Choice of Implicit Signal Matters: Accounting for User Aspirations in Podcast Recommendations

Zahra Nazari, Praveen Chandar, Ghazal Fazelnia, Catherine M. Edwards, Benjamin Carterette, Mounia Lalmas
{zahran,praveenr,ghazalf,catherinee,benjaminc,mounial}@spotify.com
Spotify

ABSTRACT

Recommender systems are modulating what billions of people are exposed to on a daily basis. Typically, these systems are optimized for user engagement signals such as clicks, streams, likes, or a weighted combination of such sets. Despite the pervasiveness of this practice, little research has been done to explore the downstream impacts of optimization choice on users, creators and the ecosystem they are offered in. We used a platform that caters recommendations to millions of people and show in practice what you optimize for can have a large impact on the content users are exposed to, as well as what they end up consuming.

In this work, we use podcast recommendations with two engagement signals: *Subscription* vs. *Plays* to show that the choice of user engagement matters. We deployed recommendation models optimized for each signal in production and observed that consumption outcomes substantially defer depending on the target used. Upon further investigation, we observed that users' patterns of podcast engagement depend on the type of podcast, and each podcast can cater to specific user goals & needs. Optimizing for streams can bias the recommendations towards certain podcast types, undermine users' aspirational interests and put some show categories at disadvantage. Finally, using calibration we demonstrate that informed balanced recommendations can help address this issue and thereby satisfy diverse user interests.

CCS CONCEPTS

• Information systems → Personalization; **• Human-centered computing →** *User models.*

KEYWORDS

recommender systems, user satisfaction, implicit signals, aspirational recommendations

ACM Reference Format:
Zahra Nazari, Praveen Chandar, Ghazal Fazelnia, Catherine M. Edwards, Benjamin Carterette, Mounia Lalmas . 2022. Choice of Implicit Signal Matters: Accounting for User Aspirations in Podcast Recommendations. In *Proceedings of the ACM Web Conference 2022 (WWW '22), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3485447.3512115

1 INTRODUCTION

Recommender Systems that provide personalized user experiences have been successfully deployed in many commercial applications. These systems are generally trained on implicit signals of user engagement, such as clicks, streams, saves, purchases. In practice, many different signals may be available from user logs—for example, "likes" and "streams" in video streaming [32]. The choice of engagement signal as optimization target is a key component in the development of recommendation systems; the main approaches to this decision either rely on heuristics or take a multi-objective approach that takes other business objectives such as revenue into account. Regardless which approach is taken, there are large implications in terms of what content users are exposed to and end up consuming. The choice of implicit signal impacts the various recommendation surfaces we interact with on a daily basis and is critical to understand, yet this topic has received little attention in the literature.

We hypothesize that different engagement signals could represent the user's interests and goals differently. For instance, users could "follow" or "subscribe" to an educational podcast on reinforcement learning with an aspirational intent to learn more about the topic but might not stream the podcast immediately. In these scenarios, optimizing for "clicks" or "streams" (common practice in industry) could suppress certain needs of the users. It remains a non-trivial task which signal could better capture a user's interest with interacted items and how to effectively combine this information.

Existing solutions to this problem generally involve identifying one type of signal that best captures success. However, Steck [25] argues that these solutions may result in unbalanced recommendations, and over time lead into "echo-chambers" or "filter bubbles". Another approach to this problem is to identify a set of objectives relevant to the problem at hand and optimize jointly to account for these objectives, often referred to as multi-objective optimization. However, there has been little effort in understanding these engagement signals and their relationships to user interests and goals. In this work, different from previous literature, we argue that different types of implicit signals could capture a distinct quality of engagement; therefore, grouping implicit signals that are assumed to be similar into objectives could be misleading.

Podcast recommendation provides an ideal domain in which to explore this problem. Podcasts are an audio medium for information and entertainment, growing rapidly in prominence since the early 2000s. There are now more than one million active podcast shows consisting of 64 million episodes.These podcasts are distributed by RSS feeds; people subscribe to automatically receive each new

WWW '22, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9096-5/22/04…$15.00
https://doi.org/10.1145/3485447.3512115

episode.But it is also common to "dip in" and just listen to single podcast episodes. Thus podcasts offer two strong engagement signals: show *subscriptions* and episode *plays*. Many information or entertainment media offer only one strong signal of engagement amongst many weaker signals.

In this work we take advantage of this distinguishing property of the podcast domain to tackle the problem of optimizing in the presence of multiple implicit engagement signals. We show that the choice of signal for training can have surprising downstream consequences. Further, we show that different signals capture different pieces of information about user preferences. This points to a need for a unified approach to recommendation. We approach this problem by asking three specific research questions. The first concerns the choice of implicit engagement signal to optimize.

RQ1: How does choice of implicit signals for optimization affect podcast recommendations and consumption patterns?

To answer this, we conduct both offline and online experiments comparing two different recommendation models trained on *subscribe-based* and *play-based* engagement for podcast recommendations. We find significant discrepancies in outcomes, leading to our next research question.

RQ2: What factors are predictors of each engagement type?

We use human-annotated labels to categorize podcasts to study the relationship between podcast characteristics and different engagement types. We conduct regression analyses using logs from Spotify, a large audio streaming platform, finding that the category and theme of a show are strong predictors of one type of engagement while inversely predicting the other. For instance, educational shows are subscribed to more often than they are listened to. Similar findings have been reported in the psychology literature [19, 23]. This raises the question of how to simultaneously optimize for different user goals, which leads to our final research question.

RQ3: How do we optimize recommendation systems to account for user goals captured across different engagement signals?

The insights from previous steps served as motivation to propose a new optimization approach that reflects our new understanding of user behavior in podcast consumption. We train models that are optimized for user engagement in the form of streams, while calibrating the recommendations to reflect the user's aspirational interests captured in subscription behavior. A by-product of this calibration results in increased diversity and coverage in terms of podcast categories. The goal of this step is not to compare the calibration approach to other multi-objective optimization solutions, but to showcase an example of how a deeper understanding of user engagement signals can help us make informed decisions on the optimization approach.

The key contributions and findings in this paper include:

- To the best of our knowledge, this work is the first to study the consequences of implicit engagement signal choice for recommendation systems both in offline and online settings.
- We provide a methodology to investigate this further that involves both human manual annotations and millions of interaction logs from a large audio streaming platform.
- Motivated by our insights from the previous step, we employ calibration methods to show that leveraging both engagement signals can result in more balanced recommendation and increased user consumption across user goals.

The rest of this paper is organized as follows. Section 2 summarizes previous work. Section 3 shows that the choice of implicit signal used for training a recommender has a large impact on both recommended items and user consumption. Section 4 shows that these differences are not due to availability factors such as show release cadence and episode length, but due to the goal a show serves. Section 5 presents a calibration method to reconcile the two different engagement signals, and we conclude in Section 6.

2 RELATED WORK

This work is related to four areas, users' preferences, bias in recommender systems, podcast consumptions and metrics.

Users' Preferences. Historically, recommender systems were developed assuming access to user preferences in the form of explicit feedback, e.g. item ratings [13]. The deployment of recommender systems in many domains, where millions of users interact with millions of items, shifted the attention from explicit feedback to implicit interaction signals [2]. Oard et al. [21] classified implicit signals into three categories. *Examination* describes engagement signals such as "plays", "views", and "purchase", where as *Retention* includes signals that imply future use such as "save", "bookmark" and "subscription". Finally, *Reference* describes social signals such as "forward". These categories correspond to different ways users engage with content. Two of these categories are also apparent when users engage with podcasts, namely Examination as Plays and Retention as Subscriptions.

Bias in Recommender Systems. Studies of analyzing bias arising from using implicit signals have mostly focused on an underlying assumption: the mere existence of an implicit signal is considered as positive feedback for a user and item pair, and considering all other items a user has not interacted with as a potential negative association pool. This is known to not hold [22, 24], and in addition can introduce severe bias into the system. Researchers have sought to address this problem in both training [7, 10, 12] and offline evaluation [29]. However, we focus on a fundamentally different bias, that is the positive labels themselves could be biased towards specific categories of items due to various reasons, such as how users engage with particular categories of items.

There has recently been more work on the optimization of recommender systems for users' long-term engagement. Delayed signals such as dwell time and revisits are either explicitly [33] or implicitly [4, 11] used to plan sequences of recommendations and optimize for longer engagement. These studies report improvements on increasing long-term engagement, but it is not clear what would be their effect on satisfying various user goals and item exposure. To optimize for a longer engagement, such algorithms end up recommending items with longer dwell time and more frequent revisits. In our work, we show that such practices could advantage specific set of items and ignore some user intents.

Podcast Consumption Goals. Podcasts are produced as shows, where each show has a specific theme and releases episodes periodically. According to [3, 17], podcasts are consumed for a variety of reasons including education, relaxation, and entertainment. In particular, the role of podcasts as an effective tool for educational purposes has been well documented in [8]. A study from [9] suggests that podcasts are used as a way to pursue knowledge and access to intellectually challenging content. A recent study on podcast recommendations showed that taking into account explicit feedback in the form of topic onboarding resulted in an increase in engagement levels by 24% [30]. In this work, we focus on large-scale applications where information about users' interests is limited to the implicit signals from their interactions. We show that depending on what goal a show serves, users engage with podcasts in a different manner. This has implications when developing podcast recommender systems, as they need to cater for different types of engagement.

Beyond Accuracy. There has been a growing interest in developing recommenders that optimize for objectives beyond accuracy such as diversity [28], novelty [26], sustainability [27], aiming at satisfying users' diverse needs. In an attempt to satisfy users' lesser-known interests, Steck et al. [25] proposed a calibration framework. Ekstrand et al. [6] call out the current approach of using implicit engagement signals as "behaviorism" and urge practitioners to include users' explicit goals. Knijnenburg et al. [15] suggest that recommender systems should not replace human decision making but support them to understand their preferences and optimize accordingly. In this work, we use manually annotating content with metadata that describes the primary goal a podcast seves and use them as proxies for user goals. We adopt a framework inspired by Steck [25] to balance different types of user goals.

3 IMPACT OF IMPLICIT SIGNAL CHOICES ON RECOMMENDATIONS

The choice of implicit engagement signals used to optimize recommender systems have a significant impact on the type of items shown to the user, which in turn, can affect what the user consumes. To demonstrate this, we conducted offline experiments on large-scale real-world datasets as well as online experiments with users of an audio content recommendation service.

3.1 Podcast Recommendations

Podcast recommendation entails matching user interests and tastes with podcasts that are most likely to satisfy them. Although podcasts are audio content like music, users interact with them differently. Podcasts contain spoken content, and are usually more informational than music. Further, podcasts are produced as shows consisting of several episodes. Often, a show has a specific theme and its episodes are released periodically.

We conduct experiments on Spotify, a large audio streaming platform that hosts both music and podcasts content. The platform recommends content to the user using a grid layout that groups related content into a carousel, where podcasts are assigned to a dedicated carousel. Users can interact with the recommended podcast shows by clicking or tapping on the show. This take them to the page of the show, which contains a description of the show along with a list of episodes. Users can decide to stream an episode or subscribe to the show. Subscribing to a show adds the show to a user's library, allowing them fast and easy access later on and does not imply monetary payment.

3.2 Recommendation Model

Our goal here is to highlight the differences observed when optimizing recommendation algorithms based on different engagement signals. We use a recommendation algorithm based on deep neural networks that has shown promising results in similar recommendation applications [20]. Inspired by the CBOW (Continuous Bag of Words) model [18], the framework casts the recommendations problem as an extreme multi-class classification task modeled by a multilayer perceptron. This provides flexibility in handling heterogeneous feature sets, and their success in recommendation applications has been widely reported [1, 5, 31]. We expect our results to hold for most machine learning-based recommendation algorithms and differ an actual empirical comparison to future work.

3.2.1 *Recommendation Algorithm.*

Given a set of features S and a representation $U(S)$ of user U, we train a neural network that maps a user to a distribution over items in the podcast domain P. Representation of sparse features could be learned through backward passes of the entire network in an end-to-end optimization, whereas dense features could be memorized through the network and concatenated to the final user representation at any stage.

We use a softmax activation layer and minimize the cross entropy loss for the true label and the negative samples. Let $i \in P$ be the label, and p_i be an N-dimensional vector representation of i. The user $U(S)$ is then shown as an N-dimensional vector u.

$$P(i|U(S)) = \frac{e^{p_i u}}{\sum_{j \in P} e^{p_j u}}$$

This optimization results in a dense vector $u \in \mathbb{R}^N$ that is closer to the item i's vector as the weights of the node i in softmax layer $p_i \in \mathbb{R}^N$ are trained.

In our case, a class corresponds to a podcast show, so we need to handle a large number of classes during training. We therefore use importance sampling, a negative sampling approach proposed in [14] that allows models to efficiently converge. The loss function is then calculated as:

$$J_\theta = -\sum_{i \in P}[\log \sigma(up_i) + \sum_{j=1}^{k} \log \sigma(-up_{ij})]$$

where k is the number of sampled negatives, $p_{ij} \in \mathbb{R}^N$ is the vector for the jth negative class sampled for label i and $\sigma(x) = \frac{1}{1+\exp(-x)}$.

The model enables us to augment side information about the user, such as demographics and their podcast interests. Finally, the model output provides us with a list of podcast recommendations.

3.2.2 *Optimization Target.*

We use one of two types of engagement signals to optimize our recommender systems. The positive labels are determined based on one of the following criteria:

- **Subscribe**: User-show pairs are assigned a positive label if the user has subscribed to a show. Users can subscribe even before listening to any of its episodes.
- **Play**: User-show pairs are assigned a positive label if the user streams at least one episode of the show.[1]

[1] We use a small time threshold to avoid labeling accidental streams as positive.

In our data, plays are three times as common as subscribes, and there is actually very little overlap between the two—less than 0.01 of all user-show pairs. This already hints at aspirations being neglected due to recommendations aimed at short-term engagement.

3.3 Dataset

We randomly sampled 800K users who had listened to podcasts in July 2020 on Spotify. For each user, we recorded the podcast shows the user streamed or subscribed to. We restrict our dataset to users in the US. The different feature sets are described next.

3.3.1 Show Features. The podcast shows studied in this analysis were the 440 top popular shows in the US and were accompanied by creator-provided metadata, such as category and sub-category information that have gone through an internal quality check. The annotators would use this metadata to annotate each show. Each podcast show was annotated as follows:

Category: Each show was categorized by an expert into one of these seven categories: Knowledge, Entertainment, Sex & Relationships, Business & Technology, Sports, Politics & Current Events, Wellness & Spirituality.

Episode length: Using average episode length, each show was put into one of the following buckets: 1 −15 minutes, 15–30 minutes, about 30 minutes, 30–45 minutes, 45–60 minutes, about 60 minutes, 60–90 minutes, 90–120 minutes, about 120 minutes, 120+ minutes.

Release cadence: To capture the frequency of episode release for each show, we bucketed each show into one of these: >1 episode per day, daily (5–7 episodes per week), 2–4 episodes per week, weekly (1 episode per week), 1 episode every other week, Ÿ 1 episode per month, 1 episode per month and 'cadence varies'.

Self-contained or serialized: A story wrapped up in one episode is annotated as self-contained; otherwise it is annotated as serialized.

Topical or evergreen: If listening to the show more than a week after release reduces its value and appeal, the show was annotated as Topical. If the show contains information that will remain correct and relevant for a long period of time, it was annotated as Evergreen.

Theme: The primary theme of a show was one of the followings: Learning (provides in depth knowledge about a topic), Stay Updated (keeps you updated on the current event), Companionship (makes you feel like you are hanging out with friends), and Sleep Aid (one of the main use cases of this show is for sleep aid).

3.3.2 User Features. We use a set of features to represent users, as used in many recommender systems, including basic demographic information such as self-reported age, gender, etc. In addition, we had access to podcast preferences (e.g. affinity scores), which are derived from user listening history and regularly updated.

3.4 Offline Experiments

We describe our training procedure for optimizing the model based on the two chosen engagement signals and report offline model performance on a held-out set.

3.4.1 Experiment Design. Our online task requires the recommendation produced by the model to be presented to the user in a horizontal list on the platform. Therefore, to reflect the online task, we use the top-K recommendations set-up to train and evaluate the models. In other words, the goal of the recommender is to accurately predict the top K recommendations that users are likely to interact with, where K=10 in our case.

Table 1: Results from offline evaluations

	Subscriptions Label		Plays Label	
Model	prec@10	nDCG@10	prec@10	nDCG@10
Subscription Model	0.036^a	0.234^b	0.031^c	0.209^d
Plays Model	0.035^a	0.231^b	0.034^c	0.211^d

We partition the log data into the training and test sets following a user-based split with 75% of the users assigned to the training set. For each data-point, we store the user history as an ordered list of their past interactions. This allows us to use part of the history as features, and avoid data leaks. We train two different models: one using subscriptions as positive labels, and another using plays as positive labels. We call the first model the "Subscription Model" and the second one the "Plays Model".

3.4.2 Offline Evaluation. Given a user from the test set[2] and a set of all podcast shows, the system is required to provide a ranked list of recommendations. The ranked list of recommendations is then augmented with relevance labels, and standard ranking metrics such as $prec@k$ and $nDCG@k$ were used to obtain an effectiveness score. Metrics that discount relevance based on rank-position are best for our use case since users typically pay more attention to the top-ranked items. The results of our offline evaluation are shown in Table 1. We report $prec@10$ and $nDCG@10$ for both the "Play" and "Subscribe" models. Not surprisingly, the model optimized on subscribe engagement performed best when using subscribe engagement as relevance labels, and similarly for plays.

Next, we report the composition of podcast categories in the top 10 recommendations for each model in Figure 1 The Knowledge category appears more than twice as much in the recommendations from the Subscription model compared to the Plays model. On the other hand, shows in Politics & Current Events and Sex & Relationship categories tend to happen almost twice as much in the Plays model compared to the Subscription model. Entertainment shows also appear about 5% more often in the Plays model.

Our offline experiments suggest that the choice of engagement type as an optimization target affects its model performance. They also showed that optimization targets have a direct impact on the composition of the recommendations, i.e., it impacts what podcasts are shown to users. Next, we turn to online experimentation to understand how this affects user consumption of podcasts.

3.5 Online Experiments

We conducted an online experiment or A/B test, as this provides a direct way to compare variants of machine learning models given a live production recommender system [16]. To compare "Subscription" and "Plays" models, we randomly assign a portion of our user population to two different variants of the podcast carousel on the platform's homepage for a period of two weeks. Each variant ended up with 800K randomly selected users. The podcast carousel

[2] As noted earlier, the timestamps enabled us to further split the test set based on time. Therefore, a portion of the user history was used as features to the model, and the rest for evaluation.

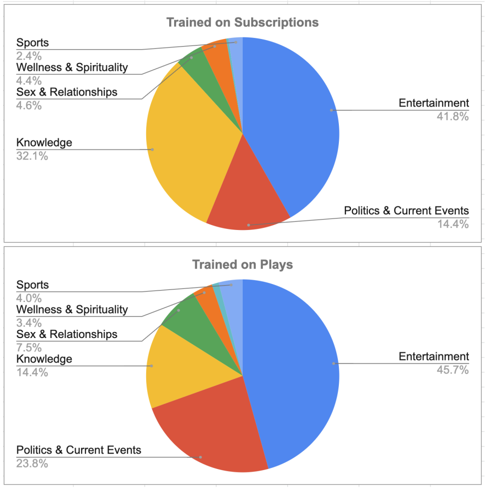

Figure 1: Category distribution of the top 10 recommendations from the Subscription model (top) compared to the Plays Model (bottom).

for each variant was powered by the “Subscription" and “Plays" models, respectively. We analyze and report show exposure and user consumption patterns for the two variants below.

We first confirm our findings from the offline experiments that the “Subscription" and “Plays" models exposed different types of podcasts to users. Figure 2 left side shows the relative difference in exposure between the two models for different podcast categories. Positive values indicate that the category was seen more by the users assigned to the “Subscription" model than the “Plays" model and vice versa for negative values. Note that Knowledge and Business podcasts are exposed more in the “Subscription" model, which is consistent with our offline findings.

Interestingly, we notice that the difference in exposure translates into a change in consumption patterns as well. Figure 2 right side shows the relative difference in consumption between the two models for different podcast categories. Similar to the exposure plots, the positive values correspond to higher consumption in the “Subscription" model. Not only does the “Subscription‘ model result in a relative increased exposure of podcasts categories such as Knowledge and Business but also an increase in consumption. This suggests that a poor choice of optimization target could potentially lead to certain content types being disadvantaged. These results provide conclusive evidence that the engagement type used to optimize machine learning models significantly impacts the category of shows that users consume.

Our offline and online results demonstrate that the choice of engagement type used to optimize recommender systems impacts user consumption patterns. This emphasizes the need for practitioners to pay close attention when choosing which feedback signals to use as proxies of engagement. Otherwise recommender systems could reinforce consumption patterns that leads to over-consumption of certain types of items and under-consumption of others. These results suggest that there exists a discrepancy between “Play" and “Subscribe" engagement types. We investigate the potential factors that could explain the observed discrepancy next.

4 UNDERSTANDING ENGAGEMENT TYPES

We carry out an observational study to gain a comprehensive understanding of various engagement types. We conduct our study on millions of users over a period of 12 months spanning across 2019 & 2020 using log data that contained user interactions on podcasts. The logs consisted of information regarding the subscription status of the user-show pair and detailed history of consumption of episodes from the show with timestamps.

Our primary goal here is to understand the relationship between characteristics of a podcast such as topic category, release cadence, and the type of engagement that users have with that podcast. We use the two engagement signals, Subscribe and Play. In addition, we expand the Play signal to include two additional categories that intend to capture different types of engagement. The four different engagement signals are:

Subscribe: If the user has subscribed to a podcast show.

Played ¡ δ mins: If the user consumed at least δ mins of any episode from a podcast show.

Played ¡ 2 Episodes: If the user has played a minimum of three episodes from a show.

Played ¡ 7 Days: If a user has returned to a show for more than seven days.

We characterize each podcast show using the manual annotations (see Section 3.3.1). We investigate the following two hypotheses:

H1: *Release cadence of a show affects how users engage with podcasts.* For example, users may subscribe to a show regardless of their average episode length, but may not end up listening to the lengthier ones. On the other hand, shows that have frequent release cycle are likely to be streamed more.

H2: *The intent of a podcast can lead to different engagement patterns.* Podcasts are created to serve different user goals. For example, a news-related show may attract more streams because users' habits might lead them to consume news on a daily basis. On the other hand, users may not listen to a knowledge show as frequently even when released daily. Nonetheless, users may be equally interested in a show even if they do not engage with them in a similar manner.

4.1 Normalization of Engagement Signals

Podcast shows have various levels of popularity, which may be a confounding factor. We therefore perform a normalization step to ensure that shows with different popularity levels and across various engagement types are comparable with one other. We denote P_i^k as the ratio of users who had the specific interaction k out of all users who had any type of interactions with the show i.

4.2 Subscription vs. Play Engagement Types

We compare the average values of P_i^k for the subscription behaviour in Figure 3 and for engagement behaviour related to plays in Figure 4 for different podcast categories. These two graphs demonstrate clear differences between show categories in the distribution of each engagement signal similar to what we observed in Section 3.5. Shows in the knowledge category are more likely to be subscribed by users, while they are less likely to have frequent episode plays or daily returns to the show. On the other end of the spectrum, sports shows tend to get engaged with at a higher cadence, as Played ¡ 2

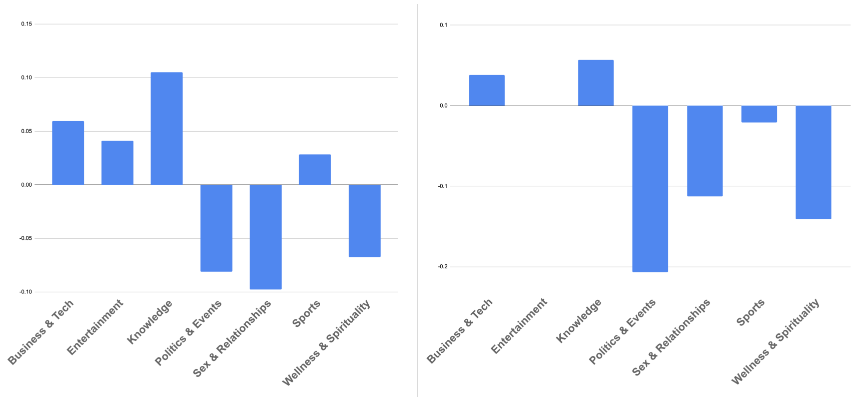

Figure 2: Relative difference in consumption (Left) and exposure (right) between Subscription and Plays model for each category. Positive values represent categories with more consumption in the Subscription Model.

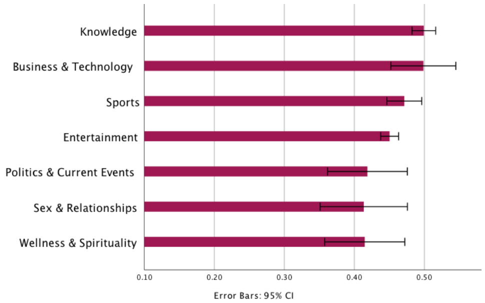

Figure 3: Subscription comparison based on show category.

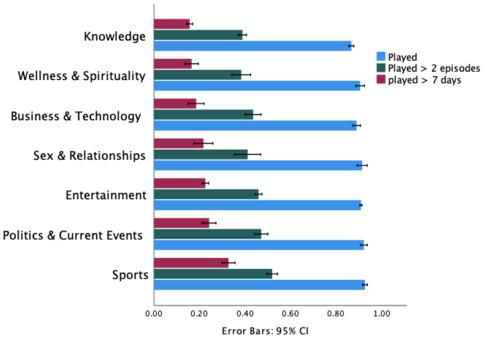

Figure 4: Engagement types comparison grouped by category.

Episodes Played ¡ 7 Days are higher for this category. We next perform regression analyses to explore the potential causes of these observed differences in engagement types.

4.3 Effect of Podcast Characteristics on Engagement

We conduct a regression analysis to investigate hypothesis **H1**. The normalized engagement score P_i^k for each engagement type k is used as the dependent variable. Our independent variables include podcast characteristics such as Category, Episode Length, Episode Cadence, Serialized, Evergreen, and Theme, obtained from the manual annotation. We followed a standard procedure in training multiple linear regression models, and since all independent variables are categorical, we used dummy variables to represent each in the model. Each categorical variable with k values was presented using k-1 numbers instead of k numbers, to avoid multicollinearity trap. Moreover, dummy variables inherently satisfy the linearity assumption. We also performed a generalized variance-inflation factors test for each independent variable and values were all below 1.5, which is small enough to rule out multicollinearity.

The results of the regression are shown in Table 2. The overall model fit (R^2) for predicting Subscriptions, Plays, >2 episodes, >7 days signals were 0.33, 0.32, 0.31 and 0.42 respectively. We report only those characteristics that were significant (p-value <0.05). The category of a show is the most important predictor across all engagement types. We also observe that shows in the knowledge category are more likely to be subscribed, whereas sports shows are more likely to be streamed periodically.

Next, we explore the effect of show category in predicting each normalized engagement score (P_i^k) when controlled for length and cadence of a show. We perform the same regression analysis, this time in two steps: (1) using only the length and cadence as independent variables, and then (2) adding show categories as additional independent variables. Results from the two steps regression are reported in Table 3. It is clear that the category of a show remains an important predictor (predicting as large as 65% of the variance in normalized engagement signal) even after controlling for length and cadence of a show (The overall model fit (R^2) for predicting Subscriptions, Plays, >2 episodes and >7 days signals were 0.27, 0.23, 0.26 and 0.38 respectively).

These results help us reject our first hypothesis that availability of a show defined by its length and cadence are the main factors in identifying engagement types with a podcast show. This suggests that podcast metadata could be useful proxies of user goals, and we look into this next.

4.4 Effect of User Goals on Engagement

We also test the hypothesis **H2** using regression. We retain only the themes annotated for each podcast to indicate the primary user goal satisfied by the podcast. Similar to how Tomkins et al. [27] use metadata to study sustainability of products, we incorporate prior knowledge about podcasts such as their primary theme as a proxy of user goals in our study. Table 4 shows the regression coefficients for the significant "themes".

There are two key findings. First, users are more likely to subscribe to shows about learning. Second, there is a lower probability for shows about learning to get "Play" related engagement types compared to the other themes. This confirms the hypothesis **H2** that goals satisfied by a podcast can lead to different interaction patterns with that podcast. Further, we argue that a lower "Play" related engagement on a podcast does not

Table 2: Feature weights in regression analysis predicting each implicit signal likelihood.

Target	Feature	Coefficient	p-value
Predicting Subscriptions			
	Category = Knowledge	0.41	0.000
	Category = Wellness	0.21	0.002
Predicting Plays			
	Category = Knowledge	-0.44	0.000
	Category = Wellness	-0.26	0.000
	Category = Business	-0.15	0.006
Predicting > 2 episodes			
	Serialized = true	0.34	0.000
	Category = Sports	0.23	0.000
	Cadence = Daily	0.10	0.005
	Cadence < 1 per month	-0.16	0.011
Predicting >7 days			
	Category = Knowledge	-0.20	0.000
	Avg length = 1-15 mins	-0.15	0.001
	Category = Sports	0.14	0.003

Table 3: Results from performing regression analysis predicting different targets in two steps

Target	Step1) Cadence and Length Features	Step2) Add Category Features	% improvement
Predicting Subscriptions	0.086	0.16	46%
Predicting Plays	0.08	0.232	65%
Predicting >2 episodes	0.07	0.19	63%
Predicting >7 days	0.24	0.33	27%

Table 4: Feature weights in regression analysis when category features are excluded.

Target	Feature	Coefficient	p-value
Predicting Subscriptions			
	Learning = True	0.26	0.000
	Evergreen = True	0.18	0.000
	Sleep Aid = True	-0.17	0.000
Predicting Plays			
	Evergreen = Topical	0.25	0.000
	Sleep Aid = True	0.21	0.000
Predicting > 2 episodes			
	Serialized = true	0.27	0.000
	Evergreen = Topical	0.15	0.001
	Sleep Aid = True	-0.162	0.000
	Learning = True	-0.13	0.003
Predicting >7 days			
	Evergreen = Topical	0.23	0.000
	Companionship = yes	0.15	0.015

necessarily mean users are not interested in them.

This shows that the podcast category and user goals play an important role in how users interact with that podcast, even after accounting for popularity, release schedules, and episode length. Factors that positively predict one type of engagement may negatively predict the other. Further, using human annotated themes for podcasts, we are able to amplify our understanding of user goals. Next, we present a method to use both types of engagements in forming recommendations.

5 RECONCILING ENGAGEMENT TYPES

The common practice in recommender system applications is to rely on a single source of ground truth. Engagement signals such as clicks, streams, or explicit feedback such as user ratings are often used as ground truth labels for training and evaluating models. However, we demonstrated in this paper that when multiple engagement types exist, they often tend to capture different user goals, i.e., how users intend to consume podcasts. Therefore, successfully incorporating different types of engagement signals becomes paramount when developing recommenders. A similar viewpoint was discussed by Ekstrand and Willemsen [6].

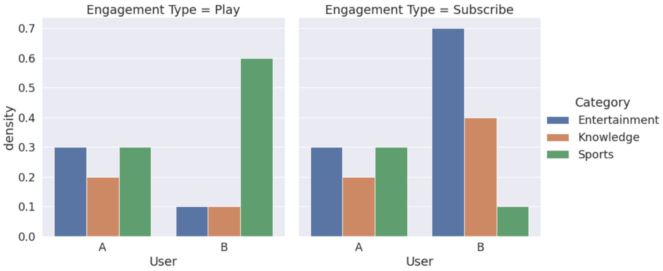

Figure 5: A toy example showing consumption patterns for two types of engagement: "Play" and "Subscribe".

Our goal in this section is to demonstrate how existing calibration frameworks can be used to optimize recommender systems for different engagement types in practice. Let us consider an example by examining the play consumption and subscription patterns of two sample users. Table 5 shows, for each sample user, the number of podcasts they have subscribed to and the podcasts they have listened to for more than δ mins for each category. $user_B$ subscribed to three different shows in the Knowledge category, and one in Sports but mostly streamed Sports-related shows, whereas $user_A$ listens to podcasts they subscribe to. We argue that the shows subscribed by the user are likely to reflect *aspirations*, whereas Play-related signals are more likely to capture their *short-term needs*. Therefore, using a "Plays" or "Subscription" model from Section 3.5 will not fully satisfy the needs of both $user_A$ and $user_B$ users. This raises the need to develop recommender systems that balance between aspirations and the short-term needs of users. Next, we briefly describe a calibration framework for simultaneously optimizing different engagement types and report its performance on our podcast recommendation use case.

5.1 Calibration

We adopt the framework proposed by Steck [25] that uses a two-stage approach. First, we train a *base* model that optimizes for a primary engagement type. Then, the recommendations obtained are re-ranked in a post-processing step to account for the secondary engagement type. Here, we use the model described in Section 3.2.1 to optimize on the primary objective and use calibration to account for the secondary engagement type.

We start with an initial recommendation list ℓ of size n generated by the model from the first step. Then, we use a weighted sum of relevance and calibration for creating the re-ranked recommendations to make sure our calibration does not harm the relevance of the recommended items. Here, relevance implies the user goals relating to primary engagement time, and calibration to secondary engagement. The final step is to pick top-k items that maximizes the marginal relevance:

$$\ell^* = \underset{\ell,|\ell|=n}{\arg\max}(1-\lambda)\cdot Rel_{primary}(\ell) - \lambda.\mathcal{JSD}_{secondary}(P(\ell), Q(S_u)) \quad (1)$$

where λ is the trade-off parameter that balances between goals relating to the primary and secondary engagement types. $Rel_{primary}(\ell)$ is the sum of the predicted scores for items in ℓ. $P(\ell)$ and $Q(S_u)$ are obtained by representing the secondary engagement type as a distribution over a pre-defined category variable such as podcast category. $P(\ell)$ provides a distribution over categories for selected items and $Q(S_u)$ for remaining candidates. The two distributions are combined using Jensen–Shannon divergence (JSD) and a final calibrated ranking is obtained by greedily ranking using Equation 1.

5.2 Experiment Setup

We compare results from the calibrated approach to models optimized for Plays and Subscribe goals, separately. We use an offline experimentation

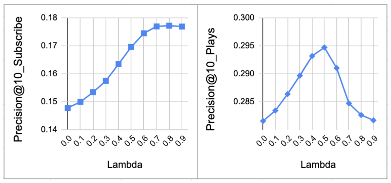

Figure 6: Effect of varying λ on $Prec@k_{Subscribe}$ (left) and $Prec@k_{Play}$ (right)

setup similar to the one described in Section 3.4.1. The *base* recommender used for the calibration approach was trained using data with podcast Play $> \delta$ mins as the positive label. We use the top 200 recommendations for the post-processing step to calibrate with the Subscribe goals.

5.2.1 Evaluation Metrics. The calibration model is expected to maximize "Subscribe" and "Play" goals simultaneously. We use Prec@k ($k = 10$ in our experiments) computed using the two signals on relevance labels. In addition, we compute show coverage – the average number of unique shows recommended for all users. We use this as a supplementary metric to measure how well the recommender system represents different podcasts.

5.3 Results and Discussion

We report the performance of the calibrated recommendation model measured based on $Prec@k$ and *coverage* metrics. A well calibrated model should maximize both the primary ($Prec@10_{Subscribe}$) and secondary ($Prec@10_{Play}$) goal.

5.3.1 Calibrated Model Comparison. Calibration aimed to address the discrepancies between Subscribe and Play engagement types. To verify if our calibration does change the distribution of exposure for various categories, we plot the top 10 recommendations grouped by the podcast category for all three models: calibrated, play, and subscribe. Figure 9 shows the results of the comparison. We notice that the exposure ratio of the calibrated model lies in between the Plays and Subscription Model for all categories except Entertainment. This further validates that the calibrated model is balancing both "Play" and "Subscribe" user goals.

5.3.2 Effect of λ on $Prec@k_{Play}$ and $Prec@k_{Subscribe}$. To understand the impact of λ, we hold all model parameters constant and vary only the value $\lambda \in [0, 1]$. Figure 6 shows the result of varying λ between 0 and 1 and report $Prec@10_{Subscribe}$ and $Prec@10_{Play}$ metrics. As expected, increasing the λ values results in an increased $Prec@k_{Subscribe}$ score. However, we observed that increasing λ also improves precision ($Prec@k_{Play}$), where the best performance is obtained at $\lambda = 0.6$. This result is encouraging since it suggests that there exists an equilibrium at which users' Play-related and Subscribe-related goals are met.

5.3.3 Effect of λ on Show Coverage. Next, we report the Show Coverage metrics when λ is varied. We observe that the overall show coverage shown in Figure 7 is improved as λ increases. Another side effect of calibration was an increase in coverage for less-represented categories. Figure 8 shows this growth in each category. "Knowledge" shows obtain more coverage, while "Entertainment" shows lose some of their coverage. Other categories more or less remain the same. These results suggest that the calibration framework used to balance between play and subscribe goals not only improves the precision metrics for the two goals but also recommends more diverse content to users for shows in less represented categories.

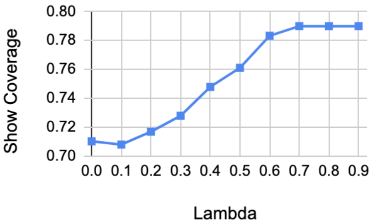

Figure 7: Effect of calibration varying λ on overall coverage

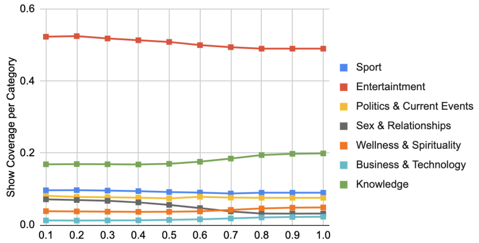

Figure 8: Effect of calibration varying λ on category coverage

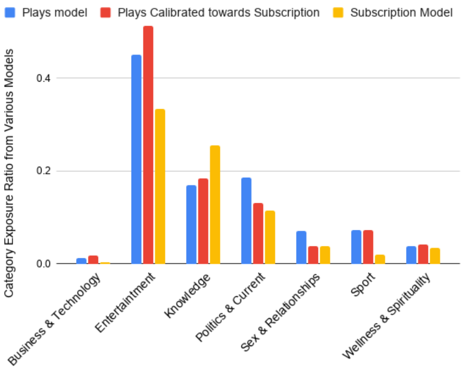

Figure 9: Comparison of the calibarated model against Play and Subscription Model for different podcast categories.

6 CONCLUSION AND FUTURE WORK

We addressed the challenges that arise from having multiple engagement signals in recommendation applications. Specifically, we focused on podcast recommendation as it provides an ideal domain to explore the dilemma faced by developers in choosing between two strong engagement signals: subscriptions and episode plays. We presented results from a randomized controlled experiment in scale to show that the choice of engagement signals can have a significant impact on not only what recommendations users are shown but also on what they consume. Based on these findings we argue that the repercussions of blindly picking an engagement signal for optimization can lead to undermining certain user goals and putting some content categories at disadvantage.

We presented an observational study that makes use of human annotations to understand the reason for the discrepancies between subscriptions and episode plays engagement signals. Insights from the study indicate that users engage with podcasts in different ways depending on the goals each show serves. For example, users tend to subscribe to shows with a "Learning" theme, but are less likely to play such shows. Optimizing a recommender based on plays can result in under-representing Knowledge-related shows. Coupled with the feedback loop problem in recommender systems, these biases can drastically suppress certain content categories over the long run. Therefore, incorporating different engagement signals is essential when building recommendation models.

Finally, we employed a simple and explainable calibration method to show that combining subscriptions and episode plays engagement signals can indeed result in significant improvements in user engagement. We show that a recommender trained on "Play" and calibrated using "Subscribe" signals would satisfy user goals and ambitions related to both signals.

REFERENCES

[1] Taleb Alashkar, Songyao Jiang, Shuyang Wang, and Yun Fu. 2017. Examples-rules guided deep neural network for makeup recommendation. In *Thirty-First AAAI Conference on Artificial Intelligence.*

[2] Xavier Amatriain and Justin Basilico. 2016. Past, present, and future of recommender systems: An industry perspective. In *Proceedings of the 10th ACM Conference on Recommender Systems.* 211–214.

[3] Monica Chadha, Alex Avila, and Homero Gil de Zúñiga. 2012. Listening in: Building a profile of podcast users and analyzing their political participation. *Journal of Information Technology & Politics* 9, 4 (2012), 388–401.

[4] Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H Chi. 2019. Top-k off-policy correction for a REINFORCE recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining.* 456–464.

[5] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems.* ACM, 7–10.

[6] Michael D Ekstrand and Martijn C Willemsen. 2016. Behaviorism is not enough: better recommendations through listening to users. In *Proceedings of the 10th ACM Conference on Recommender Systems.* 221–224.

[7] Charles Elkan and Keith Noto. 2008. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining.* 213–220.

[8] Simon Fietze. 2009. Podcast in higher education: Students usage behaviour. *Same places, different spaces. Proceedings ascilite Auckland 2009* (2009), 314–318.

[9] Pamela L Gay, Rebecca Bemrose-Fetter, Georgia Bracey, and Fraser Cain. 2007. Astronomy Cast: Evaluation of a podcast audience's content needs and listening habits. *Communicating Astronomy with the Public* 1, 1 (2007), 24–29.

[10] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining.* Ieee, 263–272.

[11] Eugene Ie, Vihan Jain, Jing Wang, Sanmit Narvekar, Ritesh Agarwal, Rui Wu, Heng-Tze Cheng, Morgane Lustman, Vince Gatto, Paul Covington, et al. 2019. Reinforcement learning for slate-based recommender systems: A tractable decomposition and practical methodology. *arXiv preprint arXiv:1905.12767* (2019).

[12] Dietmar Jannach, Lukas Lerche, and Markus Zanker. 2018. Recommending based on implicit feedback. In *Social Information Access.* Springer, 510–569.

[13] Gawesh Jawaheer, Peter Weller, and Patty Kostkova. 2014. Modeling user preferences in recommender systems: A classification framework for explicit and implicit user feedback. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 4, 2 (2014), 1–26.

[14] Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2014. On using very large target vocabulary for neural machine translation. *arXiv preprint arXiv:1412.2007* (2014).

[15] Bart P Knijnenburg, Saadhika Sivakumar, and Daricia Wilkinson. 2016. Recommender systems for self-actualization. In *Proceedings of the 10th ACM Conference on Recommender Systems.* 11–14.

[16] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. 2009. Controlled Experiments on the Web: Survey and Practical Guide. *Data Min. Knowl. Discov.* 18, 1 (2009), 140–181.

[17] Steven McClung and Kristine Johnson. 2010. Examining the motives of podcast users. *Journal of Radio & Audio Media* 17, 1 (2010), 82–95.

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems.* 3111–3119.

[19] Katherine L Milkman, Todd Rogers, and Max H Bazerman. 2008. Harnessing our inner angels and demons: What we have learned about want/should conflicts and how that knowledge can help us reduce short-sighted decision making. *Perspectives on Psychological Science* 3, 4 (2008), 324–338.

[20] Zahra Nazari, Christophe Charbuillet, Johan Pages, Martin Laurent, Denis Charrier, Briana Vecchione, and Ben Carterette. 2020. Recommending Podcasts for Cold-Start Users Based on Music Listening and Taste. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1041–1050.

[21] Douglas W Oard, Jinmook Kim, et al. 1998. Implicit feedback for recommender systems. In *Proceedings of the AAAI workshop on recommender systems*, Vol. 83. WoUongong.

[22] Rong Pan, Yunhong Zhou, Bin Cao, Nathan N Liu, Rajan Lukose, Martin Scholz, and Qiang Yang. 2008. One-class collaborative filtering. In *2008 Eighth IEEE International Conference on Data Mining.* IEEE, 502–511.

[23] Daniel Read, George Loewenstein, and Shobana Kalyanaraman. 1999. Mixing virtue and vice: Combining the immediacy effect and the diversification heuristic. *Journal of Behavioral Decision Making* 12, 4 (1999), 257–273.

[24] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).

[25] Harald Steck. 2018. Calibrated recommendations. In *Proceedings of the 12th ACM conference on recommender systems.* 154–162.

[26] Choon Hui Teo, Houssam Nassif, Daniel Hill, Sriram Srinivasan, Mitchell Goodman, Vijai Mohan, and S.V.N. Vishwanathan. 2016. Adaptive, Personalized Diversity for Visual Discovery. In *Proceedings of the 10th ACM Conference on Recommender Systems* (Boston, Massachusetts, USA) *(RecSys '16).* Association for Computing Machinery, New York, NY, USA, 35–38. https://doi.org/10.1145/2959100.2959171

[27] Sabina Tomkins, Steven Isley, Ben London, and Lise Getoor. 2018. Sustainability at Scale: Towards Bridging the Intention-Behavior Gap with Sustainable Recommendations *(RecSys '18).* Association for Computing Machinery, New York, NY, USA, 214–218. https://doi.org/10.1145/3240323.3240411

[28] Saúl Vargas and Pablo Castells. 2011. Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems* (Chicago, Illinois, USA) *(RecSys '11).* Association for Computing Machinery, New York, NY, USA, 109–116. https://doi.org/10.1145/2043932.2043955

[29] Longqi Yang, Yin Cui, Yuan Xuan, Chenyang Wang, Serge Belongie, and Deborah Estrin. 2018. Unbiased offline recommender evaluation for missing-not-at-random implicit feedback. In *Proceedings of the 12th ACM Conference on Recommender Systems.* 279–287.

[30] Longqi Yang, Michael Sobolev, Yu Wang, Jenny Chen, Drew Dunne, Christina Tsangouri, Nicola Dell, Mor Naaman, and Deborah Estrin. 2019. How intention informed recommendations modulate choices: A field study of spoken word content. In *The World Wide Web Conference.* 2169–2180.

[31] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)* 52, 1 (2019), 1–38.

[32] Zhe Zhao, Lichan Hong, Li Wei, Jilin Chen, Aniruddh Nath, Shawn Andrews, Aditee Kumthekar, Maheswaran Sathiamoorthy, Xinyang Yi, and Ed Chi. 2019. Recommending What Video to Watch next: A Multitask Ranking System. In *Proceedings of the 13th ACM Conference on Recommender Systems* (Copenhagen, Denmark) *(RecSys '19).* Association for Computing Machinery, New York, NY, USA, 43–51. https://doi.org/10.1145/3298689.3346997

[33] Lixin Zou, Long Xia, Zhuoye Ding, Jiaxing Song, Weidong Liu, and Dawei Yin. 2019. Reinforcement Learning to Optimize Long-term User Engagement in Recommender Systems. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* 2810–2818.