

딥러닝을 이용한 음악 흥행 예측모델 개발 연구

A Study on Development of a Prediction Model for Korean Music Box Office Based on Deep Learning

이도연, 장병희
성균관대학교 미디어커뮤니케이션학과

Do-Yeon Lee(lehdus2756@naver.com), Byeng-Hee Chang(mediaboy@skku.edu)

요약

본 연구에서는 콘텐츠 산업 중 음악 분야 2차 산업데이터를 활용하여 딥러닝 기법을 이용한 흥행 예측모델 구축 가능성을 살펴보았다. 본 연구를 통해 구축한 딥러닝 예측 모델은 17개 독립변인 -가수 파워, 가수 영향력, 피쳐링 가수 파워, 피쳐링 가수 영향력, 참여 가수 수, 참여 가수의 성별, 작사가 역량, 작곡가 역량, 편곡가 역량, 제작사 역량, 유통사 역량, 앨범의 타이틀 여부, 음원 스트리밍 플랫폼 좋아요 수, 음원 스트리밍 플랫폼 코멘트 수, 사전 홍보 기사 수, 티저 영상 조회 수, 초기 흥행성과를 기반으로 음원 흥행성과 -음원이 차트 내 상주하는 기간을 예측하는 구조다. 추가적으로 본 연구가 딥러닝 기법을 콘텐츠 분야에 접목시킨 초기단계 연구임을 고려하여, 콘텐츠 흥행예측 선행연구에서 요인 추출을 위해 활용하는 선형회귀분석을 통해 변인 소거 후 구축한 DNN 예측모델과 예측률 비교를 진행하였다.

■ 중심어 : | 음악 흥행 예측 | 음원차트 | 딥러닝 | 심층신경망 | DNN |

Abstract

Among various contents industry, this study especially focused on music industry and tried to develop a prediction model for music box office using deep learning. The deep learning prediction model designed to predict music chart-in period based on 17 variables -singer power, singer influence, featuring singer power, featuring singer influence, number of participating singers, gender of participating singers, lyric writer power, composer power, arranger power, production agency power, distributing agency power, title track, LIKES on streaming platform, comments on streaming platform, pre-promotion article, teaser-video view, first-week performance. Additionally we conducted a linear regression analysis to sort out factors, and tried to compare the prediction performance between the original DNN prediction model and the DNN model made of sorted out factors.

■ keyword : | Music Box Office | Music Chart | Deep Learning | Deep Neural Network | DNN |

I. 서론

기술 발전으로 인한 1인 모바일 기기 사용의 보편화는 소비자들이 온라인 플랫폼을 통해 직접 콘텐츠를 (재)생산, 저장, 공유하는 것을 가능하게 만들며 콘텐츠

산업 생태계 변화를 초래하였다[1]. 다양한 방식으로 콘텐츠 소비가 가능해지면서 영화, 게임, 캐릭터, 애니메이션, 방송, 음악 등 콘텐츠의 영향력은 더욱 커졌고 콘텐츠 산업은 지속적으로 성장하고 있는 추세다[2]. 콘텐츠 산업 생태계의 변화와 소비 패턴의 변화는 기술의

발전과 연관되어 발생할 가능성이 있기 때문에 기술의 발전, 변화하는 환경 등을 고려한 연구는 지속적으로 수행될 필요가 있다.

본 연구에서는 1인 모바일 기기 사용을 통한 시간적, 공간적 제약을 받지 않는 콘텐츠 소비 행태가 다양한 데이터 기록을 남기며 방대한 양의 데이터 분석 가능성을 제공한다는 점에 초점을 맞추었다. 구체적으로 음원 소비와 관련된 다양한 데이터 분석을 통해 소비자들의 소비 패턴을 파악하여 특정 음원의 성적을 예측하는 음악흥행 예측모델을 만드는 데 목적을 둔다. 분석기법 및 예측모델 구축에는 심층신경망 기법을 사용하였다. 컴퓨터 연산 성능의 비약적 발전과 빅데이터 처리에 유용한 함수의 개발로 인공지능망 알고리즘은 다시 주목받게 되었는데, 그 중 본 연구에서 활용하고자 하는 심층신경망은 다중의 은닉층을 포함하며 특히 비선형 관계의 모델링에 강점을 가진다.

인공지능망 알고리즘은 확률분포를 가정하지 않는 최소한의 가정에서 출발하기 때문에 다양한 변인 투입이 가능하며, 알고리즘 스스로 학습하여 변인별 중요도와 가중치를 부여하여 모델을 완성하는 것이 가능하다는 효율성을 가진다. 다만, 본 연구는 콘텐츠 분야에 딥러닝 기법을 접목한 초기 단계 연구이며 체계화된 데이터베이스의 부재에 따른 비교적 적은 양의 데이터 활용의 한계를 고려하여 예측모델 구축을 두 단계에 걸쳐 진행 및 비교하였다. 첫 번째로는 인공지능망 알고리즘 기법의 특성을 고려하여 심층신경망 알고리즘만을 활용하여 예측모델을 구축하였다. 두 번째로는 흥행예측 선행연구에서 활용하는 기법을 통해 유의미한 변인들을 선별하여 예측모델을 구축하였다. 두 단계에 걸쳐 구축된 예측모델들의 성능 비교를 진행하였다.

다양한 콘텐츠 산업 중 음악 산업에 초점을 맞추어 연구를 진행하고자 한 배경은 기존의 음악 산업 연구가 가사 분석, 아이돌 그룹 컨셉 분석 등 사례 분석 분야에서 활발히 진행되었다는 점에 있다. 콘텐츠 산업 중에서도 영화, 게임과 같이 데이터를 활용한 분석 연구를 활발히 진행한 분야에 비해 음악 산업 연구는 제한적인 수준에서 이루어졌기 때문이다. 따라서 본 연구를 통해 음악의 다양한 흥행 요인들을 딥러닝 기법을 통해 종합적으로 분석할 수 있는 체계를 구축하여 음악 산업 분

야 연구의 방법론적 확장에 기여할 수 있을 것으로 판단된다. 또한 음악 분야뿐만 아니라 다양한 콘텐츠 분야 연구의 방법론적 확장에 기여하고자 한다.

II. 이론적 논의

1. 딥러닝

딥러닝(Deep Learning)은 기계학습 방법 중 하나로 기술의 발전에 따른 컴퓨터 연산 성능의 발전과 함수의 개발로 기존 성능 한계를 극복하며 등장한 개념이다. 구체적으로는 데이터의 심층구조 모델을 학습하는데 초점을 맞춘 기법이다[3]. 콘텐츠 산업 분야에서 기계학습 기법 기반 예측을 다룬 선행연구는 예측 연구가 가장 활발히 진행된 영화 분야에서 찾아볼 수 있다. 선행 연구에서는 기계학습 기법을 적용하여 예측하였을 때, 예측 성능이 향상됨을 검증하였다[4]. 본 영화흥행 예측 선행연구에서는 인공지능망 기법을 적용하여 예측모델을 구축하였으나 본 연구에서는 더 나아가 인공지능망에 다중의 은닉층 처리에 유용한 함수를 추가하여 비선형 모델링 성능을 높은 심층신경망을 통한 예측모델 구축을 목표로 한다.

딥러닝 기법의 세부 방법론으로는 Deep Neural Network (DNN), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) 등이 있다. DNN 기법은 수치 데이터와 영상 데이터 모두에 적용할 수 있는 보편적인 방법이다. CNN 기법은 보편적으로 이미지 데이터를 처리하는 데 적용된다. LSTM 기법은 일련의 순차적 데이터(sequential data)의 관계를 학습하는 데 있어 강점을 가진다. 본 연구에서는 다양한 요인별 수치 데이터를 학습하여 하나의 수치를 예측하는 구조의 모델을 구축하고자 하였기 때문에 DNN 기법을 사용하였다. 본 연구에서 구체적으로 DNN에 적용한 함수 및 구조는 이후 연구 방법론에서 다룬다.

2. 브랜드 이론

브랜드 이론은 기술의 발전으로 상품의 질적 차이가 감소하게 되면서 마케팅 분야에서 활발히 적용되는 이론이다. 상품의 질적 차별화 한계를 넘어서기 위해 브

랜드 자산이라는 개념으로 설명할 수 있는 브랜드만의 차별화된 이미지 및 서비스에 집중 하여 소비자들의 소비를 이끌어내는 전략이다[5].

경험재적 성격이 강한 콘텐츠 산업 분야에서는 배우, 작가, 감독 등 인적 자원을 브랜드로 정의한다. 콘텐츠 분야 연구에서는 스타 배우, 스타 작가가 본인이 출연한 영화에 긍정적 영향을 미치는 것을 브랜드 효과로 정의하고 영향 관계를 검증한 선행연구가 존재한다[6]. 흥행 연구에서 스타 파워가 흥행 영향 요인으로 작용할 수 있는 것을 보았을 때 브랜드 효과가 일반 상품뿐 아니라 경험재적 상품에도 적용될 수 있는 것으로 이해할 수 있다. 따라서 본 연구에서 음악 흥행 예측을 위한 변수로 음반 제작에 참여하는 인적 자원을 브랜드로 정의하고 분석하고자 한다.

3. 노출효과 이론

노출효과 이론은 심리학 용어로 지속적인 노출에 의해 긍정적인 정서가 형성되는 현상을 설명한다[7]. 노출효과 이론은 마케팅 전략에 활발히 적용된다. 특히 직접 경험하기 전에 판단이 어려운 경험재적 특성을 가지고 있는 콘텐츠 분야 상품은 소비자들이 미디어를 통해 노출되는 상품의 정보에 기반 하여 소비 의사결정을 하고자 하는 경향이 있다. 따라서 제작자들은 콘텐츠를 시장에 내놓기 전 사전 프로모션을 통한 지속적 노출효과를 활용하고자 하며, 이와 관련하여 사전 프로모션 노출효과가 실제 소비에 미치는 영향관계를 검증하는 콘텐츠 분야 선행연구가 존재한다. 선행 연구에서는 영화 예고편 등 사전 프로모션을 통한 관객 노출 빈도가 영화 관람 동기로 이어지는지 더 나아가 관람 만족도에 영향을 미치는지 검증하였다[8].

음악 상품도 경험재적 특성을 갖고 있기 때문에 영화 예고편과 유사한 사전 프로모션이 존재한다. 음원 발매 일 전 포털 사이트를 통한 홍보 기사 배포, 영상 스트리밍 플랫폼을 통한 티저 영상 및 음원 노출, SNS를 통한 이미지 노출 등이 이에 해당한다. 따라서 본 연구에서 음원 발매 이전에 기사, 영상, 음원 등으로 소비자에게 노출되는 프로모션이 음원 소비에 영향을 미칠 것으로 보고, 흥행 예측을 위한 변수로 정의 및 분석하고자 한다.

4. 편승효과 이론

편승효과는 심리적 요인에 의해 다수의 의견을 받아들여 본인의 결정에 반영하는 개념이다. 연구자들은 편승효과에 의한 선택을 두 가지 경우로 설명한다. 첫째는 소속감이고 둘째는 다수에 의한 품질보증이다[9]. 편승효과 이론을 통해 유행 현상을 설명할 수 있는데, 사람들이 다수에 의해 많이 선택된 상품을 따라서 소비하는 경향 때문에 유행이 일정기간 지속되는 것이다.

경험재적 성격을 가지고 있는 콘텐츠 산업은 '다수에 의한 품질 보증'이라는 측면에서 편승효과의 영향을 받는다. 편승효과를 콘텐츠 소비에 접목시켜 영향관계를 검증한 연구들이 존재하는데 그 중 본 연구에서 집중하는 음악 분야 연구에서도 편승효과를 다룬 연구가 존재한다. 음원 상품의 경험재적 특성으로 사람들은 소비의 불확실성을 줄이기 위하여 많은 사람들이 듣는 음악, 차트 상위에 위치한 음원을 소비하려는 경향을 선행 연구에서 검증하였다[10]. 또한 불확실성을 줄이기 위해 다른 유경험자의 평가를 참고해 소비하고자 하는 경향, 음악의 경우 음원사이트 코멘트 수가 최종 흥행성과에 미치는 영향을 검증한 선행 연구가 존재한다[11]. 이처럼 콘텐츠의 경험재적 특성으로 인한 불확실성을 줄이기 위해 다수의 의견을 참고하는 소비자들의 행동을 흥행 영향 요인으로 정의하고 분석하고자 한다.

5. 선행 연구

음악분야 흥행예측 선행연구에서는 앞서 다룬 이론적 배경에 기반 하여 다양한 흥행 요인들을 정의하고 영향관계를 검증하였다. 본 연구에서는 선행연구에서 검증된 변인들을 추출하여 예측모델 구축을 위한 변인으로 투입하고자 하였다.

구체적으로 '음악 콘텐츠의 흥행 요인이 K-POP의 흥행성과에 미치는 영향' 연구에서는 가수 역량, 작사가 역량, 작곡가 역량, 편곡가 역량, 기획사 역량, 배급사 역량, 방송출연 횟수, 신문 기사 수, OST 영향, SNS 영향 등을 흥행요인으로 정의 및 검증하였다[12]. 본 선행 연구 결과를 바탕으로 가수파워, 가수 영향력, 피처링 가수 파워, 피처링 가수 영향력, 작사가 역량, 작곡가 역량, 편곡가 역량, 기획사 역량, 배급사 역량 변인을 브랜드 이론에 근거하여 추출하였고, 사전 기사 수 변인을

노출효과 이론에 근거하여 추출하였다.

‘대중음악 음원의 흥행요인에 관한 연구’에서는 가수 성별, 스타파워, 장르, 타이틀곡 여부, 제작사 역량, 티저 영상 조회 수, 사전 홍보 기사 수, 음원사이트 코멘트 수 등이 음원 발매 초반 성과에 영향을 미치며 음원 발매 초반 성과가 장기적인 음원 흥행성과에 영향을 미치는 것을 검증하였다[11]. 본 선행연구 결과를 바탕으로 가수 관련 하위 요소로 가수 성별 변인을 추출하였으며 추가적으로 타이틀곡 여부, 티저 영상 조회 수 변인은 노출효과 이론에 근거하여 추출, 음원사이트 코멘트 수, 음원 발매 초반 성과 변인은 편승효과 이론에 근거하여 추출하였다. 음원사이트 코멘트 수와 유사하게 편승효과 이론에 근거하여 잠재적 경험자에게 유경험자의 평가가 정보를 제공해 줄 수 있는 요인으로 음원사이트 좋아요 수를 추가하였다.

‘중국 음악시장 내 K-POP의 흥행요인에 관한 연구’에서는 특히 중국의 음원 차트 성과 데이터를 통해 참여 가수들의 성별, 가수 타입(그룹/솔로), 가수 파워, 음악 장르, 앨범 타이틀 여부, 제작사 역량, 뮤직비디오 존재 여부, 스트리밍 플랫폼 내 코멘트 수, SNS 존재 여부 등이 흥행성과에 미치는 영향을 검증하였다[13]. 본 선행연구 결과를 바탕으로 추가적으로 가수 관련 하위 요소로 가수 타입 변인을 추출하였다. 최적화된 예측모델 구축을 위한 더미 변수 투입 최소화 목적으로 가수 타입 변인은 참여 가수 수 변인으로 변경하여 투입하였다.

III. 연구 방법

1. 데이터 수집

본 연구는 예측 모델을 구축하기 위하여 ‘가온 차트’ 기반 데이터 수집을 진행하였다. ‘가온 차트’를 본 연구 데이터 구축 플랫폼으로 선정한 이유는 ‘가온 차트’가 국내 주요 음악 웹사이트와 이동통신사에서 제공하는 데이터에 가중치를 적용하여 순위를 계산한다는 점, 또한 국내 음반 및 음원 매출의 97% 이상에 해당하는 데이터를 구축하고 있다는 점에서 연구를 위한 데이터 수집 플랫폼으로의 타당성을 판단할 수 있기 때문이다

[14][15]. 또한 선행연구를 통해 ‘가온 차트’가 세계적으로 공신력 있는 차트들 - 미국의 빌보드 차트, 일본의 오리콘 차트, 영국의 UK 차트 - 만큼 공신력 있는 차트로 평가받음을 알 수 있으며, 음악 분야 연구를 위한 데이터 수집 플랫폼으로 많이 활용되는 것을 살펴볼 수 있다[11][12].

본 연구에서는 구체적으로 스트리밍, 다운로드, BGM 판매량 등에 가중치를 부여하여 집계된 순위인 ‘가온 디지털 차트’에 기반 하여 2018년 1월 1일부터 2019년 6월 30일까지 총 78주 동안 주간차트 100위에 진입한 곡들을 수집하였다. 전체 7,800개의 데이터 중 음원차트 특성 상 장기간 음원차트에 상주하는 중복 음원 데이터 제거 및 코딩 불가 데이터를 제거하고 총 776개의 데이터를 예측 모델 학습 데이터로 선정하였다. 음악 예측 모델을 구축하기 위한 구조화된 데이터 기반의 부재로, ‘가온 디지털 차트’를 기반으로 음원 흥행 성적 데이터의 틀을 마련하고 각각의 흥행 영향 요인들의 수치는 각 변수별로 플랫폼을 통해 직접 수집하였다. 각 변수별 수집은 가수 파워, 피처링 가수 파워, 작사가 역량, 작곡가 역량, 편곡가 역량, 제작사 역량, 유통사 역량, 초기 흥행 성과, 음원 흥행성과 변인은 ‘가온 차트’ 플랫폼에서 수집을 진행하였으며, 가수 영향력, 피처링 가수 영향력 변인은 SNS 플랫폼 Instagram에서 수집을 진행하였다. 참여 가수 수, 참여 가수 성별, 앨범 타이틀 여부, 음원 스트리밍 플랫폼 좋아요 개수, 음원 스트리밍 플랫폼 코멘트 개수 변인은 음원 스트리밍 플랫폼 Melon에서 수집을 진행하였으며, 사전 홍보 기사 수 변인은 포털사이트 Naver, 티저 영상 조회 수 변인은 YouTube 플랫폼에서 제공되는 수치를 수집하였다. 각 변인별 조작적 정의는 [표 1]에서 요약 제시하였다.

표 1. 변인 조작적 정의

유형	변수명	조작적 정의
DV	음원 흥행성과	가온 차트 상주 기간
IV	가수 파워	가수 참여 음원 중 자료 수집 기간 내 차트-인 한 음원 수
	가수 영향력	가수 Instagram 팔로워 수
	피처링 가수 파워	피처링 가수 참여 음원 중 자료 수집 기간 내 차트-인 한 음원 수
	피처링 가수 영향력	피처링 가수 Instagram 팔로워 수
	참여 가수 수	음원 참여 가수 인원 수
	참가 가수 성별	참가 가수 성별: 남/여/혼성 (Dummy)

작사가 역량	음반 제작 참여자 작업 음원 중 자료 수집 기간 내에 차트-인 한 음원 수
작곡가 역량	
편곡가 역량	
제작사 역량	
유통사 역량	
앨범의 타이틀 여부	음원의 앨범 타이틀 여부 (Dummy)
음원 스트리밍 플랫폼 좋아요 수	개별 음원의 음원 스트리밍 플랫폼 내 좋아요 수
음원 스트리밍 플랫폼 댓글 수	개별 음원의 음원 스트리밍 플랫폼 내 댓글 수
사전 홍보기사 수	음원 발매 2주 전부터 포털사이트 통해 보도된 사전 홍보기사 수
티저 영상 조회 수	음반 기획사 공식 티저 영상 조회 수
초기 흥행성과	음원 발매 첫 주 '가온 차트' TOP 10 진입 여부 (Dummy)

2. 데이터 전처리

예측 모델을 학습시키기 위해 활용된 데이터는 IV(독립변수)_data(776, 64), DY(종속변수)_data(776, 1)로 최종 설정되었다. 구체적으로 딥러닝 모델 구축 및 모델 성능 분석을 위해 776개 표본 중 90%인 699개의 데이터 셋을 train_data에 활용하였고, 10%인 78개의 데이터 셋을 test_data에 활용하였다.

776개 표본이 각각 64개의 수치가 투입되고 1개의 수치를 예측하는 유형으로 설정된 것은 각 독립변인별 코딩 된 수치가 1개 이상이기 때문이다. 구체적으로 가수 파워, 가수 영향력 하위 수치가 각 6개(곡 참여 가수 인원 최대 6인), 피쳐링 가수 파워, 피쳐링 가수 영향력 하위 수치가 각 5개(곡 참여 피쳐링 가수 인원 최대 5인), 참여 가수 하위 수치 1개, 참여 가수 성별 하위 수치 3개(남성/여성/혼성), 작사가 역량 하위 수치 12개(곡 참여 작사가 인원 최대 12인), 작곡가 역량 하위 수치 12개(곡 참여 작곡가 인원 최대 12인), 편곡가 역량 하위 수치 6개(곡 참여 편곡가 인원 최대 6인), 제작사 파워/유통사 파워/타이틀곡 여부/음원 사이트 좋아요 수/음원 사이트 댓글 수/사전 홍보 기사 수/티저 영상 조회 수/초기 흥행성과 하위 수치 각 1개로 총 64개의 수치가 투입되며, 최종적으로 예측하고자 하는 수치는 음원 흥행성과 1개로 구성되기 때문이다. 이에 따라 본 딥러닝 예측 모델은 17개 독립 요인; 64개 개별 수치를 통해 1개의 종속변수를 예측하는 구조로 이루어진다.

3. 심층 신경망(Deep Neural Network)

앞선 이론적 논의에서 다룬 내용과 동일하게, 본 연

구에서는 다양한 요인별 수치 데이터를 학습하여 하나의 수치를 예측하고자 하는 구조의 모델을 구축하는 것이 목표이기 때문에 딥러닝 기법의 세부 방법론 중 수치 데이터와 영상 데이터 모두에 적용할 수 있는 보편적인 방법인 DNN 기법을 활용하였다. DNN은 입력층과 출력층 사이에 여러 개의 은닉층으로 구성된 심층 신경망이다. DNN의 은닉층은 복잡한 비선형 관계의 모델링에 강점을 가진다. [그림 1]은 DNN의 구조를 보여준다.

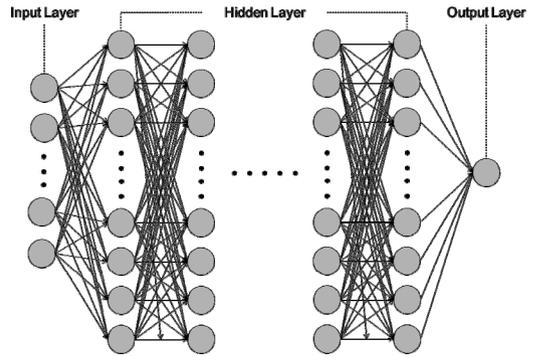


그림 1. DNN 구조

본 연구에서는 구조화된 데이터베이스의 부재라는 한계로 심층 신경망을 학습하고 시험하는데 상대적으로 적은 양의 데이터가 활용되었다. 소량의 데이터 활용은 신경망 과적합의 위험성이 높기 때문에 탐색적 접근을 위해 노드 수와 레이어 수를 적게 설정하는 것이 필요하다고 판단하였다. 따라서 본 연구의 신경망은 각각 노드 수를 달리하고 'Relu' 함수를 적용한 6개의 은닉층으로 구성하였으며, 마지막 출력층은 1개의 노드로 구성하였다. model.compile을 위한 함수로는 각각 optimizer는 Adam(lr=0.001)을 적용하였고, loss function은 mean squared error(mse)를 활용하였다. 과대적합 방지를 위하여 Dropout=0.5 함수를 추가하였다.

$$[\text{Adam}] \quad w_i^t = w_i^{t-1} - \frac{n}{\sqrt{\hat{G}_i^t + \epsilon}} \hat{m}_i^t \quad (1)$$

$$[\text{MSE}] \quad E = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2 \quad (2)$$

IV. 결과

본 연구에서는 2018년 1월 1일부터 2019년 6월 30일까지 국내 음원 차트 '가온 디지털 차트' 2차 산업 데이터 776개 표본을 활용하여 음악 흐행 예측 모델을 구축하였다. 콘텐츠 분야 흐행예측 연구에서 기계학습 분석기법을 적용한 연구가 많이 진행되지 않아 탐색적 수준에서 예측모델 구축을 시도하는 초기연구이기 때문에 두 단계에 걸쳐서 예측모델 성능을 도출하였다. 첫 번째로, 수집된 데이터들을 소거 없이 DNN 모델에 투입하여 예측모델을 구축한 후 성능을 측정하였다. 두 번째로, 흐행예측 선행 연구에서 사용하는 분석기법 선형회귀분석을 실시하여 유의확률을 확인한 후 유의미한 변인들만 DNN 모델에 투입하여 예측모델을 구축한 후 성능을 측정을 추가로 진행하였다. 두 단계 모두 DNN 예측모델 성능을 측정하기 위해 학습에 사용되지 않은 10%의 데이터 셋을 추출하여 검증하였다.

예측률 도출을 통한 DNN 모델 성능 측정에는 예측하고자 하는 변인이 '음원 차트 내 상주 기간'으로 연속형 데이터임을 고려하여 예측 데이터 값이 실제 데이터 값의 오차범위 내에 존재하는지 여부로 예측률을 계산하였다.

1. DNN 모델 구축

수집된 데이터들을 소거 없이 DNN 모델에 투입한 결과 DNN 예측 모델은 오차범위 10% 내에서 10% 수준의 예측률을 보였으며, 오차범위 20% 내에서 16% 수준의 예측률을, 오차범위 30% 내에서 21% 수준의 예측률을 보였다. [그림 2]는 오차범위에 따른 DNN 예측모델의 성능 증가 구조를 나타낸다.

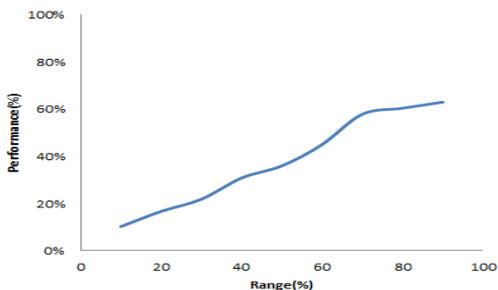


그림 2. 오차범위에 따른 DNN 예측모델 성능

2. 변인 소거 후 DNN 모델 구축

본 연구가 콘텐츠 분야 흐행예측 연구에서 초기 연구로 탐색적 수준에서 진행되는 만큼 흐행예측 선행연구에서 변인들의 영향관계를 도출하기 위해 수행하는 선형회귀분석을 선행한 후 유의확률에 기반 하여 투입 될 변인들을 선별하였다. [표 2]는 선형회귀분석을 실시한 결과이며 [표 3]은 선형회귀분석 결과에 따라 유의미하지 않은 것으로 나타난 변인들을 소거한 후 선별된 변인들을 나타낸다.

표 2. 선형회귀분석 결과

변인	유의확률	VIF
(상수)	.011*	-
가수 파워	.002**	2.704
가수 영향력	.027*	1.190
피처링 가수 파워	.368	1.447
피처링 가수 영향력	.745	1.510
참여 가수 수	.033	1.481
참여 가수 성별	.403	1.128
작사가 역량	.149	7.466
작곡가 역량	.782	7.684
평가역량	.191	1.118
기획사 역량	.228	1.243
배급사 역량	.124	1.264
앨범의 타이틀 여부	.238	1.555
음원 스트리밍 플랫폼 좋아요 수	.000***	2.281
음원 스트리밍 플랫폼 코멘트 수	.000***	2.031
사전 홍보 기사 수	.491	1.293
티저 영상 조회 수	.028*	1.367
초기 흐행 성과	.000***	1.368

R² = .754 Adjusted R² = .749 F = 136.561 p = .000

표 3. 선형회귀분석 결과 투입될 변인

변인	DNN
가수 파워	■
가수 영향력	■
피처링 가수 파워	
피처링 가수 영향력	
참여 가수 수	
참여 가수 성별	
작사가 역량	
작곡가 역량	
평가역량	
기획사 역량	
배급사 역량	
앨범의 타이틀 여부	
음원 스트리밍 플랫폼 좋아요 수	■
음원 스트리밍 플랫폼 코멘트 수	■
사전 홍보 기사 수	
티저 영상 조회 수	■
초기 흐행 성과	■

선형회귀분석 결과에 따라 유의확률 $p < .05$ 수준에서 유의한 것으로 검증된 변수들만 추출하여 DNN 예측모델 구축을 위해 투입하였다. 결과적으로 오차범위 10%에서 12% 수준의 예측률을, 오차범위 20%에서 20% 수준의 예측률을, 오차범위 30%에서 34% 수준의 예측률을 보였다. [그림 3]은 오차범위에 따른 변인 소거 후 구축한 DNN 모델 성능 증가 구조를 나타낸다.

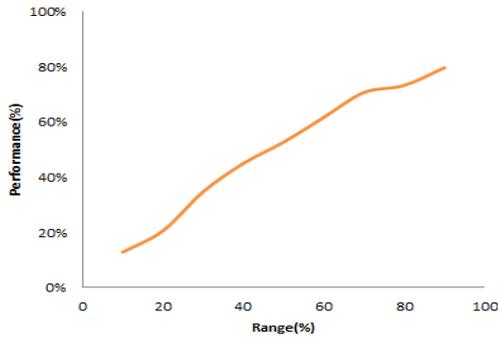


그림 3. 오차범위에 따른 유의미한 변인 투입 DNN 예측모델 성능

결과적으로 변인 소거 없이 전체 변인을 DNN 모델에 투입하여 구축한 예측모델 성능에 비해 선형회귀분석을 선행한 후 유의미하지 않은 것으로 확인된 변인 소거 후 투입한 예측모델의 성능이 상대적으로 우수하게 나타났다. [그림 4]은 각 단계별 도출된 오차범위에 따른 DNN 모델 성능 차이를 나타낸다. 각 단계별 도출된 오차범위에 따른 예측률 증가는 다음 [표 4]와 같다.

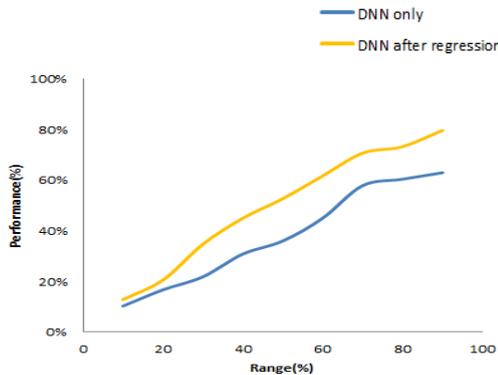


그림 4. 오차범위에 따른 예측모델 성능 비교

표 4. 오차범위에 따른 DNN 예측률 증가

오차범위	DNN 모델 예측률	유의미한 변인 투입 DNN 모델 예측률
±10%	10.256%	12.821%
±20%	16.667%	20.513%
±30%	21.795%	34.615%
±40%	30.769%	44.872%
±50%	35.897%	52.564%
±60%	44.872%	61.538%
±70%	57.692%	70.513%
±80%	60.256%	73.077%
±90%	62.821%	79.488%

V. 결론

본 논문에서는 딥러닝 기법을 음원 차트 흥행 예측에 적용하였다. 총 17개의 독립 변수를 투입하여 종속 변수인 음원 차트 내 상주 기간을 예측하는 구조로 설계하였다. 다만, 본 연구가 탐색적인 수준에서 진행되는 연구이기 때문에 선행연구에서 수행하는 선형회귀분석 기법을 추가로 접목하여 구축한 예측모델과 비교분석을 진행하였다. 딥러닝 기법만을 활용해 구축한 예측모델은 인공신경망 기법 특성상 변인소거 없이 전체 투입하여 구축하였으며, 선형회귀분석 기법을 접목하여 구축한 예측모델은 선형회귀분석 결과를 통해 유의미한 변인으로 확인된 변인들을 추출하여 딥러닝 예측모델에 투입하여 구축하였다.

결과적으로 변인 소거 없이 DNN 모델에 전체 변인을 투입하여 구축한 예측모델에 비해 선형회귀분석을 선행한 후 유의미한 것으로 확인된 변인들만 선별하여 투입한 DNN 예측모델의 성능이 더 우수한 것으로 나타났다. 이러한 결과는 인공신경망 분석이 확률분포를 가정하지 않는 최소한의 가정에서 출발하기 때문에 변인 소거 없이 최대한 많은 변인 투입이 가능하며 모델이 스스로 학습하여 최적화가 가능하다는 개념에 반하는 결과이다. 이는 다른 분야에서 활용되는 딥러닝 사례들을 살펴보았을 때 10만개 이상에 달하는 방대한 양의 표본 데이터를 다룰 때는 모델 스스로 학습이 가능하나, 본 연구에서 다룬 데이터는 기존에 구축되어 있는 데이터베이스의 부재로 상대적으로 데이터양이 부족했기 때문에 발생한 특이 사례이자 한계로 판단된다. 콘텐츠 분야 연구의 경우 방대한 양의 데이터베이

스가 구축되어 있지 않기 때문에 선행연구에서 활용하는 분석기법을 통해 변인들을 선별한 후 딥러닝 모델에 투입하는 방안이 초기단계 연구에서는 필요할 것으로 판단된다.

콘텐츠 분야에서 딥러닝 기반 접근방식의 기존 연구가 많지 않기 때문에, 본 연구는 딥러닝에 기반 한 문화산업 콘텐츠 분야 연구로서 탐색적 차원에서 행해졌다. 이러한 점에서 본 연구는 다음과 같은 함의점을 가진다. 첫째, 학술적 측면에서 본 연구는 콘텐츠 분야 연구의 폭을 넓혔다. 딥러닝 기법은 데이터가 풍부한 시대에 적합하며 주목할 만한 가치가 있는 기법이다. 하지만 콘텐츠 산업 관련 연구에는 딥러닝 기법 관련 접근이 부족한 실정이다. 따라서 본 논문은 콘텐츠 산업 분야 연구의 방법론적 확장에 기여했다고 판단된다. 둘째, 산업적 측면에서 본 연구는 음악 산업 프로세스의 전략에 기반이 될 수 있다. 딥러닝 기반 예측 모델을 통해 음악 산업 실무자들은 음원 흥행을 위해 제작 과정에 어떠한 요인들을 추가하고 고려할 지 결정할 수 있을 것이다. 높은 제작비의 리스크를 줄일 수 있는 근거로 작용할 가능성도 있다. 추가적으로, 본 연구를 통해 구축된 시스템이 더 많은 데이터 축적을 통해 체계화된다면 트렌드의 움직임을 사전에 파악할 수 있는 실직적인 척도가 될 수 있다.

하지만 본 연구는 탐색적인 수준에서 이루어졌기 때문에 향후 고려하고 발전시켜야 할 한계점이 있다. 본 연구에서 활용한 데이터의 경우 체계적으로 구축된 데이터베이스의 부재로 직접 플랫폼별로 수집을 진행했기 때문에 상대적으로 적은 양의 데이터를 활용했다는 점에서 한계점을 가진다. 심층 신경망 모델의 경우 학습을 위해 투입되는 데이터의 양이 증가할수록 모델 최적화와 모델 성능 향상에 효율적이기 때문에 투입되는 데이터양이 중요하다. 본 연구에서는 이러한 한계점을 선행연구에서 활용하는 선형회귀 분석 기법을 접목시켜 유의미한 것으로 선별된 변인들을 투입함으로써 극복하고자 하였다. 향후 콘텐츠 분야의 경우 심층 신경망 기법을 적용시킨 연구의 활발한 진행을 위해서는 데이터 활용의 효율성을 위해 체계적으로 축적이 가능한 데이터베이스 구축이 필요하다고 판단된다.

참고 문헌

- [1] 최정은, 김면, “디지털시대의 문화예술콘텐츠 플랫폼에 관한 연구 -구글 아트앤컬처를 중심으로-,” 기초조형학연구, 제20권, 제3호, pp.433-446, 2019.
- [2] 한국콘텐츠진흥원, 2019 콘텐츠산업 통계조사 보고서 (ISSN 2287-9102), 2019.
- [3] R. B. Palm, *Prediction as a candidate for learning deep hierarchical models of data*, Thechnical University of Denmark, MSc thesis, 2012.
- [4] 권신혜, 박경우, 장병희, “기계학습 기반의 영화흥행예측 방법 비교: 인공신경망과 의사결정나무를 중심으로,” 인문사회과학기술융합연구회, 제7권, 제4호, pp.593-601, 2017.
- [5] K. L. Keller, "Conceptualizing, measuring, and managing customer-based brand equity," *The Journal of Marketing*, Vol.57, No.1, pp.1-22, 1993.
- [6] A. M. Levin, I. P. Levin, and C. E. Heath, *Movie stars and authors as brand names: Measuring brand equity in experiential products*, ACR North American Advances, 1997.
- [7] A. H. Maslow, "The influence of familiarization on preference," *Journal of Experimental Psychology*, Vol.21, No.2, p.162, 1937.
- [8] 최배석, "영화 프로모션 노출이 영화 관람 만족도에 미치는 영향," 한국콘텐츠학회논문지, 제13권, 제12호, pp.92-107, 2013.
- [9] E. Van Herpen, R. Pieters, and M. Zeelenberg, "When demand accelerates demand: Trailing the bandwagon," *Journal of Consumer Psychology*, Vol.19, No.3, pp.302-312, 2009.
- [10] 임성준, 윤문수, "음반 제작주체의 역량 및 제작시스템이 흥행에 미치는 영향에 관한 연구," 전략경영연구, 제11권, 제2호, pp.131-154, 2008.
- [11] 첨애연, *대중음악 음원의 흥행요인에 관한 연구*, 성균관대학교, 석사학위논문, 2018.
- [12] 김양석, *음악 콘텐츠의 흥행요인이 K-Pop의 흥행성과에 미치는 영향*, 호서대학교, 박사학위논문, 2015.
- [13] 왕잠력, *중국 음악시장 내 K-pop의 흥행요인에 관한 연구*, 중앙대학교, 석사학위논문, 2019.
- [14] 현창민, *디지털 음원의 흥행요인에 관한 연구*, 중앙

대학교, 석사학위논문, 2014.

- [15] 김민기, 고건혁, 조슬아, “소셜웹 시대의 미디어 마케팅 전략: 디지털 음악 콘텐츠를 중심으로,” *Entrue Journal of Information Technology*, 제12권, 제1호, pp.45-58, 2013.

저 자 소 개

이 도 연(Do-Yeon Lee)

준회원



- 2018년 8월 : 성균관대학교 독어독문학과, 신문방송학과(학사)
- 2018년 9월 ~ 현재 : 성균관대학교 미디어커뮤니케이션학과 석사과정

〈관심분야〉 : 문화콘텐츠, AI, 가상현실

장 병 희(Byeng-Hee Chang)

정회원



- 2001년 5월 : 미시건주립대학교 텔레컴학과(석사)
- 2005년 4월 : 플로리다대학교 매스커뮤니케이션학과(박사)
- 2006년 3월 ~ 현재 : 성균관대학교 미디어커뮤니케이션학과 교수

〈관심분야〉 : 미디어산업, 문화콘텐츠