Understanding and Evaluating User Satisfaction with Music Discovery

Jean Garcia-Gathright
Spotify
Somerville, MA
jean@spotify.com

Brian St. Thomas
Spotify
Somerville, MA
brianstt@spotify.com

Christine Hosey
Spotify
Somerville, MA
chosey@spotify.com

Zahra Nazari
Spotify
New York, NY
zahran@spotify.com

Fernando Diaz*
Microsoft Research
Montreal, QC
diazf@acm.org

ABSTRACT

We study the use and evaluation of a system for supporting music discovery, the experience of finding and listening to content previously unknown to the user. We adopt a mixed methods approach, including interviews, unsupervised learning, survey research, and statistical modeling, to understand and evaluate user satisfaction in the context of discovery. User interviews and survey data show that users' behaviors change according to their goals, such as listening to recommended tracks in the moment, or using recommendations as a starting point for exploration. We use these findings to develop a statistical model of user satisfaction at scale from interactions with a music streaming platform. We show that capturing users' goals, their deviations from their usual behavior, and their peak interactions on individual tracks are informative for estimating user satisfaction. Finally, we present and validate heuristic metrics that are grounded in user experience for online evaluation of recommendation performance. Our findings, supported with evidence from both qualitative and quantitative studies, reveal new insights about user expectations with discovery and their behavioral responses to satisfying and dissatisfying systems.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; **Music retrieval**; • **Human-centered computing** → *User studies*;

KEYWORDS

music recommendation, discovery, user behavior, mixed methods

ACM Reference Format:
Jean Garcia-Gathright, Brian St. Thomas, Christine Hosey, Zahra Nazari, and Fernando Diaz. 2018. Understanding and Evaluating User Satisfaction with Music Discovery. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3209978.3210049

*Work completed while employed by Spotify.

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA

ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210049

1 INTRODUCTION

Designing music information access systems requires understanding the diverse needs of users and their expectations of system performance. Such needs include mood-setting, social standing, or nostalgia. The growth of streaming platforms in the recent decade, however, has reshaped users' interactions with music access systems. In streaming platforms, users are provided with access to large repositories of audio content with only a small fraction familiar to them. This necessitates a new focus on one particular need: music discovery, which we define as the experience of finding and listening to content that is previously unknown to the user.

Research has supported the critical role of discovery in streaming platforms. Mäntymäki and Najmul Islam surveyed 374 Finnish music streaming users and identified discovery as users' main motivation to continue platform subscription [22]. Brown and Krause's survey of 440 music listeners provided additional support in discovery's role in improving user retention [2]. Additional survey and interview studies also demonstrated that discovery is an important need for music listeners [16–18]. Moreover, Laplante found that music discovery requires a 'different state of mind', suggesting that discovery deserves special treatment by music information access designers [15]. Leong and Wright found through interviews that discovery results in deeper social connection [21]. Lee and Price identified several personas in music consumption that focus on discovery, demonstrating that discovery is a complex, nuanced, and personal need [19].

While existing work suggests that discovery is important, user studies in the context of recommender systems showed that novel recommendations were negatively correlated with perceived quality [3, 7, 23]. Other user studies confirmed that, when novelty is desired, approaches based on popularity lead to suboptimal performance [14]. As system designers, we need a more nuanced understanding of user expectations surrounding novelty and discovery. We have limited knowledge about how users behave in a discovery context and, moreover, how these behaviors change when systems fail to satisfy user expectations.

We investigated the following questions:

- What are users' expectations in the setting of discovery-oriented recommendation?
- Which interactions with a recommendation system for music discovery can be used to estimate user satisfaction?
- How can data-driven models of user satisfaction be used to evaluate the performance of recommendation algorithms?

We utilized a mixed methods approach that combines user research with machine learning to understand user satisfaction with music discovery. We conducted five studies that explore different aspects of user experience and user satisfaction. First, we carried out face-to-face interviews with users to help us form hypotheses about user expectations of novel recommendations and what behaviors may be correlated with satisfactory and unsatisfactory experiences. Second, we validated these hypotheses at scale by using unsupervised learning to analyze logged interaction data with a music streaming platform. Next, we deployed a survey that explicitly asked users how satisfied they were with their personalized discovery recommendations. Then, we fit a statistical model of user satisfaction based on behavioral signals identified from the first round of analysis. Finally, we proposed and validated a set of online metrics for evaluating recommendation performance.

For the purpose of this work, we gathered data from a music streaming platform, where users were provided with a fixed set of thirty personalized discovery recommendations. These recommendations were updated weekly and contained tracks similar to the taste of the user, but were unknown to them previously. Users could interact with the recommendations by playing or skipping tracks. They could also navigate to the album or artist page for any of the tracks and listen to other tracks by the same artist. Even though the content of the recommendations changed each week, users could add any of the tracks to a playlist or save them to their library for future access. Users could listen within a user playlist, library, artist page, or album page. We refer to listening to the recommendations or tracks by the same artist within these contexts as "downstream listening."

We found that user needs and behaviors were dependent upon their goals; we identified and validated four user goals in music discovery. Inferring these user goals, in addition to normalizing interaction data per-user and capturing peak interactions at the track level, were informative for estimating user satisfaction. Moreover, we developed online metrics based on these findings that can be used to compare performance between two recommendation algorithms.

The paper is organized as follows. Section 2 provides an overview of related work on novelty, discovery, and user satisfaction. Sections 3-7 summarize the methods and results for each of the five studies: user interviews, unsupervised learning, survey, statistical model, and online metrics. Sections 8-9 synthesize our findings and point to future work.

2 RELATED WORK

Researchers in the recommender systems community have realized that good user experience requires considerations beyond relevancy of the items and accuracy of the algorithms. Herlocker introduced novelty and serendipity (both related to the concept of discovery) as dimensions for evaluating recommendation quality [11]. In the recommendation context, novelty is the extent to which the recommended items are unfamiliar to the user, while serendipity adds a "surprising" effect to novelty of the recommended items. Related to novelty and serendipity is the concept of diversity, which captures the variety of recommended content. Many researchers have proposed formalisms for defining and evaluating novelty, serendipity, and diversity. Vargas and Castells developed a framework for measuring novelty and diversity in terms of discovery (the degree to which an item is familiar to a user), choice (whether the item is consumed by the user), and relevance (whether the item is enjoyed by the user) [26]. Hurley and Zhang discussed novelty and diversity and their trade-offs with system accuracy, casting this trade-off as a multi-objective optimization problem [12]. Ge *et al.* proposed an evaluation metric for serendipity [9]; Adamopoulos developed a measure of "unexpectedness" and demonstrates its use on a MovieLens dataset [1].

While there have been many studies on formalizing and measuring novelty in recommender systems, user-oriented evaluation studies are less prevalent. Schedl recognized this and argued for more user-oriented research in music information retrieval to enhance personalization and context awareness [25]. User studies on novelty in recommender systems give conflicting evidence on the value and effectiveness of novelty in this setting. Lee and Price carried out face-to-face interviews on user experience with multiple streaming music platforms over several dimensions, including novelty/serendipity [20]. Users with more "adventurous" listening habits reported positive experiences with novel or serendipitous recommendations, though "discerning" users noted that their recommendations were often not sufficiently novel. Celma and Herrera conducted a user study comparing perceived quality of recommendation for collaborative filtering vs. audio-based recommenders [3]. They found that users gave higher scores for the collaborative filtering based recommender, even though it gave less novel recommendations. Matt *et al.* found that users reported greater perceived fit and enjoyment with serendipitous recommendations, but not novel recommendations [23]. In the domain of movie recommendation, Ekstrand *et al.* showed that novelty was negatively correlated with user satisfaction, especially for new users who had not established trust with the recommender system [7].

To better understand user satisfaction in web search tasks, researchers in the information retrieval community have investigated methods for estimating satisfaction from behavioral data. For example, Fox *et al.* examined which logged user behaviors with a web search system correlated with explicit user satisfaction [8]. Since then, many studies have sought to understand and measure search success by examining user interaction data. For example, Drutsa *et al.* used search behavior data to predict engagement metrics at the individual user level for the purposes of improving sensitivity of A/B testing [6].

Related to our work that frames discovery-oriented behavior in terms of user goals, Zhang *et al.* and White *et al.* improve search retrieval by modeling tasks rather than individual behaviors [28, 29]. A recent study by Volokhin and Agichtein catalogued common goals for listening to music, finding the distribution of goals change across demographic populations, and that goals are distinct from the activity being performed [27].

Unlike the previous works described here, which explored success and satisfaction for search retrieval and ranking, we apply these methods in the domain of recommender systems, with the overarching goal of understanding and measuring user satisfaction with personalized music discovery. Given the conflicting evidence on the value of novelty in the context of recommendation, we aim to gain an understanding of users' expectations of discovery-oriented recommendations, and how their interactions with a music streaming platform may give insight to their satisfaction.

3 INTERVIEWS

We began our process with user interviews to develop hypotheses around three key areas of inquiry:

(1) How and why people listen to personalized discovery recommendations,
(2) How users define a good versus a bad experience with discovery recommendations, and
(3) What specific behaviors indicate about the quality of the user experience.

3.1 Methods

We recruited 10 participants from the Greater Boston Area who had varying levels of engagement with our recommendations over the previous 10 weeks. Highly-engaged participants had listened 9-10 weeks. Medium-engaged participants had listened 5-8 weeks. Low-engaged participants had listened 1-4 weeks. Participants ranged in age from 21 to 41, and included 5 females and 5 males. Participants received a $75 gift card.

We conducted face-to-face 45-minute interviews with each participant individually. During these interviews, we asked participants about their music tastes, music streaming habits, and experience with their personalized discovery recommendations. More specifically, participants told us how and why they used the discovery recommendations, examples of good and bad experiences with them, how the quality of the experience had affected their behavior, and a deep dive around 17 behaviors they could perform with the discovery recommendations.

We analyzed the data from the interviews using a grounded theory inspired approach [10]. To begin, one researcher went through videos and transcripts, using an open-coding system, that allowed for flexibility and refinement of codes throughout the process. To ensure we preserved the participant's holistic experience, we documented relationships between codes within a user (in grounded theory, this documentation is called a memo). Like codes were then grouped into concepts and labeled (e.g., "good experience" or "background listening"). We then organized the concepts into categories that captured how and why participants used the playlist. Categories included expectations, goals, experiences, and behaviors. From these categories, we developed hypotheses (outlined below) that we later tested through survey and analysis of user interaction data.

3.2 Results

Several important themes emerged as hypotheses to guide our exploration of the relationship between user interactions and satisfaction. First, participants generally did not expect to like every recommended track, but they hoped to find something that they loved. In other words, participants were looking to find a new favorite artist or a new favorite track and were prepared to be exposed to music they did not like in pursuit of a new favorite. This is consistent with previous work by Laplante and Downie, which found that failing to find interesting content when trying to discover music was not always evaluated as a negative experience [16]. Thus, we hypothesized that loving a track would have a larger positive effect on user satisfaction than hating a track would have a negative effect on user satisfaction.

Second, participants described a variety of possible goals for their discovery recommendations and these goals shaped how they listened and which behaviors signaled a good versus a bad experience. Specifically, four overarching goals emerged from the interviews: play new music in the background, listen to new music now and later, find new music for later, and engage with new music.

Goal 1. For participants whose goal was to play new music in the background, listening to the recommendations was not their primary focus (e.g., they were working or exercising). A positive experience for this goal was characterized by no skipping, increased saving/adding, increased listening time, and returning more times per week. A negative experience was characterized by skips (especially of the same track), switching to a different feature more quickly than normal, decreased listening time, and returning fewer times per week.

Goal 2. Participants whose goal was to listen to new music now and later expressed interest in immersing themselves in the listening experience with the intention of also curating tracks for later consumption. A positive experience for this goal was characterized by increased saving/adding, a higher percentage of tracks listened to in the recommendations, increased plays of over half the track, and increased downstream listening. A negative experience was characterized by abandonment of the recommendations with a lower percentage of tracks listened to, decreased saving/adding, increased skipping (without an accompanying save/add) before the halfway point, and decreased downstream listening.

Goal 3. For participants whose goal was to find new music for later, they were focused on quickly collecting tracks to listen to in other contexts, so they often listened to only small snippets of each track. A positive experience for this goal was characterized by increased saving/adding, increased plays (of any length), and increased downstream listening. A negative experience was characterized by decreased saving/adding, decreased plays, and decreased downstream listening.

Goal 4. For participants whose goal was to engage with new music, the discovery recommendations were seen as a springboard to find new artists or genres to explore more deeply, so their behavior often led them away from the recommendations altogether. A positive experience for this goal was characterized by increased artist and album page views and increased downstream listening. A negative experience was characterized by decreased artist and album page views and decreased downstream listening.

Given that each goal had differing patterns of behavior for positive and negative experiences, we hypothesized that understanding a user's goal each time they listened to their recommendations would be important in understanding that user's satisfaction with the recommendations.

Third, while most user interactions indicated different things in different contexts (e.g., skipping sometimes signaled dislike a track and sometimes was purely navigational), several interactions emerged as consistently signaling positive experiences: saving a track, adding a track to a playlist, viewing an artist or album page, and downstream listening. We therefore hypothesized that these interactions would be useful in predicted user satisfaction with the recommendations.

Finally, participants judged the quality of their experience based on their own previous experience. That is, saving 5 tracks could signal a particularly good week for a user who typically saves 1 track or a particularly bad week for a user who typically saves 15 tracks. Therefore, we hypothesized that metrics normalized per user would be important for predicting user satisfaction with the discovery recommendations.

4 UNSUPERVISED LEARNING

Having identified a set of four goals through face-to-face interviews, we validated these goals at scale by clustering and interpreting user interaction data. Additionally, we wanted to segment user interactions over the course of a day into types; these could then be used as features in a statistical model of user satisfaction.

4.1 Methods

We identified and gathered interaction data that user research suggested may be useful for estimating user satisfaction. Interaction types included: page views, plays, skips, saves, downstream behavior, and variants of these. In response to the hypothesis from user research that per-user normalization was important, we also derived features describing users' deviation from their average behavior over an 8-week period. For each interaction type for the last 8 weeks, we calculated the average number of interactions over the weeks during which the user interacted at all with the discovery recommendations. We then subtracted the average behavior from the current week's behaviors, giving us additional "normalized" features. There were 40 features in total (20 interaction features and 20 normalized features). The full set of features is given in Table 1.

Our dataset consisted of interaction data for a random sample of all users who interacted with the discovery recommendations over the course of a day; the sample size was approximately 140,000 users. To correct for the presence of outliers (users with extremely high engagement), we log-scale each feature prior to fitting the model.

We ran K-means clustering over the 40-dimensional data, using the "elbow test" to set the value of k. We created a column-standardized heatmap of the cluster centers with respect to each interaction feature, and manually mapped the clusters to goals identified in user interviews. We created another heatmap of the normalized features to determine users' deviation from their average behaviors.

4.2 Results

Eighteen clusters were discovered via K-means clustering. Patterns corresponding to the four user goals were identified in several of the clusters. Figure 1 shows the heatmaps for four of the clusters, each corresponding to a different goal. The first cluster of users engaged primarily in streaming, with very little skipping. We interpreted this cluster as "listen to new music in the background" (Goal 1). The second cluster was characterized by streaming, skipping, and saving behaviors; it corresponded to "listen to new music now and later" (Goal 2). In the third cluster, we observed few completed plays compared to skips; however, because we did observe saving and adding, we concluded these users' goal was to "find new music for later" (Goal 3). The fourth cluster was engaged exclusively in downstream listening at the artist and track level. These users were "exploring new music" (Goal 4). The remaining clusters also

Table 1: Set of interaction features used to cluster users by goal.

Within rec.	Description
completed plays	Tracks played to the end.
skips	Tracks skipped at any point.
quick skips	Tracks skipped within the first 30 seconds.
total plays	Total played and skipped tracks.
click skips	Tracks skipped by clicking on another track.
button skips	Tracks skipped by using the forward button.
continuations	Tracks completed by continuing to the next track.
rec. views	Visits to the discovery recommendations.
adds	Tracks added to a playlist.
saves	Tracks saved to the user's library.
Downstream	**Description**
album views	Visits to a track's album page.
artist views count	Visits to a track's artist page.
artist	
completed plays	Downstream plays of a track's artist.
skips	Downstream skips of a track's artist.
quick skips	Downstream quick skips of a track's artist.
total plays	Total downstream plays and skips of a track's artist.
track	
completed plays	Downstream plays of a track.
skips	Downstream skips of a track.
quick skips	Downstream quick skips of a track.
total plays	Total downstream plays and skips of a track.

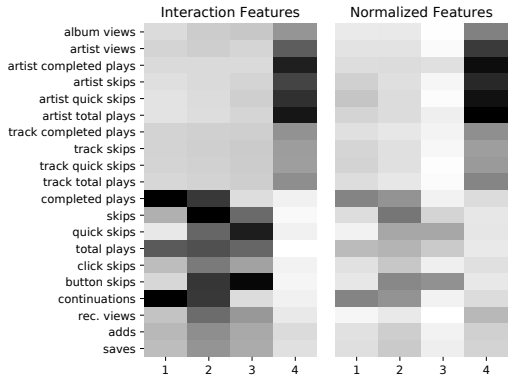

Figure 1: Heatmap of interaction features and normalized features for four behavioral clusters. User goals represented are: 1) listen in the background, 2) listen to new music now and later, 3) find new music for later, 4) explore new music.

corresponded to these four goals, but varied by level of engagement. There were also clusters depicting low engagement behaviors (such as viewing the recommendations but not playing them).

We examined the normalized features to see if they provided clues as to whether or not these users were satisfied. We observed that normalized features for the first, second, and fourth clusters showed higher than usual engagement, suggesting satisfaction. For the third cluster, we observed more skips than usual but no increase in saves, suggesting an unsatisfactory experience for the goal of finding new music for later. These observations justified the use of normalized features as inputs to our statistical model of satisfaction.

5 SURVEY

To further investigate the relationship between user goals and satisfaction, we conducted a survey to gather data directly from users. More specifically, we wanted to understand the importance of the goals uncovered in user interviews and validated at scale through unsupervised learning, along with their relationship to user satisfaction. To do so, we measured overall satisfaction with the discovery recommendations, satisfaction with the recommendations in the current week, and potential drivers for each. We analyzed a portion of the survey results for this report and are currently preparing a complete treatment of the results for future presentation.

5.1 Methods

We surveyed 18,547 users in four waves. To qualify for the survey, participants had to have streamed at least 5 minutes or 3 tracks of their personalized discovery recommendations in the week they received the survey. We chose these cutoffs to reduce the risk of users forgetting that they had listened to the recommendations that week. We emailed the survey to 100,000 users per wave, with the goal of getting at least 2,000 responses per wave. Before finalizing the survey, we ran a usability study with 4 users to ensure that we understood how respondents were interpreting and answering questions. Participants in the usability study received a $30 gift card. Survey respondents did not receive any compensation.

The survey was organized into three sections: 1) overall satisfaction and drivers, 2) user goals, and 3) current week satisfaction and drivers. Note that we included optional open-ended questions for respondents to explain their overall and current week satisfaction ratings. We did not analyze these open-ended responses.

5.1.1 Overall satisfaction and drivers. To measure and understand overall satisfaction with the personalized recommendations, we borrowed from the HaTS framework, which cites usability, usefulness, and delight as key drivers of product-level satisfaction [24]. First, we asked respondents about their overall satisfaction ("*In general, how satisfied or dissatisfied are you with your experience using [recommendations]?*"). We followed up with a question about usability ("*How easy or hard is it to use [recommendations]?*"), usefulness ("*How well or poorly does [recommendations] meet your needs?*"), and delight ("*How well or poorly does [recommendations] match your music tastes?*"). Each question used a 5 point Likert scale for response options.

5.1.2 User goals. To understand the importance of the goals uncovered in user interviews and how they relate to user satisfaction, we asked respondents why they listen to their personalized discovery recommendations and provided 7 response options. We derived the response options from the 4 overarching goals outlined in Section 3.1.2. We made some goals more granular to be more comprehensible to the respondents. Response options included "*to listen to new music*", "*to have music on in the background that won't distract*", "*to play music that fits a specific activity*", "*to add new music to playlists or library to listen to later*", "*to find new artists to explore more deeply*", "*to find new genres to explore more deeply*", and "*none of these.*" Respondents were allowed to select as many goals as they wanted and had the opportunity to add additional reasons in an open-ended follow-up. We then asked respondents to rank order the goals they had selected from most important to least important.

5.1.3 This week satisfaction and drivers. We asked respondents about their experience with their personalized discovery recommendations in the current week. To start, we asked whether or not the respondent had listened that week. If respondents selected "*no*", then the survey ended. If respondents selected "*yes*", then we asked them to think about a recent time they had listened and tell us what else was going on (e.g., where they were, what else they were doing, etc.) This was meant to refresh the respondents' memories of some specifics surrounding their experience. We then asked respondents about satisfaction with the current week ("*Thinking specifically about this week's [recommendations], how satisfied or dissatisfied were you?*"), how well the recommendations satisfied each of their goals ("*How much did this week's [recommendations] help you [insert goal]?*"), their understanding of the recommendations ("*Did you understand why specific songs were chosen for [your recommendations] this week?*"), and the extent to which the tracks fit their musical preferences ("*Did the songs on [your recommendations] this week fit your music tastes?*"). We used a 5 point Likert scale for each of these questions. Finally, we asked about annoyance with track selection ("*Did at least one song or artist annoy you on [your recommendations] this week?*"), and loving track selection ("*Did you love at least one song or artist on [your recommendations] this week?*"). Respondents could select either "*yes*" or "*no*" in response to these two questions.

5.2 Results

We collected 18,547 responses to the end-of-week survey. Figure 2 shows the levels of satisfaction with the discovery recommendations, both during the current week and overall. Respondents reported high satisfaction, with 65% of users indicating 4+ satisfaction this week and 81% of users indicating 4+ satisfaction overall.

Pearson correlations with satisfaction this week and overall are given in Table 2. We found that users were more satisfied when they felt the recommendations matched their tastes, met their needs, and helped them achieve their goals. Ease of use, understanding why songs were chosen, and reasons for use were not correlated with satisfaction.

Consistent with findings from user interviews, discovering at least one song the user loved was correlated with satisfaction this week, while being annoyed by at least one song was weakly negatively correlated. We found that 59.8% of those who were annoyed

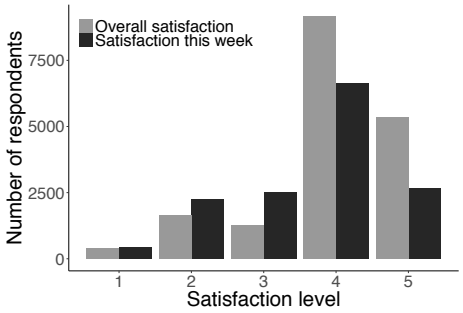

Figure 2: Survey responses for overall satisfaction and satisfaction during the current week. 1=very dissatisfied, 5=very satisfied.

Table 2: Pearson correlations with survey responses and satisfaction this week and overall.

Survey question	This week	Overall
Overall Satisfaction and Drivers		
Satisfaction	0.465	1
Ease of use	0.116	0.167
Meets needs	0.550	0.607
Fits taste in general	0.558	0.567
User Goals		
Has goal: add for later	0.016	0.069
Has goal: artist exploration	0.040	0.086
Has goal: background	0.046	0.032
Has goal: genre exploration	0.044	0.061
Has goal: new music right now	0.073	0.086
Has goal: specific activity	0.038	0.034
This Week Satisfaction and Drivers		
Satisfied this week	1	0.465
Achieved this week: add for later	0.608	0.356
Achieved this week: artist exploration	0.569	0.341
Achieved this week: background	0.354	0.187
Achieved this week: genre exploration	0.456	0.296
Achieved this week: new music right now	0.489	0.305
Achieved this week: specific activity	0.571	0.242
At least one song annoyed	-0.206	-0.152
At least one song loved	0.403	0.245
Songs fit music taste	0.632	0.407
Understands why songs are chosen	0.255	0.188

by at least one song reported satisfaction with the recommendations that week, compared to 75.6% of those who were not annoyed by a song ($t = -16.907, df = 7380$). Similarly, 74.2% of those who found at least one song they loved reported satisfaction, compared to 29.0% who did not find a song they loved ($t = 42.35, df = 3451$). The log-odds effect sizes of these two experiences on the reporting satisfaction are summarized in Table 3.

We asked each respondent to order the goals they listed for using the discovery recommendations (no ties were permitted). Separately, we asked if they achieved each of these goals that week on a five point Likert scale. We analyzed the relationship between

Table 3: Logistic regression coefficients answering either "satisfied" or "very satisfied" against having loved at least one song and being annoyed by at least one song. Coefficients are increase in log-odds.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.29	0.06	-5.27	0.0000
Annoyed	-1.01	0.05	-19.15	0.0000
Loved	2.11	0.06	38.04	0.0000

achievement of these goals this week and satisfaction with the recommendation system this week. Specifically, we modeled this using logistic regression where we coded responses as "1" if they chose a top two response and "0" otherwise. We observed that respondents provided varying numbers of goals. We grouped respondents by the number of goals provided to control for any confounding effects and estimated one model for each group. In Figure 3, we report the regression coefficients for each group. We omitted those who responded having all 6 goals due to insufficient sample size ($n = 163$).

The intercept log-odds of satisfaction for each of the groups were similar, suggesting that the number of goals did not influence the level of satisfaction when none of the goals were met. Across all groups, meeting the most important goal provided the largest boost to satisfaction compared to lower ranked goals. Amongst these secondary goals, we observed relatively comparable contributions to satisfaction. So, although all goals were important, music discovery systems must successfully identify and satisfy a user's primary goal.

6 STATISTICAL MODEL

We employed a statistical model using both user goals and interaction data to model satisfaction collected from the survey. The aim of this model was to find structure in how achieving user goals from unsupervised learning (Section 4) related to satisfaction.

6.1 Methods

From the survey, we had 17,643 respondents complete the overall satisfaction question. For each respondent, we collected the interaction data for each track in the discovery recommendations.

We wanted to choose a model that was both interpretable and had predictive power. Generally, gradient boosted regression trees have been shown to be successful in this type of optimization, while maintaining high interpretability through the tree structure. Specifically, this approach has been successful in user engagement prediction problems [6]. For this paper, we trained our tree models using the XGBoost software package [4].

Each row in the raw data set consisted of the user, track, position in the recommendations, and a daily time series for the interaction data in Table 1 for each track. Each row also contained a daily time series for the features described in Table 4, as well as a daily time series for the user's cluster assignment. Usage of the platform not specific to the discovery recommendations was also collected.

For each time series, we introduced aggregations as features in the model. For the interaction features (features that are not the cluster assignments), we computed the maximum of the time series,

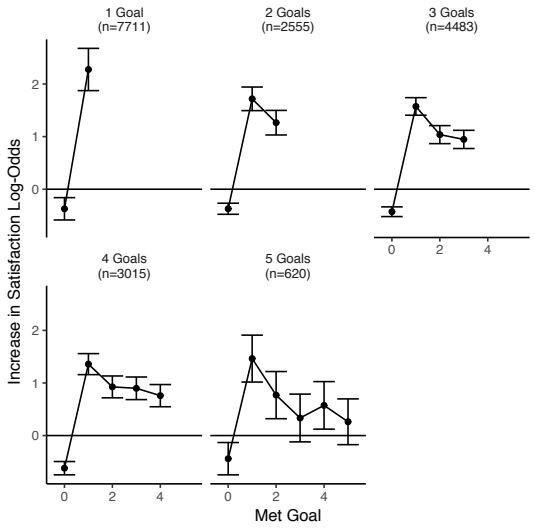

Figure 3: Logistic regression coefficients of reported satisfaction this week against success in ranked goals. The coefficients indicate the increase in log-odds of reporting satisfaction when successfully achieving the goal of corresponding rank. The zeroeth coefficient–or intercept–corresponds to being unsuccessful in all reported goals. Error bars indicate 95% confidence intervals.

the non-zero minimum of the time series (unless the value was always 0, in which case we take 0 as both the min and max), the total sum of the time series, the number of days the interaction occurred, the number of days the interaction occurred more than a threshold (which was defined per interaction type on a case by case basis), and the number of days the interaction occurred at all. We used mean squared error (MSE) as our training objective and model selection criteria.

We performed model selection in two steps. First we did feature selection by comparing against a 10% hold out and used a small grid for parameter tuning to assess model performance. Once features were selected, we performed 10 fold-cross validation to choose XGBoost parameters. We constructed hold outs by user (as opposed to user-track record).

We performed model selection against a 10% hold out of users by comparing the full feature set, combinations of feature subsets, and weekly aggregates of the features. Feature sets that included full time series tended to overfit. The winning feature set was created by taking the max, min, sum, and normalization of each interaction time series, as well as both the number of days the interaction occurred at all and over a threshold. In addition, user level features in the model were the sum of each cluster assignment and the across platform features. Goodness of fit results are in Table 5.

Through cross validation, we chose a maximum tree depth of 6 for 50 rounds, $\lambda = 1.5$, $\eta = 0.1$, $\gamma = .1$, and sampling 60% of columns by tree.

Table 4: Set of additional features for statistical model. Across platform features refer to the user's general use of the platform. Interaction Time Series features refer to user-track interactions with songs and artists on the discovery recommendations over time.

Across Platform Interactions	Description
play contexts	Number of play contexts streamed
tracks	Number of tracks streamed
platforms	Number of platforms used (e.g. iOS, Android)
plays	Number of plays
seeks forward	Number of seeks forward
seeks back	Number of seeks back
quick skips	Less than 30 second plays
shuffles	Plays in shuffle mode
skips	Skips of a track
Time Series Aggregate	**Description**
max	Maximum of daily time series
min	Non-zero minimum of daily time series
sum	Total sum of the daily time series
normalized	Difference between the observed sum and historical average sum for this user
occurred at all	Number of days interaction occurred one or more times
occurred over threshold	Number of days interaction occurred over a threshold

Table 5: Model selection through 10-fold cross validation. The full model is all time series and descriptive summaries, and total platform usage. The selected model is a subset of descriptive summaries and total platform usage.

Model	Mean MSE	Std. dev.
Intercept	0.95	0.04
Full	0.86	0.04
Selected	0.28	0.02

6.2 Results

We chose the model and training parameters through cross-validation and used the whole dataset when generating the model reported here. However, to assess model performance, we reported results fitting a 10% hold out sample of a model trained with the same parameters. The MSE of this model was 0.28. In Figure 4 we show the model residuals and marginal distributions. We saw that while our model generally scored well, it struggled to differentiate low scores (i.e. what makes a 3 vs. a 2). The match of marginal densities of high (4-5) and low (1-3) scores suggested that the tree was useful for inferring general patterns in the interaction data.

We looked at the importance of features in the regression trees through the gain (impact of the feature towards the loss function),

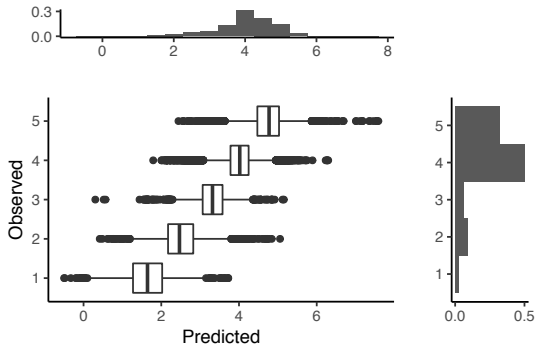

Figure 4: Performance of user-track scoring model on 10% hold out of users. The marginal densities of the observed and predicted ratings are shown on the side.

and the weight (number of times the feature is split on in the ensemble). Our model contained many features, however they generally belonged to groups in terms of the behavior and the aggregate type of the feature. The prevalence of the features are described in Table 6. The cluster models gave the largest gain as a feature type by a large margin, with the individual cluster features having a fairly flat impact. The most important types of feature by gain were the normalized in-recommendation and normalized downstream features, which captured how different the user-track interactions were from historical averages. Together they gave 13.9% of the gain in the model. The normalized, max, and min feature types all provided more gain than the sum features.

7 ONLINE METRICS

The ultimate goal of measuring user satisfaction is to be able to improve it. Online controlled experiments provide a tool that helps us make informed decisions to enhance the user experience. In this section, we introduce a suite of metrics for online evaluation of discovery recommendations and evaluate them using a corpus of a previously run online experiment.

7.1 Methods

Based on our previous analyses, we identified eight interaction signals as indicators of strictly positive user experience, irrespective of user goals, for interacting with the discovery recommendations. We used these eight metrics as our measures for online evaluation of recommended content: album views count, artist views count, library saves count, playlist adds count, downstream completed plays, album view duration, artist view duration and downstream listening.

Additionally, we used the normalized version for each metric, where the average interactions of a user in the four weeks prior to the experiment week was deducted from the values in the week of the experiment.

7.2 Results

An effective use of online experiments requires metrics that satisfy two fundamental qualities: *directionality* and *sensitivity* [5]: The direction of a good metric should align with the user's experience. In other words, a positive change in user experience should cause a change in the metrics in a consistent direction. A good metric should also be sensitive enough to capture meaningful changes in user experience. However, it should not be overly sensitive to result in false positives.

In order to evaluate directionality and sensitivity of our proposed metrics, we acquired the corpus of interaction data during a previously run A/B experiment on the discovery recommendations system. An important characteristic of this corpus was that we could confidently claim that the treatment condition led to a more desirable user experience for the users than the control condition. Therefore, we labeled the treatment condition as "enhanced" condition and the control condition as "baseline" condition.

Directionality: Since our metrics were indicators of a positive experience for the users, we expected them to be higher for users in the enhanced condition compared to the baseline condition. The results from our corpus, shown in Figure 5, confirmed this.

Sensitivity: To evaluate sensitivity, we need to consider two scenarios: 1) when there is a meaningful difference in user experience between control and treatment conditions and 2) when treatment

Table 6: Feature importance measured through feature weight and feature gain. The measures are given by percentage of the total for all features in the model. For each of the feature types, the percentage of some top performing features by gain is also given. In-recommendation: Track interactions inside the discovery recommendations. Downstream: Track interactions outside the discovery recommendations. Navigation: Album, artist, and recommendation page views. Platform: Interactions on the platform unrelated to the discovery recommendations. Cluster: Ad hoc description of cluster based on heat map of interactions.

Feature	Gain (%)	Weight (%)
cluster	70.6	54.1
background	1.3	0.7
add for later 1	1.3	0.6
add for later 2	1.2	0.4
normalized	15.4	30.7
in-recommendation	8.6	11.2
downstream	5.3	13.9
navigation	1.6	5.5
max	5.6	7.5
platform	3.0	6.4
in-recommendation	2.1	0.4
navigation	0.3	0.6
min	6.2	6.4
downstream	3.2	4.5
in-recommendation	1.6	0.3
navigation	1.0	1.3
platform	0.3	0.3
sum	2.3	1.0
in-recommendation	1.6	0.9
platform	0.8	0.1
occurred at all	0.3	0.1
downstream	0.3	0.1

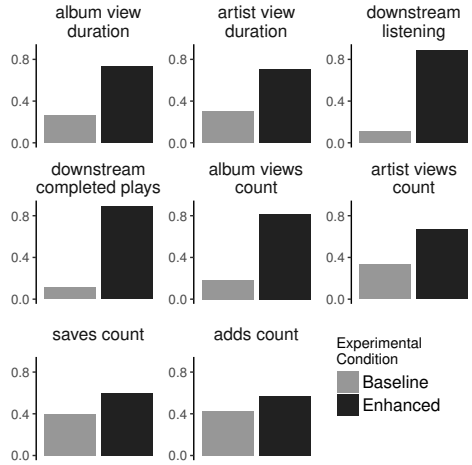

Figure 5: Density for positive interaction signals between baseline and enhanced conditions. Density used to protect user data.

and control conditions are the same. In the first scenario, we expected to see a statistically significant difference in all of our metrics between enhanced and baseline conditions, and in the second scenario, we expected to see no significant difference in our metrics when the control and treatment were the same (A/A experiments). Results from both analyses are in Table 7.

The first column contains our eight proposed metrics. The A/A and A/B tests were conducted for the eight metrics. For the normalized A/A test and normalized A/B test were run on the normalized version of the metrics. When treatment and control were different (A/B test and normalized A/B test), we saw significant ($p < 0.01$) differences across all metrics except album view duration, where significance was marginal ($p < 0.07$). When control and treatment were the same (A/A test and normalized A/A test), we saw no significant differences.

8 DISCUSSION

We found the mixed methods approach useful for gaining a holistic and nuanced understanding of user satisfaction with novel recommendations. The interviews and surveys gave insight into the user perspective on discovery and satisfaction, while examining user interaction data enabled us to validate the findings from user research and model them at scale. The findings we discuss here were supported by multiple pieces of evidence from the diverse set of studies we conducted.

First, user goals were crucial in understanding interactions with recommended content. User interviews revealed that users' behaviors changed depending on their goals; survey data showed that satisfaction was correlated with achievement of goals. This was consistent with analysis of feature importance in the tree model: cluster features gave the most information for user-track interactions. While no single cluster assignment dominated in importance, it seemed that each of the cluster types in the model provided additional context to the interaction features, allowing those interaction features to better model satisfaction.

Based on findings from the user interviews, we hypothesized that normalizing each user's interactions by their historical averages would result in meaningful improvements when modeling satisfaction. The statistical model confirmed this hypothesis, as normalized features were the second most informative type of feature (after cluster features).

Another interesting result was that interactions with a track at the positive or negative extremes were more correlated to satisfaction than total interactions. This was reflected in the survey data: loving at least one track was strongly correlated with satisfaction, whereas being annoyed by at least one track was weakly negatively correlated. From the analysis of feature importance in the statistical model, maximum features gave a higher gain than minimum features, which were both more informative than total features. Further inspection of the model showed that, for some trees, maximum and minimum features were even more important than normalized features. Taken together, this evidence suggests that in the discovery context, satisfaction judgements may be made based on the best or worst experiences with a track. We also observed this finding over the entire set of discovery recommendations, as users reported higher overall satisfaction than satisfaction this week, suggesting that overall satisfaction was based on users' best experience with the discovery recommendations. This is consistent with the peak-end heuristic, a well-established hypothesis in psychology, which demonstrates that people weight their evaluation of an experience more heavily on the extremes and the end of the experience than on the average or duration of the experience [13].

8.1 Limitations and Future Work

One challenge inherent in our statistical model design was that for each user-track pair, only a single rating (satisfaction for the entire week) was available from the user. Thus, in cases where a user loved only one track, but gave a high rating for satisfaction this week, that satisfaction rating would not be consistent with track-level satisfaction. In general, this model relies on the user interacting positively with most tracks in the discovery recommendations the same way. Because we know from insights about peak interactions that this was not always the case, we were not surprised that the scores generated from the model were fairly noisy within and between users. One area of future work would be to develop a more sophisticated behavioral model of track love, such that differences between loved, liked, and disliked tracks could be used to more accurately predict track-level satisfaction.

Another limitation of this work was selection bias: users who chose to participate in interviews or take the survey were more engaged than a typical user; they also expressed high levels of satisfaction with the discovery recommendations. A future user research study or survey could target casual or non-habitual users to better understand their reasons for low engagement with the recommendations.

For additional future work, one application of the statistical model is to use the predicted scores as an offline metric for comparing recommendation algorithms. For example, satisfaction scores

Table 7: Results from t-test ran on validation dataset for positive interaction signals.

	A/A test		A/B test		Normalized A/A test		Normalized A/B test	
Interaction Signal	p	t	p	t	p	t	p	t
album views count	0.35	-0.92	0.000	-8.0	0.53	-0.61	0.000	-5.90
artist views count	0.81	-0.2	0.000	-11.11	0.61	0.49	0.000	-9.85
saves count	0.38	-0.86	0.000	-4.94	0.99	0.01	0.000	-4.84
adds count	0.25	1.13	0.000	-3.57	0.14	1.4	0.000	-4.83
downstream completed plays	0.4	0.83	0.000	-18.912	0.90	0.11	0.000	-7.41
artist view duration (ms)	0.54	0.6	0.000	-4.03	0.44	0.77	0.000	-3.49
album view duration (ms)	0.14	1.46	0.004	-2.84	0.37	0.88	0.07	-1.75
downstream listening (ms)	0.43	0.77	0.000	-19.611	0.86	0.17	0.000	-7.29

at the user-track level could allow the set of tracks interacted with to be ranked by how they drive satisfaction; the winning algorithm produces the ordering that most closely matches the ranking. The scores could also be used for model optimization by incorporating them directly into the algorithm's loss function.

9 CONCLUSION

In this paper, we demonstrated a rigorous, user-centric approach to understanding satisfaction in discovery-oriented recommendation. Through a combination of qualitative and quantitative methods, we gained novel insight into user expectations, behaviors, and satisfaction with music discovery. For example, we identified four user goals that influence behavior: play new music in the background, listen to new music now and later, find new music for later, and engage with new music. We also learned that users expect discovery to be hit-and-miss; just one loved track is enough for a user to feel satisfied. By predicting user satisfaction from logged interactions with the discovery recommendations, we confirmed that inferring user goals, normalizing behavior with each user's historical averages, and capturing maximum and minimum interactions per track contributed significantly to satisfaction. Lastly, we presented a suite of user-oriented metrics and validated them on retrospective A/B test data, showing how these metrics can be used to evaluate recommendation performance.

ACKNOWLEDGMENTS

The authors would like to thank Jenn Thom, Marco de Sá, and the Discover Weekly team for feedback and support.

REFERENCES

[1] P. Adamopoulos. On unexpectedness in recommender systems: Or how to expect the unexpected. In *Proceedings of the Workshop on Novelty and Diversity in Recommender Systems, at RecSys*, 2011.
[2] S. Caldwell Brown and A. Krause. A psychological approach to understanding the varied functions that different music formats serve. In *ICMPC*, 2016.
[3] Ò. Celma and P. Herrera. A new approach to evaluating novel recommendations. In *RecSys*, 2008.
[4] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *KDD*, 2016.
[5] A. Deng and X. Shi. Data-driven metric development for online controlled experiments: Seven lessons learned. In *KDD*, 2016.
[6] A. Drutsa, G. Gusev, and P. Serdyukov. Future user engagement prediction and its application to improve the sensitivity of online experiments. In *WWW*, 2015.
[7] M. D. Ekstrand, F. M. Harper, M. C. Willemsen, and J. A. Konstan. User perception of differences in recommender algorithms. In *RecSys*, 2014.
[8] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *TOIS*, 23(2), 2005.
[9] M. Ge, C. Delgado-Battenfeld, and D. Jannach. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *RecSys*, 2010.
[10] B. G. Glaser and A. L. Strauss. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine de Gruyter, New York, NY, 1967.
[11] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *TOIS*, 22(1), 2004.
[12] N. Hurley and M. Zhang. Novelty and diversity in top-n recommendation–analysis and evaluation. *TOIT*, 10(4), 2011.
[13] D. Kahneman, P. P. Wakker, and R. Sarin. Back to bentham? explorations of experienced utility. *The quarterly journal of economics*, 112(2), 1997.
[14] I. Kamehkhosh and D. Jannach. User perception of next-track music recommendations. In *UMAP*, 2017.
[15] A. Laplante. Users' relevance criteria in music retrieval in everyday life: An exploratory study. In *ISMIR*, 2010.
[16] A. Laplante and J. S. Downie. The utilitarian and hedonic outcomes of music information-seeking in everyday life. *Library & Information Science Research*, 33(3), 2011.
[17] J. H. Lee, H. Cho, and Y.-S. Kim. Users' music information needs and behaviors: Design implications for music information retrieval systems. *JAIST*, 67(6), 2016.
[18] J. H. Lee and N. Maiman Waterman. Understanding user requirements for music information services. In *ISMIR*, 2012.
[19] J. H. Lee and R. Price. Understanding users of commercial music services through personas: Design implications. In *ISMIR*, 2015.
[20] J. H. Lee and R. Price. User experience with commercial music services: An empirical exploration. *JAIST*, 67(4), 2016.
[21] T. W. Leong and P. C. Wright. Revisiting social practices surrounding music. In *CHI*, 2013.
[22] M. Mäntymäki and A. K. M. N. Islam. Gratifications from using freemium music streaming services: Differences between basic and premium users. In *ICIS*, 2015.
[23] C. Matt, A. Benlian, T. Hess, and C. Weiß. Escaping from the filter bubble? the effects of novelty and serendipity on users' evaluations of online recommendations. Technical report, Darmstadt Technical University, Department of Business Administration, Economics and Law, Institute for Business Studies (BWL), 2014.
[24] H. Müller and A. Sedley. Hats: large-scale in-product measurement of user attitudes & experiences with happiness tracking surveys. In *Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: the Future of Design*, 2014.
[25] M. Schedl, A. Flexer, and J. Urbano. The neglected user in music information retrieval research. *Journal of Intelligent Information Systems*, 41(3), 2013.
[26] S. Vargas and P. Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *RecSys*. ACM, 2011.
[27] S. Volokhin and E. Agichtein. Understanding music listening intents during daily activities with implications for contextual music recommendation. In *CHIIR*, 2018.
[28] R. W. White, W. Chu, A. Hassan, X. He, Y. Song, and H. Wang. Enhancing personalized search by mining and modeling task behavior. In *WWW*, 2013.
[29] Y. Zhang, W. Chen, D. Wang, and Q. Yang. User-click modeling for understanding and predicting search-behavior. In *KDD*, 2011.