J. Inf. Technol. Appl. Manag. 26(3): 19~27, June 2019
https://doi.org/10.21219/jitam.2019.26.3.019

ISSN 1598-6284 (Print)
ISSN 2508-1209 (Online)

정보 소득율 기반의 변수 선택을 통한 영화 관객 수 예측

박현목* · 최상현**

Predicting the Number of Movie Audiences Through Variable Selection Based on Information Gain Measure

Hyeon-Mock Park* · Sang Hyun Choi**

Abstract

In this study, we propose a methodology for predicting the movie audience based on movie information that can be easily acquired before opening and effectively distinguishing qualitative variables. In addition, we constructed a model to estimate the number of movie audiences at the time of data acquisition through the configured variables. Another purpose of this study is to provide a criterion for categorizing success of movies with qualitative characteristics. As an evaluation criterion, we used information gain ratio which is the node selection criterion of C4.5 algorithm. Through the procedure we have selected 416 movie data features. As a result of the multiple linear regression model, the performance of the regression model using the variables selection method based on the information gain ratio was excellent.

Keywords : Information Gain Ratio, Machine Learning, Movie Audiences

Received : 2019. 06. 06. Revised : 2019. 06. 19. Final Acceptance : 2019. 06. 21.

* First Author, Department of Bigdata, Chungbuk National University, e-mail : qpwo002@naver.com
** Corresponding Author, Professor, Department of MIS, Chungbuk National University, 1 Chungdaero Seowongu, Cheongjusi, Chungbuk, Korea, 28644, Tel : +82-43-261-3742, e-mail : chois@cbnu.ac.kr

1. 서 론

영화산업은 대표적인 고위험-고수익(High risk-High return)산업으로 영화의 경험재적인 특성상 영화가 개봉하기 전까지는 그 품질을 알 수 없는 불확실한 상황에 놓여 있다[Kang, 2017]. 영화가 흥행에 실패할 경우 제작비를 비롯한 제반 비용은 회수가 어렵게 된다. 그러나 흥행에 성공할 경우 추가적인 수입에 따른 한계 비용(marginal cost)은 매우 낮으나 한계 수입(marginal revenue)은 비례적으로 증가하게 되어 대규모의 수익을 얻을 수 있는 구조이다. 따라서 영화 제작자는 영화의 품질을 보장하기 위해 인지도가 있는 배우나 연출력이 우수한 감독을 통해 영화의 품질을 보장하고자 한다[Choi, 2016]. 2016년 한국 영화 수익성 분석에 따르면 2016년 전체 한국 '상업영화' 수익률은 17.63%로 나타났으며 이 중 순제작비 30억 원 이상이거나 와이드릴리즈 개봉방식의 영화의 경우 21.77%로 높게 나타났다. 그러나 제작비 규모로 볼 때 50억 원 미만의 영화들은 적자 수익률을 기록하고 있는 것으로 나타났다[Son, 2018].

따라서 영화의 흥행 성과를 결정하는 관객 수를 미리 예측하는 일은 매우 중요하며 관련된 다양한 연구가 시도되었는데, 흥행에 영향을 주는 요인들을 파악하기 위한 연구들이 다수 진행되었다. 해외에서 행해진 기존 연구에서는 장르, 감독, 배우, 스크린 수, 개봉시기, 관람등급 등을 흥행 요인으로 선정하였다[Litman and Kohl, 1989; Prag and Casavant, 1994]. 이후 온라인상에서 동적으로 생성되는 변수를 추가하는 분석을 진행하여 관객 수를 예측하는 연구가 진행되기도 하였으며[Chang, 2017], 잠재적인 수요자들의 설문을 토대로 영화의 흥행성과를 예측하는 연구도 진행되었다[Jang et al., 2009].

그러나 데이터의 획득 가능시점을 고려하였을 때 영화를 제작하기 전 단계와 개봉일 직전까지의 정보만으로 투자와 개봉일 선정, 스크린 수 할당을 결정해야 한다는 점에서 변수 획득의 시점과 의사결정 시점간의 차이가 충분하게 고려되어야 한다는 문제가 있다. Kim and Kwon[2017]의 연구에서는 투자 초기에 수집 가능한 정보만을 활용하여 영화제작 전 투자에 관한 시사점을 제공하였다. 그러나 영화 투자 초기에 수집 가능한 정보만을 구성하였을 때 영화로부터 생성되는 대부분의 정보는 범주형 속성이며, 범주형 속성을 처리하고 비교하기 위한 연구는 활발하지가 않다. 따라서 본 연구에서는 개봉 전 획득이 용이한 영화 정보의 추출과 효과적으로 정성적 변수를 구분하기 위한 방법론을 제안하고자 한다. 또한 구성된 변수를 통해 데이터 획득 시점별로 관객 수 예측 모델을 구축하여 분석을 진행하였고 이를 통해 영화 산업에 대한 시사점을 제안하고자 한다.

2. 영화 관객 수 예측 선행연구

영화 관객 수 예측에 관하여 회귀모델을 비롯한 다양한 기계학습 알고리즘을 활용한 연구가 진행되었다. 우선 SF와 호러 등의 영화 장르 변수가 영화의 흥행 성과에 긍정적인 영향을 미친다는 연구 결과가 제시되었으며, 이후 연속물, P&A, 개봉 스크린 수, 공휴일 여부는 흥행 성과에 긍정적인 영향을 미친다는 연구결과가 제시되었다[Litman, 1983; Stimpert and Laux, 2008]. 또한 개인의 심리적인 속성에 관해 관람의도와 영화에 대한 인지도는 개봉 첫 주 흥행 실적에 긍정적 영향을 준다는 연구 결과가 제시되었다[Jang et al., 2009].

기계학습 알고리즘으로는 의사결정나무와 다중회귀분석을 사용한 영화 VOD 흥행 예측과 흥행 요인에 대한 연구가 진행되었다[Moon, 2017]. 나이브 베이지안 분류기 모델을 통한 연구로는 텍스트 마이닝(Text Mining)을 통한 영화 군집화와 이를 나이브 베이지안 분류기에 적용한 흥행 성과의 예측 연구가 진행 되었으며[Lee, 2017], 정적 속성 정보와 동적 속성 정보의 반영을 통한 나이브 베이지안 분류기와 인공 신경망의 비교 연구가 진행되었다[Chang, 2017]. 의사결정나무, 인공 신경망, SVM, 다항로짓델을 통해 온라인 구전효과를 반영한 예측 연구가 진행되었다[Jeon and Son, 2016]. 선행연구에 관한 내용을 표로 정리하면 〈Table 1〉과 같다.

본 연구에서는 전처리 과정에서 정보 소득율이 높은 변수와 낮은 변수에 따른 예측 성능을 비교하고자 정보소득을 계산하였다. 의사결정나무에 엔트로피(Entropy)는 클래스 C_i를 구분하기 위해 필요한 평균 정보량을 뜻한다[Quinlan, 1986; 1993]. 이를 수식으로 나타내면 식 (1)과 같다.

〈Table 1〉 Previous Studies in Movie Industry

Model	Summary	References
Multi-Linear Regression	Positive impact on sales of video/DVD sales in SF and horror genres	Litman〔1983〕
	Series, P&A, number of open screens, holidays are positive for sales performance	Stimper and Laux〔2008〕
	Intent and awareness of movies have a positive impact on the first week's performance	Jang et al.〔2009〕
Machine Learning	Analysis of Movie VOD Entertainment Factors through Decision Tree and Multiple Regression Analysis	Moon〔2017〕
	Clustering through text mining and forecasting performance by each cluster	Lee〔2017〕
	Clustering through text mining and forecasting performance for each cluster	Chang〔2017〕
	Predicting movie audience number reflecting online word-of-mouth effect	Jeon and Son〔2016〕

$$Entropy(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (1)$$

p_i는 데이터 집합 D 내 튜플이 클래스 C_i에 포함될 가능성을 뜻하며 0이 아닌 값으로 집합 내 C_i에 해당하는 튜플의 개수를 집합 내 전체 튜플의 개수로 나눈 값이다. $\log_2$는 정보를 2진수 비트화 했다는 의미이다.

독립변수 A의 평균 정보량은 데이터 집합 내에서 독립변수가 갖는 범주나 연속형 값의 역치를 통해 C_i를 분류할 때 필요한 정보량의 기댓값이다. 이를 수식으로 나타내면 식 (2)와 같다.

$$Entropy_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Entropy(D_j) \quad (2)$$

$\frac{|D_j|}{|D|}$는 데이터 집합 내에서 독립변수 A의 레벨(level) j에 해당하는 튜플의 개수를 전체 튜플의 개수로 나눈 값으로 독립변수 A의 레벨 j에 해당하는 튜플 집합에 대한 가중치이다. $Entropy(D_j)$는 독립변수 A의 레벨 j에 해당하는 데이터 집합 내에서 C_i를 분류하기 위해 요구되는 정보량의 기댓값이다. 이를 통하여 식 (1)과 식 (2)의 차를 통하여 정보 소득율을 구할 수 있으며 이를 수식으로 표현하면 식 (3)과 같다.

$$Gain(A) = Entropy(D) - Entropy_A(D) \quad (3)$$

내부 노드를 결정하는 기준은 $Gain(A)$의 값을 가장 크게 하는 독립변수 A를 선정하는 것이다.

3. 분석대상 영화에 대한 기초통계 분석

영화 관련 데이터의 수집을 위해 관객 수, 스크린 수, 장르, 개봉일 등 영화의 기초가 되는 집계 데이터는 한국영화진흥위원회 통합전산망의 데이터를 이용하였다. 영화의 개봉 전 평점 및 전문가 평점과 참여자 수 데이터는 네이버 포털 사이트를 이용하였다. 영화의 제작비 정보를 수집하기 위해 IMDB와 인터넷 신문기사 데이터를 활용하였다. 박스오피스 영화는 2014년부터 2016년까지의 데이터를 수집하였고 총 11,146편의 영화가 수집되었다. 이 중 상업적인 목적이 뚜렷한 영화를 선별하기 위하여 관객 수 10만 이상의 영화로 제한하기로 하였다. 2014년 박스오피스 영화 중 일부 영화의 경우 개봉일이 2013년 12월인 경우가 있어 최종 관객 수 정보를 통합시키는 전처리를 진행하였다.

〈Table 2〉 Collected Movie Data

Items	Contents
Period	2013. 12~2016. 12
Sites	Film Promotion Committee Integrated Network(http://www.kobis.or.kr/) Naver Movie(http://movie.naver.com/) IMDB(http://www.imdb.com/) Internet news articles
Attributes	Movie title, opening date, sales volume, number of audience, number of movie screenings, frequency of movie screenings, genre, distributor, viewing class, director, major actor, movie rating before opening, number of rated participants before opening, rating of movie critic, number of rated critic, estimated production cost

최종적인 분석을 위해 총 416편의 영화가 선정되었으며 416편 영화의 장르, 주연배우, 배급사 등의 기본 정보, 개봉 전 평점 및 참여자 수 정보 그리고 추정 제작비 정보를 수집하여 분석을 위한 데이터에 통합하였다. 데이터 수집경로와 영화에 대한 속성 정보는 〈Table 2〉와 같다.

본 연구에 사용된 416편 영화에 대한 정성적인 속성으로는 개봉년도, 계절, 제작국가, 관람등급, 장르, 추정 제작비의 속성 정보가 있다. 개봉년도의 경우 2013년 12월이 1.9%(8편), 2014년 32.5%(135편), 2015년 33.2%(138편), 2016년 32.5%(135편)로 구성되었다. 계절 별 편수의 경우 봄 21.9%(91편), 여름 23.1%(96편), 가을 24.3%(101편), 겨울 30.8%(128편)로 구성되었다. 제작국가의 경우 국내 38.5%(160편), 해외 61.5%(256)편으로 구성되었다. 관람등급의 경우 전체 관람가 19.2%(80편), 12세 관람가 30.8%(128편), 15세 관람가 34.6%(144편), 청소년 관람불가 15.4%(64편)로 구성되었다. 장르의 경우 액션 23.8%(99편), 애니메이션 17.8%(74편), 드라마 17.5%(73편), 코미디 6.5%(27편), 멜로/로멘스 6.3%(26편), 범죄 5.3%(22편), 어드벤처 4.3%(18편), 스릴러 3.8%(16편), 사극 2.6%(15편), 공포 2.6%(11편), 미스터리 2.6%(11편), SF 2.4%(10편), 판타지 1.4%(6편), 다큐멘터리 1%(4편), 전쟁 0.5%(2편), 가족 0.2%(1편), 뮤지컬 0.2%(1편)로 구성되었다.

본 연구에 사용된 416편 영화에 대한 정량적인 속성으로는 관객 수, 개봉 전영화 포털 평점, 평점 참여자 수, 편론가 평점, 참여 평론가 수, 개봉일 스크린 수, 최대 상영 스크린 수, 개봉일 스크린 수, 상영 회수, 추정 제작비가 있다. 관객 수는 평균 154,4만 명, 최댓값 1,751.3만 명, 최솟값 10만 명으로 나타났다. 개봉 전 영화 포털 평점의 경우는 평균 8.82점, 최댓값 10점, 최솟값 0점으로 나타났다. 평점 참여자 수의 경우는 평균 606.1명, 최댓값 26,665명, 최솟값 0명으로 나타났다. 평론가 평점의 경우는 평균 5.56점, 최댓값 9.5점, 최솟값 0명으로 나타났다. 최대 상영 스크린 수의 경우는 평균 588.5개, 최댓값 1,991개, 최솟값 50개로 나타났다. 개봉일 스크린 수의 경우는 평균 501.2개, 최댓값 1,864개, 최솟값 31개로 나타났다. 상영 회수의 경우는 평균 36,285.8회, 최댓값 212,675회, 최솟값 2,291개로 나타났다. 결측값 86편을 제외한 나머지 영화 330편에 대해 수집된 추정 제작비의 경우 55,137.6천 달러, 최댓값 250,000천 달러, 최솟값 90천 달러로 나타났다. 이를 표로 정리하면 〈Table 3〉과 같다.

〈Table 3〉 Quantitative Attribute Information

Attribute	Average	Max	Min
audience(1000)	154.4	1761.3	10
rating before opening	8.82	10	0
No. of participant	606.1	26,665	0
critic rating	5.56	9.5	0
No. of critic	5.7	18	0
max no. of movie screenings	588.5	1,991	50
No. of screens at opening	501.2	1,864	31
frequency of screening	36,285.8	212,675	2,291
estimated cost ($1000)	55,137.6	250,000	90

4. 데이터 분할 및 전처리

지도학습은 클래스로 이산화 된 훈련용 데이터 집합을 사용하는 기계학습 방법이다. 본 연구에서는 지도학습 기반의 기계학습 알고리즘을 활용하여 분석을 진행하기 때문에 연속형 데이터를 이산화 된 클래스로 변환하는 작업이 필요하다.

4.1 데이터 분할

사용된 기계학습 알고리즘의 적합성을 검정하기 위해 훈련용 데이터 집합과 검증용 데이터 집합으로 구분하여야 한다. 따라서 2013년 12월부터 2015년 12월까지 개봉한 영화 데이터를 훈련용 데이터 집합으로 구성하였으며, 2016년 1월부터 12월까지 개봉한 영화 데이터를 검증용 데이터 집합으로 구성하였다. 훈련용 데이터 집합은 281편의 영화가 구성되었으며 검증용 데이터 집합에는 135편의 영화가 구성되었다.

또한 연속형 데이터인 관객 수 예측을 위해 범주화하여 클래스를 부여했다. 클래스 부여의 기준은 관객 수가 10만 명 이상부터 50만 명 미만, 50만 명 이상부터 100만 명 미만, 100만 명 이상부터 250만 명 미만, 250만 명 이상부터 500만 명 미만 그리고 500만 명 이상으로 범주화하여 클래스를 부여했다. 클래스 부여 결과 훈련용 데이터 집합의 경우에는 10만 명 이상부터 50만 명 미만 49.5%(139편), 50만 명 이상부터 100만 명 미만 15.3%(43편), 100만 명 이상부터 250만 명 미만 16.4%(46편), 250만 명 이상부터 500만 명 미만 12.1%(34편), 500만 명 이상 6.8%(19편)로 구성되었다. 검증용 데이터 집합인 경우에는 10만 명 이상부터 50만 명 미만 43.7%(59편), 50만 명 이상부터 100만 명 미만 17.8%(24편), 100만 명 이상부터 250만 명 미만 17.4%(24편), 250만 명 이상부터 500만 명 미만 12.6%(17편), 500만 명 이상 8.1%(11편)로 구성되었다.

4.2 데이터 전처리

분석을 위해 각 변수들에 대한 결측값 보정, 전처리 및 변환 과정이 필요하다. 분석 대상인 416편의 영화에 대해 제작국가의 경우 14개 국가가 포함된다. 배급사의 경우 49개의 배급사 카테고리가 존재한다. 장르의 경우 16개의 장르로 구성되어 있다. 감독파워의 경우 감독파워를 구할 수 없는 영화 91편이 포함되어 있다. 배우파워의 경우에는 배우파워를 구할 수 없는 영화 49편이 포함되어 있다. 제작비의 경우 추정 제작비 정보를 구할 수 없는 영화가 86편이 포함되어 있다. 이를 분석 모델에 반영하기 위해서는 별도의 처리가 필요하며 특히 감독파워, 배우파워, 제작비의 경우 연속형 변수를 활용하는데 결측값을 0으로 부여하는 경우 모델의 신뢰도에 문제가 발생한다. 따라서 이러한 데이터들에 대해 결측치 보정이나 범주화 전처리가 필요하다. 범주화 전처리가 필요한 변수는 14개의 속성을 가지는 제작국가, 49개의 속성을 가지는 배급사, 17개 속성의 장르 등의 세 가지이다. 또한, 감독파워, 배우파워, 제작비의 경우에는 정량적 수치로 주어진 변수를 범주형으로 변환하였다. 이와 같은 범주형 또는 범주화된 변수인 제작국가, 배급사, 장르, 감독파워, 배우파워, 제작비에 대한 전처리 작업이 진행되어야 하며 새로운 특징 추출을 진행하고자 개봉일 정보를 바탕으로 개봉 후 공휴일이 존재하는지 여부를 특징으로 추가하였다. 변수의 전처리 및 특징추출을 위한 기준으로 의사결정트리 알고리즘에 활용되는 정보 소득율 값, 식 (3)을 활용하였다.

앞서 제시한 7개의 변수들에 대해서는 정보소득율을 사용하여 변수의 구분을 사용하였는데 어떻게 해당 변수의 범주 데이터를 선택할지 장르의 사례를 가지고 설명하기로 한다. 장르의 경우 17개의 장르가 존재한다. 이를 분석에 활용하기에는 각각의 장르에 해당하는 데이터가 충분하지 못해 기계학습 알고리즘의 분류 결과를 신뢰하지 못하는 문제가 발생한다. 따라서 Rasheed and Shah[2002]의 장르 구분에 따라 액션, 범죄, 어드벤처, 사극, SF, 전쟁 장르를 액션형 장르로 구분하였다. 드라마, 코미디, 멜로, 스릴러, 공포, 미스터리, 판타지, 다큐멘터리, 가족, 뮤지컬 장르를 비액션형 장르로 나누었다. 별도로 애니메이션 장르로 분류한 구분 1을 구성하였다. 그러나 애니메이션의 경우 일본형 애니메이션과 헐리우드 기반의 애니메이션이 구분된다. 따라서 액션형 장르, 비액션형 장르, 애니-비아시아, 애니-아시아로 분류한 구분 2를 구성하였다. 정보 소득율은 장르 구분 1 : 0.0461, 구분 2 : 0.0559로 구분 2로 분류한 결과 정보 소득율이 상승하였다. 분류한 결과표는 〈Table 4〉와 같다. 결국 장르 변수의 경우 정보소득율 값이 큰 구분 2로서, 액션형, 비액션형, 애니/비아시아, 애니/아시아 등의 4개의 범주를 사용하기로 하였다.

〈Table 4〉 Genre Division and Information Gain

Genre		Number	Ratio (%)	Gain
G1	action	117	41.6	0.0461
	no-action	118	42	
	animation	46	16.4	
G2	action	117	41.6	0.0559
	action	118	42	
	animation/ non asia	26	9.3	
	animation/ asia	20	7.1	

〈Table 5〉 Selected Variables

Attribute		Transformation	Gain
Target	no. of audience	100~500, 500~1000, 1000~2500, 2500~5000, above 5000 (unit : 1000)	-
Independent	No. of screens at opening	Numeric	0.4688
	estimated cost	adjusted value	0.1903
	No. of participant	Numeric	0.1696
	No. of critic	Numeric	0.1377
	director power	under 1000, 1000~5000, above 5000(unit : 1000)	0.1011
	distributor	Big4, medium, foreigner direct	0.0935
	actor power	under 1000, 1000~2500, 2500~5000, above 5000(unit : 1000)	0.0902
	country	domestic, USA, another	0.0596
	genre	action, no-action, ani-asia, ani-non asia	0.0559
	viewing class	all, above 12 years, above 15 years, no youth	0.0337
	holiday in a week	yes, no	0.0284
	season	spring, summer, fall, winter	0.0172

이와 같은 방식으로 나머지 6개의 변수들에 대해서도 정보소득율을 사용하여 구분을 선택하였다. 다음 〈Table 5〉는 최종적인 모형에 사용된 종속변수와 12개의 독립변수들을 정리하였다.

종속변수는 관객 수의 등급에 따라 5등급으로 구분하였다. 독립변수는 개봉일 스크린 수, 추정제작비, 평점 참여자 수, 참여 평론가 수 등의 4개의 정량적 변수와 감독파워 구분, 배급사 구분, 배우파워 구분, 제작국가 구분, 장르 구분, 관람등급, 개봉 후 1주일 이내 공휴일 유무, 계절 8개의 범주형 변수 등을 포함하는 12개의 변수가 사용되었다.

5. 데이터마이닝 모델 적용결과 분석

정보 소득율이 높은 변수와 초기 선정했던 변수와의 다중선형회귀 모델을 통해 예측 성능의 차이를 확인하고자 하였다. 다중선형회귀 분석을 수행하기 전 윈저라이징 기법을 통해 관객 수 500만 명 이상의 영화들은 500만 명으로 변환하였다. 제작비, 감독파워, 배우파워의 결측치를 제거한 161편의 영화에 대해 상관분석을 진행한 결과 변수 간 0.5 미만의 상관관계를 보였다. 또한 제작비 결측치를 보정하고 감독파워, 배우파워의 결측치를 제거한 206편의 영화에 대해 상관분석을 진행한 결과 변수 간 0.5 미만의 상관관계를 보였다. 상관분석 결과를 정리하면 〈Table 6〉과 같다.

〈Table 6〉 Correlation Analysis

		cost	director	actor
after mis+sing value	cost	1		
	director	0.105	1	
	actor	-0.181	0.38	1
		cost	director	actor
after cost adjust	cost	1		
	director	0.087	1	
	actor	-0.214	0.379	1

최종적으로 다중선형회귀 모델을 적용하여 훈련용 데이터 집합 내에서 10중 교차검증을 실시하여 모델을 훈련시켰다. 변수로는 정성적 변수들과 제작비를 활용하였으며 훈련 결과 초기 모델 변수에서는 비액션형 장르, 중예산/고예산 제작비, 감독파워 100만 이상의 변수가 유의미한 변수로 나타났다. 정보 소득율 선택 모델의 경우 여름, 1주 이내 공휴일 있음, 국내 제작, 중소 배급사, 제작비, 감독파워 500만 이상의 변수가 유의미한 변수로 나타났다. 수정 R제곱의 값은 초기 모델 0.266, 정보 소득율 선택 모델 0.394로 정보 소득율 선택 모델의 설명력이 높은 것으로 나타났다. 다중선형회귀 모델의 적용 결과는 〈Table 7〉과 같다.

〈Table 7〉 Result of Model Fitting

Model using information gain		
attribute	coefficient	p-value
holiday : yes*	364,724.4	0.04169
country : another	-137,067	0.602286
country : domestic***	997,288.7	0.000146
distributer : medium*	-434,280	0.029653
distributor : foreigner direct	414,205.6	0.084554
above 12	141,559.3	0.655358
above 15	467,343	0.145999
no youth	3,488.427	0.992149
genre : no action	-175,192	0.362677
genre : ani-non asia	-172,756	0.68287
genre : ani-asia	-546,684	0.184694
adjusted cost***	0.007308	0.000252
director : under 1000	111,386.8	0.593464
director : 1000~5000	308,912.7	0.158041
director : above 5000***	1,817,796	2.69E-06
actor : under 1000	-358,172	0.24498
actor : 1000~2500	-483,574	0.172184
actor : 2500~5000	-263,428	0.492502
actor : above 5000	89,6830.2	0.116565
original model	Multiple R-squared : 0.3184 Adjusted R-squared : 0.266 F-statistic : 6.074 on 18 and 234 DF, p-value : 4.8e-12	
gain model	Multiple R-squared : 0.4469 Adjusted R-squared : 0.394 F-statistic : 8.447 on 22 and 230 DF p-value : 〈 2.2e-16	

정보소득율을 사용한 모델이 사용 전 모델인 초기 모델에 대비하여 더욱 높은 R-squared 값을 가지는 것을 확인할 수 있다.

RMSE와 MAPE를 산출하기 전 모델의 예측값이 음수인 경우 예측값을 0으로 전환하여 RMSE와 MAPE를 산출하였다. 학습된 다중선형회귀 모델을 통해 2016년 1월부터 12월까지의 검증용 데이터 집합에 적용한 결과 초기 모델의 경우 RMSE는 1282156, MAPE는 231.02%로 나타났다. 정보 소득율 선택 모델의 경우 RMSE 1339819, MAPE 216.78%로 RMSE는 초기 모델이 약간 적게 나왔으나 MAPE에서는 정보 소득선택 모델이 초기 모델보다 높다. RMSE에서는 두 모델이 10% 미만의 차이를 보이나 MAPE에서는 10% 이상의 차이를 보인다. 따라서 모델의 수정 R제곱 값과 오차를 종합하여 볼 때 정보 소득율 선택 다중선형회귀 모델의 예측력이 다소 우수하다고 볼 수 있다. 이를 정리하면 〈Table 8〉과 같다.

〈Table 8〉 Accuracy Measures

measure	Original model	Gain model
RMSE	1282156	1339819
MAPE	231.02%	216.78%

6. 결 론

본 연구에서는 한국영화진흥위원회 통합전산망 DB를 활용하여 2014년 1월부터 2016년 12월까지 관객 수 10만 명 이상의 박스오피스 영화 416건에 대한 데이터를 수집하였다. 추정 제작비가 존재하지 않는 영화의 제작비를 보정하는 한편 범주형 속성의 경우 정보 소득율이 높은 속성으로 특징을 추출하는 작업을 진행하였다.

장르 등의 7개의 정성적 변수에 대해서는 범주형 변수로 변환하기 위한 보다 효과적인 범주 구분을 찾기 위해 전처리 과정에서 정보소득율을 사용하였다. 정보소득율이 높은 변수를 선택한 후 예측 효과 개선을 확인하기 위해 다중선형회귀분석을 실시하였고 모델의 수정 R제곱근 값, RMSE, MAPE를 종합하여 고려하였을 때 정보 소득율이 높은 변수를 선택한 모델의 예측력이 개선됨을 확인할 수 있었다.

본 연구는 정량적인 속성보다 정성적인 속성이 많은 영화 정보의 특성 상 정보 소득율을 통하여 정성적인 속성을 처리하는 기준을 제시하였다. 이 기준을 통하여 영화의 흥행 성과 예측을 진행하였고 이를 통한 비즈니스적 전략을 제시한다는 점에서 의의가 있다.

향후 연구에서는 대상 영화의 수를 보다 확대하여

적은 관객 수의 영화에 대한 예측할 수 있는 모델에 대한 연구가 진행될 필요가 있다. 또한 다양한 데이터마이닝 모델을 적용하여 보다 적합한 모형을 찾아보는 비교연구도 수행될 필요가 있다.

References

[1] Chang, J. Y., "An Experimental Evaluation of Box office Revenue Prediction through Social Bigdata Analysis and Machine Learning", *The Journal of The Institute of Internet, Broadcasting and Communication*, Vol. 17, No. 3, 2017, pp. 167-173.

[2] Choi, B., "Study on Competitiveness and Economic Impact of Film Industry", Korea Film Council, 2016-03, 2016.

[3] Industrial Policy Research Team, *2016 Korean film closing*, Korea Film Council, 2016.

[4] Jang, B., Lee, Y., Kim, B., and Nam, S., "Elaborating Movie Performance Forecast Through Psychological Variables : Focusing on the First Week Performance", *Korean Society for Journalism & Communication Studies*, Vol. 53, No. 4, 2009, pp. 346-371.

[5] Jeon, S. and Son, Y. S., "Prediction of box office using data mining", *Korean Journal of Applied Statistics*, Vol. 29, No. 7, 2016, pp. 1257-2170.

[6] Kang, S. J., "Analysis Box Office Success of A Movie-Focused on Commercial Film Released in 2016", *Journal of the Korean Entertainment Industry Association*, Vol. 11, No. 5, 2017, pp. 1-15.

[7] Kim, Y. and Kwon, O., "Movie Performance Indicators to Predict for Investors", *Journal of the Korean Data Analysis Society*, Vol. 19, No. 4, 2017, pp. 1963-1975.

[8] Lee, J., *A study on movie success prediction model using textmining and naive bayes*, Hongik University, Thesis for Master degree, 2017.

[9] Litman, B. and Kohl, L. S., "Predicting financial success of motion pictures : The '80s experience", *Journal of Media Economics*, Vol. 2, Issue 2, 1989, pp. 35-50.

[10] Litman, B. R., "Predicting success of theatrical movies : An empirical study", *Journal of Popular Culture*, Vol. 16, No. 4, 1983, pp. 159-175.

[11] Moon, J., *Analyzing the characteristics of movie VOD success using decision tree and multiple linear regression*, Seoul National University of Science and Technology, Thesis for Master degree, 2017.

[12] Prag, J. and Casavant, J., "An empirical study of the determinants of revenues and marketing expenditures in the motion picture industry", *Journal of Cultural Economics*, Vol. 18, Issue 3, 1994, pp. 217-235.

[13] Quinlan, J. R., "C4. 5 : Programming for machine learning", *Morgan Kauffmann*, Vol. 38, 1993.

[14] Quinlan, J. R., "Induction of decision trees", *Machine Learning*, Vol. 1, 1986, pp. 81-106.

[15] Rasheed, Z. and Shah, M., "Movie Genre Classification By Exploiting Audio-visual Features of Previews", *International Conference on Pattern Recognition, IEEE*, Vol. 2, 2002, pp. 1086-1089.

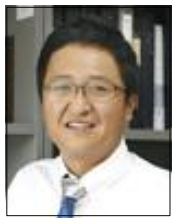
[16] Son, J., "2016 Korean movie profitability analysis", Korea Film Council, 2018.

[17] Stimpert, J. L. and Laux, J. A., "factors Influencing Motion Picture Success : Empirical Review And Update", *Journal of Business & Economics Research*, Vol. 6, No. 11, 2008, pp. 39-52.

저자소개

박 현 목

충북대학교 경영학부에서 학사, 빅데이터학과에서 석사 학위를 취득하였으며 현재 에프앤가이드 인덱스팀으로 재직중이다. 주요 관심분야는 빅데이터, 퀀트투자, 데이터마이닝, 회귀분석 등이다.



최 상 현

한양대학교 산업공학과에서 학사, KAIST 산업공학과 석사, 경영정보공학 박사 학위를 취득하였으며 현재 충북대학교 경영정보학과 교수로 재직 중이다. LG CNS 엔트루 컨설팅에서 CRM 전략 컨설팅 및 시스템 구축, 정보화 전략 계획 수립, ERP 시스템 구축 등의 IT 컨설팅을 수행하였다. 주요 관심분야는 빅데이터, 스마트팩토리, 데이터마이닝, 전략적 의사결정 시스템 등이다.