SMU Data Science Review

Volume 5 | Number 3 Article 9

2021

Spotify: You have a Hit!

Christopher E. Dawson Jr.
Southern Methodist University, dawsonc@mail.usf.edu

Steve Mann
Southern Methodist University, stevem@smu.edu

Edward Roske
Southern Methodist University, eroske@interRel.com

Gauthier Vasseur
University of California, Berkeley, gauthier.vasseur@berkeley.edu

Spotify: You have a hit!

Christopher Dawson[1], Steve Mann[1], Edward Roske[1], Gauthier Vasseur[2]
[1] Master of Science in Data Science, Southern Methodist University, Dallas, TX 75275 USA
[2] The Fisher Center for Business Analytics, Haas School of Business, University of California, Berkeley, CA 94720 USA
dawsonc@mail.usf.edu
stevenctmann8821@gmail.com
eroske@interRel.com
gauthier.vasseur@berkeley.edu

Abstract. Over 87% of the streaming music is owned by four major record labels (Jones, 2018). Yet, the songs owned by those labels account for <1% of the total amount of music created each year. These labels are historically better at identifying talent (though this talent identification is becoming more difficult). Even though Spotify has 36% of the streaming marketing share (T4, 2021), Spotify has not been profitable because of the large licensing costs paid to the large music labels. If Spotify could identify hit songs & artists before the large labels, they would sign those artists and dramatically reduce their licensing costs. Using the Spotify API, this paper will use Spotify data on over 400K songs over the last three years for exploratory data analysis, provide descriptive statistics, perform feature selection, and develop models using LASSO and XGBOOST Classification. The research determined multiple key features and predicted with over 60% accuracy songs which were going to be a hit (defined as >90% popularity).

1 Introduction

The music industry can be best described as an imperfect art. The expression of sound through instruments and voices is at its core an experimental process, with no correct answers. The lack of an absolute 'correct' is certainly frustrating to all stakeholders in the industry: artists, fans, managers, record label executives. No one likes a bad song yet everyone is subject to failed musical experiments that are quickly relegated to the dust bin of history. Time was wasted producing that failed music, and ears were (at least metaphorically) harmed by listening to it, but eventually, human listeners determined what they wanted to hear and what they didn't. Songs were sorted by pure "human power" into hits and forgotten tracks.

But what if there was a way to determine if a song would be a hit before the song was released to the world before the music video was produced before marketing dollars were spent trying to foist a song into popularity? The money would be saved for the stakeholders in the production process, artists would not release songs that resulted in commercial failure. Plus, fans would have a higher degree of enjoyment since they didn't have to sift through stacks of musical hay to find a hit needle.

Even the greatest producers in the world have less than a 1-in-18 chance (5.4%) of producing a song that becomes a number 1 hit on the Billboard charts (Pearce, 2016). These top producers have an ear for what will make a hit song, but it's not scalable: they produce by feel and not by any sort of quantifiable measure. As such, the amount of time and money wasted both on the production (songwriters, artists, producers, labels, etc.) and listening sides is staggering. It could all be minimized if there was a way to predict whether a song will achieve commercial success before it was released.

Like baseball and investment banking, the common understanding for generations was that art could not be quantified into science. Everyone was sure that only individuals with lifelong experience could decide what type of player, investment, or song would outperform the others. There was no data behind it, only a hunch proven to others by anecdote.

There is a need to move behind unquantifiable decisions. As Michael Schrage said in Harvard Business Review, "Gut feelings aren't data. Quality data deserve deference; personal experience does not" (Schrage, 2019).This research team seeks to replace those gut feelings with data.

One of the biggest problems historically has simply been a way to quantify music. Spotify has been leading the way detailing the audio features of a song and making available extensive data on recommendations and user listening preferences through their Spotify API (Skidén, 2016). Based on the completeness and ready availability of data through this Spotify API, the team has chosen to use this information for performing Hit Song Science™.

The field of Hit Song Science™ is not new (Dhanaraj & Logan, 2005), but it is still rapidly evolving. Historically, analysis has been done looking backward at popular songs[1] to see if they had anything in common. The research team felt that there was a gap: could this information be more used for *prediction*. Could data from the Spotify API be used to determine the moment a new song by an unknown artist was uploaded to Spotify if that song was going to be a hit?

While it is not the primary motivation for this analysis, information on what artists and songs will *be* popular could be used for significant monetary gain. In 2020, four major record labels provided 78% of the music listened to on Spotify: Sony Music, Merlin, Universal Music Group, and Warner Music Group (Ingham, 2021). While this is a substantial majority, that percentage has fallen from a peak of 87% just three years earlier (Ingham, 2021).

There is a battle raging between three forces: the artists, the labels, and the listening providers, which are increasingly being dominated not by traditional radio but rather by streaming (Mayfield, 2021). Artists want to keep as much of their revenue as possible, 85% of which comes from streaming (Mayfield, 2021). The labels want to sign the new artists as quickly as possible to make sure they get their piece of the pie. The streaming and radio providers want to minimize their payouts (a substantial portion of their cost of goods) to the larger music labels that control most of the music people want to stream.

Suppose the artists can stream their music on Spotify or other streaming services without going through a music label. In that case, more revenue goes to the artists, and

[1] Popular is generally defined as reaching top placements on various musical ranking charts like the Billboard Top 200.

less expense comes from the streaming service. The inevitable loser? The middle man, the large music labels, and it's doubtful they'll go gentle into that good night.

Goldman Sachs spent more than one hundred million dollars simply to reduce the timing of their stock trades by mere milliseconds (Son, 2019). Eventually, the providers and the labels will rush to see who can sign new artists as quickly as possible. The best algorithm (and then the ability to execute on the recommendations of that algorithm) could result in billions of dollars in either additional revenue or reduced expenses. As this gold rush came to investments in stocks, so will this gold rush come to investments in artists and the songs they provide the world.

What has yet to be done in this field? As stated earlier, most exploration to date in this space has been historical classification and not future prediction. It has been trying to uncover patterns in what makes songs (as a whole) popular and not using that information to determine if an artist and their recently uploaded song *will be popular* in the future.

There is also a gap in the traditional research in that it was usually applied only to a single feature area: either audio features (acousticness, danceability, etc.), lyrics (the words being sung in the song, how repetitive those words are, etc.), or metadata (genre, popularity of the artist already, length of song, etc.). The architecture proposed in this paper seeks to utilize all of those three main methodologies (audio, lyrics, metadata) together into one algorithm.

Another motivation was a need to help these streaming providers reduce their cost of sales. Until the first quarter of 2021, when they turned a fourteen million Euro profit (Goldberg, Katzen, & Jenkins, 2021), Spotify has been a money-losing company since its inception. Streaming companies have to reduce their cost of sales (which is mostly the money they pay to major music labels) if they are to survive. These streaming providers need to exist to make sure listeners have as many options as possible to control their music. If the research team's algorithm proves deployable, the streamers could license this algorithm to sign the artists before the large labels do.

From the start of 2015 to the end of 2020, Spotify has grown from 68 million users to 345 million with a more than eight-fold increase in paid subscribers (Iqbal, 2021). The company is founder-led and moving into the Podcast listenership space by storm, with over 25% of users utilizing that space. All this, including a shared love for good music, led to this research.

One might wonder, is there enough new music by enough new artists that a fast, predictive algorithm needs to be developed? According to Jeremy Erlich, Spotify's co-head of Music, as of February 2021, more than 60,000 new songs are being uploaded to Spotify every single day which is a rate of approximately one track added every 1.4 seconds (Ingham, 2021). Predicting hit songs from musical features with more than 60,000 songs loaded daily is just the first step of the new musical frontier.

2 Literature Review

2.1 Hit Song Science™

The first talent agency in America was founded in 1898 by William Morris (Grant, 1998), and the race to sign talent was born. William Morris dominated the industry for decades and continued to exist as an independent company for 109 years, signing singers including Al Jolson, Elvis Presley, Judy Garland, and more in addition to talent from other industries like TV and movies (Grant, 1998). For their 109 years, determining if a singer would become popular was either a matter of gut instinct, or it was signing an artist and then using the power of talent labels to make a singer popular. In some cases, it was a backward-looking practice: wait until the artist became popular and then sign them.

Like baseball being disrupted by analytics over the last 20 years (Birnbaum, 2021), it was only a matter of time until determining if an artist would become a star would shift from art to science. By the early 2000s, research began into an area that data scientists began calling *Hit Song Science*™ (Dhanaraj & Logan, 2005) though this term eventually became trademarked by a for-profit firm named Polyphonic Human Media Interface. This paper will be using the term HSS going forward though the trademark was allowed to lapse in 2016 (Hit Song Science European Union Trademark Information, 2021).

HSS was slow to proceed in the early days due to two problems: there was no easy to access, real-time database of which songs were popular, and there was no standard way of classifying attributes for various songs, so much of the early research focused on feature extraction which in some cases was extremely manual and based on human intuition (Dhanaraj & Logan, 2005).

Early analysts felt that the biggest issue that had to be overcome to turn HSS into a true science was the determination of "good features using feature generation techniques which have been shown to outperform manually designed features of subjectivity for even simpler musical objects such as sounds or monophonic melodies. Hit song science is not yet a science but a wide-open field" (Pachet & Roy, 2008).

The increase of MIR (Music Information Retrieval) databases has made it easier to access larger volumes of songs, labels on what is and is not popular. These together gave rise to advancements in HSS wherein research teams were able to make decent predictions about the which songs would become hits (Herremans, Martens, & Sorensen, 2014).

Interestingly, the same author (Francois Pachet) who said that HSS was not yet a science in 2008 (due to the lack of standardized feature availability mentioned above), wrote an entire chapter (chapter 10 titled "Hit Song Science") for the book *Music Data Mining* in which he expressed his feelings that HSS was progressing beyond feature generation to more interesting questions like "Are there features of popularity, for an individual or for a community, and, if yes, what are they" (Li, Ogihara, Tzanetakis, & Pachet, 2012)"?

Since 2012, the rise of streaming services allowed easier access to music through common APIs with pre-defined features such as in the Spotify API (Menten, Ng, O'Rourke, & Holmes, 2018). Since the Spotify API is well understood, has clearly defined attributes for every song in their database, and has a clear definition for popularity (streaming music counts), the research team will be using the Spotify API for this paper.

2.2 HSS Modeling

Some work has been done previously in modeling the popularity of songs both with and without the Spotify API; however, none have focused specifically on predicting the success probability of songs uploaded to Spotify in real-time.

Data from the Spotify API was used in (Middlebrook & Sheik, 2019) to predict previous Billboard top 100 hits between 1985 and 2018. The problem was modeled as a classification question to predict if a song would be a Billboard top 100 hit or not. Methods used were logistic regression, neural networks, random forest, and support vector machines. The best performing model was random forest, returning approximately 88% accuracy on a held-out test set.

What sets (Middlebrook & Sheik, 2019) apart from the research described in this paper is that the name of the artist was included in the models. As the Billboard top 100 is inundated with repeat artists, this feature was likely highly influential. As such, these models may not extrapolate well to artists previously unfeatured in the Billboard top 100. Thus, the models may misclassify songs from artists new to the Billboard top 100, leaving the opportunity for other music labels to sign them first. The models discussed in this paper did not include the artist's name as a feature, thus allowing them to extrapolate much easier to new artists, providing greater value to Spotify as a record label.

Van Der Somme (2021) also used the Spotify API differently from the researchers (Middlebrook & Sheik, 2019). This project extracted lyrics from songs and used as the only predictor for the LFM-1b dataset (a dataset of over 1 billion listening events) (Van Der Somme, 2021). The dataset was split into three (3) categorical target variables of high, medium, and low popularity. Features used were derived strictly from lyrics and included sentiment analysis, repeatability, and TF-IDF terms. K-nearest neighbors, support vector machines, and Naïve Bayes classification were used as modeling techniques, with support vector machines delivering the highest accuracy at 58%. However, this accuracy statistic is driven primarily by the accurate classification of medium and low popularity songs. The models cannot accurately determine the high popularity songs, with the highest precision rate on that class of 13%. Lyrics on their own are not enough to identify hit songs.

Yang and associates (2017) went a different route than the previous papers cited, using data from 30,000 users of the Taiwanese media company KKBOX Inc. gathered from October 2012 to September 2013, representing over 125,000 songs. This data was split into subsets of songs sung in Mandarin Chinese and songs sung in English. Ten thousand songs for each subset were selected for model building. This paper was unique

in that the data gathered from the songs were purely audial. Audio recordings from the middle 60 seconds of each song were sampled at 22 kHz and fed into a mel-spectrogram. This data was then fed into different convolutional neural networks of progressive depths to build and test models. Recall scores were emphasized in the paper (probability of correctly predicting a top 100 song), with the highest recall score being 30% on the deepest model. This proved the paper's hypothesis true that deeper neural network models will perform better than shallow models on this data. However, this is not actionable for predicting hit songs, given the low recall score.

2.3 LASSO Modeling

LASSO (Least Absolute Shrinkage and Selection Operator) is a type of regression that focuses on feature selection with a regularization component that makes models more accurate (and easier to interpret) than traditional regression methods. LASSO was first created in geophysics in 1986 (Santosa & Symes, 1986) and was independently discovered by Robert Tibshirani in the mid-1990s (Tibshirani, 1996).

In Tibshirani's article, "Regression Shrinking and Selection via the Lasso", he outlines why traditional OLS (Ordinary Least Squares) estimates result in high variance and how if bias is slightly sacrificed (by shrinking the coefficients, possibly to zero, for some of the features), a more accurate and more interpretable model will result (Tibshirani, 1996). Also, in this article, he outlines how his "Lasso" method "retains the good features of both subset selection and ridge regression" (Tibshirani, 1996).

In the Spotify API, the problem is not one of enough data (and features) but instead of too much. With the goal of being able to make a real-time predictive model, features not relevant to predicting hits should be discarded. Also, there is a significant risk of overfitting data even with just the subset of records used by the research team in creating this model. Reducing the features that aren't genuinely impactful is a hallmark of LASSO and why it was chosen for this HSS analysis.

One of the critical aspects of LASSO models is tuning the regularization parameter. Traditional methods of tuning this parameter include AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion). Still, more modern methods such as ERIC (Extended Regularized Information Criterion) can outperform for traditional AIC and BIC metrics (Francis K. C. Hui, 2015).

2.4 XGBoost Modeling

Boosting is the general process of using a series of weak learners in an ensemble to make a strong learner (Schapire, 1990). As Schapire (1990) shows, boosting can reduce both bias and variance in a learning model. Early models were not adaptive, so Schapire and Freund worked together to create the next generation of boosting, which they called AdaBoost (Freund, 1995), and thus adaptive boosting was born.

Gradient Tree Boosting was the next major evolution in the world of boosting (Friedman, 2001). Gradient tree boosting algorithms uses a differentiable loss function "from the perspective of numerical optimization in function space, rather than

parameter space" (Friedman, 2001). These models continued to evolve over the 21st century until, in the mid-2020s, an extremely scalable tree boosting system took the data science world by storm (Chen & Guestrin, 2016).

Chen and Guestrin called their open-source solution XGBoost (eXtreme Gradient Boosting). XGBoost is less of a methodology and more technological solution to provide a scalable solution in all scenarios. In their "XGBoost: A Scalable Tree Boosting System" paper in 2016, they outline how Tree Ensemble Models (powered by XGBoost) have been leading to success in machine learning competitions regardless of the type of competition:

> *"The most important factor behind the success of XGBoost is its scalability in all scenarios. The system runs more than ten times faster than existing popular solutions on a single machine and scales to billions of examples in distributed or memory-limited settings. The scalability of XGBoost is due to several important systems and algorithmic optimizations. These innovations include: a novel tree learning algorithm is for handling sparse data; a theoretically justified weighted quantile sketch procedure enables handling instance weights in approximate tree learning. Parallel and distributed computing makes learning faster which enables quicker model exploration. More importantly, XGBoost exploits out-of-core computation and enables data scientists to process hundred millions [sic] of examples on a desktop. Finally, it is even more exciting to combine these techniques to make an end-to-end system that scales to even larger data with the least amount of cluster resources."*
>
> *(Chen & Guestrin, 2016, p. 785)*

By the end of 2016, XGBoost seemed to be not only winning every data science competition (Nielson, 2016), it was also dominating the top ten places. As for the "why?" on how it became the default tool for data science applications, Nielson (2016) concludes that it is due to additive tree models being quite good at estimating complicated functional relationships (including the ability to generalize into higher-order dimensional spaces). Also, Nelson felt that adaptive tree boosting methods are extremely flexible including in smaller spaces (representing neighborhoods within the data). The flexibility and adaptation of XGBoost do an excellent job of balancing the bias-variance tradeoff.

Since this paper does not seek to establish a new method of machine learning but rather a new application of existing machine learning in a new way, the research team felt that XGBoost would be the most flexible, adaptable, and importantly, *scalable* solution for the model being developed.

2.5 Hypothesis

The hypothesis is that there are patterns and features of songs that, when considered in combination with each other in the context of a predictive model, will indicate whether that song will have commercial success, as defined by being in the top 10% of popularity on Spotify.

The algorithm will use LASSO and XGBOOST to create a classification model for hit songs. That classification model will be able to predict with over 75% Accuracy songs that will be a hit (defined as >90% popularity). Analysis of the model will also include which features are of key importance so that this information can be used to one day guide artists in making songs that appeal more to the masses.

3 Methods

3.1 Data

Data used in the analysis for this paper was obtained entirely from Spotify's developer application programming interface or API. This data within the API consists of several of Spotify's proprietary musical features derived from various attributes of an individual song some of which include:

- Popularity (the target variable: a measure derived from the number of times the song has been streamed, and how recently)
- Energy (a measure of the liveliness of a song)
- Loudness (a feature derived from the decibel measurements of a song)
- Liveness (the degree to which a song sounds like it was record live vs. studio produced)
- Danceability (how easily a song is to dance to)
- Speechiness (a measure of number of words spoken)
- Instrumentalness (a measure of instrumental sounds within the song)
- Valence (a sentiment analysis regarding the positivity of a song)
- Other, more traditional, features are included as well, such as the length of the song, the beats per minute, and the key.

While Spotify's API includes information for songs dating as far back as 1908 (Compilation, 1908), the data used for this analysis was limited to songs released in recent years, specifically 2018 – 2020, to retain relevance to the immediate future. Resulting in a final dataset of over 400,000 songs for analysis.

3.2 Design and Procedure

As outlined previously, the objective of this research paper is to identify key variables that drive the popularity of songs to create an algorithm that accurately predicts the popularity of a song as a 'hit' (>90% popularity) immediately upon upload into the Spotify library. Exploratory analysis was conducted on the dataset described above, and features that appeared correlated to the target variable, popularity, were identified. The top three features in terms of linear correlation are shown in Figure 1, Figure 2, and Figure 3 below. From here, several different modeling methods were deployed to determine the feasibility of predicting popularity.

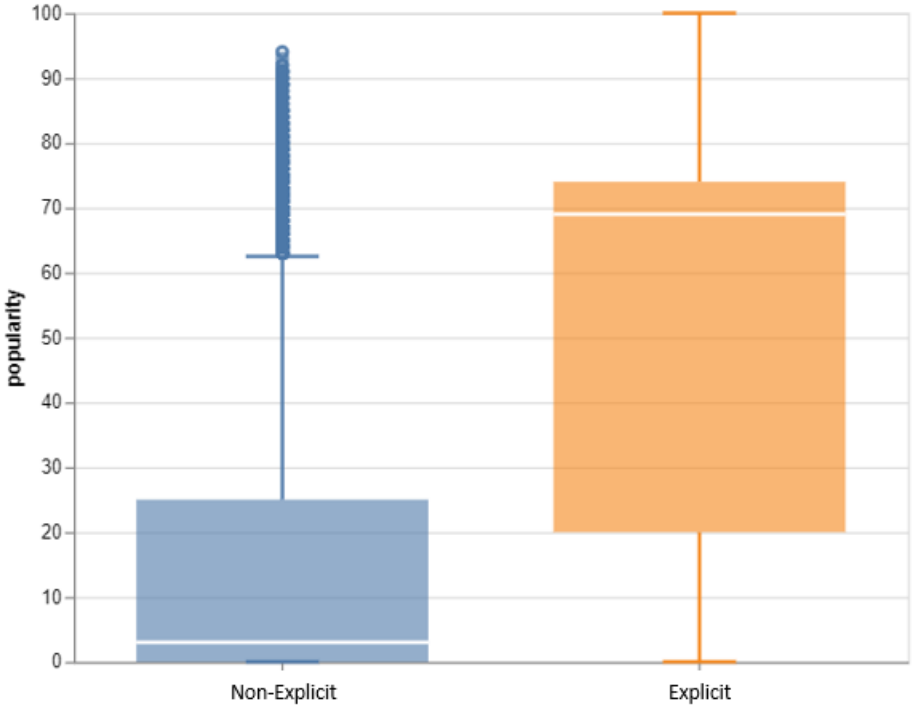

Figure 1: Distribution of Popularity by Explicitness.
Popular songs tend to have more explicit lyrics.

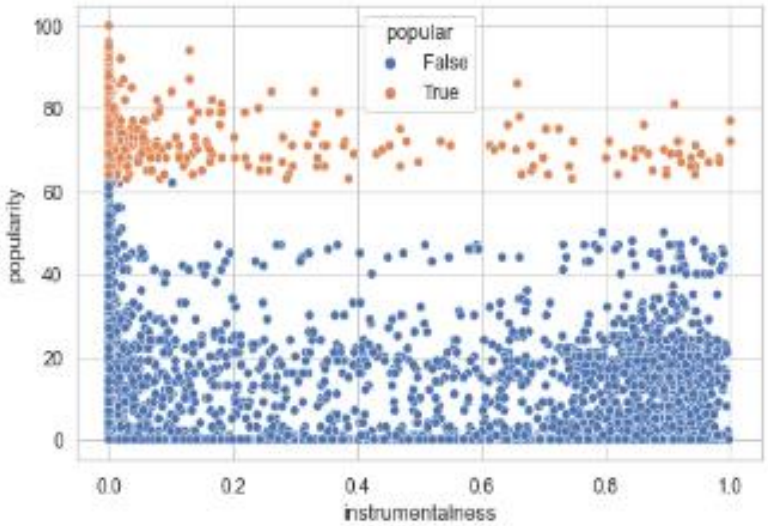

Figure 2: Distribution of Popularity by Instrumentalness
Popular songs tend to cluster around lower instrumentalness.

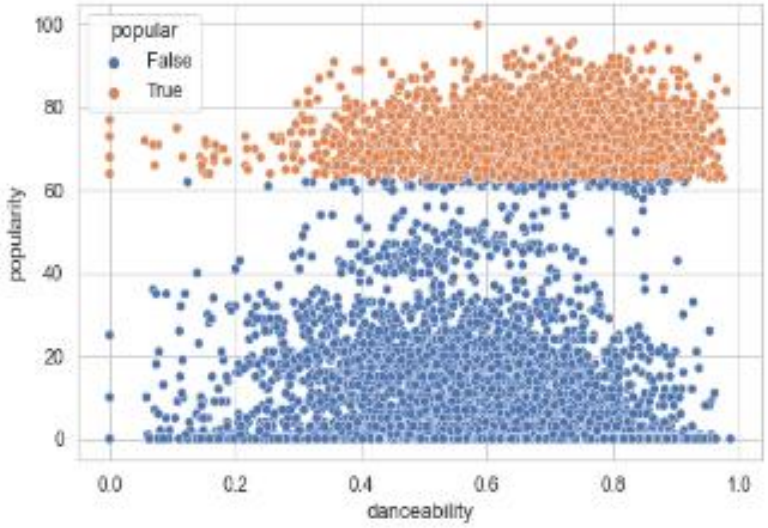

Figure 3: Distribution of Popularity by Danceability
Popular songs tend to cluster around higher danceability.

3.3 K-Nearest Neighbors

K-Nearest Neighbors, or Knn for short, is a modeling technique that can be used for both classification and regression tasks. Datapoints are plotted in a high-dimensional space, and the distance between all of these points is calculated. Then, a value of K is determined, usually through iterating through several potential values of K within the training data to uncover the optimal value, and the K nearest neighbors are looked at to determine a prediction. For regression, the target value of the K nearest neighbors are averaged together, and for classification tasks, the simple majority class of those neighbors is assigned to the prediction.

3.4 Random Forest Classification

To understand random forest classification, one must first understand a decision tree. A decision tree is a method of classification that relies on a series of descending choices, or 'branches.' Based on the decisions made down each of the branches, a conclusion is reached regarding the question of interest. A simplistic example of a decision tree regarding the classification of different classes of animals is shown below.

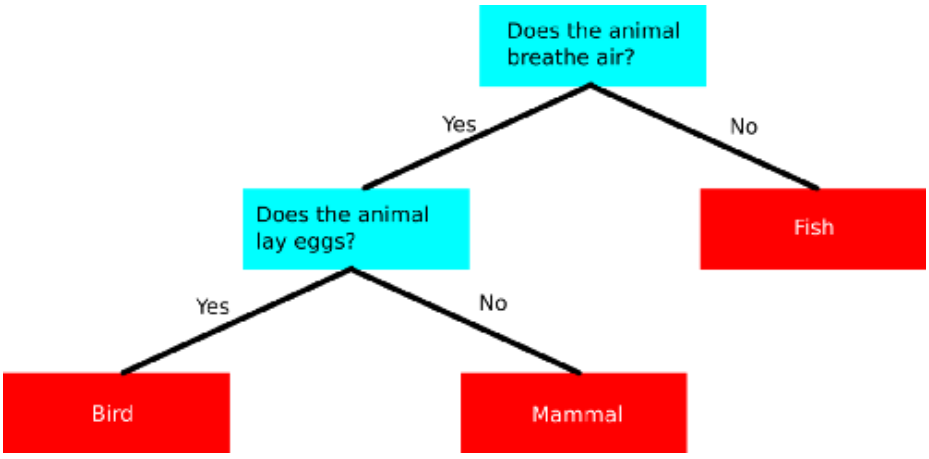

Figure 4: Simple depiction of a decision tree **(Bickerton, 2018)**.

Random forest classification is a method in which many decision trees are evaluated against the data. Each of the decision trees considers a random subset of *n* variables as determined by the researcher, where *n* is commonly the square root of the total number of variables in the final dataset. These individual decision trees determine the optimal criteria to split each subset of variables to classify the target best. Then, an average of all of the individual decision trees is calculated to derive the final algorithm.

The research team did make a decision tree for the algorithm developed in this project. It is extremely detailed, but is essentially an expansion of the basic idea seen above:

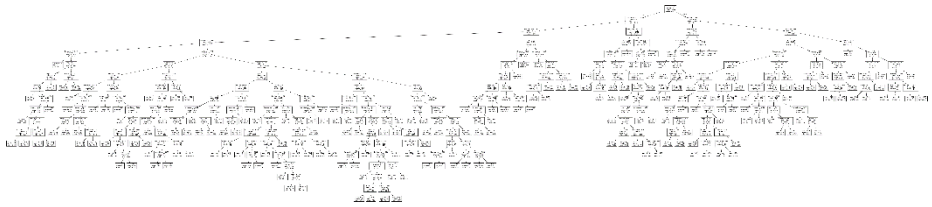

Figure 5: decision tree as output by the HSS algorithm.

3.5 Gradient Boosting

Gradient boosting is a method developed recently in the machine learning world that allows for algorithms to more highly weight individual observations that are difficult to produce a prediction for. Gradient boosting is commonly used with random forest models. Individual decision trees are built, and the forest is grown, as normal, but observations that the algorithm isn't able to accurately predict are sent back through to the model again. The model is then revised based on these higher weighted observations. This process is iterated through until no further improvements can be made (Singh, 2018).

3.6 Logistic Regression

Logistic regression is a special case of the more commonly known linear regression, used exclusively for binary classification problems. Like linear regression, coefficient weights are assigned to each variable in a logistic regression model and multiplied by the observed values of a given data observation to generate an output. However, unlike linear regression, the output of logistic regression is always a value between 0 and 1. Suppose the output value of the logistic regression model is higher than a predetermined cutoff level (also between 0 and 1). In that case, the model predicts the positive class, and vice versa for below the cutoff.

4 Results

Due to the nature of the task, the research team built the models to optimize the accuracy statistic, which is the total correct classification predictions divided by the total number of predictions made. This statistic was chosen for optimization, as the research team assumes equivalent cost of errors for misidentifying songs that will or

will not be hits. A summary of performance statistics for the models built by the research team is shown below.

Type	Method				Confusion Matrix
		Accuracy	Precision	Recall	
Classifier	*KNN w/ Downsampling*	*0.70*	*0.86*	*0.51*	*[[501 48] [288 297]]*
Classifier	Gradient Boosting Model	0.71	0.86	0.52	[[499 50] [279 306]]
Classifier	Random Forest	0.63	0.87	0.34	[[518 31] [384 201]]
Classifier	*Decision Tree*	*0.67*	*0.78*	*0.49*	*[[470 79] [297 288]]*
Classifier	*Logistic Regression*	*0.72*	*0.9*	*0.52*	*[[515 34] [283 302]]*
Classifier	SGD	0.72	0.87	0.55	[[502 47] [265 320]]

Figure 6: Model Comparison. The models are compared based on accuracy, precision, and recall.

As illustrated by the chart, logistic regression and stochastic gradient descent (SGD) resulted in the highest accuracy statistics, with logistic regression having a slight edge in the area under the curve (AUC) in a receiver operating characteristic curve (ROC Curve). These methods tend to perform well when there is a linear separation between classes in a dataset. As shown in the methods section, SGD attempts to find the widest margin between classes to draw a decision boundary. Similarly, the cutoff threshold for the logistic regression model was optimized to create the widest boundary between classes.

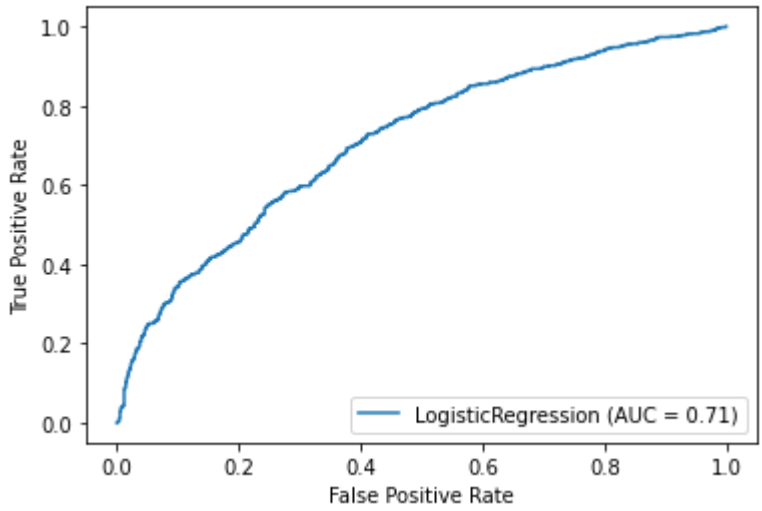

Figure 7: AUC score curve of Logistic Regression

The other modeling methods explored by the researchers produced less favorable accuracy and AUC statistics. Specifically, K-nearest neighbors (KNN) modeling was not as successful as the other models due to the nature of that algorithm. KNN models tend to perform well when there are clean, defined clusters of classes in the data. For that data in this question of interest, the target variable was more easily discriminated by a linear function rather than a distance function.

The initial hypothesis was simplify "are there patterns and features of songs that, when considered in combination with each other, indicate that a song will have commercial success?" Based on these performance statistics, the research team fails to reject the hypothesis. An algorithm not only seems possible, but using a large dataset, outperforms a "starting point" random forest model.

5 Discussion

The research conducted in this project uncovered the characteristics of songs shown in Figure 8 & 9 below were more likely to indicate commercial success. The researchers hypothesize that these variables emerged as the most indicative of popularity due to the project sampling a broad range of genres, in which pop music was the most frequent genre present in commercially successful songs. Explicit lyrics, low instrumentalness, and high danceability and loudness are all dominant features in hit pop songs of the past three years. The researchers hypothesize that the feature importance would vary when determining top 10% popularity in a specific genre. For instance, these same features might seem very out of place in a popular country or heavy metal song.

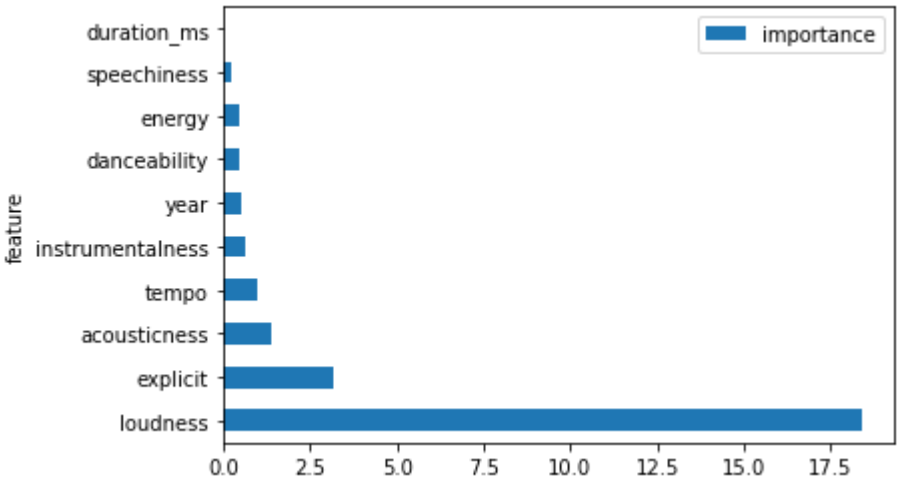

Figure 8: Feature Importance. Absolute LASSO scaled linear correlation with popularity.

Decision Tree Feature Importance	
instrumentalness	0.304
acousticness	0.126
speechiness	0.099
year	0.099
liveness	0.087
danceability	0.082
energy	0.065
tempo	0.042
long_tracks	0.04
loudness	0.03
explicit	0.027
duration_ms	0

Figure 9: Feature Importance. Absolute decision tree correlations with popularity.

As mentioned previously within this paper, it is estimated that the most popular music producers creating songs on gut feel alone have less than a 1-in-18 chance (5.4%) of producing a song that becomes a number 1 hit on the Billboard charts (Pearce, 2016). Given that statistic, it would be hard to have much confidence that any given song made by a top producer will reach commercial success, let alone an unheard-of producer or independent artist.

However, when the best algorithm developed by the research team evaluates a song and predicts it will be a hit, that algorithm is correct approximately 90% of the time – an incredible lift over gut feeling alone. Conversely, when the best algorithm predicts a song will not be a hit, it is correct approximately 65% of the time. These metrics in combination show that there can be a high degree of confidence put into the predictions that are output from this model – certainly a greater degree of confidence than gut feel alone.

Going forward, the researchers believe there is additional data that could be gathered that could further increase the efficacy of the algorithm. This project was limited to features of individual songs within the context of all songs in Spotify's library uploaded between 2018 and 2020.

The researchers hypothesize that the most impactful data that could be gathered would be the degree to which the general public has previously been exposed to the individual song or artist. This brings about a derivation from the proverbial 'If a tree falls in the woods, does it make a sound?' If someone writes a hit song, but no one hears it, is it a hit song?

There are likely several songs in Spotify's library that were used for this project's training data that have the characteristics of a hit song, but have not had the widespread exposure or distribution in order to be classified as a hit song. This effectively means the current iteration of the algorithm is biased towards songs and artists that have the backing of large labels or other distribution means to promote widespread exposure. Quantifying this datapoint would allow the model to apply more broadly, and not falsely exclude smaller, lesser-known artists as often.

From a practical standpoint, there is a clear application of this algorithm: to be utilized by artists, producers, and record labels to create music, edit and enhance songs, and identify talent at scale to simultaneously increase streaming and live performance revenue and decrease production and opportunity costs.

As it relates specifically to Spotify, Spotify would be able to use this algorithm to identify hit songs that are uploaded to their library in real time. Spotify has famously had a contentious relationship with artists, most notably Taylor Swift and Tool, due to the very small portion of streaming revenue that the individual artists receive from Spotify, which Spotify blames on the large licensing costs it must pay to record labels.

The ability to identify hit songs in real time would allow Spotify to identify up-and-coming artists before their competitors, and sign them to their own record label first, eliminating the need to pay these hefty licensing fees to 3rd party record labels, leading to more take home pay for artists, and increased profits for Spotify.

The practical application does lead to some interesting ethical questions that need to be asked:

- Will bringing a scientific approach to an artistic business kill the creative process?

- Due to models making predictions on past performance, will music cease to innovate?
- If this model, or models such as this one were owned solely by large companies, such as Spotify, would the very small chance that indie artists have at attaining commercial success further evaporate?

Each decade of the modern era has welcomed in a new popular genre. The 50's were full of crooners and heavily blues influenced musicians, like Elvis, Frank Sinatra, and Chuck Berry. The 60's saw the British Invasion into U.S. markets with The Beatles and The Rolling Stones dominating the top charts. In the 70's came classic rock – the 80's had hair and thrash metal. The 90's saw grunge come onto the scene, before boy bands like Backstreet Boys and *NSYNC took over in the early 2000's, ushering in the current era of pop dominance that still persists today.

Within all of these eras are two constant themes: artists trying to replicate others' success, and artists continuing to create and innovate. The researchers believe that the algorithm developed within this project would enhance the scale of success that artists and record labels would enjoy, but it would not destroy the innovation and creativity that is at the core of all musicians.

Additionally, record labels pushing their artists to develop songs for commercial success instead of artistic intent and innovation is nothing new. The Smashing Pumpkins' hit 'Cherub Rock' is about exactly this topic, for instance. Motley Crüe and David Bowie both launched independent labels due to disputes with their respective record companies. The rebelliousness that hallmarks some of music's most successful names gives the researchers confidence that this algorithm would not inhibit or destroy the creativity underlying the overall industry.

There is one other ethical consideration that must be brought up: if Spotify is able to sign new artists (destined, per the algorithm, to be popular), does that then turn Spotify into the middleman as it relates to other streaming services? In other words, since Spotify would doubtless put the independent artists under contract, does Spotify end up recreating the exact problem we're hoping to solve? Are we giving up one evil for another one, a more modern evil that uses algorithms initially for good but turning towards profit just like many before it?

Anytime one gives the power to a corporation to make analytics-driven decisions, one does risk the possibility that the corporations use analytics to make themselves more powerful. That said, it has to be better than the current situation where only a few powerbrokers at powerful labels use gut feel and instinct to decide who succeeds, who doesn't, and how much money they'll make in the process.

6 Conclusion

The research conducted within this paper demonstrates that underlying features of individual songs can be collected and input into statistical models that can predict with a high degree of accuracy whether that song will become a commercial success. This scientific approach can be applied to the inherently artistic music industry in order to allow for greater efficiency and scale for individual artists, producers, and record companies.

As it relates specifically to Spotify, Spotify would be able to use this algorithm to identify hit songs that are uploaded to their library in real time, and sign these artists to deals before other record companies. The elimination of these record company middlemen would reduce the licensing fees that Spotify must pay in order to stream songs of these artists. This would allow Spotify to pay the individual artists a greater share of the revenue increasing Spotify's own bottom line. Access to this algorithm would also enable Spotify to identify songs from unknown artists which have the potential to become hit songs. These songs and artists would then be promoted through their Discover feature – once again a win-win for both the individual artist and Spotify.

The research team expects the use of data in the music industry to become prevalent in the upcoming decade. Just as the *Moneyball*-led Oakland Athletics revolutionized how sports teams are assembled and managed, algorithms such as the one discussed in this paper will change the landscape of the music industry. While these algorithms will by no means replace the creative process that goes into making music, the instantaneous feedback they are able to provide will give artists, producers, and record companies that embrace their power the ability to beat their competitors.

Acknowledgments. Jacquelyn Cheun, PhD. – Capstone Professor

References

1. Anderson, I., Gil, S., Gibson, C., Wolf, S., Wolf, W., Shapiro, O., & Semerci, D. M. (2020, July 10). “Just the Way You Are”: Linking Music Listening on Spotify and Personality. *Social Psychological and Personality Science, 12*(4), pp. 561-572. doi:10.1177/1948550620923228
2. Bickerton, C. (2018, September 21). *"A Beginner's Guide to Decision Tree Classification"*. Retrieved from Towards Data Science: https://towardsdatascience.com/a-beginners-guide-to-decision-tree-classification-6d3209353ea
3. Birnbaum, P. (2021, June 11). *Guide to Sabermetrics Research*. Retrieved from SABR: Society for American Baseball Research: https://sabr.org/sabermetrics
4. Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. doi:10.1145/2939672.2939785
5. Compilation (1908). Anthologie De La Chanson Francaise - L'esprit Mortmartrois. France. Retrieved July 9, 2021, from L'esprit Mortmartrois
6. Dhanaraj, R., & Logan, B. (2005). Automatic Prediction of Hit Songs. *International Society for Music Information Retrieval* (pp. 488-491). London, UK: Queen Mary. Retrieved from https://ismir2005.ismir.net/proceedings/2024.pdf
7. Francis K. C. Hui, D. I. (2015). Tuning Parameter Selection for the Adaptive Lasso Using ERIC. *Journal of the American Statistical Association, 110*(509), 262-269. doi:10.1080/01621459.2014.951444
8. Freund, Y. &. (1995). *A decision theoretic generalization of on-line learning and an application to boosting.* Murray Hill, NJ: AT&T Bell Laboratories.
9. Friedman, J. (2001). Greedy function approximation: a gradient. *Annals of Statistics, 29*(5), 1189-1232.
10. Goldberg, B., Katzen, L., & Jenkins, D. (2021). *Shareholder Letter Q1 2021.* Luxembourg: Spotify. Retrieved from https://s22.q4cdn.com/540910603/files/doc_financials/2021/q1/Shareholder-Letter-Q1-2021_FINAL.pdf
11. Grant, T. (1998). *International DIrectory of Company Histories* (Vol. 23). Detroit, MI, USA: St. James Press. Retrieved from http://www.fundinguniverse.com/company-histories/william-morris-agency-inc-history/
12. Herremans, D., Martens, D., & Sorensen, K. (2014, February). *Dance Hit Song Prediction.* University of Antwerp, Department of Engineering Management. Antwerp, Belgium: Faculty of Applied Economics. Retrieved from http://dorienherremans.com/sites/default/files/wp_hit.pdf
13. *Hit Song Science European Union Trademark Information*. (2021, June 11). Retrieved from TrademarkElite: https://www.trademarkelite.com/europe/trademark/trademark-detail/004876462/Hit-Song-Science

14. Ingham, T. (2021, February 24). *Over 60,000 tracks are now uploaded to Spotify every day*. Retrieved from Music Business Worldwide: https://www.musicbusinessworldwide.com/over-60000-tracks-are-now-uploaded-to-spotify-daily-thats-nearly-one-per-second/
15. Ingham, T. (2021, March 1). *Slowly bur surely, the major labels' dominance of Spotify is declining*. Retrieved from Music Business Worldwide: https://www.musicbusinessworldwide.com/slowly-but-surely-the-major-labels-dominance-of-spotify-is-declining/
16. Iqbal, M. (2021, July 6). *Spotify Revenue and Usage Statistics*. Retrieved from Business of Apps: https://www.businessofapps.com/data/spotify-statistics/
17. Jones, R. (2018, August 16). *HOW SUCCESSFUL ARE SPOTIFY'S 'DIRECT' ARTIST SIGNINGS?* Retrieved from Music Business Worldwide: https://www.musicbusinessworldwide.com/how-successful-are-spotifys-direct-signings/
18. Li, T., Ogihara, M., Tzanetakis, G., & Pachet, F. (2012). *Music Data Mining* (1st ed.). (T. &. Group, Ed.) Boca Raton, FL, USA: CRC Press. Retrieved from https://www.francoispachet.fr/wp-content/uploads/2021/01/pachet-11a.pdf
19. Mayfield, G. (2021, Februaary 10). *As Streaming Dominates the Music World, Is Radio's Signal Fading?* Retrieved from Variety: https://variety.com/2021/music/news/radio-signal-fading-streaming-1234904387/
20. Menten, M., Ng, K., O'Rourke, T., & Holmes, R. (2018, December). Temporal Trends in Music Popularity - A Quantitative analysis of Spotify API data. *Temporal Music Analysis*. doi:10.13140/RG.2.2.11551.71843
21. Middlebrook, K., & Sheik, K. (2019, September 20). SONG HIT PREDICTION: PREDICTING BILLBOARD HITS USING SPOTIFY DATA. *eprint arXiv:1908.08609*. Retrieved from https://arxiv.org/pdf/1908.08609.pdf
22. Mo, S., & Niu, J. (2019, July-Sept 1). A Novel Method Based on OMPGW Method for Feature Extraction in Automatic Music Mood Classification. *IEEE Transactions on Affective Computing, 10*(3), 313-324. doi:10.1109/TAFFC.2017.2724515
23. Navlani, A., & Pandey, P. (2021, March 6). *Support Vector Machine Classification in SciKit Learn*. Retrieved from Machine Learning Geek: https://machinelearninggeek.com/support-vector-machine-classification-in-scikit-learn/
24. Nielson, D. (2016). *Tree Boosting with XGBoost: Why does XGBoost win "every" machine learning competition?* Norwegian University of Science and Technology, Mathematical Sciences. Trondheim, Norway: NTNU. Retrieved July 8, 2021, from https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/2433761/16128_FULLTEXT.pdf
25. Nijkamp, R. (2018). Prediction of product success: explaining song popularity by audio features from Spotify data. Retrieved from http://essay.utwente.nl/75422/1/NIJKAMP_BA_IBA.pdf
26. Pachet, F., & Roy, P. (2008). Hit Song Science Is Not Yet a Science. *International Conference on Music Information Retrieval* (pp. 355-360).

Philadelphia, PA, USA: Drexel University. Retrieved from https://www.cs.swarthmore.edu/~turnbull/cs97/f08/paper/pachet08.pdf

27. Pearce, S. (2016, April 12). *What are the odds of landing a pop hit?* Retrieved from The Fader: https://www.thefader.com/2016/04/12/odds-of-landing-a-pop-hit
28. Pinter, A., Paul, J., Smith, J., & Brubaker, J. (2020). P4KxSpotify: A Dataset of Pitchfork Music Reviews and Spotify Musical Features. *Proceedings of the International AAAI Conference on Web and Social Media*, *14(1)*, pp. 895-902. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/7355
29. Santosa, F., & Symes, W. W. (1986). Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing, 7*(4), 1307-1330. doi:10.1137/0907087
30. Schapire, R. E. (1990). The Strength of Weak Learnability. *MIT Laboratory for Computer Science, Machine Learning, 5*, 197-227. Retrieved from https://link.springer.com/content/pdf/10.1007/BF00116037.pdf
31. Schrage, M. (2019, July 15). *What Baseball Can Teach You About Using Data to Improve Yourself*. Retrieved from Harvard Business Review: https://hbr.org/2019/07/what-baseball-can-teach-you-about-using-data-to-improve-yourself
32. Sciandra, M., & Spera, I. C. (2020, August 10). A model-based approach to Spotify data analysis: a Beta GLMM. *Journal of Applied Statistics*. doi:10.1080/02664763.2020.1803810
33. Sharma, I. (2020). *Hit song classification with audio descriptors and lyrics.* Rutgers, The State University of New Jersey, School of Graduate Studies Electronic Theses and Dissertations. doi:10.7282/t3-14kc-nf60
34. Skidén, P. (2016, March 29). *API Improvements and U*. Retrieved from Spotify: https://developer.spotify.com/community/news/2016/03/29/api-improvements-update/
35. Son, H. (2019, August 1). *Goldman Sachs is spending $100 million to shave milliseconds off stock trades*. Retrieved from CNBC: https://www.cnbc.com/2019/08/01/goldman-spending-100-million-to-shave-milliseconds-off-stock-trades.html
36. T4. (2021, January 23). *T4.AI*. Retrieved from Music Streaming Market Share: https://www.t4.ai/industry/music-streaming-market-share#:~:text=Spotify%20market%20share%20was%2036,%2Dsupporters%20listeners%20(free).
37. Tibshirani, R. (1996). Regression Shrinkage and Selection via the lasso. *Journal of the Royal Statistical Society. Series B (methodological), 58*(1), 267-288. Retrieved July 8, 2021, from https://www.jstor.org/stable/2346178
38. Van Der Somme, S. (2021, January 28). *Automatically Predicting Popularity of Music Tracks Based on Lyrics.* Utrecht University, Faculty of Humanities. Retrieved from https://dspace.library.uu.nl/bitstream/handle/1874/401323/Sander_van_der_Somme_Thesis_6287247.pdf?sequence=1&isAllowed=y
39. Yang, L.-C., Chou, S.-Y., Liu, J.-Y., Yang, Y.-H., & Chen, Y.-A. (2017). Revisiting the problem of audio-based hit song prediction using convolutional neural networks. *2017 IEEE International Conference on Acoustics, Speech*

and Signal Processing (ICASSP), 621-625. doi:10.1109/ICASSP.2017.7952230