Master's Thesis

Detecting the next pop star: Artist breakthrough predictions based on listener characteristics

Key words: Music, Streaming, Artists, Breakthrough, Prediction, Trendsetters, Listening

14 June 2021

T. Alkemade, 464766

Coach: M. Szymanowski

Co-reader: D. Zegners

Preface

Abstract

Nowadays, large-scale music streaming provides rich insights into listening activities, listener profiles, preferred genres, similar listeners and social networks. This information opens the door to a new approach towards future star detection. This paper proposes a model which detects musical trendsetters, based on listener data from the music database 'Last.fm', over a ten-year period. Each user is rated in terms of how often he or she listened to an artist before that artist broke through: The user's trendsetting score. It is studied what characterizes the most influential trendsetters: Their age, Last.fm membership, openness to novelty, music originality and/or network strength? Based on the strongest indicators of being a trendsetter, a 'trendsetter detection model' and a 'trendsetter profile' are built. These models classify users into 'trendsetters' and 'non-trendsetters'. Based on the variables included in the trendsetter detection model, the 'star prediction model' is proposed. This model analyses an artist's listener base characteristics to determine whether that artist's listeners fit into the trendsetter profile. Based on this information, the model predicts which musical talents will break through. Various stakeholders in the music industry can use this model to target those artists with the most promising career perspective.

Contents

1) Introduction

a. Problem background

Evidence from cultural economics literature shows that it's very difficult to predict which artists will break through in the future (Aguiar et al., 2018). This is not only caused by the cultural and artistic nature of music, but also by the digitization of recorded music. Digitization makes it easier to records and distribute songs, which attracts many more artists to the music industry. The overload of music supply makes it more difficult to predict which artists will be trending in a couple of years. Digitization of music also results in enormous amounts of musical content being uploaded daily. On the one hand, this results in enormous resources with rich and useful data about listening behaviour. On the other hand, it leads to fierce competition between artists. This raises the question whether insights can be gained from platform streaming data to predict which stars will breakthrough in the future. In this study, streaming platform data is analysed to determine the role of trendsetters, influential users, in the early career of music artists.

b. Problem statement and research questions

In this paper, the follow problem statement is studied: *on streaming platforms, trendsetters can be identified who listen to artists that become successful in the future.*

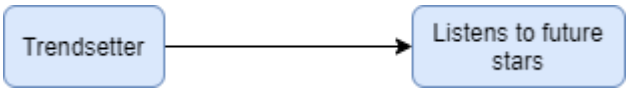


Diagram 1: Visual graph of the studied relationship

This paper studies the relationship between being labelled as a trendsetter and listening to future stars. Trendsetters are defined as "people that adopt and spread new ideas influencing other people before these ideas become popular" (Saez-Trumper et al., 2012). Applying this definition to the music streaming market, trendsetters are defined as "platform users who start

to listen to certain artists, after which other users listen to the same artist". Therefore, the independent variable 'trendsetter' reflects whether a user has proven to detect future stars in the past. In this paper, a 'future star' is defined as an 'artist which will become successful in the future'. Thus, the dependent variable 'Listens to future stars' indicates whether a user *currently* listens to an artist which will become successful in the future. A music streaming platform is either a paid or free music service which offers its users song streaming services.

To test the problem statement, the following research questions are answered in this paper:

1. Who are the most influential trendsetters in terms of star detection?
2. What characteristics distinguish trendsetters from non-trendsetters?
3. Can the characteristics of trendsetters be used to detect future stars?

c. Research approach

The trendsetter detection and star prediction models are built using a dataset obtained from Last.fm between 2005 and 2015. Last.fm is a music database to which users can connect their accounts on any music streaming platform, in order to track their streaming activities (Last.fm, 2021a). The data includes numerical and categorical data about user characteristics (*Age, Last.fm membership*) songs (*Song ID, artist, album, genre tags*), song streams (*User ID, Song ID, Time*), friends of focal users (*focal user ID, friend ID*) and users similar to the focal user (*focal user ID, similar user ID, similarity score*).

In order to identify so-called 'stars', all data is split into an early set (2005-2011) and a late set (2012-2015). For both the early and the late set, (the total number of) streams per artist is calculated. Artists who were streamed relatively often during one period are considered successful during that period, as will be discussed in detail at the 'Methodology' section. Artists who were considered unsuccessful during the early phase and successful during the late phase are considered 'stars'. All other artists are defined as a 'non-star'. A 'Future star' is defined as an artist which will become a star in the future. For clarity, the process of becoming a star is outlined in Table 1:

Early phase (2005-2011)		Late phase (2012-2015)
Unsuccessful	→	Successful
´Future star'		'Star'

Table 1: process of becoming a 'star'

The determine who are the most influential trendsetters, trendsetters are identified within the train set, consisting of 67% of all users. This is determined by means of their 'trendsetting score': how often the listened to (one of the) identified stars, before they broke through. The 7.5% users with the highest trendsetting score are classified into the trendsetter group. Remarkably, a user's absolute number of stars detected is chosen, rather than the fraction of stars among all artists listened to. This is to avoid the exclusion of users who have listened to many stars and non-stars, such that their fraction of stars is relatively low. These users did show clear trendsetter behaviour, so including them is expected to improve the models.

Next, the characteristics of trendsetters are identified: their age, membership, number of genres listened to (openness to novelty), average similarity score (preference originality) and number of friends on Last.fm (social network strength). Based on the characteristics identified within the train set, a trendsetter detection model is built. With this model, a trendsetter group is identified within the test group. Based on this, a general profile of trendsetters can be determined. With this profile, it is investigated whether more stars are detected by trendsetters compared to non-trendsetters.

To determine whether trendsetter characteristics can be used to detect future stars, the star prediction model is built. This model predicts whether an artist will breakthrough in the future, based on the characteristics of his listener base. Artists should target those listeners with the characteristics of a typical trendsetter to increase their chances of future success.

This star prediction model might also be applicable to other entertainment markets, such as the artwork, movie, book and podcast markets. When applying the model to the book market, the breakthrough of authors could be predicted by analysing their readers, i.e., who bought their books. In this way, publishers could detect talented authors early on.

d. Managerial relevance

This paper is relevant to artists, their managers, record labels, investors, listeners and streaming platforms, for various reasons. Streaming platforms become increasingly important for artists, since it provides them access to an enormous listening population and streaming royalties. Instead of only receiving a percentage of record sales or radio revenues, artists are now entitled to collect their earnings based on the number of plays they receive from each streaming service (Fly, 2021). This payment system emphasizes the need for information about which type of listeners should be targeted in order to be streamed more often and break through on a platform, which is studied in this paper. Such information is relevant to artist managers, to determine a better strategy for selecting the artists with the highest potential. It also informs them which listeners to target in the artist's first months to increase their chances of success. Similarly, record labels can use this information to detect future stars earlier and more accurately, after which they can offer them a contract at their label. Once artists are discovered by record labels, they will have more resources and opportunities to produce their music. This process makes it easier for talents to break through. However, record labels are currently losing control, with artists and consumers having the upper hand (Stafford, 2010). This suggests that artists become increasingly independent and need to develop their own targeting strategy, nowadays. Furthermore, investors would be interested in a star prediction model, since they have difficulty predicting which cultural products will be commercially successful (Aguiar et al, 2018). Music listeners would also benefit from future star detection, because it determines whether their favourite artists will break through or not. Moreover, the online music platforms have a stake in this problem, because more talent discoveries lead to more listeners, network effects, subscriptions and advertisement revenues. However, the most important contribution of this paper is that it provides artists, their managers and record labels a quantitative method to detect talents before they break through, based on their listener base.

e. Academic relevance

This paper contributes an evidence-based method to predict an artist's future performance, based on who is currently listening to that artist. These predictions are made using linear regression. The main literature streams which this paper relate to are social influence, music listening behaviour, talent detection and prediction models. A lot of research has already been

conducted on social influence and prediction models. Also, Zhang et al. (2013) and Jacobson et al. (2016) studied music listening behaviour and personalized recommendations. Furthermore, Anshel et al. (2012) studied sports talent detection methods. Little research has been done on *artist* talent detection, but within this research area, Schedl et al. (2005) tried to detect musical talents by exploring how often their names are mentioned on webpages. In addition, Haroutounian (2000) interviewed music school students to identify the characteristics of young talents. However, a new approach to talent detection is to analyse the characteristics of an artist's listener base. Thus, the prediction of future artist performance based on listener behaviour and characteristics is considered a gap in the existing literature.

2) Theoretical background

Over the past years, the music streaming market has grown tremendously. Digital music revenues now account for 54% of the global recorded music market, and streaming has for the first time become the single largest recorded music revenue source (IFPI, 2018). Spotify leads the audio market with 144 million subscribers, which is more than twice the number of subscribers of its closest competitor, Apple Music, which has 68 million users (Visual Capitalist, 2021). In June 2020, Spotify reached a $50 billion market valuation, which is the result of constantly adding more user tools and about 14 years of music data collection (DJMag, 2020). Furthermore, incorporating new media types, such as podcasts, has created tremendous opportunities for streaming platforms to expand their content offering (Li et al., 2020). Consequently, tens of thousands of pieces of music content are being uploaded at every moment (Hann, 2018). Such content is increasingly produced independently, at home instead of in an expensive studio. Especially during the corona pandemic, independent talents have been breaking through on Spotify, Twitch and Tiktok without having performed live once (Pfeiffer, 2021). The massive adoption of song streaming by a wide audience led to very large increases in the quantity and diversity of consumption (Datta et al., 2017). Such increasing music quantity and diversity provide streaming platforms enormous rich and useful data sources about listening behaviour. Meanwhile, streaming platforms increase competition between artists, with thousands of talented artists competing for exposure (Berg, 2018). All these artists compete with one another for streams in hopes of generating a higher stream share (Bender et al., 2021).

Considering the large amounts of streaming data on the one hand, and fierce competition on the other hand, it might be interesting to analyse and gain insights about listening behaviour. Such insights could be used to boost listening activities and attract more listeners, especially influential listeners. Several researchers have attempted to do this already. Zhang et al. (2013) analysed user behaviour and arrival patterns to determine the favourite times of day for Spotify users. Jacobson et al. (2016) analysed music listening behaviour to optimize music recommendations.

Other researchers studied the impact of social influence on streaming activities. According to Schedl et al. (2021), users with similar interests and with frequent correlate actions have a stronger influence on each other. Szymanowski et al. (2021A) claimed that music platforms can leverage social influence to stimulate the activities of their users. Further evidence was found that music discovery attributed to social influence and popularity both positively contribute to users' usage of the platform (Szymanowski et al., 2021B). Salganik et al. (2006) claim that increasing the strength of social influence increased both the inequality and unpredictability of an artist's success. This conclusion was drawn by creating an 'artificial market' in which users rated previously unknown songs either with or without knowledge of previous participants' choices. Innovativeness is considered to have a moderating effect in this process: Social influence between users is stronger when the innovativeness of friends is higher, and when the innovativeness of focal users and friends are more similar (Szymanowski et al., 2021A). Dewan et al. (2017) identified two types of social influence: popularity influence, driven by the total number of favourites (or 'likes') from the community as a whole, and proximity influence, due to the favouriting behaviour of immediate social network friends. The two types of influence are substitutes for one another, and proximity influence, when available, dominates the effect of popularity influence. It is clear that social influence plays a large role on music platforms, can be divided into proximity and popularity influence and is moderated by the innovativeness of friends.

The importance of social influence on streaming platforms raises an interesting question: "can information about the most influential users be used to predict an artist's future success (talent detection)?" This question is very relevant, since artist managers and record labels are searching for a data-based method for talent spotting (Hann, 2018). Potentially, streaming platforms are investing in talent detection models, but they tend to keep their strategic choices and data resources to themselves. Although Anshel et al. (2012) studied sports talent detection methods, little research has been done on *artist* talent detection. As mentioned before, within

this research area, Schedl et al. (2005) tried to detect musical talents by exploring how often their names are mentioned on webpages and Haroutounian (2000) interviewed young artists to detect the typical characteristics of musical talents. As far as one can tell from the literature, a new approach to talent detection is to analyse the characteristics of an artist's listener base.

In Table 2 below, an overview is provided of all the concepts introduced in the introduction, and other concepts introduced in the remainder of this paper.

Concept	**Definition**
Early phase	The period between 2005 and 2011
Late phase	The period between 2012 and 2015
Successful artist	An artist who managed to reach at least 200 streams in one of the phases
Unsuccessful artist	An artist who did *not* reach 200 streams in one of the phases
Future star	An artist who *will* move from being unsuccessful in the 'early phase' to successful in the 'late phase'
Star	An artist who *moved* from being unsuccessful in the 'early phase' to successful in the 'late phase'
Trendsetter	Platform users who start to listen to certain artists, after which other users listen to the same artist. In this study, it is defined as a user who has listened to relatively many future stars.
Trendsetting score	The number of times a user listened to a star, before their break-through. This score is calculated by the trendsetter detection model and determines whether a user is a 'trendsetter' or 'non-trendsetter'.
Trendsetter detection model	A model which classifies users into trendsetters and non-trendsetters based on their listener characteristics
Trendsetter profile	A profile used to characterize trendsetters, which is based on upper and lower limits on the listener characteristics included in the trendsetter detection model
Star prediction model	A model which calculates the star score of an artist, based on the average listener characteristics of his listener base. A star score within a pre-specified interval suggests that he is a future star.

Star score	The coefficients in the trendsetter detection model multiplied by the *average* values of the listener characteristics of the artist's listener base. This score is determined by the star prediction model.
Friends	Two users on Last.fm who are connected to each other
Neighbours	Two users on Last.fm who are similar to each other in terms of music preferences

Table 2: Relevant terminology

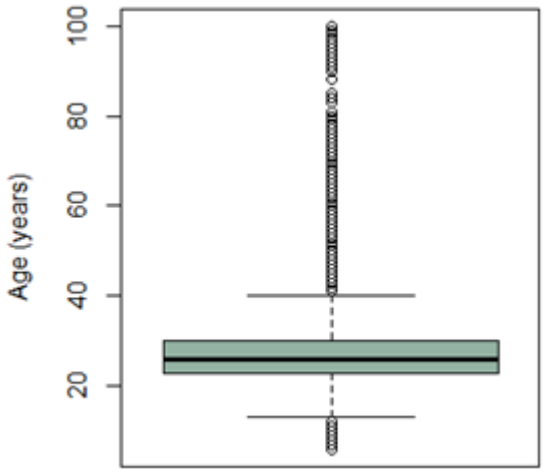
3) Data Preparation and Methodology

a. Data description

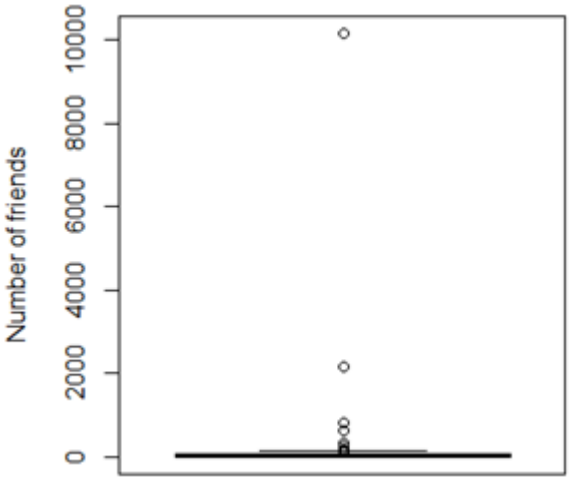
The Last.fm dataset includes several tables which include data collected between 2005 and 2015. It includes friends table (*focal user ID, friend ID*), similar users table (*focal user ID, similar user ID, similarity score*), user characteristics table (*User ID, Country, Age, Gender, Member, Registered*) and genres table (*focal user ID, Genre*). Those three tables include 150 focal users, their 50 most similar users, on average 40 Last.fm friends and all genres they have listened to. The dataset also includes a table with all songs on last.fm (*Song ID, artist, album, genre tags*) and a streams table (*User ID, Song ID, Time*), which includes all recorded streams by about 13,000 different users, of 870,830 different artists. Before any analysis is conducted, these tables are prepared for analysis. That is, an outlier analysis is conducted and only the data relevant for this study is subtracted from the data.

b. Preliminary outlier analysis of user characteristics

The first step of the data preparation is to identify any outliers in the data and determine whether they should be removed. The outliers in the user characteristics, friends, similar users and genres tables are plotted in a boxplot.

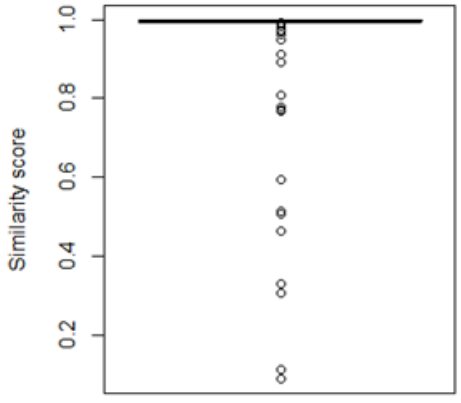


Box plot 1: Age distribution on Last.fm

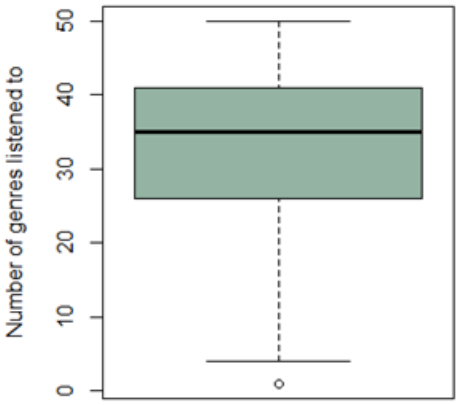


Box plot 2: Distribution of 'Number of friends' on Last.fm

Box plot 1 shows that most users have less than 40 friends, although there are numerous outliers with ages below 15 and above 40 years. However, ages on Last.fm below 6 and above 99 are removed from the user characteristics table, since people of these ages are considered unable to manage a Last.fm account. Box plot 2 indicates several users who have a few thousand friends; one user even has around 10,000 friends. Although these users are outliers, a closer investigation on Last.fm reveals that they actually do have that many friends (Last.fm, 2021c). Therefore, these values are *not* removed from the table.



Box plot 3: Similarity score distribution on Last.fm



Box plot 4: Distribution 'Number of genres listened to' on Last.fm

According to box plot 3, most users have an average similarity score just below 1.0. except for around twenty users. However, once again, these values correspond with the true similarity scores, so they are kept in the dataset. There is one user in box plot 4 with an

exceptionally low number of genres listened to. Still, listening to only a one genre is not unrealistic, so this value is *not* removed from the 'genres' table.

c. Data filtering

Only the data which is relevant for this study is subtracted from the tables in the Last.fm dataset. From the 'friends' table, the number of friends per user is computed, by summing all the user-friend combinations per user. From the 'neighbours' table, the average similarity score per user is calculated by averaging each focal users' similarity scores with his 50 most similar neighbours. From a combination of the song and genres table, the average number of times each user listened to a particular genre is calculated, by averaging all recorded user-genre combinations across the music genres in the dataset.

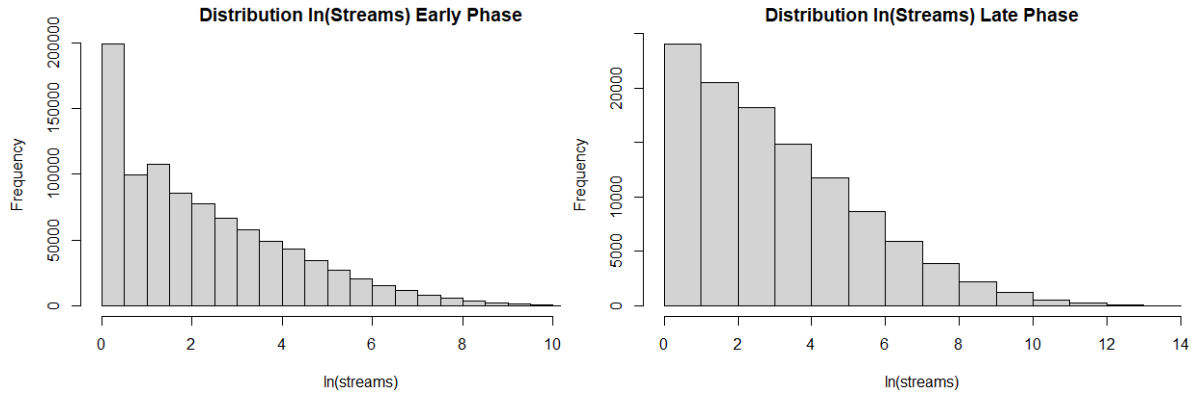
To build the trendsetter detection model, a list of streams is required, which includes which artists are streamed by which user, and when. To achieve this, the song and stream table are merged on the variable *Song ID*, to create a song-user-artist table which includes all song streams with their *song ID* and *Artist ID*. Next, a list of streams during one phase is needed to build the prediction models with. Also, a list of streams during a subsequent phase is required, which is used to predict breakthroughs. Therefore, the song-user-artist is split into an 'early phase' (all data between 2005 and 2011) and a 'late phase' (all data between 2012 and 2015), by filtering streams on their UTS timestamp. This results in a list of all streamed songs before 2012 and a list of streams after 2011. Remarkably, the early phase set was chosen to be almost twice as long as the late phase, to be able to identify trendsetters over a longer period. This is expected to lead to a larger group of trendsetters, a more accurate and precise trendsetter detection model and star prediction model.

d. Methodology

In this study, data about listening activities and user characteristics are studied to determine the influence of trendsetters on the popularity of music artists. The characteristics of these trendsetters are then used to predict star breakthroughs. The study proceeds as follows.

First, stars are being identified. Histograms 1 and 2 below show that artists who were streamed more than ln(9) times are truly exceptional, in both the early and the late phase. Therefore, it is assumed that 'successful artists' have been streamed more than ln(9), or 8,103, times during a particular period. All other artists are considered 'unsuccessful'. Artists who were considered unsuccessful during the early phase and successful during the late phase ('stars') are identified within the song-user-artist table. There are 29 artists in the stars table, whose number of streams increased with less than 50%. Those artists are filtered out, to include only artists with a substantial popularity increase and possibly increase the predictive value of the linear models. This results in a very exclusive group of 307 stars, which represent 1/2837 of all artists. This is intended to increase the accuracy of the trendsetter model.

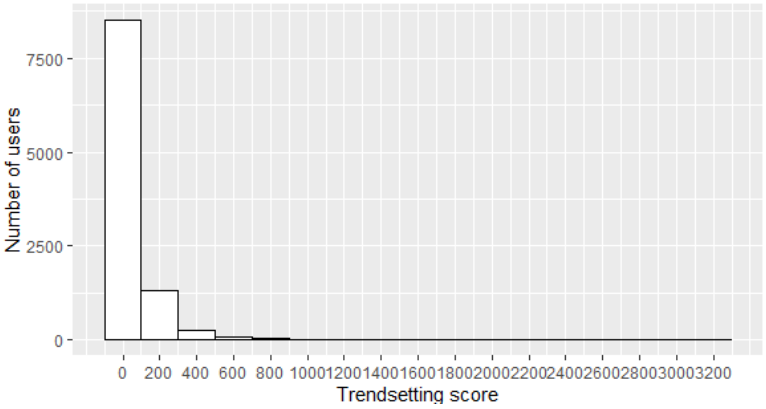
Note that the original distribution of streams was very hard to interpret, as shown in Appendix B. For that reason, the natural logarithm of streams was plotted, which was easier to interpret (cf. Histograms 1 and 2).



Histograms 1 and 2: distribution of ln(streams) in (1) the early phase and (2) the late phase

From the user table, 67% of the users is randomly assigned to the train set and 33% to the test set. This results in a train set consisting of 10,198 potential trendsetters, and a test set consisting of 5059 potential trendsetters.

The distribution of trendsetting scores of the users the train set is analysed, to determine which users are assigned to the trendsetter group.



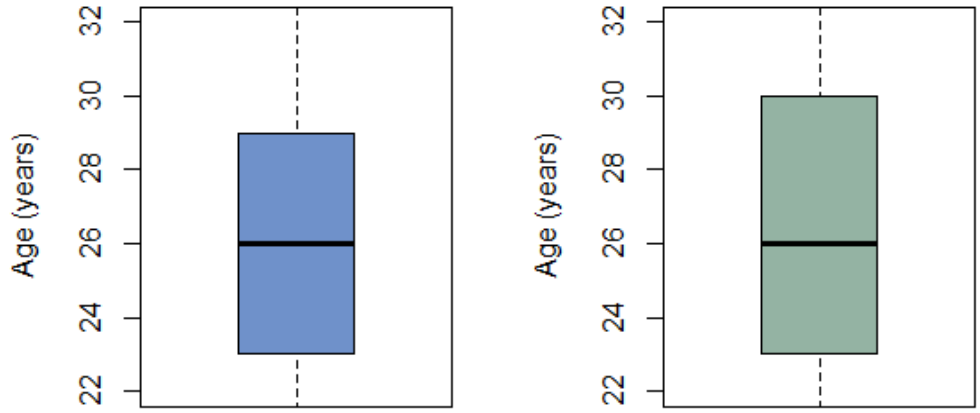
Bar graph 1: Frequency distribution of Trendsetting Scores

Bar graph 1 shows that around 8500 users, the vast majority, seems to have listened less than 200 times to a star. These trendsetters do not show clear trendsetter behaviour, so only users with a trendsetting score above 200 are assigned to the trendsetter group. This results in a group of 763 trendsetters, which represent 7.5% of the users in the train set. The remaining users are assigned to the non-trendsetter group. An overview of the trendsetter-classification is provided in Table 3:

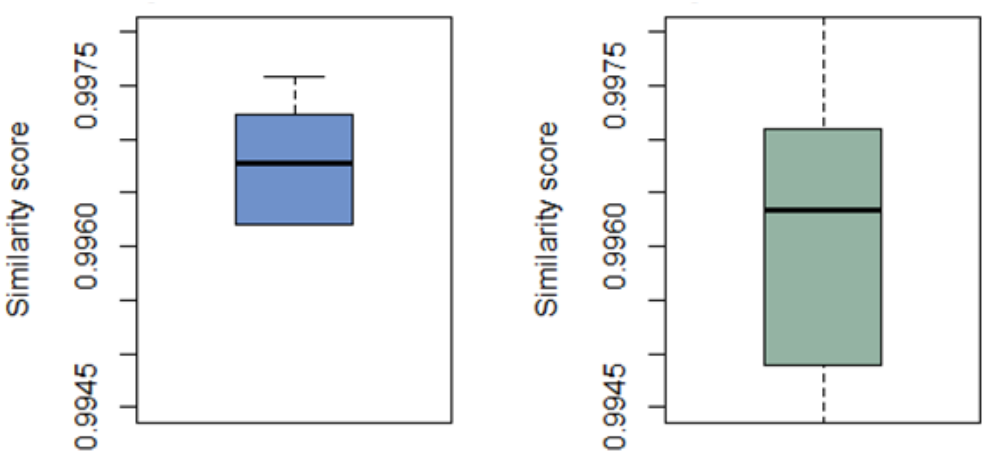
	Trendsetting score	**Classification**
Potential trendsetters (train set)	> 200	Trendsetter
	< 200	Non-trendsetter

Table 3: Overview trendsetter classification

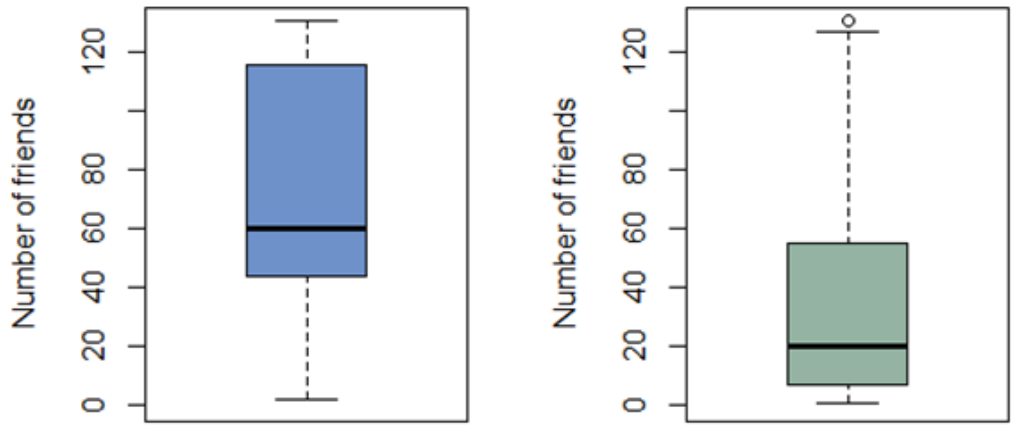
Several characteristics of the trendsetting group are analysed to determine which of these characteristics are included in the trendsetter detection model. For every characteristic, the distributions among trendsetters are compared to the distributions among *all* Last.fm users. In this way, trendsetters can be distinguished from non-trendsetters.



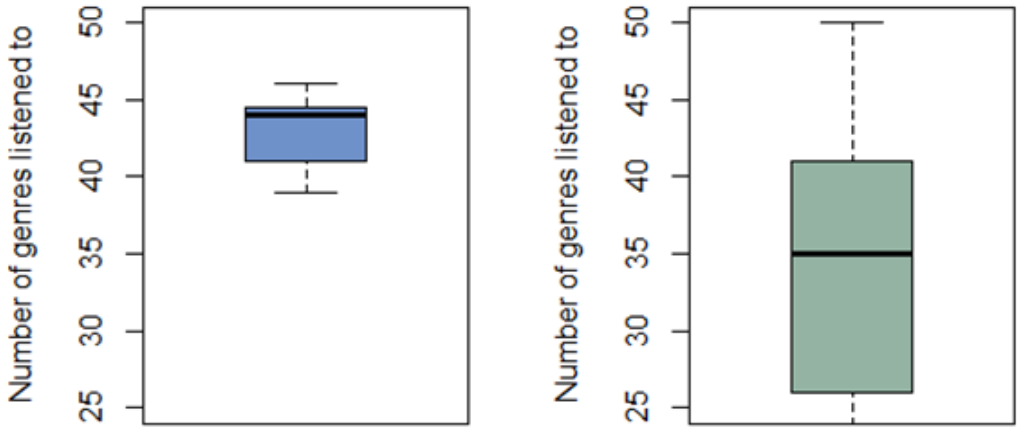
Box plots 5 and 6: Age distributions among (1) trendsetters and (2) Last.fm users in general



Box plots 7 and 8: Similarity score distributions among (1) trendsetters and (2) Last.fm users in general

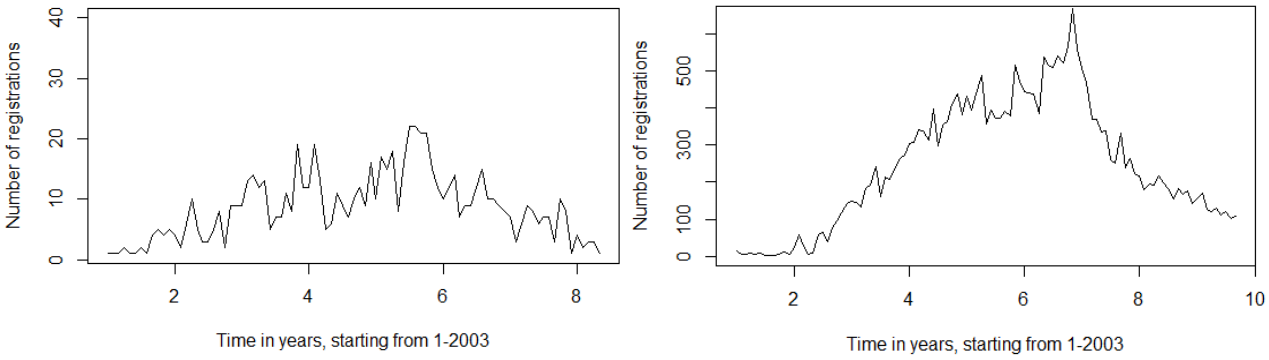


Box plots 9 and 10: Distributions of 'number of friends' among (1) trendsetters and (2) Last.fm users in general



Box plots 11 and 12: Distributions of (1) 'number of genres listened to' among trendsetters and (2) Last.fm users in general

The box plots above provide a few important insights about the identified trendsetters. Box plot 1 shows that the age distribution of the trendsetter group is similar to the age distribution of the average user, although the 3rd quartile of the trendsetters' age is 1 year above average. Most trendsetters seem to be 29 years or younger, while most Last.fm users are at most 30 years old. Box plot 2 shows that most trendsetters score higher than average on similarity with other users. According to box plot 3, trendsetters have around 40 more Last.fm friends than the average user. Lastly, box plot 4 shows that most users on Last.fm have only listened to about 25-40 genres, with an average of 35 genres. The trendsetters have listened to about 10 more genres, on average.



Time series plots 1 and 2: registrations by (1) trendsetters and (2) all Last.fm users

Time series plot 1 shows that most trendsetters registered to Last.fm between 2004 and 2010. A similar trend can be seen on Last.fm in general, although the peak of registrations between 2004-2010 was clearly higher for Last.fm users than trendsetters. Still, the time of registration

does not seem to influence the chance of being a trendsetter, so this variable is not included in the trendsetter detection model.

Several demographic variables are also evaluated as candidates for the trendsetter detection model. An analysis of the trendsetters' nationalities reveals that users from the UK are overrepresented the most in the trendsetter group, followed by the US and Mexico. Thus, being from the UK, US or Mexico increases a user's chances of being a trendsetter. However, this relationship might be very specific to the Last.fm dataset, so targeting listeners from these specific countries would reduce the generalizability of the model. So, this variable is also not included in the trendsetter detection model. Furthermore, 73.7% of the trendsetters are male, while only 62.3% of all Last.fm users are male. However, a selection of trendsetters based on gender seems undesirable, since it will exclude a group of influential, but female users. Therefore, gender is not included in the trendsetter detection model either. Lastly, 0.9% of the trendsetters is a Last.fm member, while only 0.7% of all users is a member. Membership is considered a good candidate variable for the trendsetter detection model, so it is included.

As discussed above, trendsetters distinguish themselves by means of their Age, Last.fm membership, number of friends, number of different genres listened to and similarity score. Therefore, those variables are used as input to linear model 1a, which is built as follows.

A linear regression is run on the dependent variable *Y*, which indicates whether a user is a trendsetter (Y = 1) or not (Y = 0). Variable *Y* is regressed on the variables age *A*, number of friends on Last.fm *F* (social network strength), average similarity score *S* (preference originality), number of genres listened to *G* (openness to novelty) and being a Last.fm member (M=1) or not (M=0) *M*. The age *A* is expected to have an effect on the dependent variable, since younger users are considered more likely to detect upcoming artists than older users. This is in line with Anderson et al. (2020), who argue that "As age increases, organic (user-driven) diversity of music streaming goes down". The number of friends on Last.fm *F* is considered a good measure for social network strength, since having more friends assumingly provides a user greater influence over the listening activities of other users. This assumption is based on Dewan et al.'s (2017) proximity influence theory, which claims that having a friend who has liked as song increases the likelihood of listening to a song. The variable *S* is chosen to capture preference originality, since a high similarity score means that a user has relatively similar music preferences as other users. In other words, a high similarity score is associated with a *low* preference originality. Furthermore, the assumption that a preference originality is an indicator of trendsetter behaviour, is underpinned by Schedl et al.'s (2015) theory that

users with similar interests and with frequent correlate actions have a stronger influence on each other. Moreover, the number of genres listened to *G* provides information about a user's willingness to try new music genres. Therefore, *G* measures a user's 'openness to novelty', which is defined as "An individual's intrinsic need to seek stimulation through novelty, i.e., previously unfamiliar genres or artists" (Tang et al., 2017). Lastly, by including Last.fm membership *M* in the model, it is tested whether members detect more trendsetters than non-members. This relationship is based on the assumption that members are more active on the platform and probably discover more future stars. Together, these five variables are expected to provide a good indication of trendsetting behaviour. However, only the values which increase the r-squared when added to linear model 1a are included in linear model 1b, the trendsetter detection model. Models 1a and 1b will take the following form:

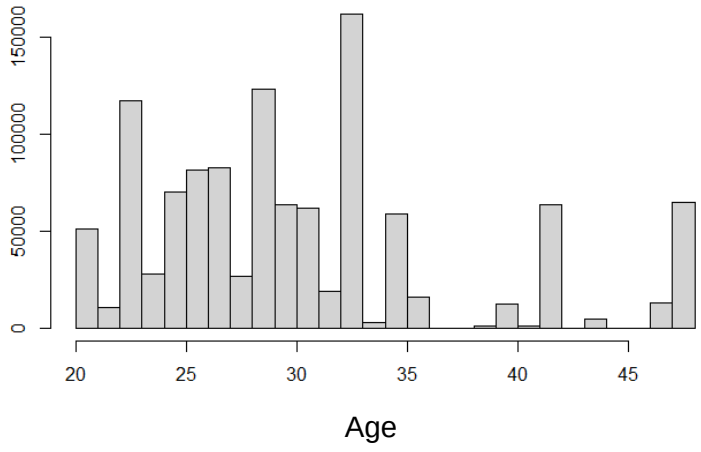
$$Y = \alpha + x_1\beta + x_2\beta \ldots + x_i\beta + \varepsilon$$

$Y \in \{0,1\}$

$\varepsilon = random\ error\ term$

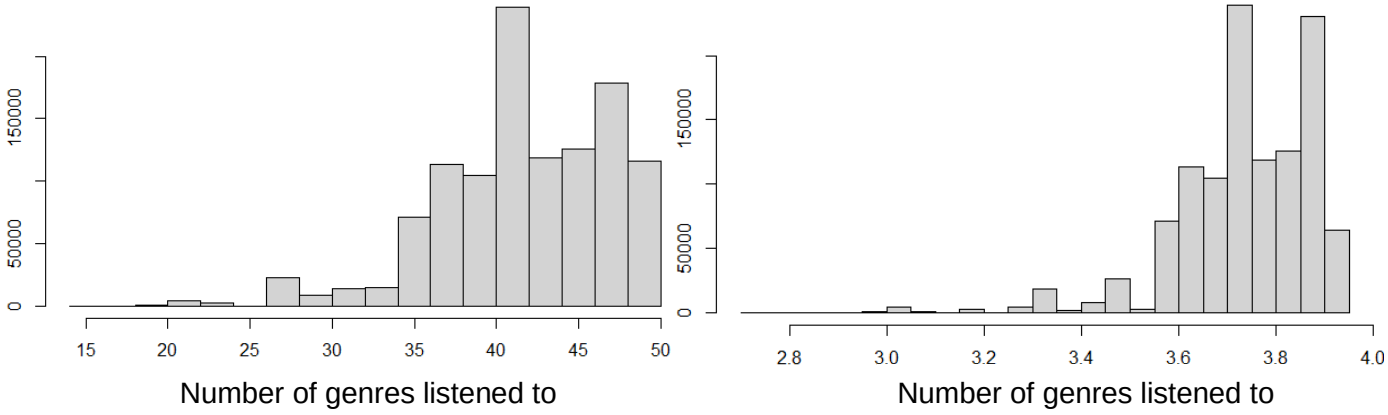
i = user characteristic

Before the linear regression is run, a histogram of the distribution of the independent variables *A*, *G*, *F* and *S* is plotted, to check whether they are suitable for linear regression or have to be log-transformed first. Note that the natural logarithm of the variables is taken, if necessary.



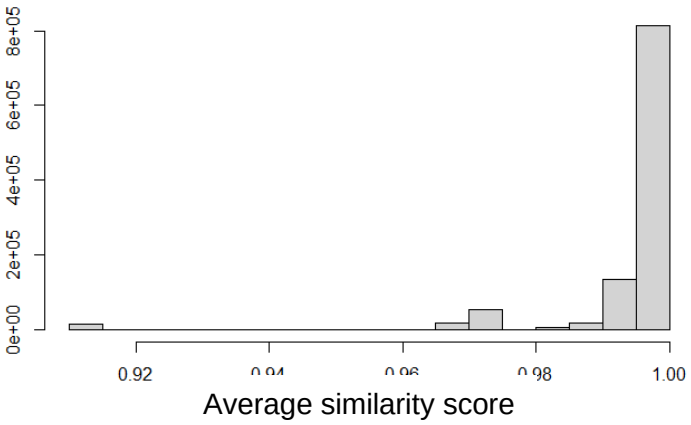
Histogram 3: distribution of Age among potential trendsetters

As shown by histogram 3, most listeners are between 20 and 35 years old. The variable *A* is distributed more or less equally between the ages 20-35. Since *A* does not approach a normal distribution, log-transformation would be inappropriate.



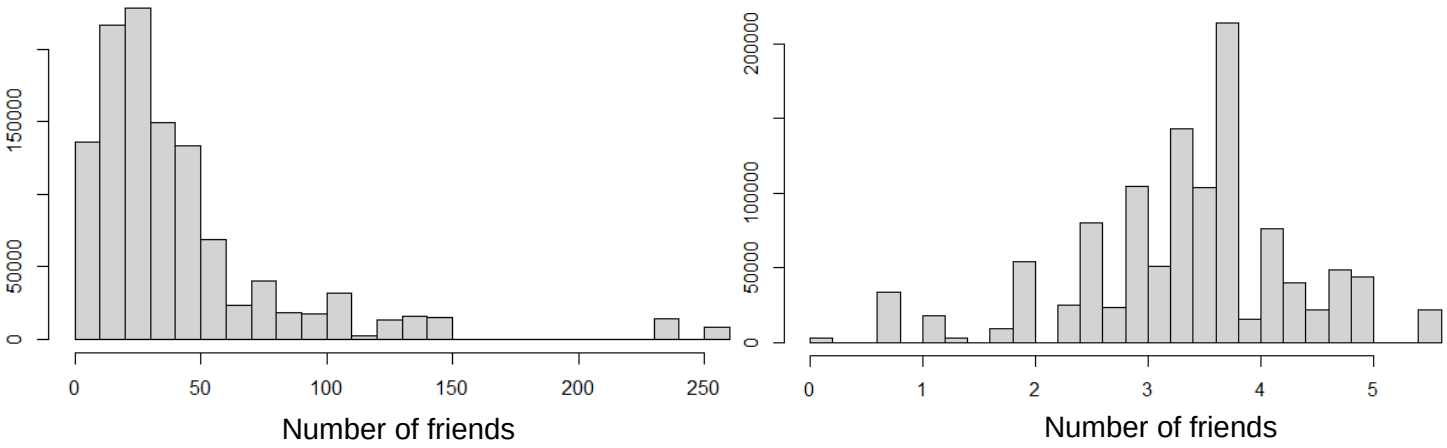
Histograms 4 and 5: distribution of (1) original G *and (2) log-transformed* G *among potential trendsetters*

According to histogram 4, the distribution of *G* is skewed to the left. However, since *G* approaches a normal distribution, it can be log-transformed. Approaching a normal distribution more accurately would increase the validity of *G* when included in the linear model. However, as histogram 5 shows, *G* is skewed to the left even more when log-transformed. Therefore, the original *G* is included in the model.



Histogram 6: distribution of S among potential trendsetters

As shown by histogram 6, almost all similarity scores approach 1. *S* is clearly not normally distributed, so this variable is not log-transformed either.



Histograms 7 and 8: distribution of (1) original and (2) log-transformed F among potential trendsetters

According to histogram 7, the distribution of *F* is skewed to the right. As an attempt to approach a normal distribution, the variable *F* is log-transformed. As histogram 8 shows, the log-transformed *F* does approach a normal distribution, so the linear model a will be run with both *F* and log(*F*) to check which model leads to the highest r-squared.

After running the regression, model 1b is checked for robustness by means of a heteroskedasticity test. Finally, the model's performance is evaluated on two metrics: its accuracy and precision. Accuracy refers to how many artists could be classified correctly as 'star' or 'non-star'. Precision reflects how many of the users classified as a 'trendsetter', are actually trendsetters. These metrics are calculated by implementing model 1b on all 107 Last.fm users about which the listener characteristics in model 1b and the true value of *Y* are known. By comparing the predicted Y-values to the true Y-values, it is determined how many trendsetters and non-trendsetters could be detected by means of model 1b. The accuracy of model 1b is compared to the number of users which could be 'classified' correctly by a randomly generated list of binary numbers: the 'random predictor'. Likewise, the precision of model 1b and the random predictor are compared. If model 1b is more accurate and/or precise than the random predictor, model 1b provides valuable information.

Based on the listener characteristics in linear model 1b, a 'trendsetter profile' is built. This profile is used to select a group of trendsetters from the *test* group, by setting lower and upper bounds to the variables included in model 1b. The next step is to determine the detection accuracy and precision of the trendsetter profile. This check is performed by determining how many trendsetters could be identified within in the test group, using the trendsetter profile.

The variables included in the trendsetter detection model are used to build linear model 2, the star prediction model. This model predicts whether an artist will breakthrough in the future, based on the values of the user characteristics in model 1b, averaged on the artist's listener base. Thus, the model is based on the assumption that if an artist's listeners have the characteristics of a typical trendsetter, that artist is more likely to become a star. The dependent variable *Y* indicates an artist's 'star score': the coefficients in linear model 1b multiplied by the *average* values of A, F, S and G of the artist's listener base. If an artist's star score falls within a pre-specified interval, the artist predicted to become a star. The model is run on all the 32,829 artists in the Last.fm dataset whose listener base characteristics are known. The model will take the following form:

$$(2)\ Y\ =\ \alpha\ +\mu x_1\beta\ +\mu x_2\beta\ \ldots+\mu x_i\beta\ +\ \varepsilon$$

$\varepsilon\ =\ random\ error\ term$

x_i = the ith user characteristic included in the model

Similar to model 1b, model 2 is checked for robustness by means of a heteroskedasticity test. Its performance is also being evaluated. To evaluate its accuracy, the number of correct classifications by model 2 is compared to how many artists were 'classified' correctly means of a list of random binary numbers. In addition, the precision of model 2 and the random predictor are compared. This refers to how many of the artists classified as a 'star', are actually stars.

To put model 2 into perspective and compare it to another prediction method, the trendsetter profile is used to predict which artists will become stars. This will be achieved as follows. A list of users about which the listener characteristics in the trendsetter detection model are known, is compiled. These user characteristics are then averaged across each artist. For each artist, the average values of his listener base characteristics are calculated. The thresholds in the trendsetter profile are used to filter out those artists whose listener bases have the characteristics of a typical trendsetter. These artists are predicted to become stars. Equivalently to model 1b and 2, the model's performance is being evaluated by comparing its performance to the performance of a random predictor.

e. Alternative approaches

Instead of choosing independent variables for the linear model by means of logical reasoning and linear regression, there would have been two alternatives to choose these variables.

First, the study could have started with a Principal Component Analysis (PCA) to identify the variables which explain the most variation in variable *Y* (being a trendsetter or not). This would reduce dimensionality and avoid overfitting to the Last.fm dataset. Based on the identified principal components, a linear regression could have been run. However, a drawback of this approach is that only a selection of the available variables is used, while listener characteristics were already scarce in the dataset. This would lead to an even more simplistic model. Therefore, a linear model was conducted without conducting a PCA first.

A second approach would have been to conduct a cluster analysis to determine whether a cluster of trendsetters exists in the train set. This approach would, once again, start with a PCA to identify the variables which explain the most variation in being a trendsetter or not. Based on these variables, users would have been classified to clusters of trendsetters and non-trendsetters. If a distinguishable cluster of trendsetters could be found, the characteristics of this cluster would be used as independent variables of the trendsetter detection and star prediction models. However, when performing both a PCA and a cluster analysis, the number of available listener variables would have been reduced even more. Therefore, a cluster analysis was also not performed.

Finally, instead of a single 67%-33% split, multiple samples could have been created by means of n-folds cross validation. Running the models on different samples would have provided more test samples and possibly an increased prediction accuracy. However, this approach was not chosen because of the difficulties of training, testing and evaluating the models on multiple samples, within a short timeframe.

4) Results

a. Trendsetter detection model

Regressing the variables *A*, *F*, *S*, *G* and *M* on the variable *Y* resulted in the coefficients in Table 4.

Variable	β (σ)
AGE	-0.004 (-0.005)
MEMBER	-
Number_of_friends	0.001 (-0.001)
Average_similarity_score	0.007 (-0.312)
Number_of_different_genres_listened_to	0.006 (-0.004)
Constant	-0.073 (0.408)
Observations	76
R2	0.046
Adjusted R2	-0.008
Residual Std. Error	0.273 (df = 71)
F Statistic	0.848 (df = 4; 71)

Table 4: Coefficients Linear Model 1a

The coefficients in Table 4 serve as input for model 1a.

$$(1a) \quad Y = -0.073 - 0.004A + 0.001F - 0.048S + 0.006G + 0M + \varepsilon$$

$Y \in \{0,1\}$

$\varepsilon = random\ error\ term$

Model 1a

According to model 1a, the variable *S* (β = -0.048, *p* = 0.371) explained the most variation in *Y*, followed by the variables *G* (β = 0.006, *p* = 0.005), *A* (β = -0.004, *p* = 0.006), *F* (β = 0.001, *p* = 0.001) and *M* (β = 0, *p* = 0). The R-squared of the model is 0.046 (cf. Table 4). So, the 'average similarity scores' *S*, 'number of different genres listened to' *G*, 'age' *A*, 'number of friends on Last.fm' *F* and 'member' *M* explained the most variation in the variable *Y*, indicating whether a Last.fm user is a trendsetter (Y = 1) or not (Y = 0). However, the variable 'member' *M* did not increase the r-squared when added to the model, so it is dropped

from the model. The log transformation of *F* also decreased the r-squared, so the original *F* is included in the model (cf. Appendix C). The resulting model is model 1b, the trendsetter detection model, which has the same R-squared as model 1a: R2 = 0.046 (cf. Table 5).

Variable	β (σ)
AGE	-0.004 (-0.005)
Number_of_friends	0.001 (-0.001)
Average_similarity_score	0.007 (-0.312)
Number_of_different_genres_listened_to	0.006 (-0.004)
Constant	-0.073 (-0.408)
Observations	76
R2	0.046
Adjusted R2	-0.008
Residual Std. Error	0.273 (df = 71)
F Statistic	0.848 (df = 4; 71)

Table 5: Coefficients Model 1b

The coefficients in Table 5 serve as input for model 1b, the trendsetter detection model:

$$(1\text{b})\ Y = -0.073 - 0.004A + 0.001F + 0.007S + 0.006G + \varepsilon$$

$Y \in \{0,1\}$

$\varepsilon = random\ error\ term$

Model 1b: The Trendsetter Detection model

According to linear model 1b, the variable *S* (β = -0.007, *p* = 0.312), explained the most variation in *Y*, followed by the variables *G* (β = 0.006, *p* = 0.004, *A* (β = -0.004, *p* = 0.005) and *F* (β = 0.001, *p* = 0.001). Notably, age *A* is negatively correlated with being a trendsetter *Y*, while the other variables are positively correlated.

	Trendsetter detection				*Performance metrics*	
	A: Correct detections	**B: Correct trendsetter detections**	**C: Number of users indicated as 'trendsetter'**	**D: Total detections**	**Accuracy (=A/D)**	**Precision (=B/C)**
Model 1b (refined)	24	10	92	106	22.64%	10.87%
Random predictor	48	3	54	106	45.28%	5.56%

Table 6: Performance evaluation models 1b and 2

When using linear model 1b to detect trendsetters in the *test* group, trendsetters and non-trendsetters could be classified with 22.64% accuracy (cf. Table 6). The precision of the model is 10.87%. In other words, among those indicated as 'trendsetter', 10.87% of the users actually was a trendsetter. The random predictor resulted in a 45.28% accuracy and 5.56% precision.

A heteroskedasticity test of linear model 1b reveals that the effects of the different standard errors on model accuracy are limited, except for the effects of 'number of different genres listened to' G, which becomes significant in the case of white heteroskedasticity standard errors ($\beta = 0.006$, $p < 0.05$) (cf. Appendix D). Overall, the model is considered robust.

As an attempt to find a more accurate trendsetter detection method, a trendsetter profile is built, with an upper bound for 'Age' *A* and a lower bound for the other three independent variables in the model. The minimum and maximum bounds in the profile are determined by the first and third quartiles of the distributions of those variables (cf. box plots 1-4). With a minimum number of friends of 44, genres listened to of 20 and similarity score of 0.9123 and a maximum age of 29, a group of trendsetters is identified with a trendsetting score which is almost double the trendsetting score of the average Last.fm user (117.14 compared to 60.68).

b. Star prediction model

Next, the coefficients in the trendsetter model 1b are used to build linear model 2, which determines an artist's star score.

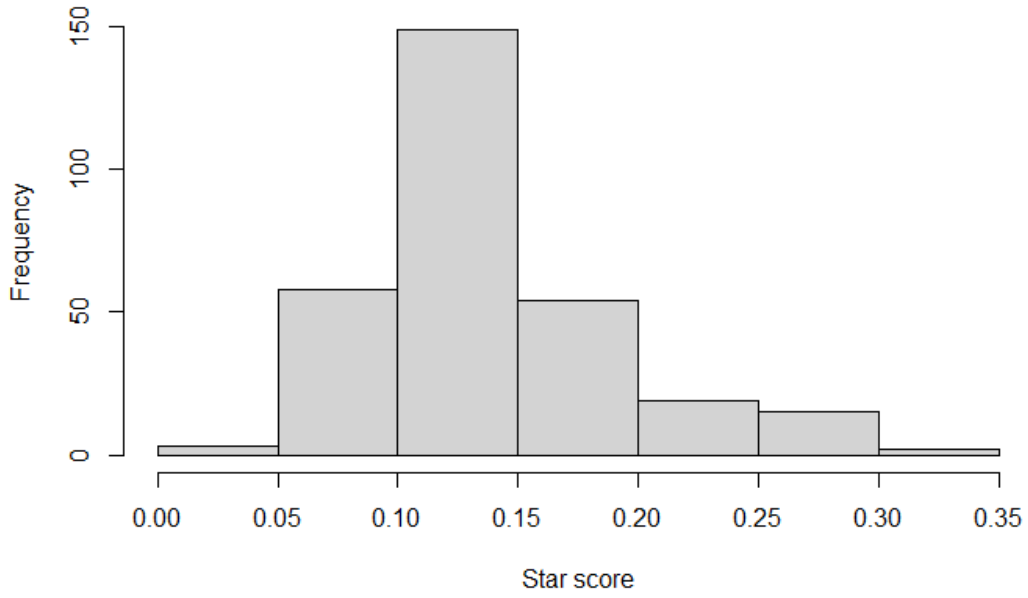
$$(2)\ Y = -0.073 - 0.004\mu A_i + 0.001\mu F_i + 0.007\mu_t S_i + 0.006\mu_t G_i + \varepsilon$$

$\varepsilon = random\ error\ term$

μX_i= average value of variable *X*, of the listener base of artist *i*

Model 2: The Star Prediction model

When using linear model 2 to detect which artists will become stars, the following distribution of star scores is retrieved:



Histogram 9: distribution of star scores among stars

As shown by histogram 9, the distribution approaches a normal distribution, but is skewed slightly to the right. The majority of the stars have a 'star score' between 0.05 and 0.2, so only artists in the test set with a star score within this interval are predicted to become stars. Out of the 32,829 artists in the dataset, 26,696 are indicated to be stars, while 300 artists are actually a star (cf. Table 7). This leads to the performance values in Table 7.

	Star prediction				***Performance metrics***	
	A: Correct predictions	**B: Correct star predictions**	**C: Number of users indicated as 'star'**	**D: Total predictions**	**Accuracy (=A/D)**	**Precision (=B/C)**
Model 2	6355	261	26696	32829	19.36%	0.98%
Random predictor	16327	158	16518	32829	49.73%	0.96%
Trendsetter profile	94	94	1925	3187	2.95%	4.88%
Random predictor	54	54	1578	3187	1.69%	3.42%

Table 7: Performance evaluation model 2 and the trendsetter profile

By means of model 2, stars and non-stars could be identified with an accuracy of approximately 19.36% (cf. Table 7). This accuracy is significantly lower than the accuracy which random binary generation would provide: an accuracy of 49.73%. However, the precision of the star prediction model (0.98%) exceeds the precision of the random predictor (0.96%) slightly. When using the trendsetter profile to analyse the listener base of a group of random artists, an accuracy of 2.95% and a precision of 1.69% was achieved, compared to an accuracy of 1.69% and a precision of 3.42% by the random predictor.

c. Main findings

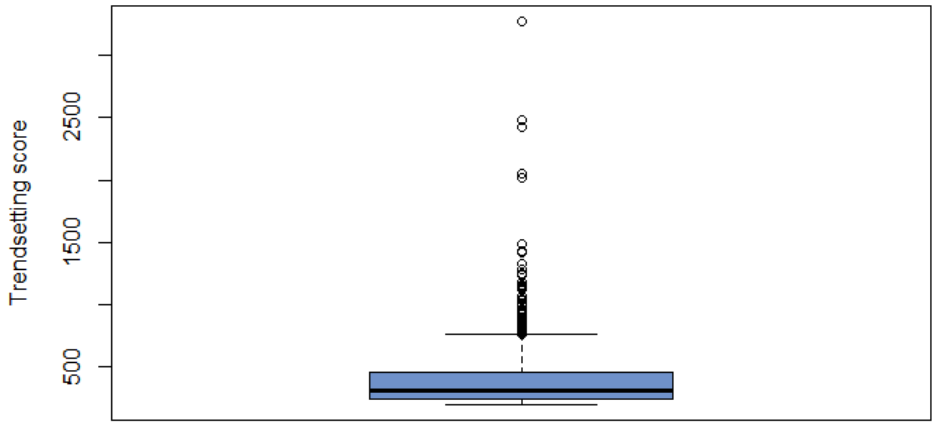
As indicated by the trendsetter detection model, the indicators of whether a user is a trendsetter or not, from most to least important, are: 'average similarity score' *S*, 'number of different genres listened to' *G*, 'age' *A* and 'number of friends on Last.fm' *F*. The effects of *G*, *A* and *F* on *Y* were positive, but *S* had a negative effect on *Y*. On the one hand, this relationship is intuitive, since users who currently have a preference for mainstream music are more likely to listen to artists who produce music similar to mainstream music, but have not broken through yet. In addition, users who are more similar have a greater influence over each other (Schedl et al., 2015). On the other hand, this finding is surprising, since listening to more mainstream music decreases a user's chances of listening to 'revolutionary' artists which currently distinguish themselves from other artists, and will break through in the future. No evidence was found that *M* has any predictive value for *Y*. *F* did have an effect on *Y*, but it was clearly the weakest indicator. The slight positive effect of *F* on *Y* might be attributed to proximity influence: the favouriting behaviour of immediate social network friends (Dewan et al., 2017). A possible explanation for the observed effect is that users with more friends are favourited by more other users, and thus have a larger social influence on the music preferences of others. Since having a friend who has liked a song increases the likelihood of listening to that song by 12%, users with many friends are more likely to influence more users and be trendsetters. As expected, *A* is also positively correlated with *Y*, which could be because young listeners stream more diverse music than old users, thereby detecting more stars (Anderson et al., 2020). Finally, the underlying theory behind the positive effect of *G* on *Y* might be that individuals with a higher *G* have a higher intrinsic need to seek stimulation through unfamiliar genres and artists, and are more likely to discover future stars in this process (Tang et al., 2017).

By using the trendsetter detection model, including the four variables above, about one out of five users was correctly classified as trendsetter or non-trendsetter. However, the random predictor was exactly twice as accurate. Still, the trendsetter detection model was twice as precise as the random predictor. This means that among the users detected as trendsetters by the detection model, relatively more users actually were trendsetters. In other words, the model performs poorly when classifying trendsetters and non-trendsetters, but if a user is identified as trendsetter, he has a relatively high chance of actually being one. Furthermore, with the trendsetter profile, it was possible to identify a group of trendsetters with almost twice the trendsetting score of the average Last.fm user.

By means of the star prediction model, about one out of five artists was correctly identified as star or non-star. However, the random predictor classified about half of the artists correctly. Clearly, the star prediction model is outperformed by the random predictor in terms of accuracy. Nevertheless, the star prediction model is slightly more precise, since a few more artists indicated by the model as 'stars', actually were stars. The reason for the low accuracy, but relatively high precision of the model is that many stars are incorrectly classified as a 'non-star', while few non-stars are incorrectly classified as a 'future star'. This means that the star prediction model is not useful for a classification between stars and non-stars. However, relatively many artists who are classified as a 'future star', are actually stars. Furthermore, the trendsetting profile was used to detect the artists with the most promising listener base. Selecting those artists whose listener base characteristics fitted to the trendsetter profile, provided both more accurate, but less precise predictions than a random predictor did.

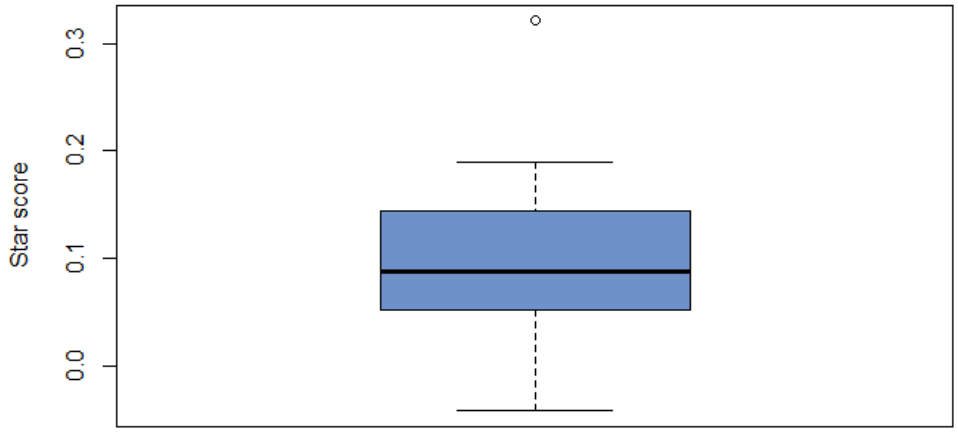
d. Outlier analysis trendsetters and stars (post-analysis)

An outlier analysis is conducted to identify outlier values among the trendsetter and star scores, and determine whether any outliers should be removed from the results.



Box plot 9: Distribution trendsetting scores among trendsetters

An outlier analysis of the trendsetting score of trendsetters reveals that there are about 47 outliers with a trendsetting score above 800. Three users (‘oroboras’, ‘TimaSliwinski’ and ‘GP0119’) have an exceptionally high trendsetting score of 3276, 2483 and 2428, resp. A closer investigation reveals that these users have listened to 1200, 16,848 and 2018 different artists, respectively (Last.fm, 2021c). In contrast, the average user has only listened to 41 different artists every week (Iqbal, 2021). This difference in listening activity and artist discovery might explain the relatively high trendsetter scores of the three outliers. Their trendsetting scores are not removed from the dataset.



Box plot 10: Distribution star scores among stars

The outlier analysis of the star scores of stars highlights one artist named ‘Active Star’, with an exceptionally high star score of 0.32 (cf. box plot 19). A detailed analysis reveals that this

artist has a relatively young listener base with many friends, genres listened to and high similarity scores (Last.fm, 2021c). Thus, the relatively high star score of this artist is attributed to the characteristics of his listener base, and he is not removed from the dataset.

5) Discussion

a. Conclusions

i. Research conclusions

This paper lays the ground for a new perspective on trendsetter detection. It shows the identification of a trendsetter group, how they distinguish themselves from other users and that they are able to detect future stars better than other users, as discussed in detail below.

This study started with retrieving a list stars, and a list of users who discovered relatively many future stars. Thereby, an answer is provided to the first research question, *´Who are the most influential trendsetters in terms of future star detection?"*. As expected, being curious, young and having a strong social network increases a user's chances of being a trendsetter. However, being more similar to others and thus having less original musical preferences *decreases* one's chances of being a trendsetter. These effects can all be (partially) explained by social influence theories. With the observed effects, the second research question, *"What characteristics distinguish trendsetters from non-trendsetters?"*, has been answered.

Using these findings, a unique approach could be taken towards the detection of future stars: It shows that characteristics of an artist's listener base can be used to predict whether that artist will break through. Although stars could not successfully be distinguished from non-stars with the star prediction model, a group of future stars could be identified with a relatively high probability of breaking through, although the model is only slightly more precise than a random predictor. These performance results provide an answer to the third research question: *"How well can the most influential trendsetters detect future stars?"*

By identifying, characterizing and evaluating the performance of trendsetters, evidence is provided that "O*n streaming platforms, trendsetters can be identified, who listen to artists which become successful in the future"*, as the problem statement states.

ii. Strategic conclusions

The findings in this paper suggest that a rising artist is more likely to break through if his listeners have the characteristics of a typical trendsetter. This means that artists should build up a listener base of teenagers and users in their twenties, who have at least 44 friends, listened to more than 20 genres and have a similarity score of at least 0.9123 on Last.fm. This is expected to increase the likeliness of becoming a star.

To determine whether their listener base currently fits to this profile, artists could use the star prediction model to calculate their 'star score'. This score provides them an indication of whether they will become a star, given their current listener base. Furthermore, artists should determine to what extent their listener base belongs to the trendsetter profile. Artist with a 'star score' outside the ideal interval and/or a listener base which does not fit to the profile, should attract more typical trendsetters. This can be achieved by attracting even younger artists, with more friends, listening to more genres, who have an even higher similarity score. Artists with an ideal star score and a listening base which belongs to the trendsetter profile are advised to target even more typical trendsetters, to increase the chance of breaking through.

b. Limitations

This study has five main limitations. The first limitation is that it is based on two very important, but strong assumptions. First of all, it is expected that listeners who have proven to detect artists before they broke through in the past, will continue to do so in the future. Secondly, it is assumed that the characteristics of an artist's listener base can be used to predict whether that artist will break through. Evidence is provided from the existing literature that the four characteristics in the proposed models play a role in social influence and the willingness to try new music. However, it has not been academically proven that the characteristics increase a user's capabilities to detect future stars.

The second limitation is the inconsistency within the linear models in terms of their unit of measurements. A, G and F are positive integers, while S is continuous and can be any number between 0 and 1. Having different units of measurement in one model generally leads to less precise coefficients (Sorzano, 2014). This limitation could have been addressed by standardizing the variables: simply subtracting the mean and dividing by the standard

deviation, for every observation. However, these transformations have not been implemented, since they would lead to a loss of the already scarce predictive information provided by the user characteristics. The drawback of keeping the variables in their original state is that a linear model with less precise coefficients was build. Therefore, the model was complemented by the trendsetter profile to identify trendsetters.

Thirdly, throughout the process of building a trendsetter profile, numerous arbitrary choices have been made regarding cut-off points. For instance, the users in the train set with a trendsetting score above 200 are assigned to the trendsetter group and the early phase is chosen to be almost twice as long as the late phase. Similarly, a few arbitrary choices have been made in the process of detecting stars. Most importantly, artists whose number of streams increased with less than 50% are not considered stars, and have been filtered out. Most of the cut-off points above were based on distribution analyses, including histograms and box plots. However, they are still influenced by the author's intuition, while they have a large effect on the research outcomes of this paper.

The fourth limitation concerns data availability issues. Due to the small number of artists in the dataset, only 307 stars could be identified. Consequently, the trendsetter group had to be identified by means of only a few hundred artists. Furthermore, the test group used to detect trendsetters is very small: There are only 106 listeners of which the values for *A, F, S* and *G* are known. Due to this scarcity, only 7 trendsetters could be detected in the test group. Still, the entire trendsetter profile was based on these 7 trendsetters. Additionally, the star prediction model was run on only 32,829 artists. This group contained only 300 stars, on which the star prediction model was based completely. This sample might be too small to be able to infer results about it to the whole Last.fm listener base, not to mention the entire music streaming market. Furthermore, the trendsetter and star prediction models might very well be overfitted to the Last.fm data.

One last limitation is that the models are trained and tested on one specific period: 2005-2015, divided in the early phase (2005-2011) and late phase (2012-2015). Therefore, the independent variables in the models are likely to be time-bound. A potential danger of this limitation is that the models are only useful for detecting future stars during the chosen period. After 2015, different independent variables might have distinguished trendsetters from non-trendsetters. To address this limitation, the period 2005-2015 could have been divided in

three rather than two phases: an early phase (2005-2008), middle phase (2009-2012) and late phase (2013-2015). The model could have been trained on the transition from the early to the middle phase, and tested on the transition from the middle to the late phase. In other words, stars in the train set would be defined as 'Artists moving from unsuccessful before 2009 to successful after 2008'. In the test set, stars would be considered 'Artists moving from unsuccessful before 2013 to successful after 2013'. Ideally, the models would perform equally well in the train and the test set. Consistent performance would imply that the results are not affected by time-bound factors: a typical trendsetter would have the same characteristics over time. However, this research approach would not have been feasible, since only 307 stars were identified in the test set, when dividing the data into two periods. If the data were divided in three periods, even less stars would be identified in the test set, which would lead to less dependable results.

c. Recommendations

Artists (and their managers) are advised to use the star prediction model to classify themselves as either a future 'star' or 'non-star'. Artists who are already defined as a future star are advised to maintain their current listener base or attract even more trendsetters. Artists who are not expected to become stars, are encouraged to analyse their listener base thoroughly. Such an analysis requires access to in-depth information about the characteristics of their listener base. One way to get achieve this information is via statistics provided by the streaming platforms. Numerous data trackers have already been developed on Last.fm, by which listeners can trace statistics such as their favourite genres and users with the most similar preferences (Last.fm, 2021b). One well-adopted tracking application is Last.fm's 'Desktop Scrobbler'. One advice to artists is to aggregate the data tracked by a random group of listeners and filter out those listeners who listened to them in the past. This would shape an image of their overall listener base. On Spotify, artists do not even have to collect all these data themselves, since the platform provides artists information about the age, gender, nationalities of the artist's listener base (Spotify for Artists, 2021). The page 'Listeners also Like' on the website 'Spotify for artists' also provides an indication of how similar the music taste of an artist's fans is to other users.

After doing a thorough analysis, artists should have a good overview of their listeners' age, social network strength, openness to novelty and preference originality. A comparison of

those characteristics to the characteristics of a typical trendsetter informs an artist which of the characteristics of his listener base do not fit in the trendsetter profile. This allows them to target and attract more listeners who do have these characteristics, which increases their chances of becoming a star. For instance, an artist might discover that his listeners are on average 24.3 years old, have 61 friends and a similarity score of 0.94, but have listened to only 14 different genres. This means that the listeners are not curious enough to fit in the trendsetter profile. Targeting listeners who try out more different genres would increase the artist's chances of breaking through.

Similarly, record labels are recommended to collect and analyse publicly available data about the listener base of talented, rising artists. The proposed star prediction model and the trendsetter profile, provide them an indication of which artists are likely to break through and which are not. This will lead to a more effective artist targeting strategy, time and cost reductions and a higher chance of signing a future star.

d. Future research

There are several ways in which the models in this paper can be improved and generalized in future research. Firstly, since the models in this study are focused on a very limited selection of user characteristics, additional user characteristics should be considered. In this paper, it is assumed that a user's Age, Similarity score, 'Number of friends on Last.fm' and 'Number of different genres listened to' determine whether that user is a trendsetter. As mentioned, the coefficients between these variables and the dependent variable are all insignificant. Also, the number of user characteristics included in the proposed trendsetter profile is very limited. Therefore, a complete set of strong indicators of being a trendsetter should be found.

Secondly, further research is needed to test whether the proposed models are actually robust across streaming platforms and over time. Currently, the proposed star prediction and trendsetter detection models are likely to be overfitted to the Last.fm data. To that end, they should be used to detect stars on different streaming platforms and between multiple periods. Consequently, the models can be trained on more data, and the predicted outcomes can be compared to more actual outcomes. Using this feedback, the star prediction model can gradually be finetuned. Ideally, future research will propose a model which is applicable to

any streaming platform, in any country, which can be used by any artist, manager and record label.

Finally, the steps taken to build, train, test and evaluate the proposed models, allow future researchers to build similar models for the artwork, movie, book and podcast markets and multiple other markets. In this study, the star prediction and trendsetter detection models are only applied to the music streaming market. However, when applying them to another market such as the book market, researchers could analyse reader characteristics to build a trendsetter profile of readers, and then analyse the characteristics of an author's audience to predict which authors will break through. As long as these steps are taken carefully and sufficient data is available, the possibilities are endless.

Appendix A

A link to the GitHub repository including all the code used for this thesis is provided here: https://github.com/TimAmade/Master-Thesis-Project.git. The code repository is structured as follows:

Data preparation
- a. Loading packages
- b. Loading datasets
- c. Data cleaning
- d. Outlier analysis (preliminary)

Section 1. Identifying stars
- a. Importing listening dataset
- b. Splitting list of streams into early and late phases
- c. Identifying (un)successful artists
- d. Identifying stars

Section 2. Identifying trendsetters in the train set
- a. Reloading listening dataset
- b. Selecting early phase streams
- c. Creating list of artist-user combinations
- d. Identifying potential trendsetters
- e. Splitting train and test sets
- f. Outlier analysis of trendsetting scores

Section 3. Analysing user characteristics in the train set
- a. Adding user characteristics
- b. Trendsetter group analysis

Section 4. Trendsetter detection model
- a. Adding independent and dependent variables
- b. Creating the linear model (model 1a)
- c. Creating refined linear model (model 1b)
- d. Robustness check

Section 5. Identifying trendsetters by means of the trendsetter detection model
- a. Creating list of star streams
- b. Identifying trendsetters
- c. Accuracy and precision evaluation

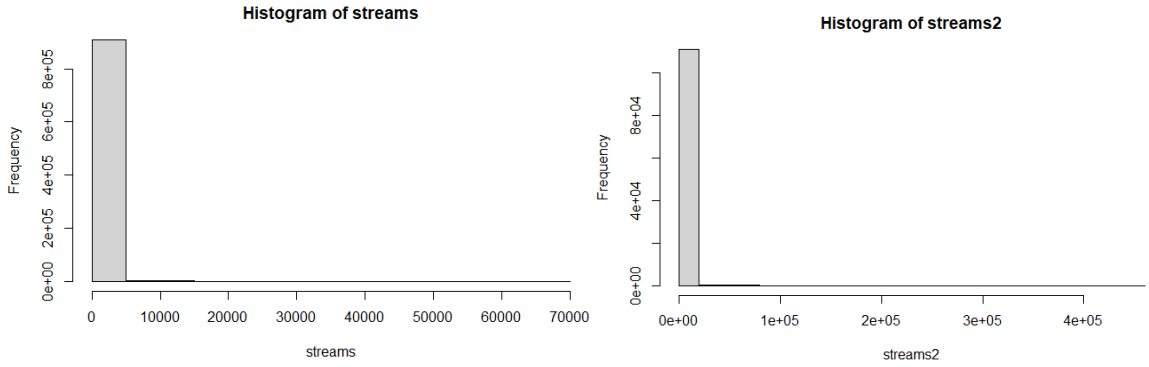
Section 6. Identifying trendsetters by means of the trendsetter profile
- a. Adding user characteristics
- b. Selecting users based on the trendsetter profile
- c. Evaluating trendsetting scores
- d. Accuracy and precision evaluation

Section 7. Detecting stars by means of the star prediction model
- a. Creating list of users with known A, F, S and G
- b. Creating list of artists with listener base characteristics

c. Adding star prediction variable
d. Accuracy and precision evaluation

Section 8. Detecting stars based on whether their users belong to the trendsetter profile
a. Selecting stars based on trendsetter profile
b. Accuracy and precision evaluation
c. Outlier analysis of stars

Appendix B



Histograms 10 and 11: distribution of streams in (1) the early phase and (2) the late phase

Appendix C

Variable	β (σ)
AGE	-0.004 (0.005)
MEMBER	-
Log(Number_of_friends**)**	0.012 (0.027)
Average_similarity_score	0.024 (0.312)
Number_of_different_genres_listened_to	0.006 (0.004)
Constant	-0.098 (0.418)
Observations	76
R2	0.040

Adjusted R2	-0.014
Residual Std. Error	0.273 (df = 71)
F Statistic	0.734 (df = 4; 71)

Table 8: Coefficients model 1a with log-transformed

Appendix D

Variables	Original	White heteroskedasticity
Constant	-0.073	-0.073
	(-0.408)	(-0.151)
AGE	-0.004	(-0.004
	(-0.005)	(-0.003)
Number_of_friends	0.001	(0.001
	(-0.001)	(-0.001)
Average_similarity_score	0.007	(0.007
	(-0.312)	(-0.121)
Number_of_different_genres_listened_to	0.006	0.006**
	(-0.004)	(-0.003)
Observations	76	76
R2	0.046	0.046
Adjusted R2	-0.008	-0.008
Residual Std. Error (df = 71)	0.273	0.273
F Statistic (df = 4; 71)	0.848	0.848

Table 9: Model 1b: White heteroskedasticity test for robustness

References

Aguiar, L., & Waldfogel, J. (2018). Quality Predictability and the Welfare Benefits from New Products: Evidence from the Digitization of Recorded Music. *Journal Of Political Economy*, *126*(2), 492-524. doi: 10.1086/696229

Anderson, A., Maystre, L., Anderson, I., Mehrotra, R., & Lalmas., M. (2020). Algorithmic Effects on the Diversity of Consumption on Spotify. *Association for Computing Machinery*, New York, USA, pp. 2155–2165. doi: https://doi.org/10.1145/3366423.3380281

Anshel, M., & Lidor, R. (2012). Talent detection programs in sport: the questionable use of psychological measures. *The Free Library*

Bender, M., Gal-Or, E., & Geylani, T. (2021). Attracting artists to music streaming platforms. *European Journal Of Operational Research*, *290*(3), 1083-1097. doi: 10.1016/j.ejor.2020.08.049

Berg, B. (2018). Planning A Marketing Strategy For Artists. *Agora Gallery*

Datta, H., Knox, G., & Bronnenberg, B. (2017). Changing Their Tune: How Consumers' Adoption of Online Streaming Affects Music Consumption and Discovery. *SSRN Electronic Journal*

Dewan, S., Ho, Y., & Ramaprasad, J. (2017). Popularity or Proximity: Characterizing the Nature of Social Influence in an Online Music Community. *Information Systems Research*, *28*(1), 117-136. doi: 10.1287/isre.2016.0654

DJMag. (2020). Spotify reaches $50 billion market valuation. Retrieved 25 March 2021, from: https://djmag.com/news/spotify-reaches-50-billion-market-valuation

Fly, B. (2021). How Does Music Consumption Impact the Music Industry and Benefit Artists? Retrieved 25 March 2021, from https://scholarworks.uark.edu/acctuht/20/

Hann, M., (2018). *Music's 'Moneyball' moment: why data is the new talent scout*. Financial Times. Retrieved 10 April 2021, from: https://www.ft.com/content/474ae18a-7f1b-11e8-bc55-50daf11b720d

Haroutounian, J. (2000). MusicLink: Nurturing Talent and Recognizing Achievement. *Arts Education Policy Review*, 101(6), pp. 12-20

IFPI. (2018). IFPI Global Music Report (2018). *IFPI.* Retrieved 30 January 2021, from https://www.ifpi.org/news/IF__PI-GLOBAL-MUSIC-REPORT-2018

Iqbal, M. (2021). Spotify Revenue and Usage Statistics. *Business of Apps*. Retrieved 17 May 2021, from https://www.businessofapps.com/data/spotify-statistics/

Jacobson, K., Murali, V., Newett, E., Whitman, B., & Yon, R. (2016). Music Personalization at Spotify. *Association for Computing Machinery*, New York, USA

Last.fm (2021a). About Last.fm. Retrieved 10 May 2021, from https://www.last.fm/about

Last.fm (2021b). Track My Music. Retrieved 10 May 2021, from https://www.last.fm/about/trackmymusic

Last.fm (2021c). User database. Retrieved 10 June 2021, from https://box.last.fm/user/username

Li, A., Wang, A., Nazari, Z., Chandar, P. and Carterette, B. (2020). Do podcasts and music compete with one another. *Association for Computing Machinery*, New York, USA, 1920–1931. doi: https://doi.org/10.1145/3366423.3380260

Pfeiffer, T. (2021). Twitch and TikTok help breakthrough artists to bypass big music. *The Seattle Times.* Retrieved 8 April 2021, from: https://www.seattletimes.com/business/technology/ breakthrough-artists-tap-tech-to-bypass-big-music-find-listeners-on-twitch-spotify-youtube/

Saez-Trumper, D., Comarela, G., Almeida, V., Baeza-Yates, R., and Benevenuto. F. (2012). Finding trendsetters in information networks. *Association for Computing Machinery*, New York, USA, 1014–1022. doi: https://doi.org/10.1145/2339530.2339691

Salganik, M. (2006). Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science, 311*(5762), 854-856. doi: 10.1126/science.1121066

Schedl M. & Hauger, D. (2015). Tailoring music recommendations to users by considering diversity, mainstreaminess, and Novelty[C]. *International ACM SIGIR conference on research and development in information retrieval*. pp 947–950

Schedl, M., Knees, P., & Widmer, G. (2005). A web-based approach to assessing artist similarity using co-occurrences. In Proceedings of the 4th International Workshop on Content-Based Multimedia Indexing. *CBMI*

Sorzano, C. (2014). A survey of dimensionality reduction techniques. *Center for Applied Scientific Computing, Lawrence Livermore National Laboratory*. p. 2

Spotify for Artists. (2021). Listeners also like. Retrieved 10 May 2021, from https://artists.spotify.com/help/article/listeners-also-like?category=audience-stats

Stafford, S. (2010). Music in the digital age: the emergence of digital music and its repercussions on the music industry. *Elon J. Undergrad. Res. Commun.* 1(2) pp. 112-120

Statista. (2021). Music streaming market share. *Statista.* Retrieved 30 January 2021, from https://www.statista.com/statistics/653926/music-streaming-service-subscriber-share/

Szymanowski, M., Deng, W., & Chen, X. (2021A). Innovativeness and Social Contagion in Music Discovery.

Szymanowski, M., Deng, W., & Chen, X. (2021B). The impact of social influence attributed discovery on music consumption.

Tang, M. & Yang, M. (2017). Evaluating Music Discovery Tools on Spotify: The Role of User Preference Characteristics. Journal of Library and Information Studies. 15(1) pp. 1-16

Visual Capitalist (2021). Which Streaming Service Has the Most Subscriptions?. Retrieved 25 March 2021, from https://www.visualcapitalist.com/which-streaming-service-has-the-most-subscriptions/

Zhang, B., Kreitz, G., Isaksson, M., Ubillos, J., Urdaneta, G. Pouwelse, J., & Epema, D. (2013). Understanding user behavior in Spotify. *Proceedings IEEE INFOCOM,*. pp. 220-224