Decision Support Systems 145 (2021) 113535

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Music intelligence: Granular data and prediction of top ten hit songs☆

Seon Tae Kim, Joo Hee Oh *

Department of Economics, Management School, University of Liverpool, UK

ARTICLE INFO

Keywords:
Business intelligence
Granular data
Digital streaming
Online platform
Spotify

ABSTRACT

In the music market, superstars significantly dominate the market share, while predicting the top hit songs is notoriously difficult. The music intelligence technology, retrieving and utilizing granular acoustic features of songs, provides opportunities to improve the prediction of top hit songs. Using data on 6209 unique songs that appeared in the weekly Billboard Hot 100 charts from 1998 to 2016, especially acoustic features provided by *Spotify*, we investigate empirically how the top-10-hit-songs likelihood prediction is improved by acoustic features. We find that some acoustic features (e.g., danceability, happiness, and some metrics of timbre and pitch) significantly improve the model's ability to predict the top-10-hit-songs probability. These results suggest that the granular data, provided by the music intelligence technology, carries a substantial predictive value in the era of online music streaming.

1. Introduction

In cultural product markets (e.g., music, art, books, etc.), superstars significantly dominate the market share [19]. Thus, identifying the success of blockbusters in cultural markets is important. But, it has been a great challenge. William Goldman, the well-known screen writer of *Princess Bride*, famously said of Hollywood, "Nobody knows anything" when predicting success under uncertainty [13]. Carrying out the artificial music market experiment (called as *Music Lab*), Salganik et al. [20] also conclude that predicting top hit songs in the music market is quite difficult. The recent development of new technologies (e.g., artificial intelligence (AI), machine learning (ML), and big data) could, however, help to overcome the difficulty of prediction in cultural markets.

In this paper, we examine how the prediction of top hit songs can be improved by the aforementioned new technologies increasingly used in the music market [1], whereas inequality in the music market has increased [5]: over time, the smaller number of unique songs reach the top ten position[1] and hold that position for a longer period (authors' calculation). More specifically, we investigate how much the prediction of the top-ten-hit-songs probability can be improved by the granular acoustic data, provided by music intelligence technologies of *Spotify* (a leading platform of digital music streaming service). *Spotify* recently acquired *The Echo Nest*, which uses "music intelligence" technologies to retrieve and analyse granular acoustic features of a song. *Warner Music Group* also acquired a tech company *Sodatone* that uses algorithms to review social, streaming and touring data to find promising songs (https://www.wmg.com/news/warner-music-group-acquires-sodatone-33396, 2018). We aim to deepen understanding of whether such new technologies can improve the prediction of top hit songs.

Using data on songs that appeared in the weekly Billboard Hot 100 charts from 1998 to 2016, we investigate determinants of the top-ten-hit-songs likelihood, including the song's genre (e.g., classic, dance, etc.) and number of the artist's previous songs (proxy of artist's fame). We estimate the logit regression model[2] over the training period prior to 2016 and test the model's predictability for the out-of-time songs in 2016 (weekly) charts. In particular, we conduct [24] test to provide strong statistical evidence about whether, and how much, the predictive ability can be improved by acoustic features [10].

We find that some acoustic features (e.g., danceability, happiness, and some metrics of timbre and pitch), alongside genre and number of artist's previous songs, are significant determinants of the top-ten-hit-songs likelihood. Importantly, these acoustic features also improve substantially the top-ten-hit-songs predictive ability. More specifically, the top-ten-hit-songs logistic regression model, estimated over the 3-

☆ We thank Sander Ouwejan for sharing data with us.

* Corresponding author at: IMC 348 WMG, University of Warwick, Coventry CV4 7AL, UK.
E-mail addresses: Seon.Kim@liverpool.ac.uk (S.T. Kim), joo.oh@warwick.ac.uk (J.H. Oh).

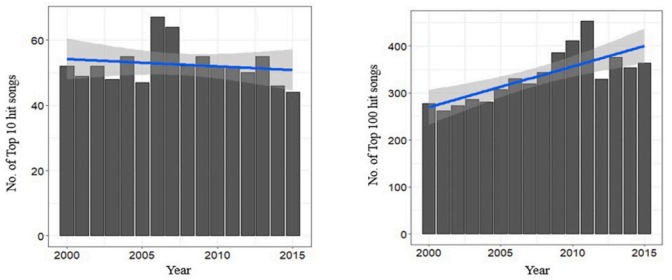
[1] According to Mulligan [17], in the music market, 1% of musical works account for 77% of all revenues.

[2] We also consider other training/estimation algorithms: random forest, deep learning, and gradient boosted machine. Main qualitative results are robust to these alternative algorithms.

https://doi.org/10.1016/j.dss.2021.113535
Received 28 July 2020; Received in revised form 18 February 2021; Accepted 18 February 2021
Available online 22 February 2021

Table 1
Key variables of billboard hot 100 dataset.

Variable	Description
Date	The date of the chart
Artist	The name of the artist of the song
Title	The title of the song
Weeks	The number of weeks the song has been in the Billboard Hot 100
Current	The current position/rank of the song in the Billboard Hot 100
SpotifyID	The song's unique identifier in the Spotify system



(a) Top Ten Hit Songs (b) Top One Hundred Songs

Fig. 1. Annual number of unique songs: billboard hot 100 chart, 2000—2015.
Note: this figure provides the time series of the annual number of unique songs ever entered the weekly list of Billboard top-ten and top-one hundred hit songs, respectively. The estimated trend (solid line) and 95% confidence interval are also provided.

year training period of 2013—2015, is used to predict top-ten-hit successes of out-of-time songs in 2016 weekly charts. Results of McNemar's test provide strong evidence that the predictive ability increases substantially when main (e.g., danceability and happiness) and auxiliary (some metrics of timbre and pitch) acoustic features are included into predictors. These findings are robust to (i) different training/prediction algorithms (e.g., gradient boosting machine, random forest, deep learning), (ii) different test/training periods (e.g., test songs in 2013, 2014, 2015 (weekly) charts, respectively, with the corresponding 3-year rolling-window training periods), and (iii) different criteria to select a threshold value above which the continuous predicted "success" probability is classified into the top-ten-hit-song category.

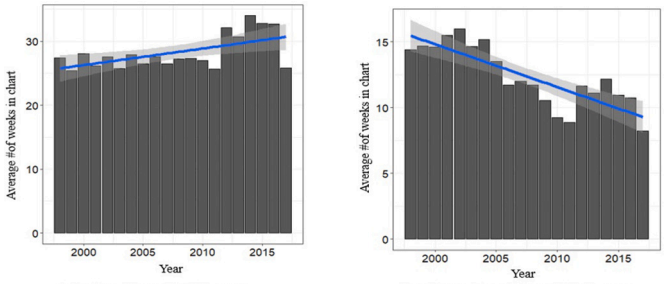
Our findings suggest that acoustic features of a song are closely related to music consumers' choices, driven by their motives to change mood and emotion, consistent with previous findings in neuroimaging science that listening to certain musical songs releases dopamine, a neurotransmitter responsible for tangible pleasures [16,21]. In short, the granular data, generated by music intelligence technologies, carries a substantial predictive value in the era of online music streaming and can help the music industry's practitioners to make better decisions.

Researchers have discussed many factors to explain top chart dynamics of songs: e,g., previous popularity and demographics of artists, brand or label information, online word-of-mouth, marketing and promotion around the song [4,8,9,12]. Askin and Mauskapf [2,3] and Interiano et al. [14] investigate roles of some acoustic features in determining the song's popularity. This paper complements previous findings, by measuring the *incremental effect* of *detailed* acoustic features on the top-ten-hit-songs *predictive ability*.

2. Data and methodology

2.1. Data

Our sample consists of two data sources: (i) *Billboard* dataset, and (ii) *Spotify* dataset, where the former provides the dynamics of an individual song's rankings over time, and the latter the song-level acoustic characteristics.



(a) Top Ten Hit Songs (b) Non-Top Ten Hit Songs

Fig. 2. Average duration: billboard hot 100 chart, 1998—2017.
Note: this figure provides time series of the average duration (number of weeks for a song to remain in the weekly list for a given year) of Billboard top-ten and non-top-ten hit songs, respectively. The solid line refers to the estimated trend.

Table 2
Main acoustic features, via *Spotify* API.

Variable	Description
Acousticness	A confidence measure (scale from 0.0 to 1.0) of whether the track is acoustic or not: 1.0 represents the highest (0.0 the lowest) level of confidence that the track is acoustic.
Danceability	Degree (scale from zero to one) to which a track is suitable for dancing, based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
Valence	A measure (scale from 0.0 to 1.0) describing the musical positiveness conveyed by a track. Valence of 1.0 means that the track sounds very positive (e.g. happy, cheerful), while valence of 0.0 means that the track sounds very negative (e.g. sad, depressed).
Duration_ms	The duration of the track in milliseconds.
Energy	A measure (scale from 0.0 to 1.0) of intensity/activity. Energetic tracks often feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude has low energy.
Instrumentalness	A confidence measure (scale from 0.0 to 1.0) of whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context (value close to 1.0). Rap or spoken word tracks are clearly "vocal" (value close to 0.0). Values above 0.5 are intended to represent instrumental tracks.
Liveness	A confidence measure (scale from 0.0 to 1.0) of whether an audience is present in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
Loudness	The overall loudness of a track in decibels (dB). Values typical range between −60 and 0.
Mode	Dummy indicating the track's modality (major or minor), from which the melodic content is derived. Value of one is assigned for major, and zero for minor.
Speechiness	Degree (scale from 0.0 to 1.0) of the presence of spoken words in a track. The closer to 1.0 the value, the more exclusively speech-like the track (e.g. talk show, audio book, poetry). Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
Tempo	Beats per minute (BPM), representing the speed or pace of a track. The larger the value, the higher the speed of a track. Values typical range between 50 and 200 (BPM).
Time_signature	A measure to specify how many beats are in each bar (where a bar is a segment of time corresponding to a specific number of beats). The time signature ranges from 3 to 7 indicating time signatures of "3/4", to "7/4".

Note: this table provides the song's main acoustic features provided by the *Spotify* Web API.

2.1.1. Billboard data

The weekly Billboard Hot 100 chart is one of the most popular charts and provides 100 most popular songs for a given week based on *Nielsen Music* data (e.g., radio airplay, sales, streaming activity, etc.). We

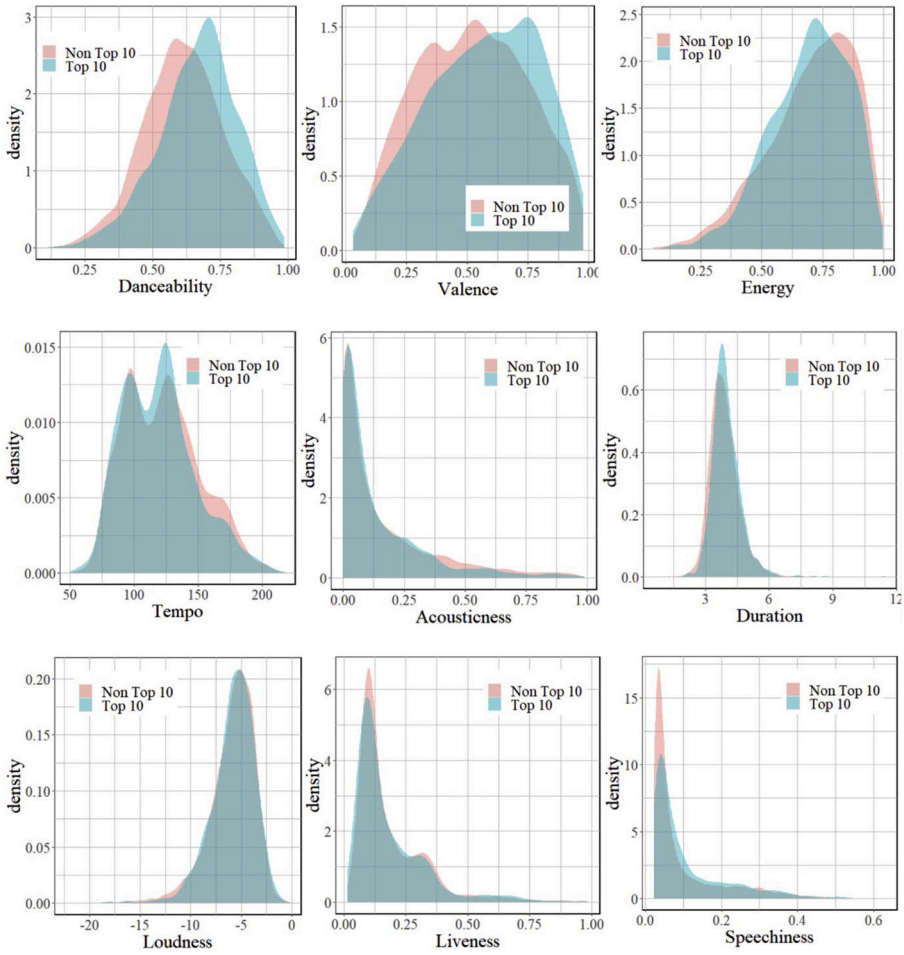


Fig. 3. Density of main acoustic features: top-Ten vs. non-Top ten hit songs.
Note: this figure provides the density of nine main acoustic features: danceability, valence, and energy (from left to right) at the top panels; tempo, acousticness, and duration in second-row panels; and loudness, liveness, and speechiness at the bottom panels, respectively; all of them by the two groups: top ten vs. non-top ten songs, respectively.

recover the Spotify ID for each individual song that Billboard published [23]. This Billboard sample is a weekly unbalanced panel and has a total of 6209 unique songs (dropping observations with missing Spotify IDs) that appeared in the Billboard Hot 100 chart over the period from May 2nd 1998 to January 1st 2017. Table 1 lists key variables of Billboard sample.

We discuss how the composition of top ten hit songs has changed over time. First, as shown by Fig. 1, the annual number of unique top-ten-hit songs (i.e., songs ever entered the weekly top ten list for a given year) has slightly decreased over the years, while that of unique top-hundred songs has increased. This implies that reaching a top ten position is *increasingly* difficult over time. Second, Fig. 2 shows that the duration (i.e., number of weeks for a song to remain in the given list for a given year) increased slightly for top ten hit songs but decreased greatly for non-top ten songs. In short, inequality among songs has increased over time [5]: the smaller number of unique songs reach the top ten position and hold that position for a longer period.

2.1.2. Spotify data: main acoustic features

We collect data on audio characteristics of songs via the *Spotify* Web API. *Spotify* is a leading platform of digital music streaming services and provides data on the song's structure and musical content, including pitch and timbre of every segment (typically under a second) in the song. Table 2 lists the main acoustic features [22].

In Fig. 3, we illustrate the differences in some of the main acoustic features between top ten hit and non-top ten songs over the entire sample period of 1998—2016 (i.e., densities of a given acoustic variable conditional on top-ten vs. non-top ten hit songs). For instance, *danceability* measures how suitable a track is for dancing (different from the genre of dance), and *valence* the degree to which a track makes the listener feel *happy*, where these two acoustic features are determined by a combination of musical elements. Both danceability and valence exhibit visible differences between top ten hit songs and the others: for top ten hit songs, distributions of both danceability and valence are skewed more to the right than for non-top ten songs (see, Fig. 3). That is, top ten hit songs are more danceable and higher in valence (i.e., 'happiness') than non-top ten hit songs. We can also see that there are some (small extent) differences in energy and tempo between top ten hit and non-top ten songs: for top ten hit songs, both of these two features seem smaller than for non-top ten hit songs. Moreover, we can see that between top ten hit and non-top ten songs, differences in acousticness, duration, loudness, liveness, and speechiness are seemingly of small magnitude.

2.1.3. Spotify data: auxiliary acoustic features

We also consider auxiliary acoustic features (i.e., multidimensional aspects of timbre and pitch), listed in Panel A, Table 3. According to Jehan [15]: "[Timbre] is a complex notion also referred to as sound color, texture, or tone quality, and is derived from the shape of a segment's spectro-temporal surface, independently of pitch and loudness." There are 12 different classes of timbre and pitch, respectively, whereas each timbre (pitch) class is a vector having values over different segments of a song; therefore, given timbre (pitch) vector's various moments over segments (e.g., average and skewness) are used.

More specifically, timbre is the quality of a musical note or sound that distinguishes different types of musical instruments, or voices, and related to sound color, texture, or tone quality of a segment of a song [22]. According to the documentation provided by Spotify [22], 12

Table 3
Auxiliary acoustic features, via *Spotify* API.

Variable	Description
Panel A: Auxiliary acoustic features	
12 Timbre vectors	For each of 12 different timbre vectors, 7 moments (average, standard deviation, minimum, maximum, median, skewness, and kurtosis) over all segments of the given song are considered. 12 different timbres (scale of unbounded values, roughly centered around 0) are best interpreted in comparison with each other. For instance, given the segment of a song, the first timbre class represents the average loudness of the segment; the second the degree of brightness; the third the flatness of a sound.
12 Pitch vectors	For each of 12 different pitch vectors, 7 moments (average, standard deviation, minimum, maximum, median, skewness, and kurtosis) over all segments of the given song are considered. For a given segment of a song, 12 pitch vectors have values (scale from 0.0 to 1.0, respectively) that measure the degree of (relative) dominance of 12 pitch classes. Suppose that at the given segment of a song, pitch 1 (C) dominates the other pitches; at such a segment, pitch 1 vector has the value of one and the other pitch vectors have zero values.
Panel B: Timbre 1 vector: 7 moment variables	
Timbre_1_avg	The average value of timbre1 vector over all segments of the song.
Timbre_1_stdev	The standard deviation of timbre1 vector over all segments of the song.
Timbre_1_min	The minimum value of timbre1 vector over all segments of the song.
Timbre_1_max	The maximum value of timbre1 vector over all segments of the song.
Timbre_1_median	The median value of timbre1 vector over all segments of the song.
Timbre_1_skewness	The skewness of timbre1 vector over all segments of the song.
Timbre_1_kurtosis	The kurtosis of timbre1 vector over all segments of the song.
Panel C: Pitch 1 vector: 7 moment variables	
Pitch_1_avg	The average value of pitch 1 (C) over all segments of the song.
Pitch_1_stdev	The standard deviation of pitch 1 (C) over all segments of the song.
Pitch_1_min	The minimum value of pitch 1 (C) over all segments of the song.
Pitch_1_max	The maximum value of pitch 1 (C) over all segments of the song.
Pitch_1_median	The median value of pitch 1 (C) over all segments of the song.
Pitch_1_skewness	The skewness of pitch 1 (C) over all segments of the song.
Pitch_1_kurtosis	The kurtosis of pitch 1 (C) over all segments of the song.

Note: this table presents auxiliary acoustic features, 12 classes for timbre and pitch, respectively, in panel A, especially 7 moments of Timbre 1 (Pitch 1) vector over all segments of the given song in panel B (panel C).

different timbre classes are measured and have meanings as follows:

> "The timbre feature is measured as 12 different metrics, labelled *timbre classes*: e.g., timbre1, timbre2, …, timbre12, where each of 12 timbre classes can have unbounded values roughly centered around 0, for a given segment of a song. For instance, given the segment of a song, the first timbre class represents the average loudness of the segment; the second the degree of brightness; the third the flatness of a sound. 12 different timbres are best interpreted in comparison with each other."

Given that a song has multiple segments, each of these 12 timbres has different values over different segments of a given song (track). Thus, for a given song, each of 12 timbres is a vector (rather than a scalar). That is, consider timbre 1, which is a vector (i.e., *function*) having values over different segments of a given song. Thus, we can consider a number of moments (over different segments) of a timbre vector 1 of a given song: e.g., average, standard deviation, and so on (presented in Panel B, Table 3). Similarly, various moments of a given timbre vector are gathered for timbre vectors 2 through 12, too.

Pitch is another important aspect of auxiliary acoustic features of a song. Each segment of a song contains a value denoting the dominance of each of the 12 pitch classes (C, C#, D, etc.) in that segment. For a given segment of a song (track), a value close to 1.0 is assigned to the relatively dominant pitch classes, and a value close to 0.0 to the other pitch classes [22]. For instance, if the 'C' pitch class is strongly dominant for a given segment of a song, then 'C' pitch class has a value of 1.0 and the other pitch class have values of 0.0, for the given segment of a song. A pitch class (e.g., 'C' pitch) has different values (scale from 0.0 to 1.0) over different segments of a song. Thus, for a given song, each of 12 pitch classes is a vector (rather than a scalar). Therefore, we consider various moments of a given pitch class of a given song, as for the case of pitch 1 vector in Panel C, Table 3.

In Fig. 4, we illustrate the differences in some of the auxiliary acoustic features between top-ten and non-top ten hit songs over the entire sample period of 1998—2016 (i.e., densities of a given acoustic variable conditional on top-ten vs. non-top ten hit songs). For instance, for top-ten hit songs, Timbre_3_median (i.e., median of Timbre 3 over segments of a given song) tends to be smaller than for non-top ten hit songs, while the opposite pattern is observed for Pitch_8_stdev (i.e., standard deviation of Pitch 8 over segments of a given song). There are also differences, though seemingly difficult for human eyes to detect, between top-ten vs. non-top ten hit songs in several other auxiliary variables: Timbre_6_skewness, Timbre_10_kurtosis, Pitch_4_min, and Pitch_10_min. A systematic statistical analysis is needed to recognize how these auxiliary acoustic variables are related to top-ten-hit-songs probability.

2.2. Methodology: prediction models for top-ten hit songs

We develop a set of prediction models to compare top-ten-hit-songs prediction results, where we consider, as baseline case, generalized linear models (GLM, i.e., logistic regression model) of the top-ten-hit-songs dummy with independent variables: (1) the artist's previous number of songs (proxy of artist's fame), (2) genre of the song (seven categories), (3) main acoustic features (12 variables listed in Table 2), and (4) main and auxiliary acoustic features, where 30 auxiliary acoustic variables (listed in Table S2.3 in appendix) are used. More specifically, we consider the four GLM models as follows: Model1 includes independent variable(s) in (1) above; Model2 those in (1) through (2); Model3 those in (1) through (3); Model4 those in (1) through (4). Such models are trained to classify the training set of songs as either a hit song or not, and then their prediction results for the out-of-time test set of songs are assessed.

2.2.1. Prediction results and classification: threshold value

ur primary goal in this paper is to assess how much the top-hit-songs prediction results can be improved by the song-level granular data on acoustic features. For this purpose, we compare the prediction results of Model2 (without any acoustic features) to those of Model3 (with main acoustic features) and Model4 (with both main and auxiliary acoustic features), respectively. As for the metric to assess the model's predictive ability, if the predicted variable is continuous, one can use the mean squared prediction error (MSPE) for the out-of-sample prediction errors. In our case, the predicted variable of our interest is, however, categorical (i.e., top-ten vs. non-top ten hit song), though the model's predicted outcome (i.e., fitted value of the "success" probability) is continuous.

This raises an issue of selecting a threshold value θ above which the model's predicted "success" probability $p_{1,\,i}$ for observation (song) i is classified as a top-ten-hit song: song i is a top-ten-hit song if $p_{1,\,i} \geq \theta$, and not otherwise. Given the threshold value θ, there are four possible prediction/classification results for song i: (i) true positive (TP) for the case in which the model's prediction is "yes" for a top-ten-hit song and correct, (ii) false positive (FP) when the model's prediction is "yes" but wrong, (iii) true negative (TN) when the model's prediction is "no" and correct, and (iv) false negative (FN) when the model's prediction is "no" and wrong.

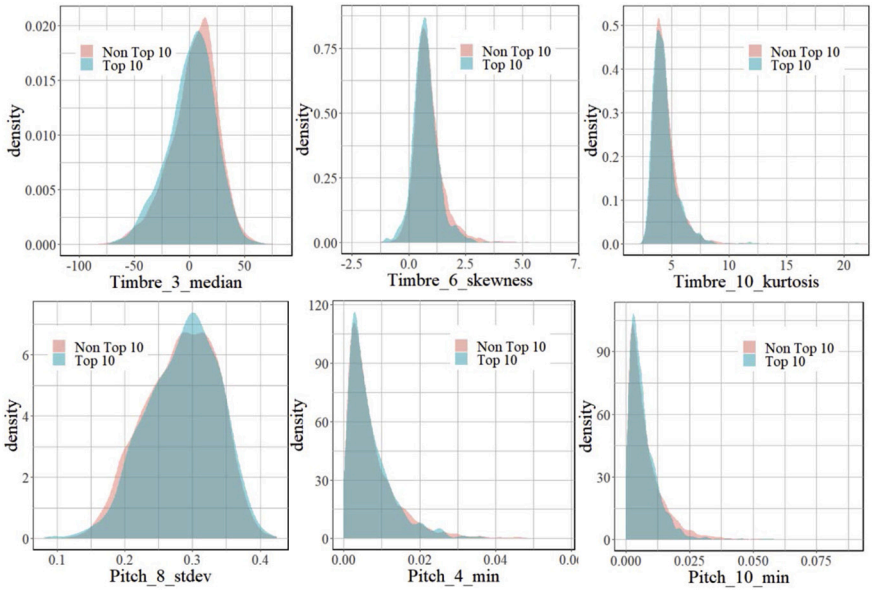


Fig. 4. Density of auxiliary acoustic features: top-ten vs. non-top ten hit songs.
Note: this figure provides the density of six auxiliary acoustic features: Timbre_3_median, Timbre_6_skewness, and Timbre_10_kurtosis (from left to right) at the top panels; Pitch_8_stdev, Pitch_4_min, and Pitch_10_min at the bottom panels, respectively; all of them by the two groups: top ten vs. non-top ten songs, respectively.

2.2.2. Assessment metrics of predictive ability

Note that there are many different metrics to assess such four possible prediction/classification results over the test set of songs. For instance, 'precision' refers to the number of true positive (TP) instances relative to that of true positive (TP) and false positive (FP) instances: *precision* = TP/(TP + FP); 'sensitivity' (also called *recall*) the number of true positive (TP) instances relative to that of true positive (TP) and false negative (FN): *sensitivity* = TP/(TP + FN); 'specificity' the number of true negative (TN) instances relative to that of true negative (TN) and false positive (FP): *specificity* = TN/(TN + FP). Here, precision is related to the type I error rate (due to FP instances, i.e., they are detected as positive but should not), and sensitivity is closely related to (one minus) the type II error rate (due to FN instances, i.e., they should have been detected as positive but not). For the case in which one is concerned about correctly predicting non-top-ten-hit songs, specificity is, similarly to sensitivity, also closely related to the type II error rate (due to FP instances, they should have been detected as negative but not).

2.2.3. Threshold-selection criteria: two combined metrics

Note that various metrics of prediction assessment are specific to the threshold value θ used in classification. Taking into consideration such dependence on the threshold value θ, we consider two rules to pick θ. More specifically, we consider two combinations of the aforementioned metrics (e.g., precision, sensitivity, and specificity) and choose the threshold value θ at which a given combination of the metrics is maximized. The first combination of metrics is written as:

$$\text{Combined metric 1} = \text{precision}^{*}\text{sensitivity}/([\text{precision} + \text{sensitivity}]/2) \quad (1)$$

and the second combined metric is written as:

$$\text{Combined metric 2} = \text{specificity}^{*}\text{sensitivity}/([\text{specificity} + \text{sensitivity}]/2) \quad (2)$$

Combined metric 1 (2) can be thought of as the squared geometric average of precision (specificity) and sensitivity, normalized by their average level. The combined metric 1 is statistic *F1*, default-setting metric used by H2O *R*-software packages, the one used in this paper, in selecting the threshold value and returning the predicted classification results. The combined metric 2 is an otherwise equivalent to the combined metric 1 except that specificity replaces precision. These two combined metrics will be used as criteria (e.g., maximize them) in selecting a particular threshold value for classification and hence considered in conducting McNemar's test, discussed soon.

Last, we also consider another metric that is not dependent on a particular value of θ. For this purpose, we consider AUC (Area under the ROC Curve (receiver operating characteristic curve)). More specifically, given a (wide) range of θ, the ROC curve plots the sensitivity against the false positive rate (FPR = FP/(FP + TN) = 1 – TN/(FP + TN)) that is equal to one minus specificity. As θ decreases, FPR tends to increase (because FP tends to increase), specificity tends to decrease, and sensitivity tends to increase, implying a trade-off between sensitivity and specificity; taking this trade-off into consideration, AUC measures the model's predictive ability in terms of both sensitivity and specificity, by considering a (wide) range of θ.

The better the model's predictive ability, the closer to one the model-implied AUC.[3] We will examine how much AUC of Model3 (with main acoustic features) and Model4 (with both main and auxiliary acoustic features), respectively, increases compared to that of Model2 (without any acoustic features), as an informal way to see whether acoustic features improve the predictive ability. One caveat is that we can not test whether such an increase in AUC is significantly different from zero. This calls for McNemar's test, as discussed below.

2.2.4. McNemar's test of predictive abilities of different models

The primary purpose of this paper is to examine whether, and how much, acoustic features improve the predictive ability. To this end, we conduct McNemar's (1947) test[4]: we test the null hypothesis that the predictive ability is the same between the two models excluding and including, respectively, acoustic features. More specifically, as in Dietterich [10], we consider a contingency table: n_{11} refers to the number of observations of the test set for the case of correct predictions of both Model2 and Model3; n_{10} for correct prediction of Model2 and wrong

[3] In the case of a perfect predictive ability, both sensitivity and specificity are one (i.e., both FN and FP being zero), and hence AUC is one. In the random prediction case (i.e., useless model), AUC is 0.5.

[4] We appreciate one of the anonymous reviewers, who suggested to use McNemar's (1947) test.

Table 4
Logistic regression (GLM) of top-ten-hit-songs dummy.

Model	Model1	Model2	Model3	Model4	Model4	Model4
Training Period	2013–2015	2013–2015	2013–2015	2013–2015	2010–2015	1998–2015
Regression	(1)	(2)	(3)	(4)	(5)	(6)
Num. Previous songs	0.0003	−0.002	0.002	0.004	−0.018***	−0.019***
Genre: Dance		4.096***	4.255***	4.171***	3.049***	2.628***
Genre: Hip hop		2.964***	3.498***	3.801***	2.821***	2.094***
Genre: Other		3.549***	3.825***	3.480***	1.525**	1.507***
Genre: Pop		3.684***	3.910***	3.912***	2.713***	2.127***
Genre: Rap		3.073***	3.540***	3.732***	1.959***	1.752***
Genre: Rock		3.203***	3.375***	3.206***	2.036***	1.896***
Acousticness			−0.146	−0.264	−0.923*	−0.547**
Danceability			0.239	2.254*	2.382***	1.646***
Valence			1.478***	1.474**	0.945**	0.705***
Duration_ms			0.000	0.000	0.000	0.000***
Energy			−1.523	−3.429**	−2.869***	−1.419***
Intrumentalness			1.576*	1.767*	0.805	−0.209
Liveness			−0.126	−0.073	0.031	−0.261
Loudness			0.134**	0.143*	0.065	0.031
Mode.minor			−0.057	−0.067	−0.025	0.068
Speechiness			−3.480**	−1.722	−0.458	−0.040
Tempo			−0.001	−0.002	0.001	0.001
Time_signature			0.448	0.522	0.153	0.199
Timbre_1_min				−0.009	−0.021*	−0.010*
Timbre_3_median				0.006	0.015***	0.006***
Timbre_6_skewness				−0.447**	−0.429***	−0.262***
Timbre_7_max				0.009	0.013**	0.001
Timbre_9_stdev				−0.082	−0.109***	−0.064***
Timbre_10_kurtosis				0.274***	0.056	0.050
Pitch_10_min				−41.58*	−27.02**	−26.40***
Pitch_4_min				19.09	29.08**	10.332
Pitch_8_stdev				2.978	2.815**	1.060
Pitch_9_max				−8.098	−9.463	−7.507**
Constant	−1.881	−5.356***	−6.320***	−0.123	9.093	3.259
Number of Obs.	1094	1094	1094	1094	2288	5806
Prob>Chi-square	0.9690	0.0000	0.000	0.0000	0.0000	0.0000
Pseudo R-square	0.0000	0.0893	0.1215	0.1806	0.1686	0.1123
AUC: Prediction accuracy (top-ten-hit songs, 2016)	0.4896	0.6553	0.6790	0.6868	0.6851	0.6845

Note: this table provides logit regression results of the top-ten-hit-songs dummy over the training period of 2013–2015 for regressions (1) through (4); 2010–2015 for regression (5); and 1998–2015 for regression (6), respectively. Model4 uses 49 independent variables, where estimates of some of 30 auxiliary acoustic-feature variables are not reported here. Full results are available in the appendix, Section 2.1 (Table S2.1) and 2.2 (Table S2.2A and S2.2B). The row headed 'AUC: Prediction accuracy' provides the prediction accuracy results (in terms of AUC) where models are tested for out-of-time sample of songs (in 2016 charts).
Standard errors are not reported here, but available in appendix: Table S2.1, Table S2.2A, and S2.2B.
***Indicates significance at the level of 1%, ** at 5%, and * at 10%.

prediction of Model 3; n_{01} for wrong prediction of Model2 and correct prediction of Model3; n_{00} for wrong predictions of both Model2 and Model3. Similarly, we also write another contingency table by replacing Model3 by Model4.

Under the null hypothesis about the equal predictive ability between Model2 and Model3 (Model4), n_{10} should be the same with n_{01}: $H_0 : n_{10} = n_{01}$. By contrast, under the alternative hypothesis that Model3 (Model4) is superior to Model2 in terms of the predictive ability, n_{01} is expected to be greater than n_{10}: it is more likely that the prediction of Model3 (Model4) is correct and Model2's prediction is wrong than the other way round. As in Dietterich [10], we calculate the continuity corrected version of the McNemar's test statistic, proposed by Edwards:

$$\chi^2 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}$$

where the distribution of χ^2 is, under the null hypothesis, approximated well by the chi-squared distribution with degree of freedom of one, as long as $(n_{01} + n_{10})$ is large enough (greater than 25), which is the case in all of our results.

2.2.5. Training vs. test periods of songs

We consider, as our baseline case, the training set of songs that appeared in weekly Billboard Hot 100 charts over the period 2013—2015, and the test set of those in 2016. The reason is as follows: In this paper, our primary purpose is to develop a top-ten-hit-songs prediction model and to examine whether, and how much, acoustic features improve the prediction so that practitioners in the music industry can improve their prediction of top hit songs. As such, the training sample, used to training the prediction model(s), should be available at the time of prediction (of future success of a song): it should not include the future observations (i.e., songs to be predicted in practice and/or tested in this paper). Thus, we choose the training and test samples such that they are separate from each other in the temporal aspect [14,18].[5]

Importantly, in the empirical aspect, we also find that our prediction results are improved more when the training period is shorter and hence closer to the test period, potentially due to that using a shorter time span for training data captures more accurately the changing tastes in music for listeners, than using a longer time horizon. This is another reason

[5] This temporal aspect of the prediction framework has been considered, in selecting training/test samples, by many authors. For instance, Interiano et al. [14] consider the one-year rolling window training sample ending just before the one-year test sample, and Ralcheva and Roosenboom [18] consider, to predict the success of equity crowdfunding, the three-year rolling window training sample and one-year test sample.

why we focus on cases of the 3-year (rolling window) training periods and corresponding test periods (including our baseline case).

For robustness check, we will also consider the four different test/training periods as follows: As alternative test sets of songs, we consider (i) songs in (weekly) Billboard 100 charts in 2016, (ii) those in 2015, (iii) those in 2014, (iv) those in 2013, where the models are trained over the corresponding 3-year (rolling window) training period ending just before the given test period (e.g., training period of 2010—2012 corresponding to the test period of 2013). Last, we also consider the alternative longer training period of either 1998—2015 or 2010—2015, keeping the test period of 2016 (of which results are available in the appendix).[6]

2.2.6. Robustness check: training algorithm

For robustness check, we consider three alternative methods/algorithms of training/estimation: (i) Gradient Boosting Machine (GBM), (ii) Random Forest (RF), and (iii) Deep Learning (DL). We use the H2O package (*R* version), which is an interface to use a distributed parallel computing system (many CPUs) over the internet, to implement all of these three algorithms: *R* package 'h2o.gbm' for GBM; 'h2o.randomForest' for RF; and 'h2o.deeplearning' for DL.[7] Note that Interiano et al. [14] use RF algorithm, in which randomly-drawn subsets of the training sample are used in multiple decision trees, of which results are combined to make a final prediction [7]. In case of GBM, differently from RF (in which subsequent decision trees are independently created), subsequent decision trees are created to minimize the loss function, depending on the prior trees [11], and hence the high prediction accuracy can be achieved with a relatively small number of trees. It turns out that in case of GBM algorithm, we obtain the most robust and strongest evidence that acoustic features improve the top-hit-songs predictive ability. Last, in case of DL, the specific algorithm to predict the final outcome is structured by using a multi-layer feedforward artificial neural network.[8]

3. Results

In this section, we discuss the main results: whether, and how much, acoustic features improve the top-ten-hit-songs predictive ability for the baseline case of the generalized linear model. Moreover, we also discuss how the main results are robust to different test/training periods and different training methods, respectively.

3.1. Main results

Table 4 provides the logistic regression (i.e., the GLM training method) results (as well as out-of-time test results) over the training period of 2013—2015, where the training sample includes 1094 observations,[9] and the test sample (songs in 2016 weekly charts) includes 403 observations: Model1 through Model4 results are presented in columns Regression (1) through (4), respectively. We also report regression results of Model4 over the alternative training periods of 2010—2015 (with 2288 training songs) and 1998—2015 (with 5806 training songs) in columns Regression (5) and (6), respectively. We have found that in case of the baseline training period of 2013—2015, coefficients on some variables are less significant than the counterparts in cases of the alternative training period (e.g., 2010—2015), mainly due to the small number of observations: 1094 for the period of 2013—2015, compared to 2288 for 2010—2015. Taking this into our consideration, when we discuss the regression results of Model4, we consider those for the baseline period of 2013—2015 as well as those for periods of 2010—2015 and 1998—2015.

Table 5
Prediction ability of top-ten-hit-songs dummy, test period of 2016 (training period of 2013—2015).

Model with acoustic features	Model3: Correct	Model3: Wrong	Model4: Correct	Model4: Wrong
Threshold criteria 1: maximizing metric 1 (=[precision*sensitivity]/[(precision + sensitivity)/2])				
Model2: Correct	218	24	210	32
Model2: Wrong	71	90	99	62
McNemar's test	$\chi^2 = 22.274$ ($p = 2.4e\text{-}06$)		$\chi^2 = 33.252$ ($p = 8.1e\text{-}09$)	
Threshold criteria 2: maximizing metric 2 (=[specificity*sensitivity]/[(specificity + sensitivity)/2])				
Model2: Correct	216	26	207	35
Model2: Wrong	57	104	80	81
McNemar's test	$\chi^2 = 10.843$ ($p = 0.001$)		$\chi^2 = 16.835$ ($p = 4.1e\text{-}05$)	

Note: this table provides results of McNemar's test of the null hypothesis is that the prediction capability of Model2 is the same with that of Model3 (Model4), respectively, where the test set includes songs appeared in Billboard 100 weekly charts in 2016 (a total of 403 unique songs) and each model is trained over the period of 2013–2015 (a total of 1094 unique songs). This table presents the number of observations for which the prediction of Model2 and Model3 (Model4), respectively, is either correct or wrong. The distribution of χ^2statistic of McNemar's test is approximated well by a chi-squared distribution, with degree of freedom equal to one.

Genre and the artist's previous number of songs are significantly (at 1% significance level) related to the top-ten-hit-songs probability. Importantly, some acoustic features are significantly related to the top-ten-hit-songs probability, too. In Model3 and Model4, both *danceability* and *valence* are significantly (at 5% significance level) and positively related to the top-ten-hit-songs probability [14]. Songs more suitable for dancing and conveying higher positiveness (i.e., feeling 'happier') are more likely to be top hit songs [2]. Results for Model4 (Regression (4), (5), and (6) in Table 4) show that the top-ten-hit-songs probability is also significantly related to some metrics of auxiliary acoustic features: median, standard deviation, and skewness of timbre vectors, and minimum, standard deviation, or kurtosis of pitch vectors. Taken together, these results suggest that acoustic features of a song are closely related to music consumers' choices, driven by their motives to change mood and emotion, consistent with previous findings in neuroimaging science that listening to certain musical songs releases dopamine, a neurotransmitter responsible for tangible pleasures [16,21].

The row headed 'AUC: Prediction accuracy' of Table 4 reports the top-ten-hit-songs prediction accuracy (in terms of AUC) for out-of-time songs in 2016 charts. We can see that adding acoustic features improves prediction results substantially: the prediction accuracy increases from 0.655 under Model2 (without acoustic features, Regression (2)) to 0.679 under Model3 (with main acoustic features, Regression (3)), and further to 0.687 under Model4 (with both main and auxiliary acoustic features, Regression (4)). These results, though not strong statistical evidence, suggest that the granular data provided by music intelligence technologies carries a substantial predictive value in the era of online music streaming and can help the music industry's practitioners to make better decisions.

3.1.1. McNemar's test

We provide more formal statistical evidence that acoustic features

[6] Results for the alternative long training periods of either 1998—2015 or 2010—2015, and test period of 2016 are provided in the appendix: more specifically, see Table S2.2A (S2.2B) for GLM (logit) regression results over the training period 1998—2015 (2010—2015); panel G (H) in Table S2.4 for the McNemar's test results over the training period 1998—2015 (2010—2015).

[7] Documentation of *R* 'H2O' packages is available at: https://cran.r-project.org/web/packages/h2o/h2o.pdf.

[8] For more details, see: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/deep-learning.html

[9] See Section 2.1 in appendix for detailed results (including standard errors of estimates); see Section 2.2 in appendix for results with the training period of either 1998—2015 or 2010—2015.

Table 6
McNemar's test of top-ten-hit-songs prediction ability between Model4 vs. Model2, billboard charts in various test periods.

Threshold criteria	1	1	1	1	2	2	2	2
Training algorithm	Generalized linear model (GLM)	Gradient boosting machine (GBM)	Random forest (RF)	Deep learning (DL)	Generalized linear model (GLM)	Gradient boosting machine (GBM)	Random forest (RF)	Deep learning (DL)
Panel A: Test period: 2016 (Training period: 2013–2015)								
n_{01} vs. n_{10}	99 vs. 32	88 vs. 33	67 vs. 69	122 vs. 36	80 vs. 35	81 vs. 37	64 vs. 76	120 vs. 40
χ^2	33.252	24.099	0.007	45.728	16.835	15.669	0.864	39.006
p-value	8.1e-09	9.2e-07	0.932	1.4e-11	4.1e-05	7.5e-05	0.353	4.2e-10
Panel B: Test period: 2015 (Training period: 2012–2014)								
n_{01} vs. n_{10}	99 vs. 25	82 vs. 20	90 vs. 53	175 vs. 36	76 vs. 25	95 vs. 28	89 vs. 56	116 vs. 40
χ^2	42.976	36.480	9.063	90.256	24.752	35.415	7.062	36.058
p-value	5.5e-11	1.5e-09	0.003	Less than 2.2e-16	6.5e-07	2.7e-09	0.008	1.9e-09
Panel C: Test period: 2014 (Training period: 2011–2013)								
n_{01} vs. n_{10}	95 vs. 23	56 vs. 28	88 vs. 60	76 vs. 26	67 vs. 33	78 vs. 34	76 vs. 62	61 vs. 30
χ^2	42.720	8.679	4.926	23.539	10.890	16.509	1.225	9.890
p-value	6.3e-11	0.003	0.026	1.2e-06	0.001	4.8e-05	0.269	0.002
Panel D: Test period: 2013 (Training period: 2010–2012)								
n_{01} vs. n_{10}	68 vs. 16	53 vs. 21	116 vs. 40	70 vs. 28	34 vs. 43	93 vs. 23	94 vs. 57	58 vs. 33
χ^2	30.964	12.986	36.058	17.153	0.831	41.043	8.583	6.330
p-value	2.6e-08	0.0003	1.9e-09	3.4e-05	0.362	1.5e-10	0.003	0.012

Note: this table provides the results of McNemar's test (with continuity correction) of the equal predictive ability between Model4 and Model2, where Model4 includes, as independent variables, the main and auxiliary acoustic features; Model2 without any acoustic features. Each model is trained over the (rolling-window) 3-year training period ending just before the beginning of the given out-of-time test period and tested for the top-ten-hit-songs prediction for out-of-time songs in (weekly) Billboard Hot 100 charts in the given test year.

improve prediction results: we test the null hypothesis that the predictive ability of Model2 (without acoustic features) is the same with that of Model3 (with main acoustic features) and that of Model4 (with both main and auxiliary acoustic features), respectively. For this purpose, as in Dietterich [10], we conduct McNemar's test (with continuity correction).

Table 5 presents the test results, especially the chi-squared test statistics (together with corresponding *p* values) at the row headed 'McNemar's test' in Table 5. The test results show that for both Model3 and Model4, the null hypothesis is rejected at the 1% significance level, implying that the predictive ability of Model4 (Model3) significantly differs from that of Model2. Moreover, we can see that Model4 makes less prediction error than Model2: n_{01} (the number of instances predicted correctly by Model4 but incorrectly by Model2) is greater than n_{10} (the number of instances predicted incorrectly by Model4 but correctly by Model2); for instance, for the case of threshold criteria 1 (threshold criteria 2), n_{01} is 99 (80), and n_{10} is 32 (35). Taken together, these test results support that Model4 has a significantly better predictive ability than Model2, and the same for Model3. In short, these findings indicate the importance of acoustic features in improving the top-ten-hit-songs predictive ability.

3.2. Results: robustness check

Table 6 presents robustness-check results for McNemar's test (with continuity correction) of the equal predictive ability (more specifically, the top-ten-hit-songs prediction) between Model4 and Model2, for different test periods, for different training methods, and for different threshold criteria. More specifically, we consider four different test periods: (i) songs in (weekly) Billboard 100 charts in 2016, (ii) in 2015, (iii) in 2014, (iv) in 2013, with the corresponding 3-year (rolling window) training period, respectively.[10] For each case of different test/training periods, we also consider the four different training methods/algorithms: (1) baseline case of generalized linear model (GLM), (2) gradient boosting machine (GBM), (3) random forest (RF), and (4) deep learning. We also consider the two different threshold-selection criteria, too.

As shown by Table 6, the prediction ability of Model4 (including both main and auxiliary acoustic features) is, for most cases, significantly better (at the 1% significance level) than that of Model2 that includes no acoustic features. These results are robust to different prediction algorithms/methods, different test/training periods, and different threshold-selection criteria, respectively. (Here, the exception is for cases of random forest method for songs in 2016 and in 2014.) Interestingly, we find that in the case of the gradient boosting machine (GBM) algorithm, the results provide the strongest and most robust evidence supporting our main findings.

4. Conclusion

Using the weekly Billboard Hot 100 charts from 1998 to 2016 and granular data on song-level acoustic features provided *Spotify*, we examine how the granular acoustic data, generated by music intelligence technologies, can improve the prediction of the top-ten-hit-songs probability. We find that some acoustic features (e.g., danceability, happiness, and some metrics of timbre and pitch), beyond genre information, are significant determinants of the top-ten-hit-songs likelihood. Importantly, these acoustic features also improve substantially the top-ten-hit-songs predictive ability. These results suggest that the granular data provided by music intelligence technologies, combined by machine learning prediction algorithms, can help the music industry's practitioners to make better decisions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dss.2021.113535.

[10] Results for different test/training periods (including the training periods of 1998—2015 and 2010—2015) are also available in the appendix (Section 2.4), especially Table S2.4.

References

[1] L. Aguiar, J. Waldfogel, Quality predictability and the welfare benefits from new products: evidence from the digitization of recorded music, J. Polit. Econ. 126 (2) (2018) 492–524.
[2] N. Askin, M. Mauskapf, Cultural attributes and their influence on consumption patterns in popular music, in: Int. Conf. on Social Informatics, Barcelona, Spain, 2014, Springer, Berlin, Germany, 2014, pp. 508–530.
[3] N. Askin, M. Mauskapf, What makes popular culture popular? Product features and optimal differentiation in music, Am. Sociol. Rev. 82 (5) (2017) 910–944.
[4] M. Benner, J. Waldfogel, The song remains the same? Technological change and positioning in the recorded music industry, Strategy Science 1 (3) (2016) 129–147.
[5] S. Bhattacharjee, R.D. Gopal, K. Lertwachara, J.R. Marsden, Stochastic dynamics of music album lifecycle: an analysis of the new market landscape, International journal of human-computer studies 65 (1) (2007) 85–93.
[7] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[8] H. Datta, G. Knox, B.J. Bronnenberg, Changing their tune: how consumers' adoption of online streaming affects music consumption and discovery, Mark. Sci. 37 (1) (2017) 5–21.
[9] S. Dewan, J. Ramaprasad, Social media, traditional media, and music sales, MIS Q. 38 (1) (2014) 101–121.
[10] T.G. Dietterich, Approximate statistical tests for comparing supervised classification learning algorithms, Neural Comput. 10 (7) (1998) 1895–1923.
[11] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. 29 (5) (2001) 1189–1232.
[12] D. Godes, D. Mayzlin, Using online conversations to study word of mouth communication, Mark. Sci. 23 (4) (2004) 545–560.
[13] W. Goldman, Adventures in the Screen Trade: A Personal View of Hollywood and Screenwriting, Warner Books, New York, 1983.
[14] M. Interiano, K. Kazemi, L. Wang, J. Yang, Z. Yu, N.L. Komarova, Musical trends and predictability of success in contemporary songs in and out of the top charts, R. Soc. Open Sci. 5 (5) (2018) 171274.
[15] T. Jehan, Analyzer documentation, The Echo Nest Corporation (2014). Retrieved from, http://docs.echonest.com.s3-website-us-east-1.amazonaws.com/static/AnalyzeDocumentation.pdf.
[16] J. Jolij, M. Meurs, Music alters visual perception, PLoS One 6 (2011), e18861.
[17] M. Mulligan, The Death of the Long Tail, MIDiA Consulting, 2014.
[18] A. Ralcheva, P. Roosenboom, Forecasting success in equity crowdfunding, Small Bus. Econ. 55 (2020) 39–56.
[19] S. Rosen, The economics of superstars, Am. Econ. Rev. 71 (5) (1981) 845–858.
[20] M.J. Salganik, P.S. Dodds, D.J. Watts, Experimental study of inequality and unpredictability in an artificial cultural market, Science 311 (2006) 854–856.
[21] V.N. Salimpoor, M. Benovoy, K. Larcher, A. Dagher, R.J. Zatorre, Anatomically distinct dopamine release during anticipation and experience of peak emotion to music, Nat. Neurosci. 14 (2011) 257–262.
[22] Spotify. Spotify.com, 2018. Retrieved from, https://developer.spotify.com/web-api/get-audio-features/.
[23] K. Subramanian, Splunking the Billboard Hot 100 With Help From the Spotify api. function1.com, 2017. Retrieved from, https://www.function1.com/2017/09/splunking-the-billboard-hot-100-with-help-fromthe-spotify-api.
[24] Q. McNemar, Note on the sampling error of the difference between correlated proportions or percentages, Psychometrika 12 (2) (1947) 153–157.

Seon Tae Kim is a lecturer in economics at University of Liverpool Management School. His research focuses on understanding the effects of digital technologies on the firm's strategic decisions and its economic implications, at the industry and aggregate level. His previous research has examined the effects of digital technology and financial shocks on the decision making processes of firms and households.

Joo Hee Oh is an Assistant Professor at WMG, University of Warwick. Her research focuses on quantifying the growth and value of data and digital economy. Dr. Oh has developed a stream of research that includes measuring the value of free goods and services on the Internet and quantifying the amount of user-generated capital for the Internet-based media industry. Dr. Oh previously was a Postdoctoral Associate at MIT Sloan Initiative on Digital Economy, managing research in Attention Economy and User Generated Capital, working with Prof. Erik Brynjolfsson at MIT Initiative on the Digital Economy. In particular, she worked with industrial partners to examine digital content market and to measure social welfare from free Internet services.