HIT SONG PREDICTION BASED ON EARLY ADOPTER DATA AND AUDIO FEATURES

Dorien Herremans
Singapore University of Technology and Design
Information Systems, Technology and Design Pillar
& Institute of High Performance Computing
A*STAR, Singapore
dorien_herremans@sutd.edu.sg

Tom Bergmans
University of Antwerp
Department of Engineering Management

ABSTRACT
Billions of USD are invested in new artists and songs by the music industry every year. This research provides a new strategy for assessing the hit potential of songs, which can help record companies support their investment decisions. A number of models were developed that use both audio data, and a novel feature based on social media listening behaviour. The results show that models based on early adopter behaviour perform well when predicting top 20 dance hits.

1. INTRODUCTION

“I’ve written about 78 top 10 songs, and I still don’t know what a hit song is.” Lamont Dozier, one of the most successful producers behind the Motown label [7]

Given the huge investments in new talent by the music industry [6], the question of predicting hits is extremely relevant. The science of hit song prediction has had a controversial history, as early studies such as [1, 9] showed that random oracles can not always be outperformed when it comes to predicting hits. More recent studies, however, have managed to developed more accurate classifiers for separating hits from non-hits [3, 8]. These successful hit prediction models use various audio features, including temporal features [3], or lyrics [2]. In this study, we not only focus on audio features, but also include social media listening behaviours to identify early adopters.

2. EARLY ADOPTERS

There are people … with some sort of sixth sense for hit songs, such that they listen to hit songs even before they actually climb to the top of the record charts. [11]

In [10], Rogers defines early adopters as individuals with ‘the highest degree of opinion leadership among the other adopter categories’. This group of adopters is typically younger and has a higher social status and is more ‘socially forward’ than others. The idea of early adopters fits in the context of today’s music listening experience and the way hit songs are spread among peers on platforms such as Last.FM. Inspired by preliminary research of [11], the models in this research will use listening behaviour and hit listings information to identify a group of early adopters and embed this knowledge in a model for future hit predictions. In contrast to the original thesis by [11], which only uses 5 songs, we will use a much larger dataset in order to draw more general conclusions. The creation of this dataset is described in the next section.

3. DATASET

3.1 Hits versus non-hits
A list of hit songs was parsed from dance charts listed on the Belgian website ‘The Ultratop 50’. The top 20 songs were used to identify ‘hits’. In order to obtain a list of non-hits, previous research has used low ranking songs from hit lists (e.g. top 30-50 in [3]). In this research, however, we were able to use ‘The Bubbling Under chart’, also from the Ultratop 50 site, which contains a list of 20 upcoming songs identified by music experts. We filtered out any song that eventually hit the charts, and only kept those that never became a hit. The data from 2/07/11 until 16/11/13 was recorded our the dataset, resulting in a total of 8,750 songs, of which 982 were unique.

3.2 Listening behaviour
Listening behaviour data for three Last.FM groups, ‘Consistently New Dance and Electronic’ (2,116 members), ‘Electronic Music’ (3,747 members) and ‘Belgium’ (1,859 members), was parsed using the Last.FM API. Listening data was retrieved for a period of 6 months (16/04/13 – 16/11/13). This data was thoroughly cleaned to compensate for irregular spelling, e.g. ‘featured’ versus ‘ft’. The resulting cleaned dataset consists of 854,060 listening records (time, date, user, song, artist). This dataset was further processed to contains a row for each song at each week (instance), and a column (feature) for each user indicating the number of times they listened to that particular song. A feature called ‘predictive’ was set to 1 if the song was not yet a hit now, but will be in the future.

1 www.ultratop.be/nl/ultratop50

help us predict top 20 hits. 

already clearly show that listening data from Last.FM can be explored in future research. The current values, however, can theoretically improve the performance of the model and should be tried using different kernels. This could potentially drastically improve the model’s performance.

When building models based on social media listening behaviour, overall higher AUC values are reached (see Table 1). The best performing model, reaching an AUC of 0.79, is logistic regression. Currently, no parameter tuning of the models and look at an embedded version model that combines both audio and social listening data to further enhance hit prediction accuracy. In order to advance this field, it would also be beneficial to have an open dataset with fixed hit definitions for benchmarking.

In this study, a large dataset of social listening behaviour gathered through the Last.FM API was created. The dataset also contained hit/non-hit information from the Belgian Dance charts Ultratop 50, and audio data gathered from EchoNest. Based on an early adopter model, the hit prediction models built in this research are able to predict upcoming top 20 hit with an AUC value of 0.79.

In future research, it would be interesting to perform parameter tuning of the models and look at an embedded version model that combines both audio and social listening data to further enhance hit prediction accuracy. In order to advance this field, it would also be beneficial to have an open dataset with fixed hit definitions for benchmarking.

5. CONCLUSIONS

In this study, a large dataset of social listening behaviour gathered through the Last.FM API was created. The dataset also contained hit/non-hit information from the Belgian Dance charts Ultratop 50, and audio data gathered from EchoNest. Based on an early adopter model, the hit prediction models built in this research are able to predict upcoming top 20 hits with an AUC value of 0.79.

In future research, it would be interesting to perform parameter tuning of the models and look at an embedded model that combines both audio and social listening data to further enhance hit prediction accuracy. In order to advance this field, it would also be beneficial to have an open dataset with fixed hit definitions for benchmarking.

6. REFERENCES