# Prediction of User Demographics from Music Listening Habits

Thomas Krismayer
Christian Doppler Lab MEVSS
Institute for Software Systems Engineering
Johannes Kepler University Linz
thomas.krismayer@jku.at

Markus Schedl
Department of Computational Perception
Johannes Kepler University Linz
markus.schedl@jku.at

Peter Knees
Institute of Software Technology
and Interactive Systems
Vienna University of Technology
peter.knees@tuwien.ac.at

Rick Rabiser
Christian Doppler Lab MEVSS
Institute for Software Systems Engineering
Johannes Kepler University Linz
rick.rabiser@jku.at

## ABSTRACT

Online activities such as social networking, shopping, and consuming multi-media create digital traces often used to improve user experience and increase revenue, e.g., through better-fitting recommendations and targeted marketing. We investigate to which extent the music listening habits of users of the social music platform Last.fm can be used to predict their age, gender, and nationality. We propose a TF-IDF-like feature modeling approach for artist listening information and artist tags combined with additionally extracted features. We show that we can substantially outperform a baseline majority voting approach and can compete with existing approaches. Further, regarding prediction accuracy vs. available listening data we show that even one single listening event per user is enough to outperform the baseline in all prediction tasks. We conclude that personal information can be derived from music listening information, which indeed can help better tailoring recommendations.

## CCS CONCEPTS

•**Computing methodologies** → **Machine learning approaches;**
•**Social and professional topics** → *User characteristics;*

## KEYWORDS

User Trait Prediction, Digital User Traces, Music Listening Habits

**ACM Reference format:**
Thomas Krismayer, Markus Schedl, Peter Knees, and Rick Rabiser. 2017. Prediction of User Demographics from Music Listening Habits. In *Proceedings of CBMI, Florence, Italy, June 19-21, 2017,* 7 pages.
DOI: 10.1145/3095713.3095722


DOI: 10.1145/3095713.3095722

## 1 INTRODUCTION

Online activities such as using social networks or microblog services or shopping and consuming media leave digital traces, that indicate products or topics the user is interested in. These traces are recorded and many services use systems to recommend new items based on items the user selected or rated in the past (e.g., the music recommender system Last.fm or the online movie streaming service Netflix) [7].

It has been shown that many of the digital traces that are left by the users can also be exploited to predict additional information about them such as predicting a person's location from their tweets [1] or predicting personality traits from Facebook likes [10]. In this work, we focus on digital traces on the social music platform Last.fm, and use various different sources of information either directly from data available via the Last.fm API or extracted from the collected data to infer personal information of users.

We consider this a highly relevant topic with respect to digital media consumption and social media usage behavior for two reasons: on one hand, gaining a better understanding of the users will help in better understanding the contents of the media they are using, and thus help in creating more "semantic" indexing techniques. On the other hand, we are interested in how much this "harmless" and therefore often unthinkingly shared information can be used to derive additional information about the users. This second aspect exhibits direct ties to concerns regarding privacy and profiling.

To explore these aspects, we formulated the following two research questions: *(RQ1) To which extent is it possible to predict the age, gender, and nationality of the users based on their listening events and related information* (e. g., how the listening behavior changes over time)? *(RQ2) In which way does prediction accuracy depend on the available user data*, i. e., number of listening events?

The results of the proposed system can be utilized to enrich the input for recommender systems (e.g., to replace missing values for collaborative filtering approaches) or directly for recommending new items (e.g., artists that are popular in the country or within the age group of the user). In further steps the system could also be used to directly predict topics (e.g., genres) or items (e.g., artists or songs) the user is interested in, thus improving the user experience.

The remainder of this paper is structured as follows. In Section 2, we discuss literature related to the prediction of user traits from digital traces. Section 3 provides a description of the dataset used in our experiments. We introduce the actual algorithm for predicting user traits in Section 4. In Section 5, we describe the experiments performed and the results gained. Finally, Section 6 wraps up the paper with a conclusion and an outlook on future work.

## 2 RELATED WORK

In this section, we discuss work on automated prediction of user traits from digital traces, structured according to the source of collected user traces.

Kosinski et al. [10] show that user traits can be predicted based on the **Facebook** Likes of a person. The predicted values include basic profile information, such as age and gender, but also highly personal attributes, such as sexual orientation, ethnicity, political views, and personality traits. The prediction is based on the Likes of 58,000 Facebook users, for which demographic profiles and psychometric tests are available. A follow-up study to [10], conducted by Youyou et al. [23], shows that personality judgments made from Facebook Likes can be even more accurate than those of close friends or family members. Golbeck et al. [2] show that the personality of Facebook users can even be predicted based only on their publicly available profile information.

The algorithm described by Cheng et al. [1] estimates the location of **Twitter** users based on the text of their tweets. The estimation is entirely content-based and does not rely on meta-data, such as profile or network information. The proposed algorithm is trained on Twitter users in continental USA whose locations are known and then predicts the user location by inferring probabilities for cities from the microblogs. In their experiment, Cheng et al. report that 51% of the users were placed within 100 miles of their actual hometown.

Most closely related to our paper is work that exploits **Last.fm** data to predict listener characteristics. Liu et al. [12] estimate the gender of Last.fm users based on their listening history. Additionally, the age is estimated in a binary form as under or above 24 years. The features for the classification are constructed purely from the listening events of the user and are based on three factors: the listening timestamps, the meta-data of the song and the artist (e.g., artist and song tags), as well as signal features of the songs. For both tasks, a support vector machine classifier (SVM) with RBF kernel is used and the average of five runs with 80% of the users as training set is reported. The accuracy for age is 71.1%; the accuracy for gender is 66.1%.

The approach described in the work by Wu et al. [22] estimates gender and age of Last.fm users based on music meta-data. Their algorithm uses the songs that the user most frequently listens to. In contrast to [12], the approach does not exploit temporal information, nor any audio-based features. The authors describe two different ways to generate features for the user: Term Frequency - Inverse Document Frequency (TF-IDF) combined with Latent Semantic Indexing (LSI) and Gaussian Super Vectors (GSV). For both tasks, SVM with RBF kernels are used in a two-fold cross validation. The reported accuracy for gender estimation is 78.87% and 78.21% for GSV and TF-IDF, respectively. For age estimation a mean absolute error of 3.69 and 4.25 is reported for the GSV and the TF-IDF approach, respectively.

In contrast to these two existing works [12, 22], our main *contributions* are: (i) we present a novel approach for the prediction of user traits from music listening habits that combines multiple sources of information and uses PCA-compressed TF-IDF-like features, (ii) we also support the prediction of user nationality, (iii) we ran our experiments with users with a very limited number of listening events, to assess performance in cold-start situations, and (iv) we compare different machine learning classification and regression algorithms.

## 3 DATASET

The dataset used in our experiments is a subset of the LFM-1b dataset [16]. It was created using the Last.fm API, which allows the collection of users' profile information (including age, gender, and country) as well as listening events for these users. Additionally we used weighted artists tags, which were also extracted using the Last.fm API and can be used to identify artists that produce similar music, for our experiments.

The LFM-1b dataset additionally includes scores describing the listening behavior of the users. These scores include novelty, i.e., percentage of new artists in a specific time period, mainstreaminess, i.e., how well the preferences of the user fit to the average preferences of all users, and different listening counts (e.g., the absolute number of distinct artists the user listened to, the average number of events per week, and the relative number of events for one specific day of the week).

Discarding from the LFM-1b dataset users with missing demographic information or less than 500 listening events, a total of 12,181 users remained for our experiments. This allows to use the same dataset for all three prediction tasks (age, gender, and country). The restriction to users with at least 500 listening events ensures that all users have the same number of listening events for the experiments with listening event subsets.

The dataset eventually contains users from 144 countries with 72.5% of them being male and the average age being 25.6 years. In terms of number of users, the top countries in our dataset are: USA (19% of all users), Russia (8.9%), Germany (8.4%), Brazil (7.9%), Poland (7.8%), Great Britain (7.8%), and the Netherlands (2.6%). This distribution is similar to the distribution among the users in the entire LFM-1b dataset.

### 3.1 Balanced Gender Dataset

Due to the high share of male users in the dataset the baseline for the accuracy of gender prediction is rather high (72.5%). Although the best classifiers perform significantly better (81.4%, cf. Section 5.4), it is difficult to assess the performance of these classifiers. To overcome this problem when investigating the first research question for gender, during all experiments for gender prediction, we created multiple datasets, for which the users are filtered by selecting all female users and randomly selecting exactly as many male users. The datasets resulting from this procedure contain a total of 6,698 users (compared to the 12,181 users of the entire dataset) with a 50% share of female users.

### 3.2 Sampling Listening Event Subsets

We sampled small random subsets from the listening histories of users with 1, 2, 5, 10, 20, 50, 100, 200, and 500 listening events per user to investigate our second research question, i. e., to what degree the accuracy of predictions depends on the number of listening events used for training.

## 4 PREDICTION OF USER TRAITS

For prediction of user traits, we developed three models, one for age, gender, and country, respectively. Each model is built individually and does not use results from the other models. Furthermore, the models are built entirely from the listening data of the users, metadata of the artists, and extracted user information. Therefore, e. g., for the prediction of age, the model does not use the gender or the country of the user.

### 4.1 Experimental Setup

The prediction models are evaluated with a 10-fold cross-validation on the dataset introduced in Section 3. All steps for the prediction pipeline (feature selection, feature vector generation, dimensionality reduction, classification/regression) were individually performed for the different user traits age, gender, and country. The calculations for all steps are based solely on the training set; this also implies that the selected features and the dimensionality reduction rules are different for each fold of the cross-validation.

### 4.2 Feature Selection

For each user, an individual feature vector is constructed containing elements from three separate sources – the first part is based on artist listening information, the second part on artist tag information, and the third part on additional user information provided as part of the LFM-1b dataset. These three parts are created independently from each other. The first two parts are vector normalized separately, for the third part this is pointless as we will explain below. Finally, the three parts are merged to create one feature vector per user ("early fusion").

The *first part* of a user's feature vector (*artist listening information*) is created as follows. 10,000 artists are selected based on the number of users that listened to them. The first half of artists that is selected are the artists that have the most different users in the overall training set that listened to them at least once. The second half of the artists is selected based on their number of different listeners in user-groups chosen for the specific task. This means the users in the training set are split into distinct groups and the artists with the most users listening to them for each of the groups are selected.

For the age prediction task the users are split into eight distinct age groups also used in [17]. These groups contain the users in the age intervals [6–17], [18–21], [22–25], [26–30], [31–40], [41–50], [51–60], and [61–100]. For the gender prediction the artists with the most male and female listeners, respectively, are selected. Finally for the country prediction task the groups comprise the countries with the most users in the training set. The dataset contains 144 different countries, however the feature selection only takes into account the 25 most common countries within the training data to concentrate on the most crucial user groups. For the whole dataset the 25 most common countries contain 88.5% of all users.

The *second part* of a user's feature vector (*artist tag information*) is created by selecting 10,000 tags in the same way as the artists for the first part of the vector. The tags with the most users that listened at least once to an artist associated with this tag (with a tag weight higher than 0) are selected. The first 5,000 tags are selected based on the overall training set, while the second half is selected based on the same user groups as for the artists.

The *third part* of the feature vector contains 42 *additional scores* for each user, comprising scores for novelty (i. e., how many new artists did the user listen to in a given time period), mainstreamness (i. e., how well do the genre preferences of the user fit to the overall genre preferences of all users in the dataset), and various listening event counts (e. g., the average number of listening events per week).

The differences in the range of the scores makes a vector normalization of the third part pointless. For instance, the novelty scores of a user are calculated in the interval [0–1], while the count values of listening events have no boundary and are often above 10,000.

### 4.3 Feature Vector Generation

The entries for the first two parts of the feature vector of a user are calculated in the form of TF-IDF values for a term $t$ (i. e., an artist or a tag) and a document $d$ (i. e., the listening history of this user) as:

$$\text{tf-idf}(t, d) = \left(1 + \log\left(f_{dt}\right)\right) \cdot \log\left(\frac{n}{f_t}\right) \tag{1}$$

where $n$ is the number of users in the training set, and $f_t$ is the number of users with at least one listening event with the artist or tag. While $f_{dt}$ for artists simply is the number of listening events with the artists, the value for tags also takes the tag weight into account:

$$f_{dt} = \sum_{e \in E} weight(a_e, t') \tag{2}$$

where $E$ is the listening history of the user, $a_e$ is the artist of listening event $e$, and $weight(a_e, t')$ is the tag weight for tag $t'$ and artist $a_e$, which is 0, if the artist is not connected to $t'$.

### 4.4 Dimensionality Reduction

The feature vectors that result from the previous step have a high dimensionality, therefore dimensionality reduction via Principal Component Analysis (PCA) [6] is performed. The PCA is performed on the combined first two parts of the feature vector (i. e., 20,000 dimensions) to ensure that correlations between artist and tag features can be resolved.

The number of features is thereby reduced from 20,000 to 450. The new number of features results from adding 50 features as long as the average variance gained per feature stays above 0.03% (i. e., 1.5% for the 50 new features). The dimensionality reduction is performed in Python using the library scikit-learn [13]. The transformation is calculated based solely on the training set. The compressed feature vectors for the test set are then constructed using the same transformation.

### 4.5 Predictions for Listening Event Subsets

For the predictions based on listening event subsets (cf. Section 3.2) only the PCA-compressed first and second part of the feature vectors is used. The third part of the vector includes information that is not available in a cold-start-like situation that is simulated with these experiments and can therefore not be used. For instance, the novelty score represents an indicator of how the listening behavior of the user changes over time – an information that the system cannot estimate for a user, who just has one single listening event.

The classification/regression algorithm is trained on the original user vectors containing all listening events for the users in the training set. Based on this model the predictions for all subsets of the test set are made.

## 5 EXPERIMENTS AND RESULTS

Based on the reduced feature vectors resulting from the dimensionality reduction, different supervised models are built. The models are constructed using a selection of diverse machine learning classifiers and regressors. For this purpose, we use the Java API of the open source library Weka [3]. In this section, we analyze the results for the individual experiments using the same evaluation methods as in [12, 22] (i. e., mean absolute error for age and accuracy for gender and nationality).

Additionally, we evaluate the performance of the best classifiers on the reduced listening event subsets and the datasets with balanced gender share. We compare the results for all tasks to a baseline to help interpret their quality.

### 5.1 Learning Algorithms

For the prediction of the results a variety of different supervised classification and regression techniques are used.

**Support Vector Machines** (SVMs) aim at separating two classes by defining a border function in a potentially higher dimensional space such that data points from the two classes lie on the different sides of the border. SVMs can also be used in regression tasks by creating a function such that all data points fall within a given maximum error margin. The values for new data points are then predicted with this function. The predictions are made using implementations of the Sequential Minimal Optimization algorithm (SMO [4, 9, 14] and SMOreg [18, 19]).

**M5P** [15, 21] is a decision tree algorithm enhanced with linear regression, which can be a decision criterion for some of the nodes within the tree. Based on this algorithm, **M5Rules** [5, 15, 21] creates a decision list that is filled with rules from decision trees built with M5P.

**Linear regression** generates a regression function as a linear combination of the features. Similarly **logistic regression** [11] predicts the class of a data point based on a linear combination of the features. We use the two logistic regression algorithms Bayesian Logistic Regression and Simple Logistic.

**Naïve Bayes** [8] and **DMNBtext** [20] use Bayes' theorem to predict the class of a new instance based on the probabilities for the different classes inferred from the training instances.

**Table 1: Mean absolute error for age prediction (best results)**

| Classifier | Settings | Mean absolute error |
|---|---|---|
| SMOreg | RBF Kernel | 4.13 |
| SMOreg | Normalized Poly Kernel | 4.17 |
| SMOreg | Poly Kernel | 4.20 |
| Linear Regression | | 4.36 |
| M5P | | 4.40 |
| M5Rules | | 4.40 |
| SMOreg | PUK | 4.71 |
| ZeroR | | 6.23 |

### 5.2 Baseline

The baseline for the given tasks represents a trivial lower bound for the results of the classifiers. For the classification tasks, the baseline used is a classifier that predicts the majority class of the training set for all instances of the test set. E. g., for country prediction the baseline is a classifier that predicts the country with the most users in the training set for all users in the test set. In case of a regression task, the classifier predicts the average value in the training set for all instances of the test set. For both cases the calculation is done with Weka's ZeroR classifier [3].

### 5.3 Age Prediction

Table 1 shows the algorithms that achieved the lowest mean absolute error for predicting the age of the users. The support vector regression (SMOreg) outperforms all other algorithms with three of the four kernels available for this task. The lowest error (4.1; achieved with the RBF kernel) is 66.3% of the error achieved with the baseline algorithm. The Linear Regression achieves a slightly better result than M5P and M5Rules. The baseline for this task is 6.2 (calculated with ZeroR).

The results for the age prediction based on the subsets of limited listening events can be seen in Figure 1. We achieved these results with the SMOreg algorithm using the RBF kernel, which produced the best results for the entire dataset. Just one single listening event is sufficient to predict the age of the user more accurately than the baseline approach (5.8 vs. 6.2). The error of the prediction decreases steadily with an increasing number of listening events. Also, the final prediction that uses all of the available listening events achieves an error even lower than the prediction based on 500 listening events per user.

### 5.4 Gender Prediction

The baseline for the gender prediction is 72.5%. As a result of each of the training folds having a majority of male users, this is the share of male users among the dataset (see Section 3). Table 2 shows the performance of the best classifiers for this task. The algorithm achieving the best results is the Bayesian Logistic Regression. This algorithm, which was developed for text categorization, benefits from the features of the feature vectors including clustered TF-IDF values, because TF-IDF weighting is an approach developed as basis for text analysis and text categorization. Both the support vector

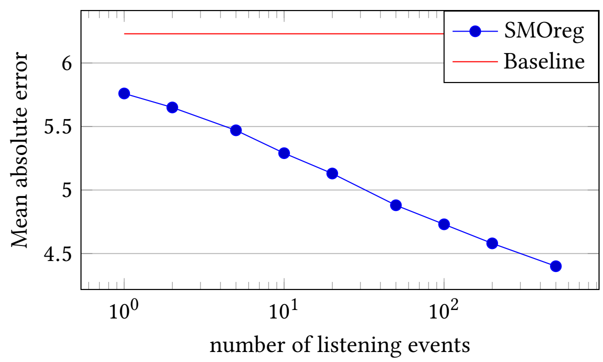


**Figure 1: Error for age prediction on listening event subsets.**

**Table 2: Accuracy for gender prediction (best results)**

| Classifier | Settings | Accuracy |
|---|---|---|
| Bayesian Logistic Regression | Gaussian Prior | 81.36% |
| SMO | Poly Kernel | 81.24% |
| Simple Logistic | | 80.43% |
| SMO | Normalized Poly Kernel | 78.06% |
| SMO | RBF Kernel | 78.33% |
| DMNBtext | | 77.22% |
| SMO | PUK | 76.31% |
| ZeroR | | 72.51% |

classifier (SMO) and the logistic regression algorithm (Simple Logistic) achieve results very close to the Bayesian Logistic Regression. The other algorithms yield far lower accuracy.

*Balanced gender dataset.* To compensate for the uneven gender distribution in the dataset, datasets with uniform gender distributions have been created, as detailed in Section 3.1. In order to ensure that the experiments on this dataset are not influenced by the listeners that are randomly picked for classification, the filtering is performed five times, the experiments are performed on each of the resulting datasets, and results are reported averaged over the five runs.

Due to the resampling of the dataset to achieve equal distribution of gender, the baseline for this task is obviously 50%. The results for the three classifiers that performed best on the whole dataset are given in Table 3. We report the average and the standard deviation over the five runs. It can be seen that all three classifiers perform between 4.2% (SMO) and 4.5% (Simple Logistic) worse than the same classifiers trained on the whole dataset (cf. Table 2), but have to be compared to a much lower baseline. The accuracy for the SMO using a poly kernel is 154.0% relative to the new baseline; for the complete dataset the Bayesian Logistic Regression achieves a relative accuracy of only 112.2% compared to the baseline.

The average results for the five runs of the balanced gender subsets for the Bayesian Logistic Regression and the SMO can be seen in Figure 2. Both classifiers achieve very similar results for all listening event subsets and are able to achieve results better than the baseline with just one single listening event (up to 54.5% with the SMO classifier). The results improve steadily with additional listening events and also improve from 500 listening events to the overall result.

**Table 3: Accuracy for gender prediction on the balanced dataset (best results)**

| Classifier | Settings | Accuracy |
|---|---|---|
| SMO | Poly Kernel | 77.01% ± 0.30% |
| Bayesian Logistic Reg. | Gaussian Prior | 76.91% ± 0.36% |
| Simple Logistic | | 75.88% ± 0.33% |
| Baseline | | 50% |

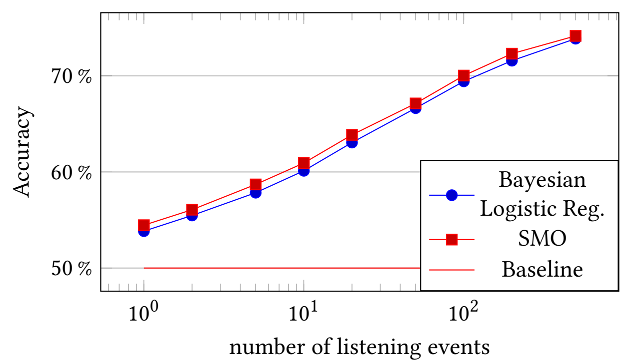


**Figure 2: Accuracy for balanced gender prediction on listening event subsets.**

**Table 4: Accuracy for prediction of countries (best results)**

| Classifier | Settings | Accuracy |
|---|---|---|
| Simple Logistic | | 69.37% |
| SMO | Poly Kernel | 69.36% |
| DMNBtext | | 63.11% |
| SMO | RBF Kernel | 59.97% |
| SMO | Normalized Poly Kernel | 59.57% |
| Naïve Bayes | | 57.39% |
| ZeroR | | 19.03% |

## 5.5 Country Prediction

Our third task is the prediction of the listeners' nationality. The baseline for this task is 19.0%, which equals the share of the most common country (USA) in the dataset. The classifiers that achieve the best results can be seen in Table 4. The two classifiers that perform best are the logistic regression algorithm (Simple Logistic) and the support vector classifier (SMO) which achieve 69.4% accuracy. The accuracy of the Simple Logistic algorithm is more than 3.6 times as high as the baseline.

The results for the reduced listening events can be seen in Figure 3, which include the results for the two best performing classifiers for the test set with all events (cf. Table 4). Similar to the predictions for age and for the balanced gender sets, both classifiers are able to beat the baseline with just one single listening

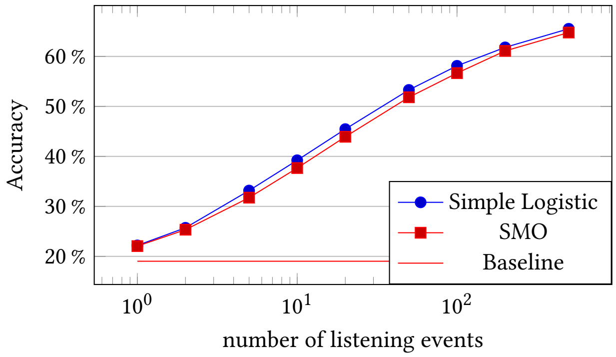


**Figure 3: Accuracy for country prediction on listening event subsets.**

event (22.2% accuracy for the SMO). and improve steadily with additional listening events.

### 5.6 Comparison with Existing Work

In the related work (cf. Section 2), the works of Liu et al. [12] and Wu et al. [22] have been introduced, which also target the prediction of user traits from music listening data.

The authors of [12] use the publicly available Last.fm 1K-users dataset to predict the gender and age of the users. This set contains users, for which user traits are missing. For the two experiments, the users, for which the respective trait is missing, are removed from the dataset. All the experiments are evaluated performing five runs with 80% of the users as training set and reporting the average of the results.

For this experiment we evaluated our approach with five-fold cross-validation, which also represents the average of five runs with 80% of the users as training set and additionally ensures that every user is part of the test set exactly once. Additional user information as in the Last.fm-1b dataset is not given and could therefore not be used.

For the gender prediction male users are removed from the dataset in order to create a set with a 50% share of female users. To lower the influence of the selected male users on the result we performed five runs of five-fold cross-validation – selecting different male users for each run – and reported the average result. The result achieved by our system is 72.9% (using Bayesian Logistic Regression with Gaussian prior), compared to an accuracy of 66.1%, which is the best result any approach in [12] achieved.

For the age prediction the authors split the user into the two classes “adolescents” (24 years and younger) and “adults” (25 years and older). Their best result achieved by [12] is 71.1%, compared to 72.4% achieved by our system (using Bayesian Logistic Regression with Laplace prior).

The authors of [22] use their own dataset to predict the age and gender of Last.fm users. Therefore it is unfortunately not possible to test our approach on their dataset; also the different number of users (96,807 vs. 12,181 users) and distribution of users (e. g., 66.2% vs. 72.5% male users) make a direct comparison of the received results pointless.

## 6 CONCLUSION AND OUTLOOK

Our experiments show that the listening history of a person allows to infer certain demographic information (RQ1). All three user traits age, gender, and country can be predicted to a substantial degree. For age the regression algorithm achieves an error that is 33.7% below the baseline error. For the balanced gender prediction and for the prediction of the nationality the increase in accuracy is 54.0% and 264.5% of the baseline, respectively. Even with a very small amount of listening events meaningful predictions can be made (RQ2). With increasing number of events the performance of the classifiers for all three user trait prediction tasks steadily increases.

Using the chosen approaches, we can indeed predict additional information about the users of online music listening services, solely from their listening histories. While the broad categorizations can help in tailoring collaborative as well as content-based recommender systems to their user groups, given the shown current limitations, however, it seems unlikely to generally predict personal information about the users that can affect their privacy.

As part of future work, we will consider additional listener- and listening-related aspects, for instance, exploiting the temporal information attached to listening events in greater depth. Also content-based features could be extracted and investigated, provided the respective audio is available. Another area that could be targeted in future work is deep learning – in addition to the learning algorithms used in our evaluation presented in this work.

## ACKNOWLEDGMENTS

This work has been supported by the Christian Doppler Forschungsgesellschaft, Austria and Primetals Technologies.

## REFERENCES

[1] Z. Cheng, J. Caverlee, and K. Lee. 2010. You Are Where You Tweet: A Content-based Approach to Geo-locating Twitter Users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. ACM, 759–768.

[2] J. Golbeck, C. Robles, and K. Turner. 2011. Predicting Personality with Social Media. In *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems*. ACM, 253–262.

[3] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations* 11, 1 (2009), 10–18.

[4] T. Hastie and R. Tibshirani. 1998. Classification by Pairwise Coupling. In *Proceedings of the 1997 Conference on Advances in Neural Information Processing Systems 10*. MIT Press.

[5] G. Holmes, M. Hall, and E. Frank. 1999. Generating Rule Sets from Model Trees. In *Proceedings of the 12th Australian Joint Conference on Artificial Intelligence*. Springer, 1–12.

[6] H. Hotelling. 1933. Analysis of a Complex of Statistical Variables Into Principal Components. *Journal of Educational Psychology* 24, 6 (1933), 417–441 and 498–520.

[7] Y. Hu, Y. Koren, and C. Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *Proceedings of the 2008 8th IEEE International Conference on Data Mining*. IEEE, 263–272.

[8] G. John and P. Langley. 1995. Estimating Continuous Distributions in Bayesian Classifiers. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, 338–345.

[9] S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, and K.R.K. Murthy. 2001. Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Computation* 13, 3 (2001), 637–649.

[10] M. Kosinski, D. Stillwell, and T. Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110, 15 (2013), 5802–5805.

[11] S. le Cessie and J.C. van Houwelingen. 1992. Ridge Estimators in Logistic Regression. *Applied Statistics* 41, 1 (1992), 191–201.

[12] J. Liu and Y. Yang. 2012. Inferring Personal Traits from Music Listening History. In *Proceedings of the 2nd International ACM Workshop on Music Information Retrieval with User-centered and Multimodal Strategies*. ACM, 31–36.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[14] J. Platt. 1998. Fast Training of Support Vector Machines using Sequential Minimal Optimization. In *Advances in Kernel Methods - Support Vector Learning*, B. Schoelkopf, C. Burges, and A. Smola (Eds.). MIT Press.

[15] R. Quinlan. 1992. Learning with Continuous Classes. In *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence*. World Scientific, 343–348.

[16] M. Schedl. 2016. The LFM-1b Dataset for Music Retrieval and Recommendation. In *Proceedings of the ACM International Conference on Multimedia Retrieval*. ACM, 103–110.

[17] M. Schedl, D. Hauger, K. Farrahi, and M. Tkalčič. 2015. On the Influence of User Characteristics on Music Recommendation. In *Proceedings of the 37th European Conference on Information Retrieval*. Springer.

[18] S.K. Shevade, S.S. Keerthi, C. Bhattacharyya, and K.R.K. Murthy. 1999. Improvements to the SMO Algorithm for SVM Regression. *IEEE Transactions on Neural Networks* 11 (1999), 1188–1193.

[19] A.J. Smola and B. Schoelkopf. 1998. *A tutorial on support vector regression*. Technical Report. NeuroCOLT2 Tech. Rep. NC2-TR-1998-030.

[20] J. Su, H. Zhang, C. Ling, and S. Matwin. 2008. Discriminative Parameter Learning for Bayesian Networks. In *Proceedings of the 25th International Conference on Machine Learning*. ACM, 1016–1023.

[21] Y. Wang and I.H. Witten. 1997. Induction of model trees for predicting continuous classes. In *Poster papers of the 9th European Conference on Machine Learning*. Springer.

[22] M. Wu, J. Jang, and C. Lu. 2014. Gender Identification and Age Estimation of Users Based on Music Metadata. In *Proceedings of the 15th International Society for Music Information Retrieval Conference*. ISMIR, 555–560.

[23] W. Youyou, M. Kosinski, and D. Stillwell. 2015. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences* 112, 4 (2015), 1036–1040.