AUTOMATIC MUSIC RECOMMENDATION SYSTEMS: DO DEMOGRAPHIC, PROFILING, AND CONTEXTUAL FEATURES IMPROVE THEIR PERFORMANCE?

Gabriel Vigliensoni and Ichiro Fujinaga
Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT)
McGill University, Montréal, QC, Canada
[gabriel,ich]@music.mcgill.ca

ABSTRACT

Traditional automatic music recommendation systems' performance typically rely on the accuracy of statistical models learned from past preferences of users on music items. However, additional sources of data such as demographic attributes of listeners, their listening behaviour, and their listening contexts encode information about listeners, and their listening habits, that may be used to improve the accuracy of music recommendation models.

In this paper we introduce a large dataset of music listening histories with listeners' demographic information, and a set of features to characterize aspects of people's listening behaviour. The longevity of the collected listening histories, covering over two years, allows the retrieval of basic forms of listening context. We use this dataset in the evaluation of accuracy of a music artist recommendation model learned from past preferences of listeners on music items and their interaction with several combinations of people's demographic, profiling, and contextual features. Our results indicate that using listeners' self-declared *age*, *country*, and *gender* improve the recommendation accuracy by 8 percent. When a new profiling feature termed *exploratoryness* was added, the accuracy of the model increased by 12 percent.

1. LISTENING BEHAVIOUR AND CONTEXT

The context in which people listen to music has been the object of study of a growing number of publications, particularly coming from the field of music psychology. Konečni has suggested that the act of music listening has vacated the physical spaces devoted exclusively to music performance and enjoyment long ago, and that music nowadays is listened to in a wide variety of contexts [13]. As music increasingly accompanies our everyday activities, the music and the listener are not the only factors, as the context of listening has emerged as another variable that influences, and is influenced, by the other two factors [11]. It has been also observed that people consciously understand these interactions [6] and use them when choosing music for daily life activities [23]. The context of music listening seems to influence the way in which people chooses music, and so music recommenders should suggest music items to fit the situation and needs of each particular listener.

Modelling the user needs was identified by Schedl et al. as one key requirement for developing user-centric music retrieval systems [20]. They noted also that *personalized systems* customize their recommendations by using additional user information, and *context-aware systems* use dynamic aspects of the user context to improve the quality of the recommendations. The need for contextual and environmental information was highlighted by Cunningham et al. and others [5, 12, 16]. They hypothesized that listeners' location, activity, and context were probably correlated with their preferences, and thus should be considered when developing music recommendation systems. As a result, frameworks for abstracting the context of music listening by using raw features such as environmental data have been proposed in the literature [16, 22]. While some researchers have reported that context-aware recommendation systems perform better than traditional ones [15, 22, 24], others have shown only minor improvements [10]. Finally, others have carried out experiments with only the most highly-ranked music items, probably leading to models biased by popularity [15, 25].

We will now discuss the impact of using listeners' demographic and profiling characteristics— hereafter referred to as *user-side* features [19]—in improving the accuracy of a music recommendation model. User-side features were extracted from self-declared demographics data and a set of custom-built profiling features characterizing the music listening behaviour of a large amount of users of a digital music service. Their music listening histories were disaggregated into different time spans to evaluate if the accuracy of models changed using different temporal contexts of listening. Finally, models based on latent factors were learned for all listening contexts and all combinations of user-side features. Section 2 presents the dataset collection, Section 3 introduces a set of custom-built features to profile listeners' listening behaviour, and Section 4 describes the experimental set up and presents the results.

Attribution: Gabriel Vigliensoni and Ichiro Fujinaga. "Automatic music recommendation systems: do demographic, profiling, and contextual features improve their performance?", 17th International Society for Music Information Retrieval Conference, 2016.

2. DATASET

We are interested in evaluating the impact of using demographic and profiling features, as well as contextual information, for a large number of people, on the prediction accuracy of a music artist recommendation model. A few publicly available datasets for music listening research provide information relating people and music items. Dror et al. presented a dataset of 1M people's aggregated ratings on music items [7]. McFee et al. introduced a dataset of song playcounts of 1M listeners [17]. Neither of these two datasets, however, provided timestamps of the music logs or demographic information about the listeners. Celma provided a dataset of playcounts with listeners' demographic data for 360K listeners and a set of listening histories with full time-stamped logs; however this last dataset only included logs for 1K listeners [3]. Cantador et al. presented another small dataset with song playcounts for 2K users [2]. Finally, EMI promised a dataset of 1M interviews about people's music appreciation, behaviour, and attitudes [9], but only partial information was made available.

None of the aforementioned datasets provide, at the same time and for a large amount of listeners, access to full music listening histories as well as people's demographic data. This means it is not possible to extract all of the user-side features that we were interested in, and so we decided to collect our own dataset made with music listening histories from Last.fm. Last.fm stands out from most online digital music services because it not only records music logs of songs played back within its own ecosystem, but also from more than 600 media players.

Next, we will present the criteria and acquisition methods used to collect a large number of music listening histories from the Last.fm service.

2.1 Data criteria, acquisition, and cleaning

Aggregating people's music listening histories requires collapsing their music logs into periods of time. In order to obtain data evenly across aggregated weeks, months, seasons, or years, we searched for listeners with an arbitrary number of at least two years of activity submitting music logs since they started using the system, and also with an average of ten music logs per day. These two restrictions forced our data-gathering crawler to search for listeners with a minimum of 7,300 music logs submitted to the Last.fm database. Also, these constraints assured us that we would collect listening histories from active listeners with enough data to perform a good aggregation over time.

Data acquisition was performed by means of using several machines calling the Last.fm API during a period of two years (2012–2014). We collected listening histories by using the Last.fm's API method `user.getRecentTracks()`. This API call allowed us to obtain full listening histories. Along with this data, we also stored all available metadata for each listener, including the optional self-declared demographic features: *age*, *country*, and *gender*.

We performed several processes of data filtering and cleaning in order to avoid noisy data. For example, we realized that there were listeners with numerous duplicated music logs (i.e., same timestamp for many music item IDs), and listening histories with a great deal of music logs that were too close in time (i.e., less than 30 seconds apart, which is the minimum that Last.fm requires to consider a played track as a valid music log). Hence, we decided to filter out all duplicated logs as well as logs that were less than 30 seconds apart in time.

2.2 Dataset demographics

Our dataset consists of 27 billion music logs taken from 594K users' music listening histories. This large repository of music listening records accounts for the interaction of listeners with more than 555K different artists, 900K albums, and 7 million tracks. There are music listening histories from people in 239 self-declared different countries, with listeners from all time zones represented. However, listeners from Africa, South Asia, and East Asia are under-represented in our dataset. In fact, the 19 "top countries" combined account for more than 85 percent of the total number of listeners in the dataset. Table 1 summarizes some of the overall and demographic characteristics of users in the dataset.

Items	No.	Demographic	%	Age groups	%
Logs	27MM	Age	70.5	15–24	57.5
Tracks	7M	Country	81.8	25–34	35.8
Albums	900K	Gender	81.6	35–44	5.5
Artists	555K			45–54	1.2
Listeners	594K				

Table 1. Dataset summary (Demographic: the percentage of people who provided demographic information)

Table 1 shows that large proportion of listeners self-declared their age, gender, and country. Previous research on online profiles concluded that people usually want to be well typified by their online profiles [4], and so we assumed there is a high degree of truth in these demographic features. Listeners from all ages are not equally represented in the dataset. The age distribution is biased towards young people, with an average age of 25 years old.

3. FEATURES FOR LISTENER PROFILING

We hypothesized that by better understanding the listening behaviour of people, we will be able to more accurately model the user needs. Hence, the recommendation can be tailored to each listener and the prediction accuracy will probably improve.

A set of computational features that attempt to describe some aspects of music listening behaviour in relation to musical artists was already proposed in previous research [21]. However, the ranking of the music items was not take into consideration and feature values were binned into categories. In our approach we try to represent similar characteristics of listening behaviour but we also consider the position of the music items within each listener's ranking as well as using normalized feature values to express the precise value of a certain listening behaviour characteristic in relation to a music item.

3.1 Feature design

We restricted ourselves to designing three novel features to describe listener behaviours: *exploratoryness*, *mainstreamness*, and *genderness*. Values for these features were computed for the three types of *music items* in the dataset: *tracks*, *albums*, and *artists*. Therefore, each listener's listening behaviour was described by a vector of nine values.

We will describe the goals behind each one of these features, give details about their implementation, visualize data patterns, and provide some analysis about the results.

3.1.1 Exploratoryness

To represent how much a listener explores different music instead of listening to the same music repeatedly we developed the *exploratoryness* feature.

For each user x's listening history, let L be the number of submitted music logs, S_k be all submitted music items of type k, where k={tracks, albums, artists}, $s_{k,i}$ be the number of music logs for the given music item key k at ranking i. We computed the exploratoryness $e_{x,k}$ for listener x on a given music item of type k as:

$$e_{x,k} = 1 - \frac{1}{L}\sum_{i=1}^{S_k} \frac{s_{k,i}}{i} \tag{1}$$

Exploratoryness returns a normalized value, with values closer to 0 for users listening to the same music item again and again, and values closer to 1 for users with more exploratory listening behaviour.

3.1.2 Mainstreamness

With the goal of expressing how similar a listener's listening history is to what everyone else listened to, we developed the *mainstreamness* feature. It analyses a listener's ranking of music items, and compares it with the overall ranking of artists, albums, or tracks, looking for the position of co-occurrences.

For each user x's listening history, let N be the number of logs of the music item ranked first in the overall ranking, L be the number of submitted music logs, S_k be all submitted music items of type k, where k={tracks, albums, artists}, $s_{k,i}$ be the number of music logs for the given music item key k at ranking i, and $o_{k,i}$ be the number of music logs in the overall ranking of music item type k ranked at position i. We defined the mainstreamness feature $m_{x,k}$ for listener x on a given music item of type k as:

$$m_{x,k} = \frac{1}{NL}\sum_{i=1}^{S_k} s_{k,i} o_{k,i} \tag{2}$$

Listening histories of people with a music item's ranking similar to the overall ranking receive mainstreamness values closer to 1. Listeners' mainstreamness whose ranking differ more from the overall ranking receive values closer to 0.

3.1.3 Genderness

With the aim of expressing how close a listener's listening history is to what females or males are listening to, we developed the *genderness* feature. The genderness feature computation basically relies on mainstreamness, but instead of computing just one overall ranking from all listeners, it uses two rankings: one made with music logs from listeners self-declared as female, and another one from male data.

For each user x's listening history and music item of type k, let $m_{x,k,male}$ be the mainstreamness computed with the male ranking, $m_{x,k,female}$ be the mainstreamness calculated with the female ranking.

We defined the feature genderness $g_{x,k}$ for listener x on a given music item of type k as:

$$g_{x,k} = m_{x,k,male} - m_{x,k,female} \tag{3}$$

3.2 Profiling listeners

To illustrate how the features we developed can be used to profile listeners, we calculated exploratoryness, mainstreamness, and genderness of users in our dataset. In order to not violate the homogeneity of variance we binned listeners into four age groups with balanced number of samples for each group. To obtain balanced groups, we drew a random sample of 100 people of each age, and created 10-year groups with 1000 people each. We then bootstrapped these groups with 1000 replications of the original sample and calculated 95 percent CI error bars. Although we quantified these characteristics in the relation of listeners with *artists*, *albums*, and *tracks*, and their interaction with listeners' age group, preliminary tests indicated that the interaction with artists was most significant. Therefore, for the rest of the paper we present only the results of the interaction between listeners and artists.

Figure 1 shows feature means by age group as well as 95 percent CI bars. In terms of artist exploratoryness, Figure 1(a) shows that while younger listeners in our dataset tend to listen more often to the same performers than adults, older listeners tend to explore more artists. Also, the rise in exploratoryness tends to stabilize in the mid-thirties. Figure 1(b) shows that while younger people listen more to the same artists that everyone is listening to, older people tend to listen to less common performers. This effect could be generated by the behaviour of older people or the fact that there are fewer older people in the original dataset, and so the artists they listen to are less represented in the overall ranking. Figure 1(c) shows artist genderness by age and gender. Listeners self-declared as male tend to listen more to music that is ranked higher in the male ranking, in all age groups, however their preference for the male ranking diminishes with age. Females, on the contrary, listen more to artists ranked higher in the female ranking when they are young, but adult women listen more to artists ranked higher in the male ranking. Overall, men and women have opposite trends of *genderness* in the different age groups, which seem to stabilize as they mature.

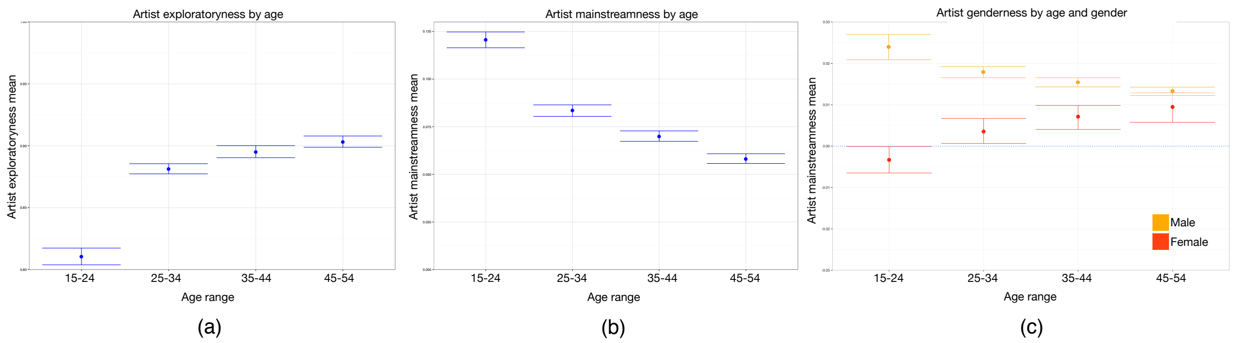

Figure 1. Feature means and 95% CI bars for a random group of listeners in our dataset. Each age group has 1K listeners. Error bars were calculated by taking 1K populations replicated from the original sample using bootstrap. (a) Artist exploratoryness by age group of listeners, (b) artist mainstreamness by age group of listeners, and (c) artist genderness by listeners' age group and gender.

We hoped that the aforementioned features captured some information about people's listening behaviour and will help to improve the accuracy of a music recommender model. However, as genderness was derived directly from mainstreamness, we did not use it in the experimental procedure for evaluating a music recommendation model with user-side data.

4. EXPERIMENTAL PROCEDURE

Our goal is to evaluate if demographics, behavioural profiles, and the use of observations from different contexts improve the accuracy of a recommendation model. Our sources of data involve a matrix of user preferences on artists derived from implicit feedback, a set of three categorical demographic features for each user: *age*, *country*, and *gender*, and a set of two continuous-valued features for describing people's listening behaviour: *exploratoryness* and *mainstreamness*. Preference matrixes were generated by considering full week of music listening histories data, as well as data coming from music logs submitted on weekdays and weekends only.

We followed a similar approach to Koren et al., in which a matrix of implicit feedback values expressing preferences of users on items is modelled by finding two lower dimensional matrixes of rank f $X_{n\times f}$ and $Y_{m\times f}$, which product approximates the original preference [14]. The goal of this approach is to find the set of values in X and Y that minimize the RMSE error between the original and the reconstructed matrixes. However, this conventional approach of matrix factorization for evaluating the accuracy of recommendation models using latent factors does not allow the researcher to incorporate additional features, such as *user-side* features. In order to incorporate latent factors as well as user-side features into one single recommendation model, we used the *Factorization Machines* method for matrix factorization and singular value decomposition [18]. In this approach, interactions between all latent factors as well as additional features are computed within a single framework, with a computational complexity that is linear to the number of extra features.

In order to perform a series of experiments with different sets of model parameters and user-side features in a timely fashion, we randomly sampled 10 percent of per-user music listening histories in the dataset, and we split this new subset into two disjoint sets: training (90 percent) and testing (10 percent) datasets. The training dataset had more than 60M observations from 59K users on 432K artists, with a density of observations of about 0.24 percent. We aggregated each dataset of listening histories by creating `<user, artist, playcounts>` triples. Then, we transformed the number of playcounts in each triple into a 1–5 Likert scale value by means of calculating the complementary cumulative distribution of artists per listener [3]. Hence, artists in each distribution quintile were assigned with a preference value according to how much each user listened to them.

In order to learn the best set of parameters of the recommendation model, we performed a grid search on the λ regularization parameter as well as the f number of latent factors with no user-side data, just using plain matrix factorization for the matrix of preferences of users on artists. Finding a good λ value helps to avoid overfitting the observed data by penalizing the magnitudes of the learned parameters. Finding the best f number of factors helps to obtain a better recommendation accuracy while also providing a set of to-be-interpreted latent factors. We used the Graphlab Create framework [1] to search over the number of latent factors in the range $[50, 200]$, and regularization values in the range $[1\times10^{-5}, 1\times10^{-8}]$. The best combination of parameters was achieved for $\lambda=1\times10^{-7}$ and $f=50$ latent factors. We used the Adaptive Stochastic Gradient Descent optimization algorithm [8] and set the maximum number of iterations at 50.

4.1 Demographic and profiling features

With these model parameter values, we evaluated the recommendation accuracy in the testing dataset of models

[1] `https ://pypi.python.org/pypi/GraphLab-Create`

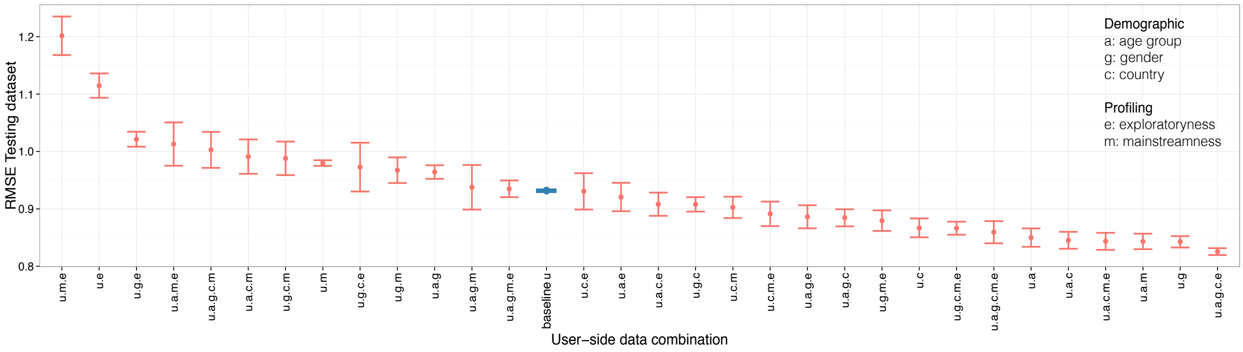

Figure 2. Root mean square error means and 95 percent CI bars for learned models evaluated in the testing dataset, with 32 combinations of the user-side features: *age*, *gender*, *country*, *exploratoryness*, and *mainstreamness* , ranked in decreasing order. Feature combinations are labelled according to the first letter of the word they represent. Baseline for comparison is combination *u*: user's preferences only, without any user-side features.

learned from the training data for all combinations of user-side demographic and profiling features. Since we had five user-side features: age, gender, country, exploratoryness, and mainstreamness, there were 32 different combinations.

Learning a model using an optimization algorithm can sometimes cause the results to converge into local minima instead of the global minimum. We informally evidenced that the variance in results of the optimization algorithm was larger than the variance in using different samples of the dataset. Hence, we repeated the process of learning and testing the accuracy of the learned models 10 times for each user-side feature combination. Using this procedure, we also wanted to compare and evaluate if results in model error were similar throughout several trials. The experiment baseline was established as the approach in which plain matrix factorization was used to estimate the recommendation accuracy of the learned models by just using the matrix of preferences of listeners on artists, without any user-side feature combination. By using this approach, we will be able to compare if the use of any feature combination resulted in a decrease in RMSE error, thus indicating an increase in the accuracy of the model.

Figure 2 summarizes the results of all trials. It shows all feature combination means, ranked in decreasing order, with 95 percent CI error bars generated from a bootstrap sample of 100 replications of the original sample. Feature combinations are labelled according to the first letter of the word they represented. For example, user preference data with age, gender, and exploratoryness is labelled *u.a.g.e*; user data with no user-side feature combinations is just labelled *u*. It can be seen that *u*, the baseline without user-side features, achieved an average RMSE value of .931 and exhibited a small variability, indicating that models in this setup were stable across all trials. All feature combinations to the right of the *u* show a smaller RMSE error, thus providing evidence for an increase in the learned accuracies of those models. Several feature combinations achieved better accuracy than the baseline. In particular, those combinations using just one of the demographic features: country (*u.c*), age (*u.a*), or gender (*u.g*) achieved improvements of about 7, 8, and 9 percent, respectively. Also, the combination of only demographic features (*u.a.g.c*), and all demographic and profiling features (*u.a.g.c.e.m*) improved the baseline model by almost 8 percent. However, the combination of features that achieved the best result was all demographic features together, plus the listener profiling feature of exploratoryness (*u.a.g.c.e*), exhibiting an improvement of about 12 percent above the baseline. The small variability in the model error of this combination throughout all trials suggested that models based on this user-side feature combination were quite stable. On the other hand, the combination of profiling features (*u.m.e)* achieved the worst performance, with a 29 percent increase in error, and a large variability in the estimated model error throughout trials. The variability in the results with these features suggests that the data topology using only profiling features is complex, probably making the iterative process of optimization converge into non-optimal local minima in the data.

4.2 Listening preferences in the contexts of entire week, weekdays only, and weekends only

We hypothesized that if people listen to different music during the weekdays than on weekends, we could create more accurate models by using data from only the weekdays or weekends, respectively. To test this hypothesis, we performed the same experimental approach that we carried out with the full-week dataset. However, this time we created two additional preference matrices of listeners for artists. The first additional matrix was made by using only music logs submitted during weekdays, and the second matrix was made by using only weekend music logs. Therefore, two extra sub-datasets were created using the original full-week dataset: *weekday* and *weekend* datasets. We then followed the same procedure described before: we split the data into training and testing datasets, we learned models from the training dataset for all 32 possible combinations of user-side features, and evaluated the accuracy of those models in the testing dataset. The number of observations, listeners, and artists, and also each of the matrix densities are shown in Table 2.

Dataset	Observations	Listeners	Artists	Density
Full-week	61M	59K	432K	0.237%
Weekdays	54M	59K	419K	0.216%
Weekends	35M	59K	379K	0.154%

Table 2. Number of observations, listeners, artists, and density for each context-based preference matrix.

As expected, the number of observations decreased in the datasets with partial data in relation to the full-week dataset. The number of listeners remained constant, which implies that most listeners in the dataset submitted music logs during weekdays as well as on weekends. Interestingly, the total number of artists for which there were submitted music logs on weekdays and weekends decreased between 3 and 12 percent in relation to the full week data, which implies that many artists in the dataset were only listened during one of the two weekly periods.

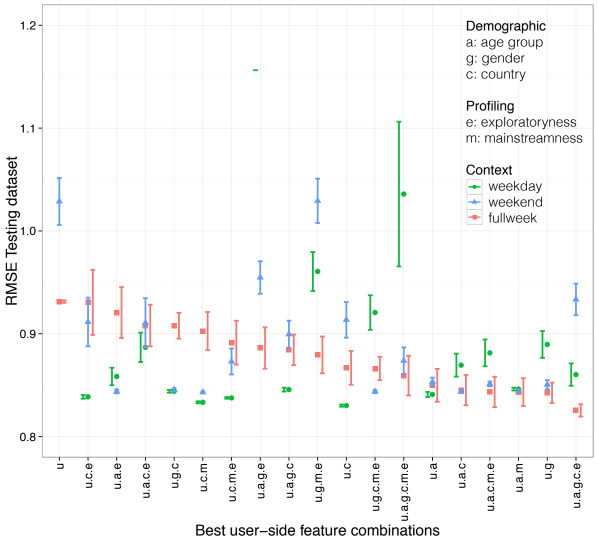

Figure 3. Root mean square error means and 95 percent CI bars for learned models with weekday, weekend, and full-week data. Only those feature combinations with a better RMSE value than the baseline for full-week data are shown.

Figure 3 summarizes model accuracies obtained using music log data from the three aforementioned contexts. Many of the models made with weekly-split listening data achieved better performance than those using full-week data. For example, models learned with weekday as well as weekend data using feature combinations *u.a.e*, *u.g.c*, and *u.c.m* achieved improvements in accuracy of about 7 percent in comparison to the model created with full-week data. They also showed smaller variability meaning more stability in model estimation. However, while the best RMSE value was obtained using the user-side feature combination *u.a.g.c.e* with full-week data, the same feature combination achieved worse performances by using listening data from weekdays and weekends only.

5. CONCLUSIONS AND FUTURE WORK

We have evaluated the impact of listeners' demographic and profiling features as well as basic forms of listening context, namely weekday and weekend versus full-week listening, on recommendation accuracy. We described our requirements for a dataset of music listening histories, explaining why none of the available datasets met our needs and how we ended up collecting our own data. We then formalized a set of profiling features that account some aspects of music listening behaviour. We also explained how we split our dataset of listening histories into weekdays and weekend listening histories to evaluate if having data from different sets of listening histories improved the accuracy of recommendation. Finally, we described how we set experiments that evaluated all combinations of user-side data features in the different contexts of listening. We found that the feature combination that achieved the smallest error was all demographic features together plus listener's exploratoryness, obtaining 12 percent improvement over the baseline of not using any user-side feature data. Although for some feature combinations the use of split listening data improved the recommendation, the best combination of features did benefit from having full-week data.

The results, in particular the many low RMSE values for several feature combinations using split listening data, seem to indicate that these error values are close to the limit in the minimum achievable error. This characteristic has already been described in the literature as a "magic barrier" in recommender systems design [1], referring to the upper bound in rating prediction accuracy due to inconsistencies in user's ratings. However, since we are mapping the number of submitted music listening logs into ratings, we do not see how these inconsistencies can explain this barrier. It would be interesting to perform a narrower grid search in order to investigate if we are actually hitting a wall in accuracy, or if there is a better set of model parameters that allows us to create more stable models and better performances throughout many trials. In comparison with previous research [21], the results are not comparable since different metrics are used. Also, our experiment directly integrated the profiling features into the matrix factorization algorithm. Finally, although these results show an improvement in the accuracy of a recommendation model based on listeners' past listening histories, we might require an online, user-centred study to measure people's actual satisfaction with the learned model.

6. ACKNOWLEDGEMENTS

This research has been supported by BecasChile Bicentenario, Comisión Nacional de Ciencia y Tecnología, Gobierno de Chile, and the Social Sciences and Humanities Research Council of Canada. Important parts of this work used ComputeCanada's High Performance Computing resources. The authors would like to thank Emily Hopkins for her thorough proofreading of this paper.

7. REFERENCES

[1] A. Bellogín, A. Said, and A. P. de Vries. The magic barrier of recommender systems—no magic, just ratings. In *User Modeling, Adaptation, and Personalization*, pages 25–36. Springer, 2014.

[2] I. Cantador, P. Brusilovsky, and T. Kuflik. Last.fm web 2.0 dataset. In *2nd Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec 2011)*, RecSys 2011, Chicago, IL, 2011.

[3] Ò. Celma. *Music Recommendation and Discovery: the Long Tail, Long Fail, and Long Play in the Digital Music Space*. Springer, 2010.

[4] S. Counts and K. B. Stecher. Self-presentation of personality during online profile creation. In *Proceedings of the Third International ICWSM Conference*, pages 191–4, 2009.

[5] S. Cunningham, S. Caulder, and V. Grout. Saturday night or fever? Context aware music playlists. In *Proceedings of the Audio Mostly Conference*, pages 1–8, Piteå, Sweden, 2008.

[6] T. DeNora. *Music in Everyday Life*. Cambridge University Press, 2000.

[7] G. Dror, N. Koenigstein, Y. Koren, and M. Weimer. The Yahoo! music dataset and KDD-cup'11. *Journal of Machine Learning Research*, 18:3–18, 2011.

[8] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–59, 2011.

[9] EMI. EMI Million Interview Dataset, 2012. `http://musicdatascience.com/emi-million-interview-dataset/`. Accessed 18 February 2016.

[10] M. Gillhofer and M. Schedl. Iron Maiden while jogging, Debussy for dinner? an analysis of music listening behavior in context. In X. He, S. Luo, D. Tao, C. Xu, J. Yang, and M. Abul Hasan, editors, *MultiMedia modeling*, volume 8936 of *Lecture Notes in Computer Science*, pages 380–91, Switzerland, 2015. Springer.

[11] D. J. Hargreaves, R. MacDonald, and D. Miell. How do people communicate using music. In D. Miell, R. MacDonald, and D. J. Hargreaves, editors, *Musical Communication*, chapter 1, pages 1–25. Oxford University Press, 2005.

[12] P. Herrera, Z. Resa, and M. Sordo. Rocking around the clock eight days a week: an exploration of temporal patterns of music listening. In *1st Workshop on Music Recommendation and Discovery*, Barcelona, Spain, 2010.

[13] V. J. Konečni. Social interaction and musical preference. In D. Deutsch, editor, *The Psychology of Music*, pages 497–516. Academic Press, New York, NY, 1982.

[14] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–7, 2009.

[15] D. Lee, S. E. Park, M. Kahng, S. Lee, and S. Lee. Exploiting contextual information from event logs for personalized recommendation. In R. Lee, editor, *Computer and Information Science*, pages 121–39. Springer-Verlag, 2010.

[16] J. S. Lee and J. C. Lee. Context awareness by case-based reasoning in a music recommendation system. In *Ubiquitous Computing Systems*, pages 45–58. Springer, 2007.

[17] B. McFee, T. Bertin-Mahieux, D. P. W. Ellis, and G. R. G. Lanckriet. The million song dataset challenge. In *Proceedings of the 21st International Conference on World Wide Web*, pages 909–16, Lyon, France, 2012.

[18] S. Rendle. Factorization machines. In *IEEE 10th International Conference on Data Mining*, pages 995–1000, Sydney, Australia, 2010. IEEE.

[19] S. Rendle, Z. Gantner, C. Freudenthaler, and L. Schmidt-Thieme. Fast context-aware recommendations with factorization machines. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 635–44, Beijing, China, 2011.

[20] M. Schedl, A. Flexer, and J. Urbano. The neglected user in music information retrieval research. *Journal of Intelligent Information Systems*, 41(3):523–39, 2013.

[21] M. Schedl and D. Hauger. Tailoring music recommendations to users by considering diversity, mainstreaminess, and novelty. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 947–50, 2015.

[22] D. Shin, J. Lee, J. Yeon, and S. Lee. Context-aware recommendation by aggregating user context. In *2009 IEEE Conference on Commerce and Enterprise Computing*, pages 423–30, 2009.

[23] J. A. Sloboda, A. Lamont, and A. Greasley. Choosing to hear music: motivation, process and effect. In S. Hallam, I. Cross, and M. Thaut, editors, *The Oxford Handbook of Music Psychology*, chapter 40, pages 431–40. Oxford University Press, Oxford, UK, 2009.

[24] J. Su, H. Yeh, P. Yu, and V. Tseng. Music recommendation using content and context information mining. *IEEE Intelligent Systems*, 25(1):16–26, 2010.

[25] X. Wang. *Interactive Music Recommendation: Context, Content and Collaborative Filtering*. PhD thesis, National University of Singapore, 2014.