사재기 의혹 음원 특징 분석과 순위 예측

정해린, 김도영, 정현정, 김성경, 김현희
동덕여자대학교 정보통계학과
haerin395@gmail.com, rlaehdud159@naver.com, jhj33931787@gmail.com,
su7789@naver.com, heekim@dongduk.ac.kr

Feature analysis and ranking prediction of music suspected of being abused

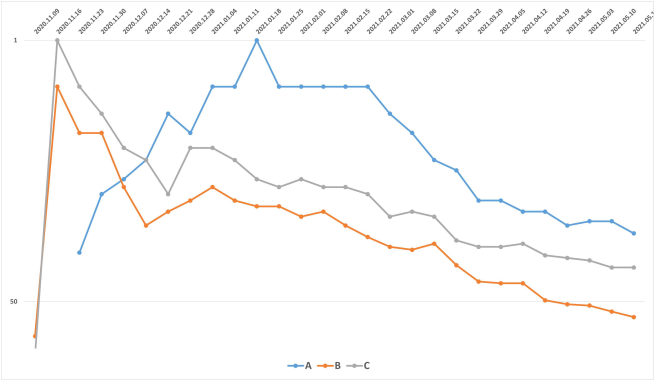
Hae Rin Cheong, Do Young Kim, Hyeon Jeong Jeong, Seong Gyeong Kim,
Hyeon Hee Kim
Department of Statistics and Information Science,
Dongduk Women's University

요 약

온라인 음원 스트리밍 서비스가 확대되면서 음원 사재기가 빈번해지고 있다. 본 논문에서는 사재기로 의심할 수 있는 음원의 특징을 분석하고, 사재기가 이루어지지 않았을 경우의 음원 순위를 예측한다. 그 결과, 랜덤 포레스트를 통해 앨범 평점이 낮은 음원, 장르가 인디나 발라드인 음원, 특정 발매사의 음원일 때 사재기로 의심할 수 있었다. 또한, 딥러닝을 통한 순위 예측 실험 결과, 사재기의 영향으로 실제 순위와 예측 순위에 큰 차이가 있는 것으로 나타났다.

1. 서론

음원 사재기란, 온라인 음원 스트리밍 서비스 차트에서 더 높은 순위를 차지하기 위해 타인의 계정을 도용하여 음원 스트리밍이나 다운로드를 행하는 불법 행위이다. 현재 대부분의 음악 시장이 오프라인에서 온라인으로 전환되었으며, 온라인 음원 스트리밍 서비스 시장은 날로 성장하고 있다.[1] 이러한 온라인 음원 스트리밍 서비스에서 대부분의 비중을 음원 차트가 차지하고 있다. 음원 순위가 중요한 평가 지표로 여겨지면서 사재기 또한 빈번해지고 있다. 사재기는 엄연한 불법 행위이며, 대중을 기만하는 행위이다. 특히 2018년에서 2019년에는 가요계에서 연쇄적으로 사재기 의혹이 제기되었다.[2]



(그림 1) 사재기 의혹 음원의 순위 변동 그래프

그림 1은 2020년 사재기 의혹이 불거진 음원의 A, B, C 음원 스트리밍 서비스 순위 변동 그래프이다. 이 그래프를 보면 해당 음원은 2020년 11월 첫째 주에 B, C 음원사이트에서 순위가 100위 이하였지만, 다음 주인 11월 둘째 주에는 단숨에 상위권으로 올라가는 이상 추이를 보였다.

이러한 이상 추이는 역주행 음원에서도 나타난다. 역주행이란, 과거에 발매된 음원이지만 특정 사유에 의해 이슈가 되면서 음원 차트 상위권에 오르는 현상이다. 비정상적인 추이로 음원 차트 상위권에 오른다는 점에서 사재기와 역주행은 구분하기 어렵다. 이 때문에 사재기 의혹 음원들이 역주행의 탈을 쓰고 시장을 혼란스럽게 하고 있다. 따라서 사재기 의혹이 있는 음원만이 가진 특징을 분석해야 한다.

본 연구에서는 사재기 의혹 음원의 특징을 분석하고 그 음원들에서 사재기 영향을 제거한 정상 최고 순위를 예측한다. 사재기 의혹 음원의 특징을 분석함으로써 역주행 음원과 구분하여 사재기 의혹이 있는 음원을 파악할 수 있다. 또한, 그 음원들로부터 정상적인 최고 순위를 예측함으로써 음원들의 원래 자리를 찾고, 더욱 공정한 차트를 만드는 데 일조할 수 있다.

본 논문은 다음과 같이 구성된다. 제2장에서는 연구에 사용된 데이터의 수집과 전처리 과정을 다룬

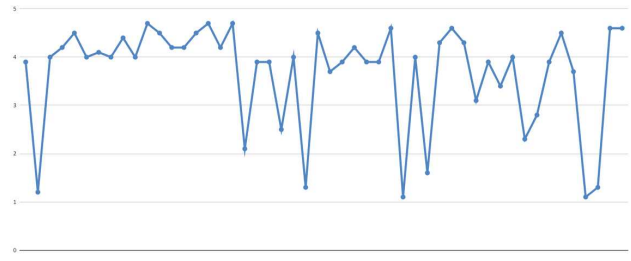
다. 제3장에서는 사재기 의혹 음원에서 사재기 영향을 제거한 최고 순위를 예측하는 모델을 생성한다. 제4장에서는 제3장에서 생성한 모델의 실험 결과와 성능 평가를 기술하며 제5장에서는 결론 및 향후 연구를 제시한다.

2. 데이터 수집 및 전처리

본 연구에서 사용된 데이터는 대표적인 음원 스트리밍 서비스 3사의 2020년부터 2021년 5월까지 주간 차트 데이터로, 크롤링을 진행하였다. 또한, 시장 분석 서비스 '와이즈앱'의 조사에 따라 가장 많은 사용자를 보유한 음원 스트리밍 서비스 A에 대해서는 2018년부터 2021년 6월까지 주간 차트 데이터를 추가로 크롤링하였다. 그 결과, 약 19,200개 음원의 날짜, 순위, 음원 명, 장르 등의 정보를 수집하였다.

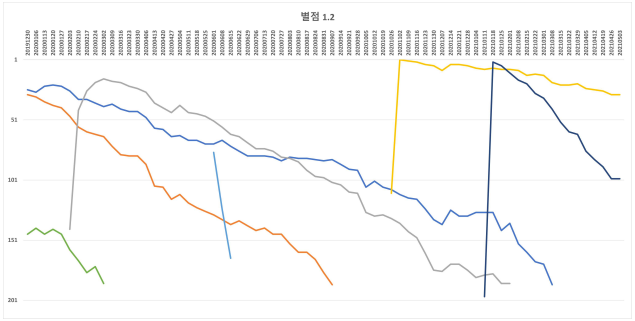
주간 차트 데이터에 대한 전처리는 다음과 같이 진행했다. 앨범 장르에 대해서는 각각 발라드, 댄스, 랩/힙합, R&B/Soul, POP, 인디, 록/메탈, 국내 드라마로 세분화하여 처리하였다. 예를 들어 장르가 발라드인 음원은 발라드 레이블 값을 1로 주고 나머지 7개의 레이블은 0으로 처리하였다. 여러 장르를 포함한 앨범은 해당 레이블 값을 모두 1로 주었다. 앨범 발매일부터 주간 차트 날짜까지의 차이와 수집 대상 기간 동안 해당 음원의 최고 순위를 계산하여 max_rank로 추가하였다.[3]

그림 2는 A사의 2020년부터 2021년 5월까지 주간 차트를 크롤링한 데이터로 순위에 따른 앨범 평점을 그래프로 나타낸 결과이다. 높은 순위를 기록하고 있음에도 눈에 띄게 평점이 낮은 음원의 존재를 확인하였다. 따라서 낮은 평점의 음원을 사재기로 의심할 수 있었다.

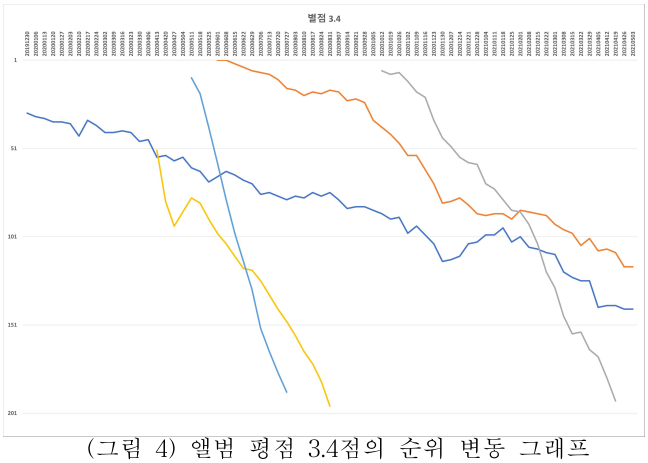


(그림 2) 순위에 따른 평점 그래프 (1위-50위)

그림 3은 C사의 주간 차트 크롤링 결과로 비슷한 평점의 음원을 묶어 주차 별 순위 변동에 대한 그래프를 그린 것이다. 상대적으로 평점이 높은 그림 4와 비교했을 때 그림 4의 음원들은 높은 순위로 진입 후 순위가 하락하는 일반적인 추이를 보이지만, 그림 3의 음원들은 낮은 순위로 진입 후 갑작스레 순위가 상승하는 이상 추이를 확인할 수 있다. 이를 통해 사재기와 앨범 평점 사이의 관계가 있음을 확인했다.

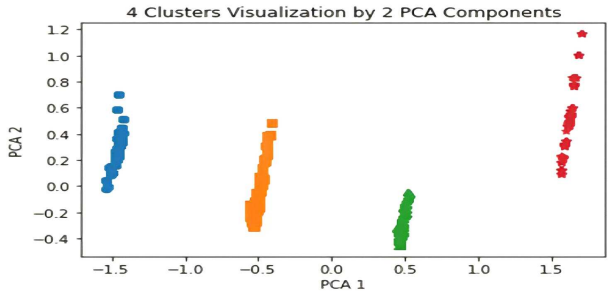


(그림 3) 앨범 평점 1.2점의 순위 변동 그래프



(그림 4) 앨범 평점 3.4점의 순위 변동 그래프

2018년부터 2021년 6월까지의 음원 차트 크롤링 데이터를 군집 분석하였다. k 값을 4로 지정한 후 kmeans 클러스터링을 수행한 결과, 팬층이 두꺼운 그룹, 사재기로 의심되는 그룹 등 4가지 군집으로 뚜렷하게 나누어졌다.[4][5]

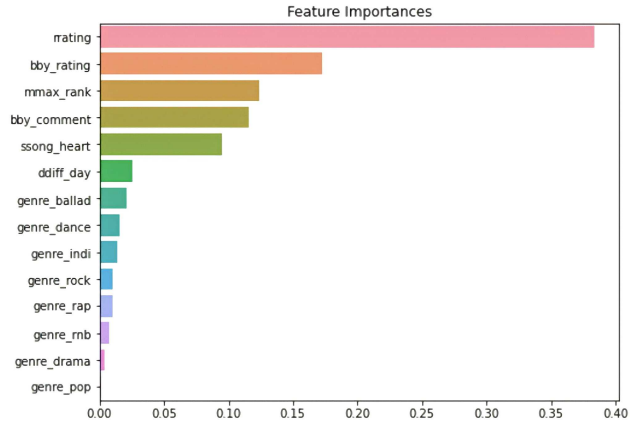


(그림 5) 2018년 군집분석 결과

군집 분석 결과를 통해 1,696개의 음원 중, 사재기 의혹 군집의 158개 음원은 레이블을 1로, 나머지 1,538개의 음원은 0으로 처리하였다. 변수의 측정 단위를 통일하기 위해 MinMaxScaler로 정규화한 후 랜덤 포레스트를 적용해 사재기 판별에 있어서 변수

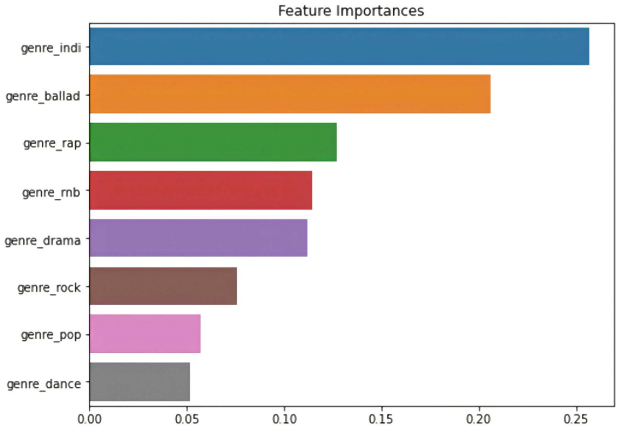
중요도를 분석해보았다.

그림6은 랜덤 포레스트 결과를 나타낸 그래프이다. 랜덤 포레스트 결과 정확도는 0.95이다. 변수 중요도 그래프를 보면 사재기 의혹 음원을 분류하는데 가장 중요한 특징으로 앨범 평점(rrating)이 나타났다. 다음으로 앨범 평점을 매긴 사람 수(bby_rating), 그다음으로 기간 내 최고 순위(mmax_rank)가 중요한 요인이었다. 따라서 앨범 평점이 가장 중요 변수이고, 차례대로 평점을 매긴 사람 수, 기간 내 최고 순위가 사재기를 판별하는 데 있어서 주요 변수로 나타났다.



(그림 6) 랜덤 포레스트 결과

어떤 장르가 사재기 판별에 있어서 중요하게 고려되는지 분석하기 위해 랜덤 포레스트를 적용하였다.



(그림 7) 장르 랜덤 포레스트 결과

그림7은 랜덤 포레스트를 적용한 변수 중요도 그래프이다. 그래프를 보면 가장 중요한 특징으로 장르가 '인디'인 음원(genre_indi)이 나타났고, 그다음으로 장르가 '발라드'인 음원(genre_ballad), 장르가 '랩/힙합'인 음원(genre_rap)이 중요한 변수로 나타났다. 따라서 음원의 장르가 인디와 발라드이면 사재기를 의심할 수 있다.

랜덤 포레스트 결과를 종합해보면, 앨범 평점이 일반적인 음원들에 비해 많이 낮은 음원과 장르가 인디나 발라드인 음원, 그리고 특정 발매사인 경우 사재기로 의심할 수 있다.

3. 모델 생성

전처리한 데이터를 바탕으로 사재기 영향을 제거한 정상적인 최고 순위를 예측하는 모델을 생성하였다. 랜덤 포레스트 결과를 토대로, 앨범 평점이 3점 미만이면서 장르가 인디, 발라드인 음원, 특정 발매사의 음원인 경우 사재기 의혹이 있다고 판단했다. 사재기가 의심되는 음원과 그렇지 않은 음원으로 분류한 후, 사재기 의혹이 없는 음원들을 최고 순위 예측 모델의 train, test 데이터로 사용하였다. 실험 모델로는 딥러닝을 선택하였다. 사재기가 의심되는 음원들은 순위 예측 모델에 적용해 각 음원의 정상 최고 순위를 예측하였다.

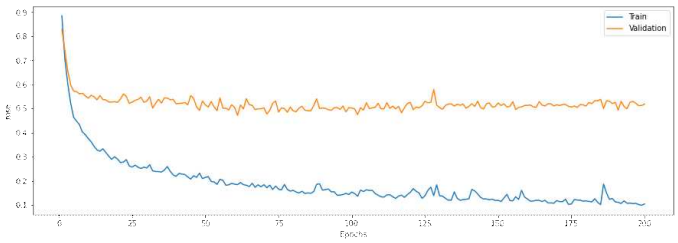
본 연구에서 사용된 딥러닝의 구조는 다음과 같다. 모델의 입력은 음원 정보 데이터를 MinMaxScaler에 적용한 실수 데이터 1538개이고, 모델의 출력은 순위를 예측한 실수 데이터이다. 1538개의 데이터를 2:1로 분할하여 train 값 1030개, test 값 508개를 사용하였다. 입력 데이터는 activation='relu' 이고 노드가 각각 128, 64, 32, 16개인 4개 층을 거친다. 마지막으로 activation='linear'를 사용하여 순위 예측 결과를 도출한다. 성능 평가에는 optimizer='adam', loss='mse', metrix=['mae']를 사용하였다. 또한, batch_size=32로 epochs=200 진행하였다.

실험 환경은 다음과 같다. 구글 Colab의 3.7.13 버전 Python을 사용하였으며, Intel(R) Xeon(R) CPU @ 2.20GHz 환경에서 진행되었다. 사용 라이브러리는 Pandas, NumPy, random, TensorFlow, matplotlib, Scikit-learn이다.

4. 실험 결과

딥러닝 모델로 사재기 의혹 음원의 정상 최고 순위를 예측하였다. 먼저 사재기 의혹이 없는 음원으로 순위 예측 모델을 생성하였다. train, test 값을 모두 표준화하여 모델에 입력하였다. 그림8은 딥러닝 예측 결과 실제 순위와 예측 순위를 비교한 결과이다. 딥러닝의 성능은 loss가 0.40, mae가 0.47로 평가되었다. 그림 8은 딥러닝 실행 결과 성능 그래프

이다.



(그림 8) 딥러닝 성능 그래프

생성한 순위 예측 모델로 사재기 의혹 음원의 정상 최고 순위를 예측하였다.

<표 1>은 딥러닝으로 예측한 사재기 의혹 음원들의 정상 순위 결과 중 일부이다. 음원 (가)는 최고 순위가 1위였지만, 평점, 앨범 좋아요 수 등을 기반으로 한 예측 모델에 적용한 결과, 최고 순위는 52위였다.

<표 1> 사재기 의혹 음원의 최고 순위 예측 결과

음원	최고 순위	예측 결과
(가)	1	52
(나)	6	53
(다)	3	65
(라)	6	62
(마)	1	49

5. 결론

본 연구에서는 랜덤 포레스트를 이용하여 사재기 의혹 음원의 특징을 분석하고, 딥러닝을 이용하여 사재기 영향을 제거한 정상 최고 순위를 예측하였다. 앨범 평점이 낮은 음원, 장르가 인디나 발라드인 음원, 그리고 특정 발매사인 경우 사재기 의혹 음원으로 분류하였다.

딥러닝의 평균 절대 오차는 0.47로 평가되었다. 이는 표준화한 순위의 절대 오차이며, 역 표준화 과정을 거치면 약 48로 계산된다. 음원 차트가 보통 100위, 많게는 200위까지 순위를 매기는 것을 보면 48이라는 오차는 매우 크다. 따라서 현재 음원 차트에서 사재기 영향이 크게 나타나는 것을 확인하였다.

그러나 온라인 스트리밍 서비스에서 차트 반영 기준은 스트리밍과 다운로드 수이다. 따라서 장르, 별점, 댓글 수, 앨범 좋아요 수 등으로만 순위를 예측하기에는 한계가 있었다. 또한, 크롤링이 진행된 시점에서는 이미 사재기 의혹 음원들이 사재기로 높은 순위를 차지한 후, 앨범 좋아요 나 댓글 수 같은 부분에서 사재기 의혹이 없는 음원들과 비슷한 수치를 보였다. 이미 대중들의 관심이 쏠려 사재기 의혹이 없는 음원들과 인기도가 비슷하기 때문에 정확한 순위 예측이 어렵다. 스트리밍과 다운로드 횟수를 고려하여 예측을 진행할 수 있다면 더욱 정확한 예측을 할 수 있으리라 기대한다.

참고문헌

[1] 이민아, “음악산업에 대한 정부개입의 한계와 필요성:음원사재기 현상을 중심으로”, 문화정책논총, Vol. 34 No.1, pp. 5-33, 2020

[2] 김지하, “음원 사재기 브로커를 만났습니다”..기획자의 고백[인터뷰], TV Daily, 2019.11.27, https://entertain.v.daum.net/v/20191127161603210

[3] 이도연, 장병희, ”딥러닝을 이용한 음악흥행 예측모델 개발 연구, 한국 콘텐츠 학회 논문지, Vol. 20 Issue 8, pp. 10-18, 2020

[4] 김평화, ““K팝, 건강한 생태계 돕겠다”…데이터 분석으로 음원 어뷰징 의혹 잡은 블랙멜론“, IT 조선, 2020.01.02, http://it.chosun.com/site/data/html_dir/2020/01/02/20200010200347.html

[5] 유인진, 박도형, “시계열 군집분석을 통한 디지털 음원의 순위 변화 패턴 분류”, Journal of Intelligence and Information Systems, Vol. 26 Issue. 3, pp. 171-191, 2020