

Recommending Podcasts for Cold-Start Users Based on Music Listening and Taste

Zahra Nazari^{1*}, Christophe Charbuillet¹, Johan Pages¹, Martin Laurent¹, Denis Charrier¹,
Briana Vecchione², Ben Carterette¹

¹ Spotify, New York, NY

² Cornell University, Ithaca, NY

ABSTRACT

Recommender systems are increasingly used to predict and serve content that aligns with user taste, yet the task of matching new users with relevant content remains a challenge. We consider podcasting to be an emerging medium with rapid growth in adoption, and discuss challenges that arise when applying traditional recommendation approaches to address the cold-start problem. Using music consumption behavior, we examine two main techniques in inferring Spotify users preferences over more than 200k podcasts. Our results show significant improvements in consumption of up to 50% for both offline and online experiments. We provide extensive analysis on model performance and examine the degree to which music data as an input source introduces bias in recommendations.

CCS CONCEPTS

• **Information systems** → **Collaborative filtering**; **Personalization**; **Recommender systems**; • **Human-centered computing** → **Collaborative filtering**.

KEYWORDS

Podcast Recommendations; Cold start Recommendations; Cross-domain Recommendations

ACM Reference Format:

Zahra Nazari, Christophe Charbuillet, Johan Pages, Martin Laurent, Denis Charrier, Briana Vecchione, 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 (SIGIR '20)*, July 25–30, 2020, Virtual Event, China. ACM, NY, NY, USA, 10 pages. <https://doi.org/10.1145/3397271.3401101>

1 INTRODUCTION

Recommender systems provide an important mode of discovery for users, with more and more content served by platforms via recommendation engines. Furthermore, platforms are increasingly diversifying, adding videos, podcasts, and other types of media to

the content they have traditionally served. These platforms face the challenge of recommending this new content to users who have never sampled it on that platform and therefore have no history from which to infer their preferences.

Consider podcasts, a relatively new media that has seen rapid growth in recent years: in the US alone, there were *three times* as many adults listening to podcasts in 2019 as in 2006, and the number of podcasts available for easy access on platforms like iTunes has grown to an estimated 500,000 in 2018, with over 18 million individual episodes in over 100 different languages. Podcasts now comprise an estimated 33% of audio-only media. In addition, more and more news and entertainment outlets are turning to podcasts for content delivery. They are indisputably a major channel for users' consumption online.

Many of the platforms that users turn to for podcasts are also used for other media, most notably music. Moreover, many of the users of those platforms have long histories listening to music (or consuming other content). It may therefore be possible to leverage consumption history for other media to recommend podcasts in the absence of both content analysis and listening history. This is known as *cross-domain recommendation*.

The first question we explore in this paper is whether cross-domain recommendation, specifically from music listening history and preferences to podcasts, can be effective. It is not clear that this would be the case, as the two mediums differ in many important ways: instrumentation versus spoken word, relatively short songs versus episodes ranging in length from minutes to hours, topical content, and so on. Thus our first research question is:

RQ1: What is the effectiveness of cross-domain models for recommending podcasts to cold-start users based on music preferences?

We investigate this by implementing several different models for podcast recommendation on a popular platform for streaming music and podcasts. Our best models show up to 50% improvement over simple popularity-based recommendation models in both offline and online tests with real listeners.

We also wish to investigate different representations within the source domain and how they affect the quality of the recommendations. Users' music listening, taste, and preferences can be represented in a myriad of ways, from simple counts of artist plays to complex embedding models. Our second research question is:

RQ2: What is the most efficient and effective way to represent the user's preferences in the source music domain?

We consider three different ways to represent users in the source domain, with increasing fidelity and complexity of computation.

* Corresponding author: Zahra Nazari, zahran@spotify.com.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR '20, July 25–30, 2020, Virtual Event, China

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-4503-8016-4/20/07...\$15.00

<https://doi.org/10.1145/3397271.3401101>

Interestingly, simpler approaches seem to work as well as complex ones.

This leads into our final question, about what we can explain about recommendations in the destination domain.

RQ3: What can we learn about how music preferences influence podcast recommendations?

Specifically, how do our models introduce or propagate bias due to using music as a source? Simpler representations turn out to be ideal for investigating such questions.

The rest of this paper is organized as follows: in Section 2 we briefly summarize previous work on podcast recommendations, cold-start recommendation, and cross-domain recommendation. In Section 3 we present our models based on transferring representations of music taste to podcasts. Section 4 describes our offline experiments and results, with additional analysis in Section 5. We conclude in Section 6.

2 PREVIOUS WORK

Here, we summarize previous work on podcast recommendation, cold-start recommendation, and cross-domain recommendation.

2.1 Podcast Recommendations

While podcast consumption has grown rapidly in recent years, there is comparatively little research on podcast recommendation. One of the few papers is recent work by Yang et al., which compares the effects of intention-informed recommendations with classic intention-agnostic systems and shows that a recommender can boost a user’s aspiration-based consumption [25].

On the volume of previous work available, the Yang et al. paper cited above itself notes that only one other prior work seemed to be relevant: that of Tsagkias et al. that “predicted users’ podcast preference using hand-crafted preference indicators” [22].

Thus, we claim that this is one of the first studies directly on the topic of podcast recommendation.

2.2 Cold-start Recommendations

Podcasts are produced as shows, where each show has a specific theme and releases episodes periodically. Contrary to other domains such as news, where items have a short life span, podcast shows usually last for longer times. This allows systems to gather enough interaction signals for each podcast to match it’s audience base. Therefore, the main focus of this paper is to examine solutions for user cold-start rather than item cold-start problem. The user cold-start problem [1, 20] refers to the challenge of making recommendations for users with little to no prior history with the service or medium, or for items that have little historical interaction to draw on. It is an important problem because cold-start recommendations often form a user’s first impression of a recommendation service; if they are poor, the user may be permanently put off. It remains a difficult challenge, however.

Xu et al. categorize cold-start collaborative filtering (CF) approaches into three classes [24]:

- (1) Onboarding, by which a new user is initially presented with some set of items and asked to express their preferences, often by ratings. This solution goes back to work by Rashid

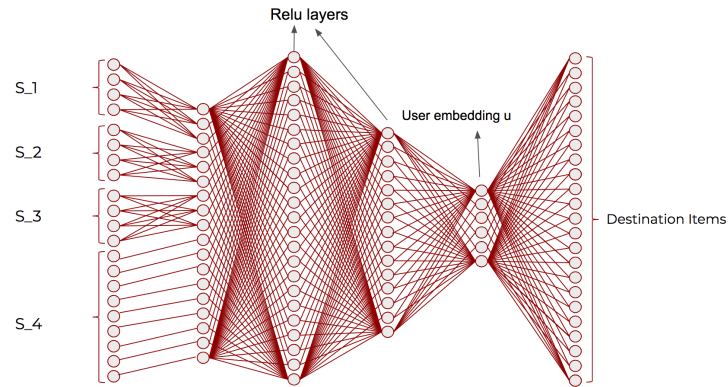


Figure 1: General model for source to destination mapping. The model could include both sparse features (S_1, S_2, S_3) as well as dense ones (S_4). Sparse features are learnt through the same end-to-end loss function.

et al. [17]. Onboarding provides necessary data to start making recommendations, but imposes a burden on the user to devote time and energy to this process.

- (2) Using features of the user [26], the user/item pair, social relationships that can be crawled or scraped from public information [11], etc. Such information can be valuable for making recommendations, but, as we shall see below, is probably not sufficient for making the best possible recommendations. Furthermore, methods that rely on this information will completely fail when it is not available (due to privacy regulations, omitted data, invisible social media profiles, etc).
- (3) Dynamically updating user representations in response to user ratings as they appear. Methods in this class generally try to provide a “good enough” recommendation slate for a new user, then quickly update the user representation as the user starts rating items. Example approaches include the incremental singular value decomposition (iSVD) method [19] and the incremental matrix factorization method [18, 21]. However, neither method can solve the problem of providing “good enough” initial recommendations, nor are they usable for users that do not provide ratings.

In this paper, we introduce a novel approach specifically for podcast recommendation based on users’ prior history with other media available on a recommendation platform. Similar to the latter two classes above, our method relies on the availability of historical interaction data, but offers the advantage that nearly all users seeking podcast recommendations will have already used the platform for music. In this way, we are able to utilize a wealth of information about their music preferences which are potentially applicable for podcast recommendation.

2.3 Cross-domain Recommendations

Cross-domain recommendation refers to the problem of using representations of users and items in one domain to make recommendations in another domain. In our case, we show how to use music preferences to make podcast recommendations.

Fernandez-Tobias et al. [7] provide a formal description of the problem in cross-domain recommendation tasks and summarize the existing methods into two categories of content-based and collaborative filtering-based.

In another work, Fernandez-Tobias et al. [8] introduce a matrix factorization model that jointly exploits user ratings and item meta-data to make recommendations on books, movies, and music. Their approach can provide music recommendations to users who have only rated movies and so forth, but it does rely on having some source of ratings.

Liu et al. [12] use a neighborhood-based transfer learning to leverage app installation behavior to make recommendations for news. They show significant improvements over popularity-based news recommendations, though overall performance is quite low.

Mirbakhsh and Ling [14] construct a rating matrix that includes available ratings on both of the domains. They use biased matrix factorization to map this matrix into a lower-dimensional latent space, then use k-means to categorize users and items.

Wongchokprasitti et al. [23] propose using a user model that can be shared amongst recommendation services, with a system that maintains the user models based on user interactions with each of these services. If a single service that provides recommendations in multiple domains fits this description, this is most similar to our proposed approach.

Elkahky et al. [6] introduce a content-based approach to cross-domain recommendation, specifically by modeling users across domains. Finally, Lian et al. [10] use content features to boost their performance in cross-domain recommendations.

2.4 Our Contribution

First, as suggested in Section 2.1, this work is one of the few on the problem of podcast recommendations in general, and as far as we are aware the first on using music as a source for cross-domain recommendation for the purpose of cold-start podcast recommendation. Our results—50% improvement over popularity-based recommendations in both offline and online tests—show the potential for future research in this area. Furthermore, our work provides insight into the representation of music tastes for the purpose of podcast recommendation, in particular for explaining recommendations and for investigating bias in them.

3 RECOMMENDING PODCASTS BASED ON MUSIC TASTE

Although podcasts and music are both consumed auditorily, they are consumed by users in different ways. Because the most important content within a podcast is spoken, podcasts as a consumption mechanism may be likened more to books than to music. In addition, patterns of consumption vary considerably between music and podcast listening: the number of tracks an active music listener typically listens to during a day is around 20 times more than the similar number for podcasts. However, we consider podcast listening to be a more significant time investment due to the fact that a typical podcast is around 10 times longer than a song. We also find the phenomenon of repeat listening to be a common behavior in music, whereas a single podcast (like a book) is rarely consumed more than once by the same user.

Ways of modeling and representing music have been studied for a long time, not just by computer scientists but by musicologists, sociologists, and others. From these studies there are a wide variety of possible features that could potentially be used to represent a user's preferences: audio features, classification features such as genres, co-listening behavior features, co-occurrences on user-made playlists, etc. Thus we need our model to accept a variety of heterogeneous features in the source domain.

Some of these features would be very sparse. Take, for example, genre: some catalogs define more than 3,000 different distinct genres, with any given song falling into only one genre. Few listeners regularly listen to more than a small fraction of genres, so a one-hot encoding of genre would result in a very sparse matrix. We require a model that could handle this as well.

Other representations of user taste could be very dense. We describe one such representation based on playlist co-occurrence below. Our model also needs to be able to take a dense representation as a feature.

Finally, podcasts, as a new medium, have received comparatively much less study. We do not know of any work that defines “standard” features for podcast representation, nor do we know of many features that correspond to those mentioned above for music. Rather than develop novel podcast features, we simply use the fact of a user following a podcast as the signal to train to. This is therefore an extreme multi-class classification problem, with each podcast show being a possible class.

3.1 Cross-Domain Transfer

As we discussed in Section 2.3, cross-domain recommendation concerns making recommendations in a destination domain D based on representations of users and items in a source domain S .

Based on the requirements emerging from the discussion above, and inspired by the CBOW model [13] originally proposed for language modeling, we use the framework of recommendations as an extreme multi-class classification task modeled by a multi-layer perceptron (MLP). Previously reported successes of MLPs in various recommendation domains [2, 4] and in large scale applications [3, 5, 27], along with their flexibility with heterogeneous sets of features, make this method an appropriate candidate for our problem.

Given a source domain S and a representation $U(S)$ of the user in S , we train a neural net that maps a user to a distribution over items in the destination domain D . A representation of sparse features (e.g., 1 , S_2 and S_3 in Figure 1) could be learnt through an end-to-end training, while dense features (e.g., S_4 in Figure 1) could be memorized through the network and concatenated to the final user representation at any stage.

We use a softmax classifier to minimize the cross entropy loss for the true label and the negative samples. Let $i \in D$ be the label, and d_i be an N -dimensional vector representation of i . A user's preferences, $U(S)$, are represented as an N -dimensional vector u .

$$P(i|U(S)) = \frac{e^{d_i u}}{\sum_{j \in D} e^{d_j u}}$$

Optimizing the network this way would result in the dense vector $u \in \mathbb{R}^N$ being closer to the item i 's vector as the weights of the node i in softmax classifier: $d_i \in \mathbb{R}^N$.

Because each podcast show is a class, we need to be able to handle a large number of classes in training. We use importance sampling, a negative sampling approach proposed by Jean et al. [9] to have the model converge in efficient time. The loss function is then calculated as:

$$J_{\theta} = - \sum_{i \in D} [\log \sigma(ud_i) + \sum_{j=1}^k \log \sigma(-ud_{ij})]$$

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

Given this modeling framework, we now consider music as the source domain, with S_1, S_2, S_3, S_4 in Figure 1 illustrating various sparse and dense features of a user's listening habits. We use podcasts as the destination. Training the network learns a mapping from music to podcast preferences.

3.2 Representing Music Taste

In order to efficiently represent a user's source profile for cross-domain recommendation, we consider a range of approaches, starting from simple demographic-based heuristics to more complex pre-trained representations of music taste. These approaches are based on features that are available on a popular music streaming service.

3.2.1 Demographics. The simplest user representation uses basic demographic information about a user to infer their music taste. This information may include country, age, and gender, all of which a user can choose to self-report on many platforms.

3.2.2 Metadata. We can generate a better representation of music preferences using music metadata information such as artist and genre. We represent users as a simple aggregation over their music consumption metadata. For example, users could be represented by their top listened artists and genres. We also use manual annotations for three levels of genres: meta-genre, a high-level music category such as "folk" or "rap"; genre, which includes more specific types of music such as "blues" or "hip hop"; and micro-genre, which defines niche subgenres and includes labels such as "Texas blues" or "East coast hip hop". Overall, our data has more than 1.3 million distinct artists, 40 genres and 3855 micro-genres with which to represent users.

3.2.3 Latent Representation. In a more complex context, we can take a modelling approach and use contextual and collaborative information to embed users into a latent representation in their source domain. Either CF, content filtering, or a hybrid approach could be sufficient to calculate these representations. In order to obtain this, we use a rich dataset of user-created playlists.

First, we represent all tracks in our catalog using a high-dimensional vector trained based on co-occurrence of tracks in a playlist. The vectors were trained using an embedding model inspired by word2vec [16] which was originally introduced for learning word embeddings. More specifically, we used the Word2vec Skip-gram model, a shallow neural network with a single hidden layer that takes a track as input and predicts the rest of the tracks in the playlist. Let's consider a playlist containing $song_A$, $song_B$, and $song_C$ as an example. The shallow neural network is then trained using $song_B$ as the

input with $song_A$ and $song_C$ as the target. The network exploits the co-occurrence of songs in a playlist to learn a high dimensional vector representation for each track, where similar tracks will have weights that are closer together than tracks that are unrelated. A user representation is then calculated as recency-based weighted average over their listening behavior. The resulting output is a dense vector that could directly be used as the user embedding.

4 EXPERIMENTS AND RESULTS

We now turn to experiments. In this section, we describe a set of offline experiments and one online experiment for cross-domain podcast recommendations based on music.

4.1 Data

Our data consists of 17 million users who follow at least one podcast on Spotify. On average, each user follows 2.9 podcasts. In addition to the podcasts they follow, each user has a list of songs (referred to as "tracks") that they have listened to on the platform along with annotated information for each track, including artist and genre.

Specifically, the data includes:

4.1.1 Demographics. For each user, we have access to self-reported information on country, gender and age. In this group of users, 17 different self-reported countries of origin are reported. Gender is classified and reported as male, female, neutral or none. Users' age is partitioned into eight buckets including "unknown".

4.1.2 Metadata. For each user, we have a list of tracks that user has listened to on the platform during the previous 90 days. Each track is associated with various metadata including artists, genres, meta-genres, and micro-genres. These genres are manually tagged by music experts.

4.1.3 Music Taste Embeddings. In addition to the user's music behavior, we use a dataset of nearly 700 million playlists curated by users. Each playlist consists of 10 tracks on average, each of which has been put together by one or more users on platform. These playlists are used in a CF approach using word2vec to create 40-dimensional vector representations of tracks as described above. A user embedding can be calculated simply as a recency-weighted average over track embeddings.

Note that these embeddings are very costly to compute. In practice it takes several hours of computation time, and they need to be recomputed frequently due to changes in the catalog and in user playlists. Furthermore, they are opaque: they provide no intuition for what a value on any given dimension represents, and thus do not explain anything about recommendations.

4.1.4 Podcast Interests. Spotify gives users the option to follow their favorite podcast shows. For each user we have a list of podcast shows that they follow.

4.2 Offline Experiments

4.2.1 Training and test splits. We split the data by user ID, with about 200,000 unique users in the test set and the rest in the training set. The training set contains one instance for each individual user/podcast follow, so if one user follows 10 podcasts there will be 10 separate instances for that user reflecting listening history,

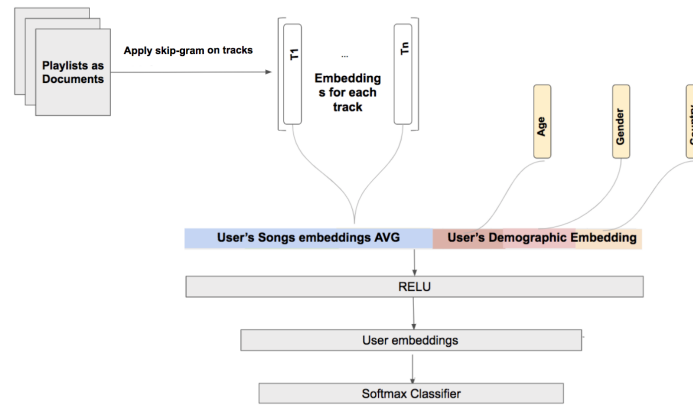


Figure 2: Learning from Music Representation

demographics, and an indication that they follow a specific podcast. By comparison, the test set contains one instance per individual user. After training, models rank a set of podcasts for each user in the test set, and that ranking is evaluated by retrieval metrics such as precision and recall (see Section 4.2.3 below).

4.2.2 Models. We compare nine different models for podcast recommendation. Two are baselines that rank popular podcasts, three are cross-domain CF models, and the remaining four combine CF models with demographic data.

- (1) **Popularity by country:** This is a simple popularity-based model that recommends the most popular podcasts in the user's country of origin.
- (2) **Popularity by country and demographic:** This refines the previous model to recommend popular podcasts for the user's country of origin as well as demographic (gender and age group).
- (3) **Cross-domain CF using logistic regression:** Since we define our problem as a pure cold-start task, we use popularity per demographic as an optimal baseline. In order to motivate the need for an MLP approach, we additionally evaluate a logistic regression model as a second baseline. This model uses as inputs the embeddings described above to predict a podcast follow for each user. In order to ensure fair comparison, we used the same softmax layer and negative sampling approach used by all other models in the study.
- (4) **Cross-domain CF:** This approach introduces our first MLP based model, which uses the embeddings described above and podcast follows in the training data to fit podcast vectors. Podcast recommendations are produced from the k nearest neighbors to the user embedding in the test set.
- (5) **Cross-domain CF with metadata:** This model uses a user's top- m listened artists, genres, meta-genres, and micro-genres to fit podcast vectors.

- (6) **Demographics + cross-domain CF:** This model uses the user embeddings along with vectors representing demographic features to fit podcast vectors. Figure 2 shows the complete model.
- (7) **Demographics + cross-domain metadata CF:** Similarly, this model uses top- m entities in metadata categories along with demographic features.
- (8) **Cross-domain CF + metadata:** This model combines the user embedding with the metadata categories, but does not include demographic features.
- (9) **Demographics + all music data:** This model combines the user embedding, metadata, and demographic features into a single model.

4.2.3 Metrics. As mentioned above, each model produces a ranking of podcasts for each user in the test set. Given these rankings, the gold standard is whether that user actually follows the podcasts suggested in the ranking. This is evaluated using the standard retrieval metrics precision, recall, and nDCG.

Specifically, we look at precision at rank 1, defined as the proportion of user requests in the test set that are recommended a podcast they follow at rank 1. We consider the limitation that this metric may be too coarse, as it can only be zero or one. Because of this, we also look at precision at rank 10 to investigate the proportion of podcasts a user follows that we are able to correctly recommend. Please note that since on average each user follows 2.9 podcasts, the upper bound of precision at rank 10 is about 0.29.

nDCG weights each item by a function of the rank at which it appears. This is important for the context of podcast recommendations because users typically see only a few recommendations in the interface and must scroll or click to see more. The effort this takes means that fewer users see lower-ranked recommendations, so relevant recommendations not initially seen by the user are down-weighted by rank to capture this. nDCG discounts relevance by $\log(\text{rank} + 1)$.

4.2.4 Hyper-parameter tuning. We use a portion of our training data as the validation set to tune a number of hyper-parameters: the embedding size for each of the input features, the number of fully connected layers, the final layer size for user embeddings, and the number of negative samples. For demographic features, we experimented with 5, 10 and 15 as the embedding size, and found using 10 for each feature worked best. Increasing the hidden fully connected layers from one to two increased the accuracy, but having three layers did not improve the results. For user embedding layer size, we observed that increasing the size from 20 to 40 improved the results, while any size larger than 40 dropped the accuracy. For number of nodes in each hidden layer, we ran experiments with 128, 256, 512, 1024. Although increasing the number of nodes in each hidden layer resulted in slightly better results, we decided to use 512 to keep our training time under 24 hours. The last parameter was the number of negative samples, where we experimented with 256, 512, 1024, 2048, 4096, and found 2048 to perform best.

4.3 Results

Table 1 summarizes offline results for all eight models, with models grouped as described in Section 4.2.2. We used Tukey’s Honest Significant Differences to perform statistical significance testing with correction for multiple comparisons; all differences are statistically significant except for those noted by superscripts.

It is clear that all of our music cross-domain models perform significantly better than popularity-based models, even when those models are tuned to country of origin and demographic. This directly answers our first research question, that indeed music-based cross-domain models are effective in recommending podcasts.

The demographic+country popularity model is 20% worse than the cross-domain CF model, while the country-only popularity model is 36% worse. This trend is seen across all measures, and while the % differences vary, it is always the case that both popularity models are substantially worse than the CF model. Moreover, the country popularity model is substantially worse than the country+demographic popularity model. All of these differences are statistically significant.

Another insight from our evaluations is that although a logistic regression model over user embedding vectors improves the popularity baselines, the results motivate the need for having non-linear models to have better results. This goal is achieved using multi-layer perceptrons in the next sets of models.

We continue to compare other models based on CF, starting with the two MLP based cross-domain models. The metadata model is the clear winner, outperforming the CF model by 11% on nDCG@10 and similar margins on other metrics. Though both of these methods are based on CF, the metadata model uses more granular information about music tastes. This appears to benefit podcast recommendation. Combining these two methods gives superior results than either do independently. The CF + metadata model shows a 12% improvement over the base CF model, though less than 1% over the metadata model, suggesting that the metadata model captures most of what is necessary about music tastes to recommend podcasts, without the need to perform the word2vec algorithm on the playlist data.

Adding demographic information to these two models continues to boost effectiveness, though at a lesser rate. For the base CF model,

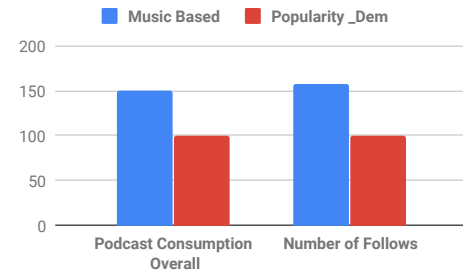


Figure 3: Results from an online experiment comparing the country + demographic popularity model to the music-based cross-domain CF + metadata model. Note that results are scaled such that the baseline is 100% in order to directly reflect the percent improvement.

the gain in nDCG@10 is about 8%; for the metadata model, the gain is only about 3%. We see similar gains across other metrics.

Finally, the best results in an absolute sense are achieved when CF is combined with metadata-based CF and demographic information. Note, however, that adding the CF does not present significant gain in effectiveness—only the difference in nDCG@50 is statistically significant. This suggests that the complexity and computation time of the vector representation is not worth it. Metadata-based representations are far easier to compute and provide essentially the same benefit. This helps answer our second research question regarding efficient and effective ways to represent user’s preferences.

4.4 Online Experiments

The ultimate objective of this work is to be able to match new users to relevant podcasts without any prior knowledge about their podcast affinities. Therefore, we implemented our best performing model—the demographic + cross-domain CF + metadata model—in production on Spotify and compared it with our best performing baseline. The test was performed on 800,000 users with no history of podcast consumption on the platform. The users were randomly split into two groups: the control group received a slate of 10 podcast recommendations generated by the country + demographic popularity model, and the treatment group received a slate of 10 podcast recommendations generated by the demographic + cross-domain CF + metadata model. Over the course of one week, we measured podcast consumption in minutes and the number of shows newly followed by these users.

Figure 3 shows the results. The group exposed to music-based recommendations listened to nearly 50% more podcast minutes than the control group and followed over 50% more shows. This provides validation of our offline experiment and offers striking confirmation of our first research question: that music tastes can predict podcast interest.

5 ANALYSIS

In this section, we first examine models’ effectiveness across various user cohorts. Then, we present diversity-related effects of using music as a training source for prediction. Lastly, we describe several anecdotal examples where newer models perform better than

Model	ndcg@10	ndcg@50	precision@1	precision@10	recall@10
country popularity	0.12256	0.17641	0.07722	0.04188	0.19027
country + demo popularity	0.15310	0.20817	0.10447	0.04814	0.21850
cross-domain CF Logistic Regression	0.16610	0.21690	0.11779	0.04974	0.23268
cross-domain CF	0.19029	0.24605	0.13939	0.05680	0.26231
cross-domain metadata CF	0.21212	0.26933	0.16056	0.06274	0.28710
demo + CF	0.20568	0.26399	0.15197	0.06129	0.28199
CF + metadata	0.21327	0.27058	0.16173	0.06294	0.28877
demo + metadata	0.21938 ^a	0.27762	0.16631^b	0.06486 ^c	0.29658 ^d
demo + CF + metadata	0.22009^a	0.27866	0.16623 ^b	0.06503^c	0.29763^d

Table 1: Results of offline experiments comparing podcast recommendation using models based on popularity, CF, CF in metadata, and combinations of the latter two with demographic information. Most differences are statistically significant. Pairs marked ^a, ^b, ^c, and ^d are *not* significantly different.

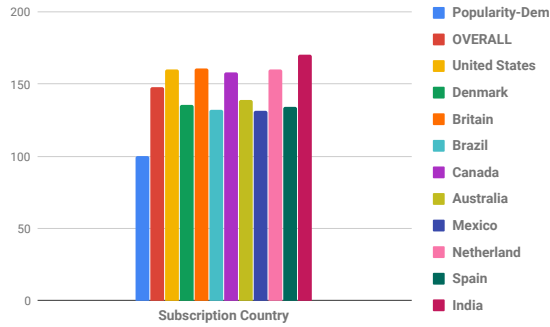


Figure 4: Podcast follow prediction performance of our best model compared to popularity baseline across countries. The blue and red bars are overall results for this group of countries. Each subsequent bar shows the relative improvement over the country + demographic popularity model for that country.

the baseline. These analyses shed insight into the model’s learned associations between music tastes and podcasts, our third research question.

5.1 Model Performance in User Cohorts

In this section, we analyze model performance across the following four dimensions of interest:

- **Country of origin:** One of the main motivations of this work is to be able to recommend relevant spoken content without going through the costly process of understanding content across a variety of languages. Therefore, performance across different countries and languages is an important target. Figure 4 shows the podcast follow results for our largest user base countries. We improve the performance by at least 30% across all countries.
- **Time on platform:** The amount of music listening history required for high model performance is important to accurately recommend podcasts of interest. The age of each user’s account is a good proxy for this; Figure 5 shows that the best

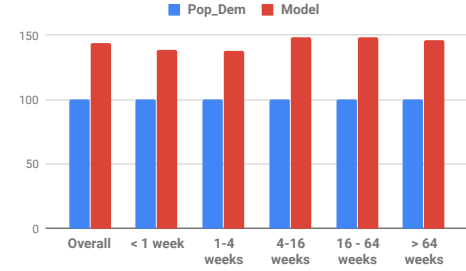


Figure 5: Podcast follow prediction performance of our best model compared to popularity baseline across account ages.

performing model is able to outperform the popularity-based model using less than a week of music consumption information. In fact, users do not need to have long listening histories in order to receive good podcast recommendations.

- **Age bucket:** We are interested in evaluating performance across all age ranges in order to see if some age ranges were being underserved by the model. Figure 6 shows that our model performs best in the [25-29] age bucket with a 50% improvement over the baseline, but performs well across all age buckets.
- **Gender** Finally, our model’s performance for self-reported genders are shown in Figure 7. We did not observe any significant difference between male and female categories.

In addition to the positive results shown here, this also shows that a demographic representation is suitable for asking further questions about the effectiveness of recommendations.

5.2 Diversity Analyses

We evaluated various dimensions of recommendations produced by a subset of models in order to understand the extent to which our input data may propagate bias into our recommendations. Specifically, we assess the models for instances of *popularity bias*, which occurs when recommenders prioritize popular items much more highly than other items in a long tail distribution regardless of user

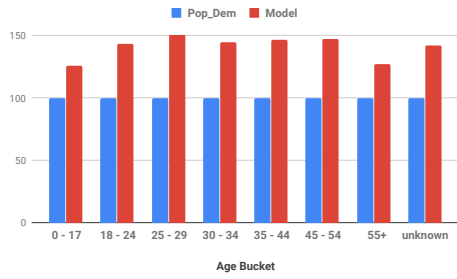


Figure 6: Podcast follow prediction performance of our best model results compared to popularity baseline across user age buckets.

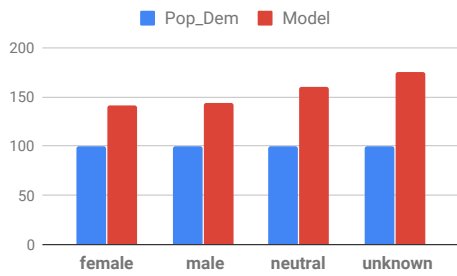


Figure 7: Podcast follow prediction performance of our best model results compared to popularity baseline across user genders.

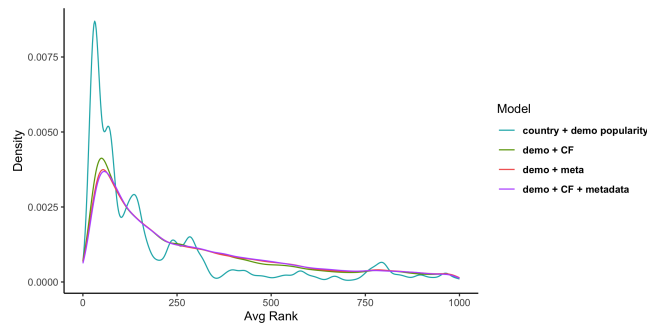


Figure 8: Popularity distribution for podcast recommendations across models.

taste [15]. The goal here is to assess whether popularity bias is necessary to achieve better performance.

We explored the extent of heterogeneous recommendations according to their overall popularity rank for models: “demo + CF”, “demo + metadata”, and “demo + CF + metadata”. We compare these recommendations against the baseline model, “country + demo popularity.” Figure 8 shows the average rank of top 10 podcast recommendations across models for a random sample of users. The baseline shows a long tail distribution skewed towards recommendations from top-ranked positions, while models with improved

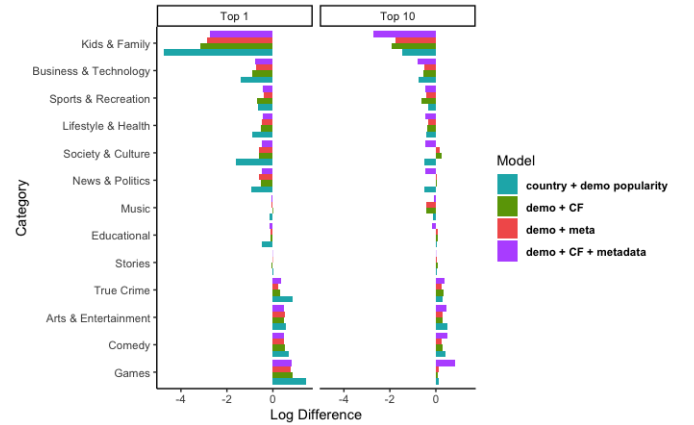


Figure 9: Podcast category differences compared to user follows across models for top 1 and top 10 recommendations.

performance show considerably less homogeneity. It is notable to mention that our best performing model suggests the least homogeneous set of recommendations on average. This implies that we are able to simultaneously show users more niche and accurately-tailored content while also improving performance.

Next, we assessed the extent to which our models introduce *behavioral bias*, which we define as the difference between organic user consumption behavior and model recommendations. Because models are trained using music listening data, we wanted to explore whether models bias users to consume higher rates of podcasts annotated with a music label. We use annotated podcast categories as our dimension to approximate organic behavior. Since each podcast may be labeled with up to ten categories in no order of importance, we weight categories and aggregate the percentage of recommendations each category comprises. We calculate the log difference for each podcast category, which is defined as $\log(x_2/x_1)$ where x_1 represents the percent of organic user follows for each category and x_2 represents the percent each category is recommended across models. Figure 9 shows the difference between categories that users follow organically on the platform compared to the top n recommendations suggested across models. Educational content is the most followed category while content labeled as kids & family is followed the least. We see that more popular categories tend to show smaller differences while less popular categories are recommended less often to users. The figure shows that our best performing model is able to reflect organic user behavior as well as serve users with a more equal distribution of podcast categories compared to the popularity baseline. Overall, the models perform in accordance with organic user behavior and do not show immediate suggestions of behavioral bias related to category.

5.3 Qualitative Analyses

We examined some individual podcast metadata in order to gain insight about the associations between music taste and podcasts being generated by our models. Specifically, we selected for US users due to our better understanding around both content as well as accuracy of genre annotations. Figure 10 shows some podcast

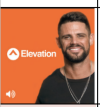




	Podcast	Podcast Description	Top Genres
	Elevation Church with Steven Furtick	Elevation with Steven Furtick is a weekly podcast featuring the most recent sermons and exclusive content from Elevation Church.	contemporary christian worship christian music anthem worship christian alternative rock
	Se Regalan Dudas's podcast	A Spanish Podcast created by Lety Sahagún and Ashley Frangie created this podcast to open a space where they invite experts, friends and people who admire, who know and do not know all that we should be talking about.	latin latin hip hop latin pop rock en espanol tropical
	The Joe Budden Podcast	A top music podcast and go-to listen for hip-hop and rap culture enthusiasts.	gangster rap vapor trap atl hip hop urban contemporary underground hip hop
	Fantastic Geeks and Where to Find Them	Join best friends and YouTubers Anna Brisbin and Tessa Netting as they cover all things nerdy including TV shows, books, movies, animation, and pop culture. Listen in as these two proud Ravensclaws debate ship wars, play games like Would You Rather, and sort characters into Hogwarts	show tunes broadway electropop indie popitism Hollywood
	Something Scary	Do you wanna hear something scary? Join Sapphire every week as she brings you the creepiest ghost stories, urban legends, and folktales.	electropop indie popitism otacore underground pop punk

Figure 10: Some examples of podcasts that online users followed when recommended by the demographic + cross-domain CF + metadata model, along with the top genres preferred by those users.

examples and their descriptions. These podcasts were selected at random from the pool of users for whom our model was correctly recommending podcasts when the baseline model was failing to do so. For the sake of clarity, the top five most popular genres are removed from the list.

Some of these examples show interesting connections between podcasts a user subscribes to and music genres they listen to. For example, *Elevation Church with Steven Furtick* is a religious podcast, so users who are correctly receiving this podcast as a recommendation have genre tastes including contemporary Christian, worship, Christian music, anthem worship, and Christian alternative rock. Although language is not an explicit input to these models, language seems to be learned by the model, e.g. Spanish podcasts are recommended to users who have an affinity for Latin music but may not necessarily be located in a Spanish-speaking country. We also noticed that a strong music-related theme in a podcast could be seen in the type of music their subscribers are listening to. Although we show genre information in Figure 10 for the sake of explainability, the models that do not use genre information, i.e. demo + CF, correspond with generally equivalent results. This leads us to believe that our models indirectly pick up music genres and do not need fine-tuned annotations to perform well.

6 CONCLUSION

In this paper, we examined the viability of using past music listening and representations of musical taste as a source domain to recommend items in the target domain of podcasts at Spotify. Due to differences in media types and consumption patterns, it appears there should be little reason for music taste to be able to accurately predict podcast listening. Yet, even with no prior information about

users' interest in podcasts, we are able to increase both minutes listened and podcasts followed by 50% in online experiments, answering our first research question positively and decisively.

We also compared two main approaches in leveraging rich interaction signals in a source domain for recommendations in a destination domain. We showed that although an approach that includes both an item-based CF model and meta-data signals would do best, the meta-data based method alone could achieve comparable results that are both more explainable and more efficient. This answers are second research question about representations.

Our analysis shows that these models work well in all countries, across all ages, for all genders, and do not require a great deal of listening history to perform well for podcast recommendation. Further, these models are able to decrease popularity bias without over-representing music-related content. This analysis is suggestive towards our third research question regarding the effect of music preferences on podcast recommendations, though we believe there is still more work to be done.

For future work, we plan to test this model for warm-start users as well. In this task, we expect to recommend new podcasts to users who already follow one. Additionally, we plan on incorporating podcast features as well, such as language, distributor, and audio embeddings. It seems intuitive that both of these directions will significantly impact user engagement with podcasts. We also note in Section 4.2.2 that as a side-effect of our models, we obtain vector representations of artists and genres which encode the relationship between a user's taste in music, their favorite artists and genres, and the podcasts they follow. We plan to investigate whether these vectors could be useful for larger-scale cross-domain recommendation as well as recommending artists, shows, and more.

REFERENCES

- [1] A Merve Acilar and Ahmet Arslan. 2009. A collaborative filtering method based on artificial immune network. *Expert Systems with Applications* 36, 4 (2009), 8324–8332.
- [2] 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*.
- [3] Cen Chen, Peilin Zhao, Longfei Li, Jun Zhou, Xiaolong Li, and Minghui Qiu. 2017. Locally connected deep learning framework for industrial-scale recommender systems. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 769–770.
- [4] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishikesh Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Isipir, et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*. ACM, 7–10.
- [5] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*. ACM, 191–198.
- [6] Ali Mamdouh Elkahky, Yang Song, and Xiaodong He. 2015. A multi-view deep learning approach for cross domain user modeling in recommendation systems. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 278–288.
- [7] Ignacio Fernández-Tobías, Iván Cantador, Marius Kaminskas, and Francesco Ricci. 2012. Cross-domain recommender systems: A survey of the state of the art. In *Spanish conference on information retrieval*. sn, 1–12.
- [8] Ignacio Fernández-Tobías, Iván Cantador, Paolo Tomeo, Vito Walter Anelli, and Tommaso Di Noia. 2019. Addressing the user cold start with cross-domain collaborative filtering: exploiting item metadata in matrix factorization. *User Modeling and User-Adapted Interaction* 29, 2 (2019), 443–486.
- [9] 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).
- [10] Jianxun Lian, Fuzheng Zhang, Xing Xie, and Guangzhong Sun. 2017. CCCFNet: A content-boosted collaborative filtering neural network for cross domain recommender systems. In *Proceedings of the 26th international conference on World Wide Web companion*. 817–818.
- [11] Jovian Lin, Kazunari Sugiyama, Min-Yen Kan, and Tat-Seng Chua. 2013. Addressing cold-start in app recommendation: latent user models constructed from twitter followers. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 283–292.
- [12] Jixiong Liu, Jiakun Shi, Wanling Cai, Bo Liu, WeiKe Pan, Qiang Yang, and Zhong Ming. 2017. Transfer Learning from APP Domain to News Domain for Dual Cold-Start Recommendation. In *RecSysKTL*. 38–41.
- [13] 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.
- [14] Nima Mirbakhsh and Charles X Ling. 2015. Improving top-n recommendation for cold-start users via cross-domain information. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 9, 4 (2015), 33.
- [15] Yoon-Joo Park and Alexander Tuzhilin. 2008. The long tail of recommender systems and how to leverage it. In *Proceedings of the 2008 ACM conference on Recommender systems*. 11–18.
- [16] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.
- [17] Al Mamunur Rashid, Istvan Albert, Dan Cosley, Shyong K Lam, Sean M McNee, Joseph A Konstan, and John Riedl. 2002. Getting to know you: learning new user preferences in recommender systems. In *Proceedings of the 7th international conference on Intelligent user interfaces*. ACM, 127–134.
- [18] Steffen Rendle and Lars Schmidt-Thieme. 2008. Online-updating regularized kernel matrix factorization models for large-scale recommender systems. In *Proceedings of the 2008 ACM conference on Recommender systems*. ACM, 251–258.
- [19] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2002. Incremental singular value decomposition algorithms for highly scalable recommender systems. In *Fifth international conference on computer and information science*, Vol. 27. Citeseer, 28.
- [20] J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. 2007. Collaborative filtering recommender systems. In *The adaptive web*. Springer, 291–324.
- [21] Gábor Takács, István Pilászy, Bottyán Németh, and Domonkos Tikk. 2008. Investigation of various matrix factorization methods for large recommender systems. In *2008 IEEE International Conference on Data Mining Workshops*. IEEE, 553–562.
- [22] Manos Tsagkias, Martha Larson, and Maarten De Rijke. 2010. Predicting podcast preference: An analysis framework and its application. *Journal of the American Society for information Science and Technology* 61, 2 (2010), 374–391.
- [23] Chirayu Wongchokprasitti, Jaakko Peltonen, Tuukka Ruotsalo, Payel Bandyopadhyay, Giulio Jacucci, and Peter Brusilovsky. 2015. User model in a box: Cross-system user model transfer for resolving cold start problems. In *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 289–301.
- [24] Jingwei Xu, Yuan Yao, Hanghang Tong, Xianping Tao, and Jian Lu. 2017. R a P are: A Generic Strategy for Cold-Start Rating Prediction Problem. *IEEE Transactions on Knowledge and Data Engineering* 29, 6 (2017), 1296–1309.
- [25] 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 Web Conference*. ACM.
- [26] Mi Zhang, Jie Tang, Xuchen Zhang, and Xiangyang Xue. 2014. Addressing cold start in recommender systems: A semi-supervised co-training algorithm. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 73–82.
- [27] 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.