# SYSTÈMES DE RECOMMANDATION DE MUSIQUE AUTOMATIQUE: LES CARACTÉRISTIQUES DÉMOGRAPHIQUES, DE PROFILAGE ET CONTEXTUELLES AMÉLIORENT-ILS LEUR PERFORMANCE?


## ABSTRACT

Les performances traditionnelles des systèmes de recommandation automatique de musique reposent généralement sur la précision des modèles statistiques tirés des préférences antérieures des utilisateurs sur les éléments musicaux. Cependant, d'autres sources de données telles que les attributs démographiques des auditeurs, leur comportement d'écoute et leurs contextes d'écoute codent des informations sur les auditeurs et leurs habitudes d'écoute qui peuvent être utilisées pour améliorer la précision des modèles de recommandation musicale. Dans cet article, nous présentons un vaste ensemble de données d'histoires d'écoute musicale avec des informations démographiques sur les auditeurs, et un ensemble de caractéristiques permettant de caractériser les aspects du comportement d'écoute des personnes. La longévité des histoires d'écoute collectées, couvrant plus de deux ans, permet de retrouver des formes élémentaires de contexte d'écoute. Nous utilisons cet ensemble de données dans l'évaluation de la précision d'un modèle de recommandation d'artiste musical tiré des préférences antérieures des auditeurs sur des éléments musicaux et de leur interaction avec plusieurs combinaisons de caractéristiques démographiques, de profilage et contextuelles. Nos résultats indiquent que l'utilisation de l'âge, du pays et du sexe déclarés par les auditeurs améliore la précision de la recommandation de 8%. Lorsqu'une nouvelle fonctionnalité de profilage appelée exploration a été ajoutée, la précision du modèle a augmenté de 12%.


## 1. COMPORTEMENT D'ÉCOUTE ET CONTEXTE

Le contexte dans lequel les gens écoutent de la musique a fait l'objet d'études sur un nombre croissant de publications, en particulier dans le domaine de la psychologie musicale. Koneˇcni a suggéré que l'acte d'écouter de la musique a vidé les espaces physiques consacrés exclusivement à la performance musicale et au plaisir depuis longtemps, et que la musique est aujourd'hui écoutée dans une grande variété de contextes [13]. Comme la musique accompagne de plus en plus nos activités quotidiennes, la musique et l'auditeur ne sont pas les seuls facteurs, car le contexte d'écoute est apparu comme une autre variable qui influence et est influencée par les deux autres facteurs [11]. Il a également été observé que les gens comprennent consciemment ces interactions [6] et les utilisent lorsqu'ils choisissent de la musique pour les activités de la vie quotidienne [23]. Le contexte de l'écoute de la musique semble influencer la façon dont les gens choisissent la musique, et ainsi les recommandateurs de musique devraient suggérer des articles de musique pour adapter la situation et les besoins de chaque auditeur particulier. La modélisation des besoins des utilisateurs a été identifiée par Schedl et al. comme une exigence clé pour le développement de systèmes de récupération de musique centrés sur l'utilisateur [20].

Ils ont également noté que les systèmes personnalisés personnalisent leurs recommandations en utilisant des informations utilisateur supplémentaires, et que les systèmes contextuels utilisent des aspects dynamiques du contexte de l'utilisateur pour améliorer la qualité des recommandations. Le besoin d'informations contextuelles et environnementales a été souligné par Cunningham et al. et d'autres [5, 12, 16]. Ils ont émis l'hypothèse que l'emplacement, l'activité et le contexte des auditeurs étaient probablement corrélés avec leurs préférences, et devraient donc être pris en compte lors de l'élaboration de systèmes de recommandation de musique. En conséquence, des cadres pour l'abstraction du contexte de l'écoute musicale en utilisant des caractéristiques brutes telles que des données environnementales ont été proposés dans la littérature [16, 22]. Alors que certains chercheurs ont rapporté que les systèmes de recommandation contextuelle sont plus performants que les systèmes traditionnels [15, 22, 24], d'autres n'ont montré que des améliorations mineures [10].

Enfin, d'autres ont réalisé des expériences avec les seuls éléments musicaux les mieux classés, conduisant probablement à des modèles biaisés par la popularité [15, 25]. Nous allons maintenant discuter de l'impact de l'utilisation des caractéristiques démographiques et de profilage des auditeurs - ci-après dénommées caractéristiques de l'utilisateur [19] - pour améliorer la précision d'un modèle de recommandation de musique. Les caractéristiques côté utilisateur ont été extraites de données démographiques auto-déclarées et d'un ensemble de caractéristiques de profilage personnalisées qui caractérisent le comportement d'écoute musicale d'un grand nombre d'utilisateurs d'un service de musique numérique. Leurs histoires d'écoute de musique ont été désagrégées en différentes périodes afin d'évaluer si la précision des modèles a changé en utilisant différents contextes temporels d'écoute. Enfin, des modèles basés sur des facteurs latents ont été appris pour tous les contextes d'écoute et toutes les combinaisons de caractéristiques côté utilisateur. La section 2 présente la collection d'ensembles de données, la section 3 introduit un ensemble de caractéristiques personnalisées pour profiler le comportement d'écoute des auditeurs, et la section 4 décrit la configuration expérimentale et présente les résultats.

## 2. DATASET

Nous sommes intéressés à évaluer l'impact de l'utilisation de caractéristiques démographiques et de profilage, ainsi que des informations contextuelles, pour un grand nombre de personnes, sur la précision de prédiction d'un modèle de recommandation d'artiste musical. Quelques ensembles de données accessibles au public pour la recherche sur l'écoute de la musique fournissent des informations sur les personnes et les éléments musicaux. Dror et al. présenté un ensemble de données sur les évaluations agrégées des personnes de 1M sur des articles musicaux [7]. McFee et al. présenté un jeu de données de compteurs de chansons de 1M auditeurs [17]. Cependant, aucun de ces deux ensembles de données n'a fourni d'horodatage des journaux musicaux ou des informations démographiques sur les auditeurs. Celma a fourni un jeu de données de compteurs avec les données démographiques des auditeurs pour les auditeurs de 360K et un ensemble d'historiques d'écoute avec des journaux à temps plein; Cependant, ce dernier ensemble de données ne comprenait que des journaux pour les auditeurs 1K [3]. Cantador et al. présenté un autre petit jeu de données avec des compteurs de chansons pour les utilisateurs 2K [2]. Enfin, EMI a promis un ensemble de données d'interviews 1M sur l'appréciation de la musique, le comportement et les attitudes des gens [9], mais seulement des informations partielles ont été mises à disposition. Aucun des jeux de données susmentionnés n'offre, en même temps et pour un grand nombre d'auditeurs, l'accès à des historiques d'écoute musicale complète ainsi qu'à des données démographiques. Cela signifie qu'il n'est pas possible d'extraire toutes les fonctionnalités qui nous intéressent, et nous avons donc décidé de collecter notre propre ensemble de données avec les histoires d'écoute de musique de Last.fm. Last.fm se démarque de la plupart des services de musique numérique en ligne parce qu'il enregistre non seulement les journaux musicaux des chansons lues dans son propre écosystème, mais aussi de plus de 600 lecteurs multimédias. Ensuite, nous présenterons les critères et les méthodes d'acquisition utilisés pour collecter un grand nombre d'histoires d'écoute musicale du service Last.fm.

### 2.1 Critères de données, acquisition et nettoyage

L'agrégation des historiques d'écoute musicale des personnes nécessite la fusion de leurs journaux musicaux en périodes de temps. Afin d'obtenir des données uniformes sur des semaines, des mois, des saisons ou des années agrégées, nous avons recherché des auditeurs avec un nombre arbitraire d'au moins deux ans d'activité soumettant des journaux musicaux depuis qu'ils ont commencé à utiliser le système. journaux par jour. Ces deux restrictions ont forcé notre robot de recherche de données à rechercher des auditeurs avec un minimum de 7 300 enregistrements musicaux soumis à la base de données Last.fm. De plus, ces contraintes nous ont assuré que nous collecterions des historiques

d'écoute d'auditeurs actifs avec suffisamment de données pour effectuer une bonne agrégation au fil du temps.
L'acquisition des données a été réalisée au moyen de l'utilisation de plusieurs machines appelant l'API Last.fm pendant une période de deux ans (2012-2014). Nous avons collecté des historiques d'écoute en utilisant la méthode API de Last.fm user.getRecentTracks (). Cet appel API nous a permis d'obtenir des historiques d'écoute complets. Parallèlement à ces données, nous avons également stocké toutes les métadonnées disponibles pour chaque auditeur, y compris les caractéristiques démographiques auto-déclarées facultatives: âge, pays et sexe. Nous avons effectué plusieurs processus de filtrage et de nettoyage des données afin d'éviter les données bruitées. Par exemple, nous nous sommes aperçus qu'il y avait de nombreux enregistrements musicaux dupliqués (c'est-à-dire le même horodatage pour plusieurs identifiants musicaux) et des historiques d'écoute avec beaucoup de journaux musicaux trop rapprochés (à moins de 30 secondes d'intervalle) , qui est le minimum que Last.fm nécessite pour considérer une piste jouée comme un journal musical valide). Par conséquent, nous avons décidé de filtrer tous les journaux dupliqués ainsi que les journaux distants de moins de 30 secondes.

### 2.2 Dataset demographics

Notre ensemble de données se compose de 27 milliards de journaux musicaux extraits de l'historique d'écoute musicale de 594K utilisateurs. Ce vaste répertoire de disques d'écoute musicale rend compte de l'interaction des auditeurs avec plus de 555 000 artistes différents, 900 000 albums et 7 millions de titres. Il y a des histoires d'écoute musicale de personnes dans 239 pays différents auto-déclarés, avec des auditeurs de tous les fuseaux horaires représentés. Cependant, les auditeurs d'Afrique, d'Asie du Sud et d'Asie de l'Est sont sous-représentés dans notre ensemble de données. En fait, les 19 premiers pays représentent plus de 85% du nombre total d'auditeurs dans l'ensemble de données. Le tableau 1 résume certaines des caractéristiques générales et démographiques des utilisateurs de l'ensemble de données.

| 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.** Résumé du jeu de données (démographiques: pourcentage de personnes ayant fourni des informations démographiques)

Le tableau 1 montre qu'une forte proportion d'auditeurs ont déclaré eux-mêmes leur âge, leur sexe et leur pays. Des recherches antérieures sur les profils en ligne ont conclu que les gens veulent généralement être bien caractérisés par leurs profils en ligne [4], et nous avons donc supposé qu'il y avait un haut degré de vérité dans ces caractéristiques démographiques. Les auditeurs de tous âges ne sont pas représentés de manière égale dans l'ensemble de données. La répartition par âge est biaisée vers les jeunes, avec un âge moyen de 25 ans.

## 3. CARACTÉRISTIQUES POUR LE PROFILAGE D'ÉCOUTE

Nous avons émis l'hypothèse qu'en comprenant mieux le comportement d'écoute des personnes, nous serons en mesure de modéliser plus précisément les besoins des utilisateurs. Par conséquent, la recommandation peut être adaptée à chaque auditeur et la précision de la prédiction s'améliorera probablement. Un ensemble de caractéristiques computationnelles qui tentent de décrire certains aspects du comportement d'écoute de la musique par rapport aux artistes musicaux a déjà été proposé dans des recherches antérieures [21]. Cependant, le classement des éléments musicaux n'a pas été pris en compte et les valeurs des caractéristiques ont été regroupées en catégories. Dans notre approche, nous essayons de représenter des caractéristiques similaires de comportement d'écoute mais nous considérons également la position des éléments musicaux dans le classement de chaque auditeur ainsi que les valeurs de caractéristiques normalisées pour exprimer la valeur précise d'une certaine caractéristique d'écoute par rapport à un élément musical.

### 3.1 Conception de fonctionnalités

Nous nous sommes limités à la conception de trois nouvelles fonctionnalités pour décrire les comportements des auditeurs: l'exploration, l'intégration et la gendrerie. Les valeurs de ces fonctionnalités ont été calculées pour les trois types d'éléments musicaux de l'ensemble de données: pistes, albums et artistes. Par conséquent, le comportement d'écoute de chaque auditeur a été décrit par un vecteur de neuf valeurs. Nous décrirons les objectifs derrière chacune de ces caractéristiques, donnerons des détails sur leur implémentation, visualiserons les modèles de données et fournirons une analyse des résultats.

#### 3.1.1 Exploratoryness

Pour représenter à quel point un auditeur explore différentes musiques au lieu d'écouter la même musique à plusieurs reprises, nous avons développé la fonction exploratoire. Pour l'historique d'écoute de chaque utilisateur x, soit L le nombre de journaux musicaux soumis, soit tous les éléments musicaux soumis de type k, où k = {pistes, albums, artistes}, sk, i le nombre de journaux musicaux pour le donné touche d'élément musical k au classement i. Nous avons calculé l'exploratoire ex, k pour l'écouteur x sur un élément musical donné de type k comme:

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

Exploratoryness renvoie une valeur normalisée, avec des valeurs plus proches de 0 pour les utilisateurs qui écoutent le même élément musical encore et encore, et des valeurs plus proches de 1 pour les utilisateurs ayant un comportement d'écoute plus exploratoire.

### 3.1.2 Mainstreamness

Dans le but d'exprimer à quel point l'histoire d'écoute d'un auditeur est semblable à ce que tout le monde a écouté, nous avons développé la fonction d'intégration. Il analyse le classement des éléments musicaux d'un auditeur et le compare au classement général des artistes, des albums ou des pistes, en recherchant la position des cooccurrences. Pour l'historique d'écoute de chaque utilisateur x, soit N le nombre de journaux de l'élément musical classé en premier dans le classement général, L soit le nombre de journaux musicaux soumis, Sk tous les éléments musicaux soumis de type k, où k = {pistes, albums, artistes}, sk, je suis le nombre de journaux musicaux pour la clé d'élément musical donné k au classement i, et ok, je suis le nombre de journaux musicaux dans le classement général du type d'élément musical k classé à la position i. Nous avons défini la caractéristique de mainstreamness mx, k pour l'écouteur x sur un élément musical donné de type k comme:

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

Les histoires d'écoute de personnes avec un classement d'élément musical similaire au classement général reçoivent des valeurs d'approche plus proches de 1. L'intégration des auditeurs dont le classement diffère davantage du classement général reçoit des valeurs plus proches de 0.

### 3.1.3 Genderness

Dans le but d'exprimer à quel point l'histoire d'écoute d'un auditeur est proche de ce que les femmes ou les hommes écoutent, nous avons développé la fonction de genderness. Le calcul de la fonctionnalité de la gendresse repose essentiellement sur l'intégration, mais au lieu de calculer un seul classement général de tous les auditeurs, il utilise deux classements: l'un fait avec les journaux musicaux des auditeurs auto-déclarés comme féminins et l'autre des données masculines. Pour l'histoire d'écoute de chaque utilisateur x et l'élément musical de type k, soit mx, k, male soit le mainstreamness calculé avec le classement masculin, mx, k, female soit le mainstreamness calculé avec le classement féminin. Nous avons défini la caractéristique genderness gx, k pour l'écouteur x sur un élément musical donné de type k comme:

$$g_{x,k} = m_{x,k,male} - m_{x,k,female} \qquad (3)$$

## 3.2 Profiling listeners

Pour illustrer comment les caractéristiques que nous avons développées peuvent être utilisées pour profiler les auditeurs, nous avons calculé l'explorabilité, l'intégration et la gendrité des utilisateurs dans notre ensemble de données. Afin de ne pas violer l'homogénéité de la variance, nous avons regroupé les auditeurs en quatre groupes d'âge avec un nombre équilibré d'échantillons pour chaque groupe. Pour obtenir des groupes équilibrés, nous avons tiré un échantillon aléatoire de 100 personnes de chaque âge, et créé des groupes de 10 ans avec 1000 personnes chacun. Nous avons ensuite amorcé ces groupes avec 1 000 répétitions de l'échantillon original et calculé des barres d'erreur de 95%. Bien que nous ayons quantifié ces caractéristiques dans la relation des auditeurs avec les artistes, les albums et les morceaux, et leur interaction avec le groupe d'âge des auditeurs, les tests préliminaires ont indiqué que l'interaction avec les artistes était la plus significative. Par conséquent, pour le reste de

l'article, nous ne présentons que les résultats de l'interaction entre les auditeurs et les artistes. La figure 1 montre les moyennes des caractéristiques par groupe d'âge ainsi que les barres à IC à 95%. En ce qui concerne l'exploration exploratoire, la figure 1 (a) montre que, bien que les jeunes auditeurs de notre jeu de données tendent à écouter plus souvent les mêmes artistes que les adultes, les auditeurs plus âgés ont tendance à explorer davantage d'artistes. En outre, la hausse de l'exploration tend à se stabiliser au milieu des années 30. La figure 1 (b) montre que si les jeunes écoutent davantage les mêmes artistes que tout le monde écoute, les personnes plus âgées ont tendance à écouter des artistes moins communs. Cet effet pourrait être généré par le comportement des personnes âgées ou par le fait qu'il y a moins de personnes âgées dans l'ensemble de données original, et donc les artistes qu'ils écoutent sont moins représentés dans le classement général. La figure 1 (c) montre la gendrure des artistes selon l'âge et le sexe. Les auditeurs auto-déclarés en tant qu'hommes ont tendance à écouter davantage la musique qui est classée plus haut dans le classement des hommes, dans tous les groupes d'âge, mais leur préférence pour le classement des hommes diminue avec l'âge. Les femmes, au contraire, écoutent davantage les artistes se classer plus haut dans le classement des femmes quand elles sont jeunes, mais les femmes adultes écoutent plus les artistes se classer plus haut dans le classement des hommes. Globalement, les hommes et les femmes ont des tendances opposées de la gendreté dans les différents groupes d'âge, qui semblent se stabiliser à mesure qu'ils vieillissent.

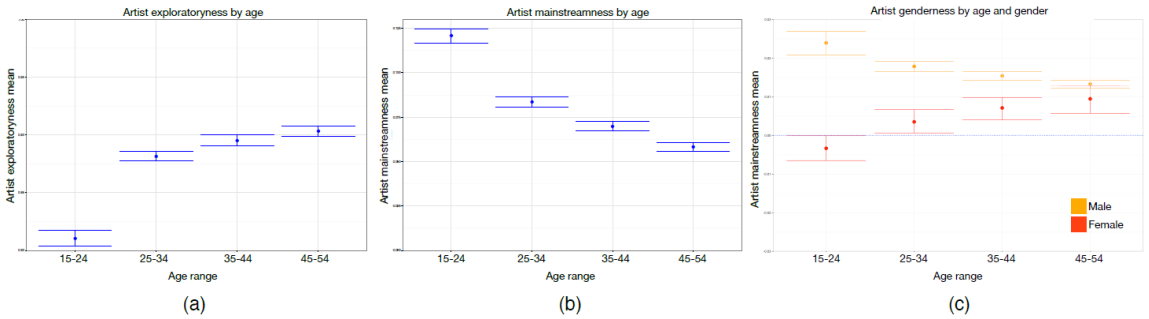


**Figure 1.** Moyens de caractéristiques et barres IC à 95% pour un groupe aléatoire d'auditeurs dans notre ensemble de données. Chaque groupe d'âge a 1K auditeurs. Les barres d'erreur ont été calculées en prenant des populations 1K répliquées à partir de l'échantillon original en utilisant bootstrap. (a) L'exploration de l'artiste par groupe d'âge, (b) l'intégration de l'artiste par groupe d'âge, et (c) la gendrité de l'artiste selon le groupe d'âge et le sexe des auditeurs.

Nous espérions que les fonctionnalités susmentionnées capturaient certaines informations sur le comportement d'écoute des personnes et contribueraient à améliorer la précision d'un modèle de recommandation de musique. Cependant, comme la gendrerie provenait directement du mainstream, nous ne l'avons pas utilisé dans la procédure expérimentale pour évaluer un modèle de recommandation de musique avec des données côté utilisateur.

## 4. EXPERIMENTAL PROCEDURE

Notre objectif est d'évaluer si les données démographiques, les profils comportementaux et l'utilisation d'observations provenant de contextes différents améliorent la précision d'un modèle de recommandation. Nos sources de données impliquent une matrice de préférences utilisateur sur les artistes dérivées de la rétroaction implicite, un ensemble de trois caractéristiques démographiques catégoriques pour chaque utilisateur: âge, pays et sexe, et un ensemble de deux caractéristiques continues pour décrire le comportement d'écoute des personnes: exploratoryness et mainstreamness. Les matrices de préférences ont été générées en tenant compte de la semaine complète des données d'écoute musicale, ainsi que des données provenant des journaux musicaux soumis en semaine et en fin de semaine seulement. Nous avons suivi une approche similaire à Koren et al., Dans laquelle une matrice de valeurs de rétroaction implicites exprimant les préférences des utilisateurs sur des items est modélisée en trouvant deux matrices de rang inférieur de rang f Xn × f et Ym × f, qui s'approche de la préférence originale. [14]. Le but de cette approche est de trouver l'ensemble des valeurs dans X et Y qui minimisent l'erreur RMSE entre l'original et les matrices reconstruites. Cependant, cette approche conventionnelle de la factorisation matricielle pour évaluer la précision des modèles de recommandation utilisant des facteurs latents ne permet pas au chercheur d'incorporer des caractéristiques supplémentaires, telles que des caractéristiques côté utilisateur. Afin d'incorporer des facteurs latents ainsi que des caractéristiques côté utilisateur dans un seul modèle de recommandation, nous avons utilisé la méthode des factorisations factorielles pour la factorisation matricielle et la décomposition de la valeur singulière [18]. Dans cette approche, les interactions entre tous les facteurs latents ainsi que les fonctionnalités supplémentaires sont calculées dans un cadre unique, avec une complexité de calcul qui est linéaire par rapport au nombre d'entités supplémentaires. Afin d'effectuer une série d'expériences avec différents ensembles de paramètres de modèle et de fonctionnalités côté utilisateur, nous avons échantillonné au hasard 10% des historiques d'écoute de musique par utilisateur dans l'ensemble de données, et nous avons divisé ce nouveau sous-ensemble en deux (90%) et tester (10%) des jeux de données. L'ensemble de données d'entraînement comptait plus de 60 millions d'observations provenant de 59 000 utilisateurs sur 432 000 artistes, avec une densité d'observations d'environ 0,24%. Nous avons regroupé chaque ensemble de données d'écoute en créant des triplets <user, artist, playcounts>. Ensuite, nous avons transformé le nombre de compteurs de chaque triplet en une valeur d'échelle de 1-5 Likert en calculant la distribution cumulative complémentaire des artistes par auditeur [3]. Par conséquent, les artistes de chaque quintile de distribution ont été assignés à une valeur de préférence en fonction de la quantité d'écoute de chaque utilisateur. Afin d'apprendre le meilleur ensemble de paramètres du modèle de recommandation, nous avons effectué une recherche de grille sur le paramètre _ de régularisation ainsi que le nombre f de facteurs latents sans données côté utilisateur, en utilisant simplement la factorisation matricielle pour la matrice de préférences des utilisateurs sur les artistes. Trouver une bonne valeur permet d'éviter d'overfitter les données observées en pénalisant les grandeurs des paramètres appris. Trouver le meilleur nombre de facteurs permet d'obtenir une meilleure précision des recommandations tout en fournissant un ensemble de facteurs latents à interpréter. Nous avons utilisé le framework Graphlab Create 1 pour rechercher le nombre de facteurs latents dans la plage [50, 200] et les valeurs de régularisation dans la plage [1 × 10-5, 1 × 10-8]. La meilleure combinaison de paramètres a été obtenue pour _ = 1 × 10-7 et f = 50 facteurs latents. Nous avons utilisé l'algorithme d'optimisation adaptative de la descente de gradient stochastique [8] et défini le nombre maximal d'itérations à 50.

## 4.1 Caractéristiques démographiques et de profilage

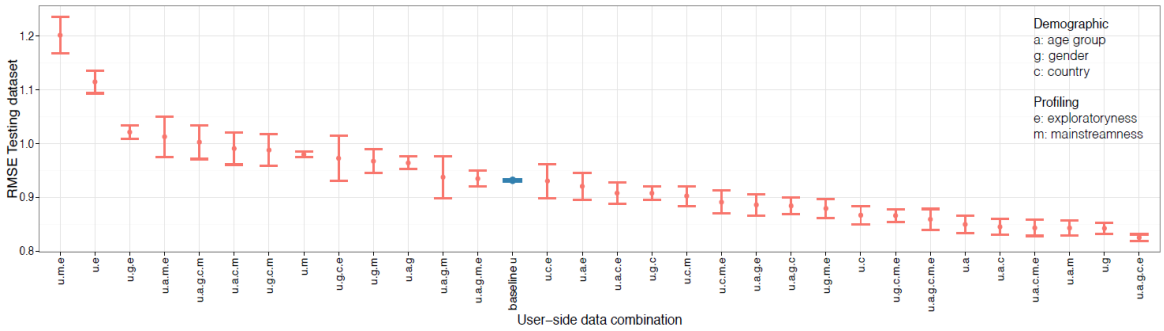


**Figure 2.** L'erreur quadratique moyenne signifie des barres d'IC à 95% pour les modèles appris évalués dans l'ensemble de données de test, avec 32 combinaisons des caractéristiques côté utilisateur: âge, sexe, pays, caractère exploratoire et mainstreamité, classés par ordre décroissant. Les combinaisons de caractéristiques sont étiquetées en fonction de la première lettre du mot qu'elles représentent. La ligne de base pour la comparaison est la combinaison u: les préférences de l'utilisateur uniquement, sans aucune fonctionnalité côté utilisateur.

Avec ces valeurs de paramètres du modèle, nous avons évalué la précision de la recommandation dans l'ensemble de données de test des modèles appris à partir des données d'apprentissage pour toutes les combinaisons de caractéristiques démographiques et de profilage. Comme nous avions cinq caractéristiques côté utilisateur: l'âge, le sexe, le pays, l'exploration et l'intégration, il y avait 32 combinaisons différentes. L'apprentissage d'un modèle utilisant un algorithme d'optimisation peut parfois amener les résultats à converger en minima locaux au lieu du minimum global. Nous avons démontré de manière informelle que la variance dans les résultats de l'algorithme d'optimisation était plus grande que la variance dans l'utilisation de différents échantillons de l'ensemble de données. Par conséquent, nous avons répété le processus d'apprentissage et de test de la précision des modèles appris 10 fois pour chaque combinaison de caractéristiques côté utilisateur. En utilisant cette procédure, nous avons également voulu comparer et évaluer si les résultats dans l'erreur du modèle étaient similaires tout au long de plusieurs essais. La base de l'expérience a été établie comme l'approche dans laquelle la factorisation matricielle simple a été utilisée pour estimer la précision des recommandations des modèles appris en utilisant simplement la matrice des préférences des auditeurs sur les artistes, sans aucune combinaison de caractéristiques côté utilisateur. En utilisant cette approche, nous serons en mesure de comparer si l'utilisation de toute combinaison de caractéristiques a entraîné une diminution de l'erreur RMSE, indiquant ainsi une augmentation de la précision du modèle. La figure 2 résume les résultats de tous les essais. Il montre tous les moyens de combinaison de caractéristiques, classés par ordre décroissant, avec des barres d'erreur d'IC à 95% générées à partir d'un échantillon bootstrap de 100 réplications de l'échantillon original. Les combinaisons de caractéristiques sont étiquetées en fonction de la première lettre du mot qu'elles représentent. Par exemple, les données sur les préférences de l'utilisateur avec l'âge, le sexe et le caractère exploratoire sont appelées u.a.g.e; les données utilisateur sans combinaisons de caractéristiques côté utilisateur sont simplement étiquetées u. On peut voir que u, la référence sans les fonctionnalités utilisateurs, a atteint une valeur RMSE moyenne de .931 et a montré une petite variabilité, indiquant que les modèles dans cette configuration étaient stables dans tous les essais. Toutes les combinaisons de caractéristiques à la droite de l'u montrent une erreur RMSE plus petite, fournissant ainsi la preuve d'une augmentation de la précision apprise de ces modèles. Plusieurs combinaisons de caractéristiques ont atteint une meilleure précision que la ligne de base. En

particulier, ces combinaisons utilisant seulement une des caractéristiques démographiques: pays (u.c), âge (u.a), ou sexe (u.g) ont obtenu des améliorations d'environ 7, 8 et 9 pour cent, respectivement. En outre, la combinaison de caractéristiques démographiques (u.a.g.c) et de toutes les caractéristiques démographiques et de profilage (u.c.e.m.) a amélioré le modèle de référence de près de 8%. Cependant, la combinaison des caractéristiques qui ont obtenu le meilleur résultat était l'ensemble des caractéristiques démographiques, plus la caractéristique d'exploration du profil d'écoute (u.a.g.c.e), montrant une amélioration d'environ 12 pour cent au-dessus de la ligne de base. La faible variabilité de l'erreur modèle de cette combinaison dans tous les essais a suggéré que les modèles basés sur cette combinaison de caractéristiques côté utilisateur étaient assez stables. D'autre part, la combinaison des caractéristiques de profilage (u.m.e) a obtenu la pire performance, avec une augmentation de 29% des erreurs et une grande variabilité de l'erreur estimée du modèle tout au long des essais. La variabilité des résultats avec ces caractéristiques suggère que la topologie de données utilisant uniquement des fonctionnalités de profilage est complexe, faisant probablement converger le processus itératif d'optimisation en minima locaux non-optimaux dans les données.

### 4.2 Préférences d'écoute dans les contextes de semaine entière, de semaine seulement et de fin de semaine seulement

Nous avons émis l'hypothèse que si les gens écoutent de la musique différente pendant les jours de semaine que le week-end, nous pourrions créer des modèles plus précis en utilisant des données provenant uniquement des jours de la semaine ou des week-ends, respectivement. Pour tester cette hypothèse, nous avons effectué la même approche expérimentale que nous avons réalisée avec l'ensemble de données de la semaine complète. Cependant, cette fois nous avons créé deux matrices de préférence d'auditeurs pour les artistes. La première matrice supplémentaire a été faite en utilisant uniquement les journaux musicaux soumis pendant les jours de la semaine, et la deuxième matrice a été faite en utilisant uniquement les journaux de musique du week-end. Par conséquent, deux sous-jeux de données supplémentaires ont été créés à l'aide de l'ensemble de données d'une semaine entière: les ensembles de données en semaine et les week-ends. Nous avons ensuite suivi la même procédure décrite précédemment: nous avons divisé les données en formation et en testant des jeux de données, nous avons appris des modèles du jeu de données de formation pour les 32 combinaisons possibles de fonctionnalités côté utilisateur et évalué la précision de ces modèles. Le nombre d'observations, d'auditeurs et d'artistes, ainsi que chacune des densités matricielles sont présentés dans le tableau 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.** Nombre d'observations, d'auditeurs, d'artistes et de densité pour chaque matrice de préférence contextuelle.

Comme prévu, le nombre d'observations a diminué dans les ensembles de données avec des données partielles par rapport à l'ensemble de données de la semaine complète. Le nombre d'auditeurs est resté constant, ce qui implique que la plupart des auditeurs de l'ensemble de données ont soumis des journaux musicaux pendant les jours de la semaine ainsi que le week-end. Fait intéressant, le nombre total d'artistes pour lesquels des journaux musicaux ont été soumis les jours de semaine et les week-ends a diminué de 3 à 12% par rapport aux données de la semaine complète, ce qui signifie que de nombreux artistes ont été écoutés pendant l'une des deux périodes hebdomadaires.

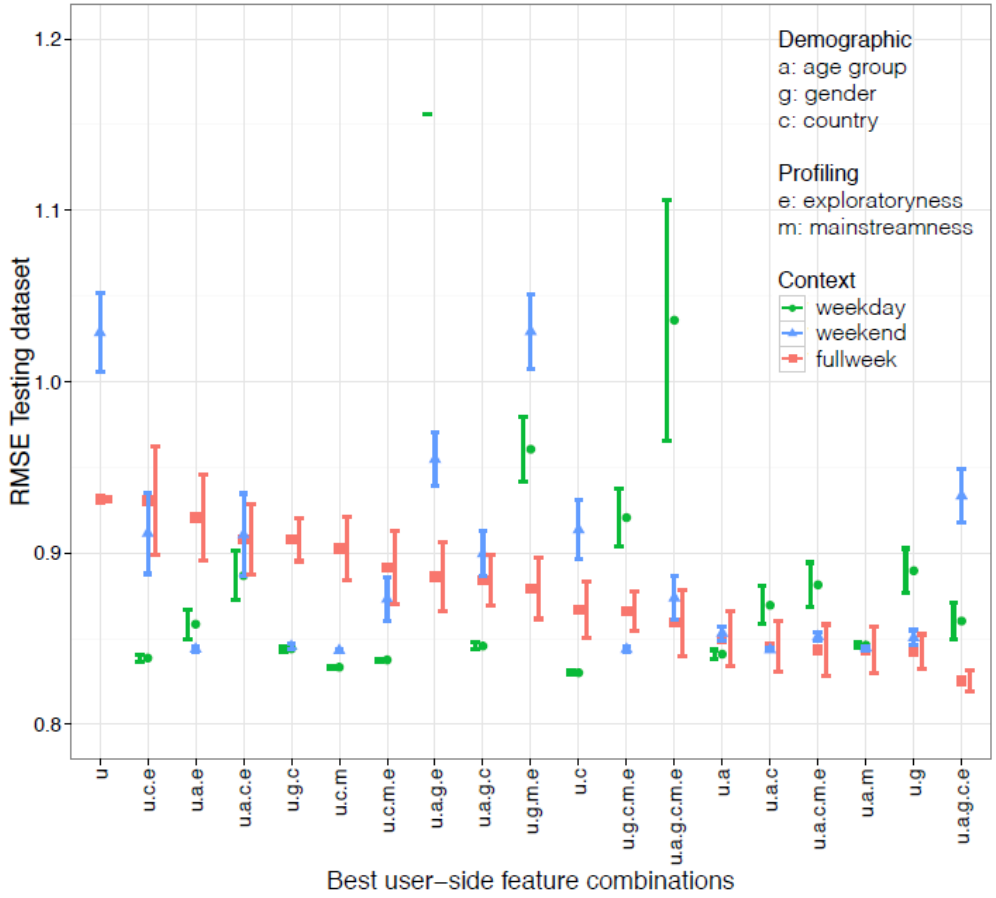


**Figure 3.** L'erreur quadratique moyenne signifie des barres d'IC à 95% pour les modèles appris avec des données de semaine, de fin de semaine et de semaine complète. Seules les combinaisons d'entités ayant une meilleure valeur RMSE que la référence pour les données d'une semaine complète sont affichées.

La figure 3 résume les précisions des modèles obtenues à l'aide des données du journal musical des trois contextes susmentionnés. La plupart des modèles réalisés avec des données d'écoute hebdomadaires ont obtenu de meilleures performances que ceux utilisant des données à la semaine complète. Par exemple, les modèles apprises avec des données de semaine et de week-end en utilisant des combinaisons de caractéristiques u.a.e, u.c.c. et u.c.m ont obtenu des améliorations de précision d'environ 7% par rapport au modèle créé avec des données de pleine semaine. Ils ont également montré une plus faible variabilité, ce qui signifie plus de stabilité dans l'estimation du modèle. Cependant, alors que la meilleure valeur RMSE a été obtenue en utilisant la combinaison de caractéristiques côté utilisateur avec u.a.g.c.e avec des données à la semaine, la même combinaison de caractéristiques a obtenu de moins bonnes performances en utilisant des données d'écoute uniquement en semaine et en week-end.

## 5. CONCLUSIONS AND FUTURE WORK

Nous avons évalué l'impact des caractéristiques démographiques et de profilage des auditeurs ainsi que des formes de base du contexte d'écoute, à savoir l'écoute en semaine et le week-end par rapport à l'écoute en semaine, sur l'exactitude des recommandations. Nous avons décrit nos exigences pour un ensemble de données sur les histoires d'écoute musicale, en expliquant pourquoi aucun des ensembles de données disponibles ne répondait à nos besoins et comment nous avons fini par recueillir nos propres données. Nous avons ensuite formalisé un ensemble de caractéristiques de profilage qui tiennent compte de certains aspects du comportement d'écoute de la musique. Nous avons également expliqué comment nous divisons notre ensemble de données d'histoires d'écoute en histoires d'écoute en semaine et en week-end pour évaluer si l'obtention de données provenant de différents ensembles d'histoires d'écoute améliorait la précision de la recommandation. Enfin, nous avons décrit comment nous avons mis en place des expériences évaluant toutes les combinaisons de caractéristiques de données côté utilisateur dans les différents contextes d'écoute. Nous avons constaté que la combinaison de caractéristiques ayant obtenu la plus petite erreur était l'ensemble des caractéristiques démographiques ainsi que l'exploration exploratoire, obtenant une amélioration de 12% par rapport au niveau de référence de ne pas utiliser de données de caractéristiques côté utilisateur. Bien que, pour certaines combinaisons de caractéristiques, l'utilisation de données d'écoute fractionnées ait amélioré la recommandation, la meilleure combinaison de fonctionnalités a bénéficié des données de la semaine complète. Les résultats, en particulier les nombreuses valeurs RMSE basses pour plusieurs combinaisons de caractéristiques utilisant des données d'écoute séparées, semblent indiquer que ces valeurs d'erreur sont proches de la limite de l'erreur minimale réalisable. Cette caractéristique a déjà été décrite dans la littérature comme une «barrière magique» dans la conception de systèmes de recommandation [1], se référant à la limite supérieure de la précision de prédiction d'évaluation en raison des incohérences dans les évaluations de l'utilisateur. Cependant, étant donné que nous comparons le nombre de journaux d'écoute musicaux soumis à des évaluations, nous ne voyons pas comment ces incohérences peuvent expliquer cet obstacle. Il serait intéressant d'effectuer une recherche de grille plus étroite afin de déterminer si nous atteignons un mur avec précision ou s'il existe un meilleur ensemble de paramètres de modèle qui nous permet de créer des modèles plus stables et de meilleures performances au cours de nombreux essais. Par rapport à la recherche précédente [21], les résultats ne sont pas comparables puisque différentes métriques sont utilisées. De plus, notre expérience a directement intégré les caractéristiques de profilage dans l'algorithme de factorisation matricielle. Enfin, bien que ces résultats montrent une amélioration de la précision d'un modèle de recommandation basé sur les antécédents d'écoute des auditeurs, nous pourrions avoir besoin d'une étude en ligne centrée sur l'utilisateur pour mesurer la satisfaction réelle des personnes face au modèle appris.

## 6. ACKNOWLEDGEMENTS

Cette recherche a été soutenue par BecasChile Bicentenario, Comisión Nacional de Ciencia y Tecnología, Gobierno de Chile et le Conseil de recherches en sciences humaines du Canada. Des parties importantes de ce travail ont utilisé les ressources de calcul haute performance de ComputeCanada. Les auteurs aimeraient remercier Emily Hopkins pour sa relecture approfondie de cet article.