산업융합연구 제20권 제4호 pp. 31-37, 2022
ISSN 2635-8875 / e-ISSN 2672-0124
DOI : https://doi.org/10.22678/JIC.2022.20.4.031

디지털 빅데이터를 이용한 영상컨텐츠 수요예측모형 개발

송민구
예원예술대학교 교양학부 교수

Development of Demand Prediction Model for Video Contents Using Digital Big Data

Min-Gu Song
Professor, Faculty of Liberal Arts, Yewon Arts University

요 약 영화 시장에서 흥행을 기록하는데 어떤 요인들이 영향을 미치는지에 대한 연구는 관련 산업의 리스크를 줄이고 영화 산업을 발전시키는데 매우 중요하다. 본 연구에서는 영화흥행에 영향이 있는 독립변수들의 상관의 정도를 찾아내기 위해서 먼저 AHP 기법을 이용한 영화전문가들에 대한 설문조사를 실시하여 측정요인별 중요도를 평가하였다. 또한, 스마트폰 보급과 사용의 증가로 검색 포털 및 SNS 관련 빅데이터에서 도출된 요인이 영화흥행에 영향을 미칠 것이라는 가설을 설정하였다. 그리고 앞에서 언급한 전문가 서베이 정보와 빅데이터를 모두 반영한 예측모형을 제안하였다. 제안한 모형의 예측의 정확도를 알아보기 위해 실 데이터를 가지고 검증한 결과 기존모형보다 향상됨(10.5%)을 확인하였다. 따라서 제안한 모형은 영화제작사 및 배급사들의 의사 결정에 도움이 될 것이라 판단된다.

키워드 : AHP 모형, 영화흥행예측, 포털 검색량, 구전효과, SNS 빅데이터, 분류행렬표

Abstract Research on what factors affect the success of the movie market is very important for reducing risks in related industries and developing the movie industry. In this study, in order to find out the degree of correlation of independent variables that affect movie performance, a survey was conducted on film experts using the AHP method and the importance of each measurement factor was evaluated. In addition, we hypothesized that factors derived from big data related to search portals and SNS will affect the success of movies due to the increase in the spread and use of smart phones. And a prediction model that reflects both the expert survey information and big data mentioned above was proposed. In order to check the accuracy of the prediction of the proposed model, it was confirmed that it was improved (10.5%) compared to the existing model as a result of verification with real data.Therefore, it is judged that the proposed model will be helpful in decision-making of film production companies and distributors.

Key Words : AHP Model, Prediction of Movie Ticket Sales, Portal Search Volume, WOM Effect, SNS Bigdata, Classification Matrix

1. 서론

미국의 영상 콘텐츠를 공급하는 넷 플릭스는 자체 제작 콘텐츠인 '하우스 오브 카드'를 만들 때 빅데이터를 활용하여 이 영상콘텐츠 작품의 소비자를 비교적 신뢰성 있게 예측함으로써 자사의 가입자, 매출, 주가 등에서 가파른 성장을 주도하였다. 이것은 문화예술 콘텐츠 분야에서 빅데이터 활용의 중요성을 보여준 사례라 할 수 있다. 문화예술 콘텐츠 중에서 영상산업은 수요의 불확실성이 가장 높고, 생산과정에서 이미 투입되어 회수할 수 없는 매몰비용의 비중이 매우 크다는 특징이 있다

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea(NRF-2019S1A5A8037916).
*Corresponding Author : Min-Gu Song(minsong3@naver.com)
Received March 31, 2022 Revised April 14, 2022
Accepted April 20, 2022 Published April 28, 2022

[1]. 이러한 사실은 영상 분야 콘텐츠 산업의 발전에 걸림돌이 되고 있다. 따라서 이것을 해결하기 위해서는 영상콘텐츠 제작 초기 단계에서 소비 수요를 예측하여 제작에 소요되는 비용의 리스크를 최소화하는 즉 흥행 실패로 인한 리스크 줄여야 한다.

본 연구에서는 영화흥행 수요예측에 영향을 미치는 지금까지의 흥행 요인들을 무엇인지를 선행연구를 통해 알아보고 그것의 문제점과 대안을 제시하고자 한다. 선행 연구를 살펴보면 스마트폰 보급의 확산 전과 후에 분명히 영화흥행 수요예측에 영향을 주는 변수가 차이가 있음을 발견할 수 있다. 특히 스마트폰의 확산으로 포털 검색량의 증가와 SNS사용의 증가로 관련 빅데이터가 영화흥행 예측에 영향을 줄 것이라는 가설을 수립할 수 있다. 이러한 가설을 규명하기 위해서 먼저 영화분야에 오랫동안 종사한 전문가들의 경험적 지식을 파악하여 활용하고자 한다. 델파이 및 AHP 조사 기법을 사용하여 전문가들에게 설문조사를 실시하고 그것을 분석함으로써 그들의 객관화된 경험적 지식의 확보가 가능하다. 확보된 경험적 지식과 실증데이터를 모두 활용하여 흥행 예측 모형을 구축하고자 한다. 즉 실증데이터를 탐색적 자료 분석을(Exploratory Data Analysis) 및 AHP 조사 기법을 통해 얻은 측정 변수들의 중요도를 함께 활용하면 예측의 신뢰성을 높이는데 도움이 될 것이라 판단된다[2]. 실증데이터는 2012년 후 개봉한 영화 중 누적관객수 500만 이상의 영화 29편을 사용한다. 분석 데이터의 시작 기준을 2012년으로 정한 이유는 이 시점이 스마트폰 대중화의 분기점으로 간주하기 때문이다. 먼저 학습용 데이터(Training Data)를 활용하여 예측모델을 만들고 검증 데이터(Validation Data)를 활용하여 정합성을 검증한다[3]. 제안한 예측모형의 실효성을 파악 하기위해서 새로운 데이터인 2018년 7월부터 2019년 12월 사이에 개봉한 영화 중 누적관객수 500만 이상의 영화 16편을 선정하여 구축된 모델에 적용하여, 영화흥행 예측여부를 확인 하고자 한다.

본 논문의 전체적인 구성은 1장은 서론부분으로 전체의 로드맵을 제시한다. 2, 3장에서는 선행연구와 분석방법론을 구체적으로 언급하였다. 4장은 본 연구에서 제안한 모형의 개발방안을 기술하였으며 5장은 연구의 효과와 결론을 도출하였다.

2. 선행연구

영화산업이 급속도로 전문화되면서 영화흥행연구 또한 증가하는 추세이다. 영화는 수요예측이 어려운 만큼 수익에 대한 위험요소도 높기 때문에 투자사나 제작 그리고 배급사들 역시 영화흥행 예측은 매우 중요할 수밖에 없다. 이에 따른 선행연구로는 배우캐스팅과 감독 선정, 구전 마케팅, 개봉시기와 스크린 수와 같은 메타데이터들을 이용하거나 최근 SNS가 중요한 요소로 작용하면서 비정형 데이터들의 분석들이 활발하게 진행되고 있는 실정이다[6,7]. 영화의 메타데이터뿐만 아니라 평점, 댓글 수와 같은 관객 반응을 고려한 변수들을 함께 이용하여 영화의 수요를 예측하고 실제 영화 흥행 성적과 비교하는 연구도 있었다[4]. 그리고 전국 관객수를 기준으로 각 년도별 상위 100위의 영화를 분석 대상으로 한정하였는데, 주요 변수는 영화의 장르, 영화의 등급, 국적, 감독, 주연배우, 속편여부, 배급사, 스크린 수, 제작비, 개봉시기, 전문가 및 관객의 평점을 설정하는 연구도 있었다[4]. 김연형은 2011년 국내에서 개봉한 영화를 대상으로 흥행에 영향을 미치는 요인을 영화의 제작, 배급, 상영 단계별로 분석하였다. 영화 흥행에 관한 변수는 국적, 장르, 등급, 감독, 배우, 배급사, 스크린 수, 포털평점 등을 이용하여 예측 모형을 구축하였다[5]. 우종필은 스마트폰 보급률에 따른 검색 포털데이터 등이 영화흥행 수요예측에 영향을 미칠 것이라 생각하고 관련 데이터를 흥행예측변수로 사용하였다[11]. 김세윤은 영화의 메타데이터 중심인 감독, 배우, 장르, 등급, 제작사, 배급사가 흥행 예측에 영향을 줄 것이라 판단하고 예측 모형을 구축하였다[4]. 이상윤은 본격적인 4차 산업혁명시대에 진입에 발맞추어 인공지능 기반형 빅데이터 정보시스템의 구축 방안을 제시하였다[13]. 또한, 김진욱은 2019년에 개봉한 영화 '기생충'이 어떠한 요인들로 인해 천만 영화에까지 이르렀는지 살펴보고 연구문제를 설계하는 연구도 있었다[14]. 선행연구를 분석해본 결과 스마트폰의 대중화 이전시기에 영화 흥행에 영향을 주는 변수와 스마트폰의 대중화로 모바일 검색량과 SNS 사용량이 증가한 시기의 흥행 변수는 차이가 있음을 확인했다. 영화 관련 모바일 검색 및 SNS 사용의 폭발적인 증가는 영화 흥행예측에 SNS 관련 빅데이터의 중요성이 강조됨과 동시에 관련 새로운 독립변수의 영향력이 크질 것이라는 가설을 수립할 수 있고 그것에 대하여 검증하는 연구의 필요성이 제기된다. 이것을 위해

서 본 연구에서는 먼저 영화분야 전문가들은 영화의 흥행을 좌우하는 요인이 어떤 것인지를 델파이와 AHP 기법을 활용하여 설문조사를 실시한다. 설문조사로 확보된 데이터를 AHP 기법을 사용하여 측정요인별 중요도 즉 가중치를 부여한다. 또한, 측정요인별 중요도와 본 연구에서 수립한 가설하의 새로운 독립변수들과 상관의 정도를 파악한다. 상관의 정도를 고려하여 영화흥행예측 모형을 제안하고 실증데이터를 적용하여 모델의 신뢰성을 검증하고자 한다.

3. 분석방법론

3.1 분석에 필요한 데이터 및 변수

3.1.1 분석에 필요한 데이터

본 연구에서 분석에 사용되는 빅데이터는 아래와 같다.

•영화진흥위원회에서 제공하는 영화정보, 일일 흥행 데이터, 영화관 입장권 통합전산망데이터 등

•웹검색 사이트인 구글트렌드 데이터, 네이버 데이터랩 데이터, 네이버 뉴스 검색결과 데이터(전언론사, 공중파 TV, jTVC), 네이버 네티즌 평점 등

•서베이(Survey) 데이터 : 델파이 및 AHP기법을 사용하여 조사한 영화분야 전문가 데이터

3.1.2 분석에 필요한 변수

본 영화흥행 수요예측 모형에 사용되는 변수의 종류를 살펴보면 Table 1과 같다.

3.1.3 영화 흥행의 기준

영화 흥행의 기준을 삼는 방법은 몇 가지가 있지만 본 연구에서 영화 흥행의 기준은 누적고객수 1,000만으로 하겠다. 영화의 속성에 따라 차이는 있겠지만 일반적으로 상업영화에서 천만을 달성하면 흥행에 성공적이라 평가한다[11].

3.2 AHP분석 모형

3.2.1 전문가, 수요예측 변수 및 계층구조

가) 전문가 선정

영화 누적 관객 수 수요예측을 위하여 내적요인과 외적요인에 관련된 정보에 대한 전문가들의 지식(가중치)을 도출하기 위하여 설문조사를 통하여 각 속성에 대한 쌍별 비교를 실시하였다. 영화산업에 종사하고 있는 전문가(감독(5), 교수(5), 연기자(10), 제작사임원(5), 평론가(5) 등) 30명의(설문지 응답자) 전문가 집단을 구성하였다[12].

나) 관객 수 수요예측 변수 선정 및 계층구조

영화 흥행예측을 위해 관련 변수들을 내적요인과 외적요인으로 구분하였다. 내적요인은 감독, 주연배우, 영화의 장르, 관람 등급, 제작사, 배급사로 외적요인은 영화평점, 구글트렌드 검색량, 네이버트렌드 검색량, 네이브 뉴스 검색량, TV 방송 보도량, 스크린 수로 구성하였다.

Table 1. Types of variables used in prediction models

Classification		Variable	Prediction Method	Variable type
Dependent variable		Number of Customers	Cumulative Number of Customers by the Film Promotion Committee	Interval
		10 Million or Not	Whether or not to exceed 10 million	Nominal
Independent variable	Movie properties	Director	Average number of movie goers mobilized for three years before the release of a director's work	Interval
		Lead Actor	Average number of movie audiences mobilized for the three years before the release of the lead actor's work	Interval
		Genre	Drama, Comedy, Action, Thriller, Melodrama/Romance	Nominal
		Movie rating	12 years old, 15 years old, 19 years old	Nominal
		Producer	Average number of movie goers in the 3 years before the release of the production company	Interval
	WOM Effect	Movie rating	Average score (out of 10) evaluating the public's movie expectations on portal sites	Interval
		GTSV	Google Trends Search Volume	Interval
		NTSV	Naver Trend Search Volume	Interval
		NNSV	Naver News Search Volume	Interval
		TVBC	TV broadcast coverage(KBS+MBC+SBS+ jTBC	Interval
	Competitive Factor	Number of Screens	Number of screens in the first week of opening	Interval
		Distributor	Average number of movie goers mobilized in the three years prior to the release of the film by the distributor	Interval

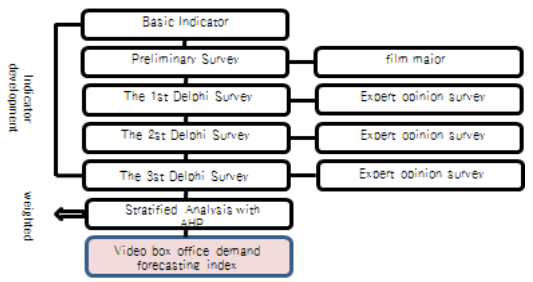

Fig. 1. Movie box office demand forecasting process using AHP

Fig. 1은 델파이와 AHP기법을 사용하여 영화전문가들에게 설문조사를 실시하여 흥행예측에 영향을 주는 변수들의 중요도를 알아내는 프로세스를 나타낸 것이다. 그리고 내적정보와 외적정보에 대한 12개의 속성으로 구성된 수요예측 모형은 Fig. 2와 같이 3개의 계층으로 구성하였다.

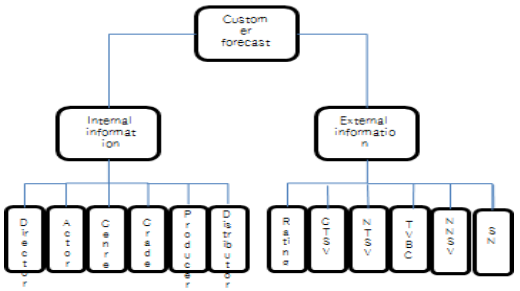

Fig. 2. Hierarchical structure of movie box office prediction model based on AHP

Table 2를 살펴보면 영화분야의 전문가들은 내적요인 보다는 외적요인이 흥행에 더 큰 영향을 주는 것으로 나타났다. 이는 서론에서 언급한 스마트폰 보급의 획기적인 확대로 모바일 포털 검색량 및 SNS사용량의 증가로 빅데이터 중심의 새로운 독립변수의 중심인 외적변수들의 영향력이 증가 때문으로 볼 수 있다.

Table 2. High factor importance assessment results

Category	CR	Importance	Priority
Internal Factor	CR=0.05	.357	2
External Factors		.643	1

Where : CR(Consistency Ratio)

Table 3. Importance of internal factor and evaluation results

Category	CR	Importance	Priority
Movie Director	CR=0.03	.114	5
Lead Actor		.237	1
Genre of Film		.164	4
Movie Grade		.092	6
FPC		.168	3
Distributor		.225	2

Table 4에서 나타난 것처럼 영화의 외적속성에서 흥행에 영향을 미치는 중요한 요인은 공중파 TV의 보도건수, 네이버 뉴스 건수, 구글트렌드 검색량 이다. 이것은 서론에서 언급한 스마트폰 사용으로 포털검색량 증가와 SNS 활성화로 인한 관련 빅데이터의 증가로 영화 흥행에 영향을 주는 새로운 독립변수의 영향력이 크질 것이라는 가설을 입증하는 것이라 볼 수 있다.

Table 4. Importance of external factors and evaluation results

Category	CR	Importance	Priority
Movie Rating	CR=0.03	.095	5
GTSV		.201	3
NTSV		.152	4
NNSV		.221	2
TVBC		.247	1
Number of Screens		.084	6

Table 5. Overall factor importance analysis

Main Category	Import-ance	Category	Import-ance	Total Weight	Ranking
Internal Factor	.357	Movie Director	.114	.019	11
		Starring Actor	.237	.113	4
		Movie Genre	.164	.071	8
		Movie Grade	.092	.003	12
		FPC	.168	.089	7
		Distributor	.225	.101	5
External Factors	.643	Movie Rating	.095	.030	9
		GTSV	.201	.143	3
		NTSV	.152	.093	6
		NNSV	.221	.146	2
		TVBC	.247	.165	1
		Number of Screens	.084	.027	10

전문가들이 생각하는 영화 흥행에 영향을 미치는 변수들의 가중치와 순위를 Table 5에 표시하였다. 내용을 살펴보면 전반적으로 내적요인 보다는 외적요인이 외적요인 중에서도 Table 4에 나타난 바와 같이 스마트폰의 활성화로 인해 발생한 빅데이터의 영향력이 증가함을 알 수 있다. 이러한 사실은 향후 영화흥행 여부를 예측하는 변수 선택에 참고가 될 것이다.

4. 분석 모형의 개발

4.1 모형 개발 절차

가. 탐색적 자료분석(EDA)와 AHP 기법을 사용하여 후보 변수를 확정한다.

나. 학습용 데이터(Training Data)와 검증용 (Validation Data) 분리한다.

다. 학습용 데이터를 이용해 모형을 개발한다.
라. 검증용 데이터를 이용해 모형을 검증한다.

4.2 분석 모형 개발 방안

4.2.1 모형 개발을 위한 분석 및 관찰 영역정의

January 2012 ····· June 2018(29 Films)	
Analysis area(Training Data):70%	Observation area(Validation Data):30%

Fig. 3. Define analysis and observation areas

일반적으로 관찰 영역과 분석 영역의 설정은 예측모형 개발에 있어 중요한 역할을 하게 된다. 이것을 설정할 때는 업무 전문가들의 의견을 청취하고 각 영역에 해당하는 데이터의 속성도 고려하여 설정하여야 한다. 따라서 전술한 상황을 종합하여 관찰 영역과 분석 분석영역을 Fig. 3과 같이 설정하였다[3].

4.2.2 분석 대상의 선정

분석 대상은 2012년 1월부터 2018년 6월 까지 개봉한 영화중에서 500백만 이상이 관람한 29편의 영화를 사용한다. 분석 데이터 테이블 셋(Master Table)형태는 가로축은 Table 1에서 제시한 독립 및 종속변수로, 세로축은 29편의 종류별 영화를 나열하고 가로축과 세로축의 각각의 셀에 해당하는 정보를 3.1.1장에서 언급한 데이터베이스에서 검색하여 할당하였다. 한편, 2011년 이전의 데이터는 전국 극장의 데이터를 가지고 있지 않거나 배급사를 통해 확인된 데이터이기 때문에 데이터의 정합성에 문제가 있다. 그리고 서론에서도 언급한 것처럼 스마트폰의 보급 확대의 분기점이 2012년으로 스마트폰 사용의 증가로 발생한 SNS 빅데이터는 영화흥행예측에 영향을 미칠 것으로 판단하여 데이터 사용의 기준을 2012년으로 정하였다. 분석영역의 학습용 데이터를 이용하여 모형을 만들고 관찰영역의 검증 데이터를 활용하여 모형을 검증한다. 일반적으로 모형을 만드는데 필요한 학습용(Training) 데이터로 70%를 사용하고 모형을 검증하는데 필요한 검증용(Validation) 자료로 30%를 사용한다[3].

4.2.3 분석 알고리즘

본 연구에서 제안한 영화흥행여부를 예측하는데 사용하는 알고리즘은 로지스틱 알고리즘, 의사결정나무(Decision Tree)알고리즘, 신경망(Neural Net)알고리즘 사용한다. 일반적으로 예측모형을 구축할 때 사용되는 알고리즘은 예측요인들의 원인규명이 필요한 경우에는 로지스틱 알고리즘이나 의사결정나무 알고리즘을 사용한다. 신경망 알고리즘을 사용했을 경우에는 예측변수들의 원인을 찾기가 용이하지 않기 때문이다.

4.3 분석 모형 구축

Table 5에서 제시한 바와 같이 전문가들도 영화의 흥행에 미치는 변수를 영화의 내적 요인 보다는 외적 요인인 구글 트렌드 검색량, 일평균 TV보도 건수, 네이브 트렌드 검색량 등 빅데이터 분석서비스의 제공으로 생성된 새로운 독립변수에 비중을 두고 있다. 본 연구에서는 이러한 사실을 바탕으로 2가지 모형을 제시하고 그것의 특징을 파악하고자 한다. 실증데이터를 내적변수 위주로 설계한 모형 1과 내적변수와 외적변수를 혼합하여 설계한 모형 2(제안한 모형)에 적용하여 각각의 모형에 영향을 미치는 변수의 특징을 파악하고 동시에 모형의 신뢰성을 검증하고자 한다[4].

모형 1(기존 모형) : 누적관객 수(천만관객관람 여부)
=감독+주연배우+장르+관람등급+제작사+배급사

모형 2(제안 모형) : 누적관객 수(천만관객관람 여부)
=감독+주연배우+장르+관람등급+제작사+배급사+영화평점+구글 트렌드 데이터+네이버 트렌드 데이터
+일평균 뉴스건수+일평균 TV건수+ 스크린 수

4.4 분석 모형의 평가

4.4.1 분류 행렬표(Classification Matrix)

이 행렬 표는 분류모형이 특정 데이터 집합에 대해 수행한 정분류와 오분류의 요약정보를 보여주며, 정오분류표의 행과 열은 각각 실제집단과 예측집단이 대응된다. 앞에서 언급한 29편의 영화데이터로 만들어진 데이터셋을 로지스틱, 의사나무결정나무 그리고 신경망 알고리즘을 적용하여 천만관객관람 여부를 예측한 결과는 Table 5 및 Table 6과 같다. 그런데 분류 행렬표에서 주목해야 할 부분은 천만관람을 예측하지 않았는데 실제로 관람한 경우의 오분류 예측보다 천만관람을 할 것이라 예측했는데 실제로 관람을 하지 않는 오분류 %를 줄이는 것이 중요하다. 왜냐하면 천만관람을 하지 않을

것이라 예측했는데 실제로 관람을 한 경우는 처음에 계획했던 것보다 수입이 증가하지만 반대의 경우에는 수익이 계획보다 감소하므로 영화제작사 입장에서는 여러 문제에 봉착할 가능성이 높기 때문이다. 모형 1 과 모형 2(제안한 모형)의 예측 정확도를 살펴보면 모형 2의 정확도가 모형 1보다 높게 나타남을 알 수 있다. 이것은 스마트폰 사용의 활성화로 축적된 빅데이터로 구성된 독립변수들이 예측모형에 유의미한 영향을 미친 것이다.

또한, 전문가 집단의 경험적 노하우도 모형에 유의적으로 작용함을 알 수 있다. 한편, 신경망 알고리즘은 흥행예측변수들의 원인을 찾기가 용이하지 않기 때문에 예측변수들의 원인을 찾아 마켓팅에 활용하는 경우에는 사용이 곤란하다.

Table 6. Classification matrix table of whether or not 10 million customers are viewed (Model 1)

Division	LR	A		DT	A		NN	A	
		Yes	No		Yes	No		Yes	No
B	Yes	10	2	Yes	10	1	Yes	11	1
	No	1	16	No	2	16	No	1	16
	True	26	89.65%	True	26	89.65%	True	27	93.10%
	Error	3	10.34%	Error	3	10.34%	Error	2	6.89%
	Sum	29		Sum	29		Sum	29	

Where A:Whether or not 10 million customers viewed(Predicted value)
B:Whether or not 10 million customers viewed(Real value)
LR:Logistic Regression, DT:Decision Tree, NN:Neural Network

Table 7. Classification matrix table of whether or not 10 million customers are viewed (Model 2)

Division	LR	A		DT	A		NN	A	
		Yes	No		Yes	No		Yes	No
B	Yes	11	1	Yes	10	2	Yes	12	0
	No	1	16	No	1	16	No	0	17
	True	27	93.10%	True	26	89.65%	True	27	100%
	Error	2	6.89%	Error	3	10.34%	Error	2	0%
	Sum	29		Sum	29		Sum	29	

Table 8. Classification matrix table of whether or not 10 million customers are viewed (Model 1)-Apply a new dataset

Division	LR	A		DT	A		NN	A	
		Yes	No		Yes	No		Yes	No
B	Yes	2	1	Yes	3	2	Yes	3	1
	No	3	10	No	2	9	No	2	10
	True	12	75.00%	True	12	75.00%	True	13	81.25%
	Error	4	25.00%	Error	4	25.00%	Error	3	18.75%
	Sum	16		Sum	16		Sum	16	

Table 9. Classification matrix table of whether or not 10 million customers are viewed (Model 2)-Apply a new dataset

Division	LR	A		DT	A		NN	A	
		Yes	No		Yes	No		Yes	No
B	Yes	3	1	Yes	4	1	Yes	4	0
	No	2	10	No	1	9	No	1	11
	True	13	81.25%	True	14	87.50%	True	15	93.75%
	Error	3	18.75%	Error	2	12.50%	Error	1	6.25%
	Sum	16		Sum	16		Sum	16	

영화 흥행예측 모형 1과 모형 2의 실효성과 활용성을 파악하기 위해서 새로운 데이터인 2018년 7월부터 2019년 12월 사이에 개봉한 영화 중 누적관객수 500만 이상의 영화 16편을 선정하여 구축된 모델에 적용하여, 영화흥행여부를 예측한 결과를 Table 8.과 Table 9.에 표시하였다. 예측 알고리즘의 종류에 관계없이 제안한 모형 2의 정확도가 모형 1보다 평균 10.5% 높게 나타났다. 이것은 스마트폰 보급률의 증가로 포털 검색량 및 SNS 사용의 증가로 발생한 빅데이터 관련 새로운 독립변수의 유의성을 재 입증한다고 볼 수 있다. 향후 영화 흥행예측에 관련한 빅데이터의 사용으로 예측모형의 정확성을 높이는데 기여할 것이다.

5. 결론

사물인터넷의 활성화와 모바일 컨텐츠의 발전은 앞으로 스마트폰 채널을 통한 소셜미디어의 사용의 의존율이 더욱 심화될 것으로 전문가들은 전망하고 있다. 이로 인하여 SNS 데이터를 비롯한 디지털 빅데이터의 기하급수적 생산은 그것을 활용하여 의사결정 정보를 생산하는 기회를 증가시킬 것이다. 본 연구에서는 영화 흥행에 미치는 요인들을 도출하고 그것들의 영향력의 정도를 파악하는데 객관적인 정형, 비정형 데이터뿐만 아니라 영화분야 전문가의 경험적인 지식과 식견을 포함하여 예측 모형의 신뢰도를 높이는데 활용하였다. 본 연구에서 제안한 흥행예측모형의 특징을 살펴보면, 첫째, 델파이와 AHP 기법을 이용하여 확보한 전문가들의 식견과(Table 5 참고) 기존의 영화흥행 메타데이터뿐만 아니라 빅데이터 관련 새로운 독립요인들의(구글 트렌드 검색량, 일평균 TV 보도 건수, 네이버 뉴스 검색건수 등)정보를 모두 사용하여 모형의 신뢰성도 제고하였다. 둘째, 모형의 실효성을 살펴보기 위해 제안한 모형에 최근에 개봉한 영화 16편을 활용하여 검증한 결과 제안한 모형이 기존 모형보다 정확도가 평균 10.5%의 향상을

보였다. 따라서 제안한 영화흥행예측 모형2는 영화제작사 및 배급사들의 의사 결정에 도움이 될 것이라 판단된다. 또한 영화 시장에서 흥행을 기록하는데 어떤 요인들이 영향을 미치는지에 대한 연구는 영화 산업의 구조적인 측면의 이해를 통해서 이 산업을 이해하는 초석이 될 것이며 스크린 쿼터나 영상산업 관련 정책을 세우는데 활용가치가 있다 하겠다. 다만 영화전문가들의 주관적인 경험과 식견을 최대한 객관화 하는 방법과 다양한 SNS 채널에서 확보된 방대한 비계량 빅데이터를 영화 흥행예측 모형에 반영하는 방안에 대한 후속연구는 계속적으로 필요하다 하겠다.

REFERENCES

[1] Variety. (2013) Big data : Media Embracing the Most Detailed Information about You yet. Retrieved from: http://variety.com/2013/biz/news/big-data-media- embracing- the-most- detailed-information-about-you-yet-12.

[2] Saaty. T. L. (2003). Decision-making with the AH P : Why is the Principal Eigen Vector Necessary. *European Journal of Operational Research, 145(1),* 85-91.

[3] M. G. Song & S. B. Kim. (2013). A Study on the Improvement of the Reliability of a Predictive Model Using Big Data Analysis. *Digital Policy Research, 11(6),* 03-112.

[4] S. O. Kim. (2018). *Predicting of Financial Success Using Data Analysis for Korean Movies.* Master's Thesis. Soongsil University, 1-23.

[5] Y. H. Kim & J. H. Hong. (2011). A Study for the Development of Picture Box-office Prediction Model. *Proceedings of the Korean Statistical Society, 18(6),* 859-869.
DOI: 10.5351/CKSS.2011.18.6.859

[6] S. H. Lee. (2015). A Study on Predicting Movie Performance Using Text Mining. *Journal of Korean Data Information Science Society, 26(6),* 1259-1269.

[7] O. J Lee. (2014). Analysis of Movie Box Office Success Factors Using Social Big Data. *Journal of the Korean Contents Association, 14(10),* 527-538.

[8] S. Y. Park. (2012). Effect of Word of Mouth Effect on Movie Box Office through SNS - Focusing on Sunny's Case. *Journal of the Korean Contents Association, 12(7),* 40-53.

[9] M. G. Song. (2016). The Suggestion of Big Data Platform for the Strengthening of Privacy and Enabled of Big Data. *Journal of Digital Convergence, 14(12),* 155-164.
DOI: 10.14400/JDC.2016.14.12.155

[10] S. J. Oh & C. H. Kim. (2016). Web Drama Analysis and Suggestion Using Social Media Big Data Mining and Opinion Mining Techniques. *Korean Society of Cartoon and Animation Studies, (44),* 285-306. DOI: 10.7230/KOSCAS.2016.44.285

[11] J. P. Ou & O. H. Lee. (2018). A Model of Predictive Movie 10 Million Spectators through Big Data Analysis. *Journal of the Korean Big Data Society, 203(1),* 63-71. DOI: 10.36498/kbigdt.2018.3.1.63

[12] H. L. Lee & G. H. Cho. (2013). A Study on Developing the Design Quality Indicator for School Building-Using Delphi Survey Method and AHP. *Journal of the Korean Institute of Architecture, 28(5),* 69-77.

[13] S. Y. Lee & H. J. Yun. (2019). A Study on Data Information System Based on Artifical Intelligence. *Journal of the KICES,* 14(2), pp. 377-388.

[14] J. W. Kim. (2020). Analyzing Factors of Success of Film Using Big Data. *Journal of Korean Entertainment Industry Association, 14(4),* 145-153.

송 민 구(Min-Gu Song) [정회원]

· 1988년 2월 : 동국대학교 통계학과 (이학학사)
· 1991년 8월 : 동국대학교 일반대학원 통계학과 응용통계학 전공(이학석사)
· 1997년 8월 : 동국대학교 일반대학원 통계학과 전산통계학 전공(이학박사)
· 1994년 9월~2007년 2월 : 동국대학교 대우 및 겸임교수
· 2002년 10월~2015년 8월 : 현대정보기술 BI 센터장(상무)
· 2016년 3월~현재 : 예원예술대학교 교양학부 교수
· 관심분야 : 빅데이터 분석, 디지털 화상처리, BI, 등
· E-Mail : minsong3@naver.com, P3172@office.yewon.ac.kr