallah, Aidan C Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Predicting user satisfaction with intelligent assistants. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 45–54.

[18] Bart P Knijnenburg, Niels JM Reijmer, and Martijn C Willemsen. 2011. Each to his own: how different users call for different interaction methods in recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. ACM, 141–148.

[19] Bart P Knijnenburg and Martijn C Willemsen. 2009. Understanding the effect of adaptive preference elicitation methods on user satisfaction of a recommender system. In *Proceedings of the third ACM conference on Recommender systems*. ACM, 381–384.

[20] Bart P Knijnenburg, Martijn C Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 441–504.

[21] Christoph Kofler, Martha Larson, and Alan Hanjalic. 2016. User intent in multimedia search: a survey of the state of the art and future challenges. *ACM Computing Surveys (CSUR)* 49, 2 (2016), 36.

[22] Joseph A Konstan and John Riedl. 2012. Recommender systems: from algorithms to user experience. *User modeling and user-adapted interaction* 22, 1-2 (2012), 101–123.

[23] Jennifer L Krull and David P MacKinnon. 2001. Multilevel modeling of individual and group level mediated effects. *Multivariate behavioral research* 36, 2 (2001), 249–277.

[24] Dmitry Lagun, Mikhail Ageev, Qi Guo, and Eugene Agichtein. 2014. Discovering common motifs in cursor movement data for improving web search. In *Proceedings of the 7th ACM international conference on Web search and data mining*. ACM, 183–192.

[25] Dmitry Lagun, Chih-Hung Hsieh, Dale Webster, and Vidhya Navalpakkam. 2014. Towards better measurement of attention and satisfaction in mobile search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 113–122.

[26] Yiqun Liu, Ye Chen, Jinhui Tang, Jiashen Sun, Min Zhang, Shaoping Ma, and Xuan Zhu. 2015. Different users, different opinions: Predicting search satisfaction with mouse movement information. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 493–502.

[27] 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, 737–746.

[28] Rishabh Mehrotra, Prasanta Bhattacharya, and Emine Yilmaz. 2016. Deconstructing complex search tasks: a bayesian nonparametric approach for extracting sub-tasks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 599–605.

[29] Rishabh Mehrotra, Imed Zitouni, Ahmed Hassan Awadallah, Ahmed El Kholy, and Madian Khabsa. 2017. User Interaction Sequences for Search Satisfaction Prediction. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 165–174.

[30] 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 (WSDM '18)*. ACM, New York, NY, USA, 547–555. https://doi.org/10.1145/3159652.3159714

[31] 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, 547–555.

[32] Peter Sunehag, Richard Evans, Gabriel Dulac-Arnold, Yori Zwols, Daniel Visentin, and Ben Coppin. 2015. Deep Reinforcement Learning with Attention for Slate Markov Decision Processes with High-Dimensional States and Actions. *arXiv preprint arXiv:1512.01124* (2015).

[33] Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miro Dudik, John Langford, Damien Jose, and Imed Zitouni. 2017. Off-policy evaluation for slate recommendation. In *Advances in Neural Information Processing Systems*. 3632–3642.

[34] Nava Tintarev and Judith Masthoff. 2007. Effective explanations of recommendations: user-centered design. In *Proceedings of the 2007 ACM conference on*

Recommender systems. ACM, 153–156.

[35] Sergey Volokhin and Eugene Agichtein. 2018. Towards Intent-Aware Contextual Music Recommendation: Initial Experiments. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 1045–1048.

[36] Sergey Volokhin and Eugene Agichtein. 2018. Understanding Music Listening Intents During Daily Activities with Implications for Contextual Music Recommendation. In *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval*. ACM, 313–316.

[37] Ryen W White, Wei Chu, Ahmed Hassan, Xiaodong He, Yang Song, and Hongning Wang. 2013. Enhancing personalized search by mining and modeling task behavior. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 1411–1420.

[38] Ryen W White and Diane Kelly. 2006. A study on the effects of personalization and task information on implicit feedback performance. In *Proceedings of the 15th ACM international conference on Information and knowledge management*. ACM, 297–306.

[39] Kyle Williams, Julia Kiseleva, Aidan C Crook, Imed Zitouni, Ahmed Hassan Awadallah, and Madian Khabsa. 2016. Detecting good abandonment in mobile search. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 495–505.

[40] Xiaohui Xie, Yiqun Liu, Maarten de Rijke, Jiyin He, Min Zhang, and Shaoping Ma. 2018. Why People Search for Images Using Web Search Engines. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)*. ACM, New York, NY, USA, 655–663. https://doi.org/10.1145/3159652.3159686

[41] Xing Yi, Liangjie Hong, Erheng Zhong, Nanthan Nan Liu, and Suju Rajan. 2014. Beyond clicks: dwell time for personalization. In *Proceedings of the 8th ACM Conference on Recommender systems*. ACM, 113–120.

[42] Yuchen Zhang, Weizhu Chen, Dong Wang, and Qiang Yang. 2011. User-click modeling for understanding and predicting search-behavior. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1388–1396.